

**PREDICTION OF NOVEL BIOCHEMICAL CLASS,  
DISEASE RELATED PROTEINS AND MICRORNAS BY  
MACHINE LEARNING APPROACH**

**ZHANG HAILEI**  
**(B.Sc. & M.S., Dalian University of Technology)**

**A THESIS SUBMITTED  
FOR THE DEGREE OF DOCTOR OF PHILOSOPHY  
DEPARTMENT OF PHARMACY  
NATIONAL UNIVERSITY OF SINGAPORE  
2008**

## ACKNOWLEDGEMENTS

Foremost, I would like to present my sincere thanks to my supervisor, Professor Chen Yu Zong, for his excellent guidance, invaluable advices throughout my PhD study.

I would like to thank Professor Cao Zhiwei and Professor Ji Zhiliang for their insightful suggestions to my work on the prediction of disease related protein and multifunctional enzymes.

My sincere gratitude also goes to BIDD group members, especially Dr. Lin HongHuang, Dr. Han Lianyi, Dr. Zheng Chanjuan, Dr. Cui Juan, Dr. Wang Rong, Ms. Tang Zhiqun, Mr. Xie Bin, Ms. Ma Xiaohua, Miss Jia Jia, Miss Liu Xin, Miss Shi Zhe, Miss Wingyee, Mr. Zhu Feng, Mr. Liu Xianghui, Ms. Ong Serene etc. I am really thankful for their valuable suggestions and support in my project, as well as enjoy the close friendship among us.

Last, but not the least, I am eternally grateful to my parents and my husband for supporting and encouraging me throughout my life.

Zhang Hailei

April 2008

## TABLE OF CONTENTS

ACKNOWLEDGEMENTS .....	I
TABLE OF CONTENTS.....	II
SUMMARY .....	IV
LIST OF TABLES .....	VII
LIST OF FIGURES .....	X
LIST OF ACRONYMS .....	XIII
1. Introduction.....	1
1.1. Introduction to multifunctional enzymes (MFEs).....	2
1.2. Introduction to disease related proteins .....	4
1.2.1. Antimicrobial proteins .....	4
1.2.2. Antibiotic resistance proteins.....	5
1.2.3. Cancer associated proteins.....	7
1.3. Introduction to microRNAs .....	9
1.4. Overview of computational methods for biological function prediction.....	12
1.4.1. Sequence similarity method.....	12
1.4.2. Motif based methods.....	13
1.4.3. Machine learning approach.....	15
1.5. Scope and objective .....	15
2. Methods.....	18
2.1. Machine learning methods.....	18
2.1.1. Support Vector Machine (SVM).....	19
2.1.2. K-Nearest Neighbors (KNN) .....	27
2.1.3. Neural Networks (NN).....	29
2.1.4. Decision Tree (DT).....	30
2.2. Feature selection .....	32
2.3. Performance evaluation .....	34
2.4. Construction of feature vectors.....	35
2.4.1. Protein feature vectors .....	35
2.4.2. MiRNA feature vectors.....	39
3. <i>In silico</i> search and characterization of multifunctional enzymes .....	41
3.1. Selection of MFEs and non-MFEs.....	41
3.2. Evaluation and discussion.....	43
3.2.1. Structural preference of MFEs.....	43
3.2.2. Characteristics of MFEs from pathway and evolution perspective .....	45
3.2.3. Identification of novel MFEs.....	56
3.2.4. Contribution of physicochemical properties in the classification of MFEs.....	57
3.3. Server for identification of multifunctional enzyme (SIME) .....	58
3.4. MFEs database .....	61
3.5. Summary .....	64
4. Prediction of disease related proteins by support vector machine.....	66
4.1. Prediction of antimicrobial proteins.....	66
4.1.1. Selection of antimicrobial proteins and non-antimicrobial proteins .....	66
4.1.2. Prediction performance for antimicrobial proteins.....	68

4.1.3.	Prediction of novel antimicrobial proteins.....	69
4.1.4.	Contribution of feature properties.....	76
4.1.5.	Server for antimicrobial protein identification (SAPI).....	76
4.2.	Prediction of antibiotic resistance proteins .....	77
4.2.1.	Selection of ARPs and non-ARPs.....	78
4.2.2.	Prediction performance .....	79
4.2.3.	Prediction of novel ARPs.....	80
4.2.4.	Scanning bacteria genomes.....	81
4.2.5.	Contribution of feature properties to the classification of ARPs.....	82
4.2.6.	Server for antibiotic resistance protein identification (SARPI).....	82
4.3.	Prediction of cancer associated proteins .....	84
4.3.1.	Data preparation .....	84
4.3.2.	Overall prediction accuracies and performance evaluation .....	85
4.3.3.	Contribution of feature properties to the classification of cancer associated proteins .....	86
4.3.4.	Analysis of individual feature contribution by feature selection.....	87
4.3.5.	Cancer associated protein identification server (CAPIS) .....	88
4.4.	Comparison with other statistical learning methods.....	90
4.5.	Summary .....	91
5.	Prediction of microRNAs by machine learning methods .....	93
5.1.	Data preparation.....	93
5.1.1.	Retrieval of precursor miRNAs and non-precursor miRNAs.....	93
5.1.2.	Retrieval of mature miRNAs and non-mature miRNAs.....	94
5.2.	Evaluation and discussion.....	95
5.2.1.	Prediction performance for precursor miRNAs and mature miRNAs.....	95
5.2.2.	Screening non-coding RNAs within four representative genomes .....	97
5.2.3.	Comparison with other statistical learning methods.....	97
5.3.	MiRNA prediction server .....	99
5.3.1.	Comparison with other microRNA prediction servers.....	99
5.4.	Summary .....	104
6.	Conclusion and future work.....	105
6.1.	Major findings.....	105
6.2.	Limitation of methods applied in this work.....	108
6.3.	Future studies.....	109
	BIBLIOGRAPHY .....	110
	APPENDICES .....	123
	LIST OF PUBLICATIONS .....	157

## SUMMARY

Proteins and functional RNAs are important components of biological organisms, which play essential roles in biological systems. Therefore, the identification of functional proteins and RNAs is of great importance for understanding biological processes, discovering new therapeutic targets, and accelerating drug development. This thesis describes my work of applying machine learning methods to facilitate the identification of multifunctional enzymes, disease related proteins and microRNAs.

Multifunctional enzymes (MFEs) are enzymes that perform multiple catalytic activities. The identification and characterization of MFEs would provide valuable insights into molecular mechanisms underlying the crosstalk between different cellular processes. In this study, a total number of 3120 experimentally verified MFEs were collected from various sources. A support vector machine (SVM) based classifier was then developed to distinguish MFEs from non-MFEs. The classifier was also applied to search against ExPASy ENZYME database to identify potential novel MFEs. Moreover, we also investigated the mechanism of multiple catalytic properties, as well as their evolutionary basis. Our results suggest that MFEs are non-evenly distributed in different species, but no solid evidence suggests complex life forms like human prefer more MFEs than simple life form like yeast. Further KEGG ontology (KO) analysis indicated that MFEs most likely evolve from ancestor enzymes in primitive life forms. From structural perspective, the alpha and beta fold topology seems to be most favored for MFEs. The analysis of physiochemical properties indicated that four properties, including charge, polarizability, hydrophobicity, and solvent accessibility, are most important for the characterization of MFEs.

Another objective of this work is to identify disease related proteins which hold promise for discovering new therapeutic targets. Three groups of disease related proteins were studied, including antimicrobial proteins, antibiotic resistance proteins and cancer associated proteins. Corresponding SVM based prediction systems were developed to identify these proteins based on their primary sequences. Independent data sets that were not included in model development were then used to evaluate the performance of classification system, showing that prediction accuracies for members and non-members of these disease related proteins are in the range of 81.8%~97.5% and 99.2%~99.9% respectively. In addition, most of non-homologous antimicrobial proteins and antibiotic resistances were correctly predicted. These results suggest the usefulness of SVM method for facilitating the identification of disease related proteins, especially for non-homologous functional proteins.

The other objective of this work is to identify microRNAs (miRNAs) from sequence derived physicochemical properties by four machine learning methods, including decision trees (DT), k-nearest neighbors (KNN), probabilistic neural networks (PNN), and support vector machines (SVM). SVM was found to reach the best performance, with prediction accuracies of precursor miRNAs and mature miRNAs at 92.2% and 94.8%, and the accuracies for non-precursors miRNAs and non-matures miRNAs at 98.4 and 99.5% respectively. Screening non-coding RNA sequences within four representative genomes, including *Homo sapiens*, *Mus musculus*, *Drosophila melanogaster* and *Saccharomyces cerevisiae*, identifies 2.2%~5.6% of non-coding RNAs as potential precursor miRNAs, which contains fewer false positives than previous studies. These findings indicate that our prediction system is capable of

identifying miRNAs with relatively high accuracy. Similar strategy can be ideally applied to the prediction of other functional RNA classes.

Beyond in-house prediction models, we also developed a series of online prediction tools to serve scientific community to identify novel functional proteins and RNAs.

Our prediction systems could be accessed at following links.

SIME            <http://jing.cz3.nus.edu.sg/cgi-bin/sime.cgi>

SAPI            <http://jing.cz3.nus.edu.sg/cgi-bin/sapi.cgi>

SARPI           <http://jing.cz3.nus.edu.sg/cgi-bin/sarpi.cgi>

CAPIS           <http://jing.cz3.nus.edu.sg/cgi-bin/capis.cgi>

MiRDetector    <http://ang.cse.nus.edu.sg/cgi-bin/mirna/mirna.cgi>

## LIST OF TABLES

Table 2-1 Example of training data for decision tree .....	32
Table 2-2 Division of amino acids into 3 different groups by different physicochemical properties.....	37
Table 2-3 List of features for proteins .....	37
Table 2-4 Characteristic descriptors of cellular tumor antigen p53 (Swiss-Prot AC P04637). The feature vector of this protein is constructed by combining all of the descriptors in sequential order. ....	38
Table 2-5 Division of nucleotides into different groups for different physicochemical properties.....	39
Table 2-6 List of features for miRNA.....	40
Table 2-7 Example of computed descriptors of miRNA precursor (cel-mir-243). The feature vector of this precursor is constructed by combining all the descriptors in sequential order. ....	40
Table 3-1 Statistics of the datasets and prediction accuracy of individual class of MFE and that of all MFEs (6=21).....	42
Table 3-2 Distribution of known and predicted enzymes of multiple catalytic domains in different kingdoms and in top 20 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 databases. ....	52
Table 3-3 Distribution of known and predicted enzymes with single multi-catalytic domain in different kingdoms and in top 20 host species.....	53
Table 3-4 Orthologs of multifunctional enzymes (MFEs) in <i>S. cerevisiae</i> and <i>H. sapiens</i> species. 36.7% (22 out of 60) MFEs in <i>H. sapiens</i> had their orthologs in <i>S. cerevisiae</i> , while 56.8% (21 out of 37) MFEs in <i>S. cerevisiae</i> had their orthologs in <i>H. sapiens</i> . ....	55
Table 4-1 Distribution of AMPs in top 10 host species.....	67
Table 4-2 Statistics of the datasets and prediction accuracy of individual class of AMPs The predicted results are given in TP, FN, TN, FP, sensitivity $SE=TP/(TP+FN)$ , specificity $SP=TN/(TN+FP)$ , positive prediction value $PPV=TP/(TP+FP)$ and overall accuracy $Q=(TN+TP)/(TP+FN+TN+FP)$ . The number of members and non-members in the testing and independent evaluation sets is TP+FN or TN+FP respectively. ....	67



Table 4-3 Statistics of prediction accuracy of antimicrobial proteins measured by 5-fold cross validation.....	69
Table 4-4 Prediction results of novel antimicrobial proteins by SVM-Prot, where “+” represents proteins correctly predicted as antimicrobial proteins, and “-” represents proteins incorrectly predicted as non-antimicrobial proteins. ..	70
Table 4-5 List of prediction results of 177 antimicrobial proteins in AMPer database (“+” represents proteins correctly predicted as antimicrobial proteins, and “-” represents proteins incorrectly predicted as non-antimicrobial proteins) .....	72
Table 4-7 Distribution of ARPs in top 10 bacteria species.....	79
Table 4-8 Statistics of the datasets and prediction accuracy of ARPs ( $\sigma=18$ ).....	79
Table 4-9 Statistics of accuracy for SVM prediction of antibiotic resistance proteins evaluated by using 10-fold cross validation.....	80
Table 4-10 Prediction results of novel ARPs.....	81
Table 4-11 Statistics of datasets and prediction accuracy of cancer associated proteins .....	84
Table 4-12 Distribution of cancer associated proteins in top 10 bacteria species .....	85
Table 4-13 Features important for characterizing cancer associated proteins as selected by recursive feature elimination method.....	87
Table 4-14 Comparison of prediction performance of all AMPs and non-AMPs with different machine learning methods.....	91
Table 4-15 Comparison of prediction performance of antibiotic resistances and non-antibiotic resistances with different machine learning methods.....	91
Table 4-16 Comparison of prediction performance of all CAPs and non-CAPs with different machine learning methods.....	91
Table 5-1 Distribution of precursor miRNAs in top 10 host species.....	94
Table 5-2 Statistics of the datasets and prediction accuracy for precursor miRNAs and mature miRNAs .....	95
Table 5-3 Location of predicted and validated rhesus miRNAs within putative precursor sequences. Sequences in <i>italic</i> denote those predicted by MiRDetector while those with <u>underline</u> denote experimentally validated miRNAs. ....	96
Table 5-5 Screening results of non-coding RNAs from four representative genomes	97

Table 5-6 Comparison of prediction performance of precursor miRNAs and non-precursor miRNAs with different machine learning methods.....	98
Table 5-7 Comparison of prediction performance of mature miRNAs and non-mature miRNAs with different machine learning methods .....	98
S1 Scanning results of <i>E. coli K12</i> genome (# indicates that data were not included in our model development) .....	123
S2 Scanning results of <i>S. aureus Mu50</i> genome (*indicates functional classification by SVMProt followed by probability of correct characterization P-value, while # indicates the data are not included in our model data set) .....	134
S3 Prediction result of potential precursor miRNAs (“+” and “-” indicates that the RNA is predicted as precursor miRNA and non-precursor miRNA, respectively).....	144

## LIST OF FIGURES

Figure 1-1 MiRNA biosynthesis. MiRNA is produced from precursor microRNA (pre-miRNA), which in turn is formed from a miRNA primary transcript (pri-miRNA). .....	11
Figure 2-1 Architecture of support vector machines .....	21
Figure 2-2 Different hyperplanes could be used to separate examples .....	22
Figure 2-3 Mapping input space to feature space .....	24
Figure 2-4 Schematic diagrams illustrating the process of the training and prediction of the functional class of proteins by using SVM. Sequence-derived feature $h_i$ , $p_i$ , $v_i$ ... represents such structural and physicochemical properties as hydrophobicity, polarizability, and volume. Feature $d_i$ , $s_i$ , $m_i$ , ..., represents properties such as domain information, subcellular localization, and post-translational (PT) modification profiles etc.....	26
Figure 2-5 Example of k-nearest neighbors (squares and triangles represent training samples and the star symbol indicates an unknown sample).....	27
Figure 2-6 Architecture of a simple three-layer neural network.....	30
Figure 2-7 Example of a decision tree classifier.....	31
Figure 2-8 The sequence of a hypothetical protein for illustration of derivation of the feature vector* .....	38
Figure 3-1 Top 10 Pfam families for known enzymes of single multi-catalytic domain (SMAD-MFEs). It is noted that about 38% of SMAD-MFEs contain ArgJ domain, and majority of them are involved in Urea cycle and metabolism of amino groups pathway (amino acid metabolism map00220).....	44
Figure 3-2 Top 10 Pfam families of known enzymes of multiple catalytic domains (MCD-MFEs).....	44
Figure 3-3 Distribution of known and predicted putative MFEs (enzymes of single multi-catalytic domain SMAD-MFEs, enzymes of multiple catalytic domains MCD-MFEs) in SCOP fold families. It is noted that 42% of MCD-MFEs and 69% of SMAD-MFEs belong to the alpha and beta fold class (a/b). .....	45
Figure 3-4 Statistics of known MFEs according to the number of biological pathways they anticipated in. Totally 1,293 known enzymes of multiple catalytic domains (MCD-MFEs) and 285 known enzymes of single multi-catalytic domain (SMAD-MFEs) were employed in this study. ....	48

Figure 3-5 Statistics of known and predicted enzymes of multiple catalytic domains (MCD-MFEs) with KEGG ontology (KO). MCD-MFEs are involved in 4 level one, 17 level two, and 74 level three pathways. Majority of them anticipate in carbohydrate metabolism (CAR), lipid metabolism (LIP), nucleotide metabolism (NUC), amino acid metabolism (AAC) and metabolism of cofactors and vitamins (COF). Number with “*” denotes the number of predicted MCD-MFEs. ....	49
Figure 3-6 Statistics of known enzymes of single multi-catalytic domains (SMAD-MFEs) in KEGG ontology (KO). SMAD-MFEs are involved in 3 level one, 10 level two and 52 level three pathways. Majority of them anticipate in the carbohydrate metabolism (CAR), amino acid metabolism (AAC) and metabolism of cofactors and vitamins (COF). Number with “*” denotes the number of predicted SMAD-MFEs. ....	50
Figure 3-7 Distribution of MFEs in different kingdoms. Totally, 2,551 known enzymes of multiple catalytic domains (MCD-MFEs), 4,075 predicted MCD-MFEs, 537 known enzymes of single multi-catalytic domain (SMAD-MFEs), and 245 predicted SMAD-MFEs were included in the statistics. It is noted the dominance of bacteria in both known and predicted MCD-MFEs and SMAD-MFEs in total enzyme number. ....	51
Figure 3-8 Statistics of currently known MFEs and predicted MFEs by screening the ExPASy Enzyme database. Totally there are 3,120 currently known MFEs, including 2,279 enzymes of multiple catalytic domains (MCD-MFEs), 572 known enzymes of single multi-catalytic domain (SMAD-MFEs). Totally, 2,641 novel MFEs with prediction probability >50% (4,320 with probability >80%), including 2,515 MCD-MFEs (4,075 with probability >80%) and 126 SMAD-MFEs (245 with probability >80%) were identified from 91,140 enzymes of ExPASy Enzyme database. ....	57
Figure 3-9 SIME interface. The sequence of a protein, in RAW format and containing no non-amino acid letters, can be input in a window provided. ....	59
Figure 3-10a Result page of SIME showing that a query sequence is predicted as a multifunctional enzyme with multiple catalytic domain .....	60
Figure 3-10b Result page of SIME showing that a query sequence is predicted as a multifunctional enzyme with single catalytic domain .....	60
Figure 3-10c Result page of SIME showing that a query sequence is predicted as non multifunctional enzyme .....	61
Figure 3-11 Graphical searching interface of MFEs database. ....	62
Figure 3-12 Graphical user interface of MFEs database. ....	62
Figure 3-13 Graphical searching interface of MFEs database. ....	63
Figure 3-14 Biological analysis results interface of MFEs. ....	63

Figure 4-1 Graphical user interface for SAPI .....	77
Figure 4-2 Result page of SAPI showing that a query sequence is an antimicrobial protein. ....	77
Figure 4-3 Interface for SARPI.....	82
Figure 4-4 Result page of SARPI showing that the query sequence is not antibiotic resistance protein .....	83
Figure 4-5 CAPIS interface. The sequence of a protein, in RAW format and containing no non-amino acid letters, can be input in a window provided. ....	89
Figure 4-6 Result page of CAPIS showing that the query sequence is a proto-oncogene. ....	89
Figure 5-1 Graphical user interface of MiRDetector. The sequence of a query sequence, in RAW format and containing non-AU(T)GC characters, can be input in a window provided.....	102
Figure 5-2 Result page of MiRDetector showing that a query sequence is a potential precursor miRNA.....	103
Figure 5-3 Result page of MiRDetector showing the location of the predicted mature miRNA within the precursor.....	103

## LIST OF ACRONYMS

AMP	Antimicrobial Protein
ARP	Antibiotic Resistance Protein
CAP	Cancer Associate Protein
CAPIS	Cancer Associated Protein Identification Server
DT	Decision Tree
FN	False Negative
FP	False Positive
IHA	Inter-base hydrogen bonds donor
IHD	Inter-base hydrogen bonds donor
KNN	K-Nearest Neighbors
MCC	Matthews correlation coefficient
MCD-MFEs	MFEs with multiple catalytic domains
MFE	Multifunctional Enzyme
MFP	Multifunctional Proteins
MiRDetector	MicroRNA Detector
MicroRNA	miRNA
ncRNAs	non-coding RNAs
NMFEP	non-MFE proteins
NN	Neural Networks
ORFs	Open Reading Frames
PNN	Probabilistic Neural Network
PSI-BLAST	Position Specific Iterative-Basic Local Alignment Search Tool
QP	Quadratic Programming

RFE	Recursive Feature Elimination
rRNA	ribosomal RNA
SAPI	Server for Antimicrobial Protein Identification
SARPI	Server for Antibiotic Resistance Protein Identification
SIME	Server for Identification of Multifunctional Enzyme
SMAD-MFEs	MFEs with single multi-activity domain
SVM	Support Vector Machine
TN	True Negative
TP	True Positive
tRNA	transfer RNA

# 1. Introduction

Proteins are important components of biological systems and essential to any life form. They participate in almost every biological process, such as catalyzing chemical reactions, providing structure rigidity to cells, and transmitting signals and nutrients. A number of proteins are involved in different disease related pathways, and dysfunction of these proteins accounts for most of human diseases. For example, over expression of oncogenes would cause cancers, while mutations in antimicrobial proteins may reduce their capacity to defend against microbial infection. Therefore, identification of these proteins and understanding of their mechanisms would be of great importance to discover novel therapeutic targets and develop new drugs to treat diseases.

Besides proteins, RNAs are also well recognized as important components of biological systems. According to central dogma of molecular biology, RNAs are responsible to transcribe gene information storing in DNA, and then translate them into protein sequences. However, since the late 1990s, a number of non-coding RNAs have been identified by experimental or computational methods. They are not to be translated into proteins; instead, their role in biological systems remains at the RNA level. In particular, a group of smallest non-coding RNAs, called microRNAs (miRNAs), have attracted intensive interests. It is estimated that one third of human genes are regulated by miRNAs, which open a new door to controlling the expression of desirable genes, and may profoundly influence current drug discovery process.

Since the sequencing of phage  $\phi$ X174 in 1977, a tremendous amount of genomic information of organisms have been decoded and deposited into varieties of database.



Up to April 2008, more than 360,000 proteins have been collected in a curated protein database, Swiss-Prot, and the number is continuing increasing rapidly. On the other hand, however, low and non-homologous proteins with unknown function constitute a substantial part (up to 20%~100%) in Open Reading Frames (ORFs), in many newly sequenced genomes. Although wet-lab experiments are still the most effective methods to determine functions of proteins and RNAs, they are, however, still costly and time consuming for annotating such tremendous amount of data. Therefore, there is a need to explore other methods including computational approach for facilitating the identification of protein and RNA function to complement web-lab experimental methods.

In this thesis, I will introduce my work on the application of machine learning to the prediction of multifunctional enzymes, disease related proteins, and miRNAs.

### 1.1. Introduction to multifunctional enzymes (MFEs)

It has been noticed for a long time that some enzymes are able to perform multiple functions [1-4], which are called multifunctional enzymes (MFEs). An increasing number of such enzymes are being discovered in recent years. MFEs are found to be beneficial to living systems and provide competitive survival edges in a variety of ways. They are able to employ alternative approaches to coordinating multiple activities and regulate their own expression [1], which demonstrates evolutionary advantage as part of a clever strategy for generating complexity from existing proteins without expansion of the genome [3, 5, 6]. Combination of multiple functions enables an enzyme to act as a switch point in biochemical or signaling pathways so that a cell can rapidly respond to changes in surrounding environment [7]. Multifunctionality

seems to be a common mechanism of communication and cooperation between many different functions and pathways within a complex cellular system or between cells [2].

Identification of MFEs and subsequent investigation of their mechanistic and structural basis of multifunctionality is important for studying biological roles of enzymes [3, 7] and for the exploration of multiple activities in protein engineering [8] and inhibitor design [9]. Studies of sequences, structures and components of MFEs have demonstrated that useful information can be derived for facilitating the understanding of the mechanism of actions [10], organizational and evolutionary features [11], and assembly patterns [12] of MFEs. In-depth study on comprehensive collection of MFEs is expected to provide a more complete picture about the functional, evolutionary, and structural features of multifunctional enzymes.

A recent study indicates that current sequence analysis algorithms (alignment, clustering and motif approaches) are capable of disclosing individual functions of MFEs [13]. Algorithms based on remote homology, like PSI-BLAST (Position Specific iterative-Basic Local Alignment Search Tool) [14], have been found to give good performance for finding alternative functions of MFEs [13]. However, in some cases, it is difficult to determine whether the predicted multiple functions by these methods are due to true multifunctionality or false identification [2-4]. Thus it is highly desirable to develop a method to determine the multifunctionality of proteins. MFEs have certain common structural and physicochemical characteristics in spite of the diversity of their sequences and structures, which can be potentially exploited for determining whether enzymes are multifunctional or not. Active sites of enzymes with

multiple catalytic activities are inherently reactive environments packed with nucleophiles, electrophiles, acids, bases and cofactors [3]. Special structural features are present in some MFEs to enable them to bind to different substrates [3]. The surface of some MFEs allows the formation of complexes with different proteins or substrates at different cellular environments [2, 7].

Proteins of multiple functions are known to have high sequence and structure diversity but none-the-less possess common structural and physicochemical features to perform common functions. Such characteristics make it difficult to identify MFEs by homology-based approaches. Thus it is desirable to explore other methods to identify MFEs.

## 1.2. Introduction to disease related proteins

### 1.2.1. Antimicrobial proteins

Microbes, such as bacteria, viruses and fungi, are responsible for a number of human or other organisms' diseases, such as acute bacterial meningitis [15], human immunodeficiency virus (HIV) [16] and latent tuberculosis infection [17]. On the other hand, host organisms have also developed a variety of sophisticated mechanisms to fight against the invasion of microbes, among which antimicrobial peptides play an important role. Antimicrobial peptides are able to induce both innate and adaptive immune responses in host organisms [18, 19]. They usually take effects by insertion into microbial membrane to either disrupt the physical integrity of the bilayer or translocate across the membrane and act on internal targets [18]. Due to their broad-spectrum antimicrobial properties, antimicrobial peptides are increasingly used

as molecular therapies [19]. A number of databases have also been developed to collect and characterize antimicrobial peptides [20-23].

Antimicrobial peptides are derived from antimicrobial proteins (AMPs) upon bacterial attack [24, 25]. Therefore knowledge of AMPs would be helpful to identify novel therapeutic targets and invent new antimicrobial agents to treat diseases caused by bacteria. The characterization of AMPs to date mainly relies on kinds of experimental approaches such as NMR [26], electron microscopy [27], and fluorescent dyes [28]. However, many of them generally require a purified or semi-purified target of interest, and usually time consuming, which limit their application to identify antimicrobial peptides in large scale [29]. Therefore, alternative approaches including computational methods would be helpful to the identification of AMPs.

### 1.2.2. Antibiotic resistance proteins

Antibiotics are believed to be one of the greatest medical inventions in the 20<sup>th</sup> century, which have significantly extended human life expectancy by 10 years [30, 31]. Antibiotics have been widely used to treat various diseases caused by bacteria, such as tuberculosis, pneumonia and leprosy, which were lethal diseases before the invention of antibiotics. Antibiotics take effect through inhibiting or killing bacteria while causing little or no harm to the host. Various mechanisms are used by antibiotics to achieve this selective effect. For instance, some antibiotics are able to inhibit the synthesis of key proteins that play critical roles in bacterial growth and proliferation [32], whilst others may disrupt bacterial membrane structure and result in bacterial death [33].

However, the widespread usage of antibiotics also applies selective pressure on bacteria [34]. Antibiotic resistance began to emerge almost as soon as the first clinical use of penicillin. The emergence of highly virulent and multi-drug resistant bacterial strains has presented a serious challenge to traditional therapies of infectious diseases [35]. Antibiotic resistance accounts for a number of treatment failures, and it could be fatal to those critically sick patients who rely on antibiotics to fight against bacteria [34]. To make the situation even worse, resistant bacteria could spread widely, posing more serious problems for infection control [36].

Antibiotic resistance is a consequence of natural selection or programmed evolution. Multiple mechanisms contribute to antibiotic resistance, such as drug modification by enzymatic mechanisms, mutation of drug targets, enhanced efflux pump expression, and altered membrane permeability [36]. A number of proteins have been found responsible for antibiotic resistance. For instance, many multi-drug resistance efflux systems can pump out antibiotics from the cell surface by a collection of membrane associated proteins [37]. Specific mutations in antibiotic targets may hinder the binding and thus the effectiveness of certain antibiotics [38, 39]. In addition, resistance determinants borne on plasmids, bacteriophages, transposons and other mobile genetic elements can be transferred to naive recipients [36, 40]. Therefore, antibiotic resistance proteins may come from different sources which diversify from DNA gyrase, topoisomerase, to mutated enzymes, or gene duplication and over-expression of certain carrier proteins.

Recognizing these proteins is critically important to study the evolution of antibiotic resistance, which will facilitate the design of novel drugs to control potential spread of

antibiotic resistance [40]. As part of the efforts for understanding and identifying these proteins, two antibiotics resistance protein databases, ARGO [41] and MvirDB [42], have been developed to collect and characterize ARPs. Various experimental methods have been explored for the identification of antibiotic resistance proteins (ARPs) [43-46].

However, these methods are usually costly, time consuming, and resource intensive, which is a particular problem because of the fluidity of the microbial genomes can further increase the burden. Therefore, it would be helpful to explore alternative methods including computational approach to identify ARPs.

### 1.2.3. Cancer associated proteins

Cancer is the second leading cause of death in western world, just slightly inferior to cardiovascular diseases. Intense efforts have been devoted to the study of cancer genesis, progression, and therapeutic implication. Normal growth-control mechanisms have no effect on cancer cells. Cancer refers to a group of diseases. Cancer cells, unlike normal cells that respond to growth control mechanism, are capable of growing indefinitely and will invade healthy tissue nearby. Moreover, cancer cells can also migrate and proliferate in other places through metastasis, which accounts for 90% of human cancer deaths.

The induction of cancer involves accumulation of multiple genetic alternations. A wide variety of chemical agents and physical agents can cause mutations in normal cells and induce malignant transformation which leads to final development to cancer. For instance, extensive exposure to UV radiation may lead to the mutation and

inactivation of p53 [47, 48], which plays important roles to suppress tumor. Another important cause of tumor is induced by DNA or RNA viruses, which may integrate their genomes into host chromosomes and result in malignant transformation in virus-infected cells. HIV-1 [49] could reverse transcribe their RNA into DNA and integrate to human genome, which may lead to malignant transformation.

Within a normal tissue, cellular proliferation and cell death is carefully regulated by a number of signals. A number of genes responsible for the malignant transformation have been identified in the past three decades [50]. The growth and death of normal cells are sophisticatedly maintained by two categories of cancer related genes: proto-oncogenes and tumor suppressors. Proto-oncogenes are normal genes whose mutations, called oncogenes, code for proteins causing cancer [51-53]. Proto-oncogenes are converted to oncogenes by mutations or genetic rearrangement. Some oncogenes are responsible for the over production of growth factor leading to uncontrolled cell growth. Some other oncogenes perturb parts of the signal cascade [54]. On the other hand, tumor suppressors are responsible for regulating cell proliferation or initiating apoptosis of cells, which reduce the possibility that a cell developing to a tumor cell [55, 56]. For example, the inactivation of mutated retinoblastoma gene results in unregulated tumor proliferation.

Identification of cancer associated proteins will facilitate efforts to understand the mechanism of cancer development and therefore helpful to discover novel pharmaceutical agents and therapeutic targets to fight against cancer. The characterization of cancer-related proteins to date mainly relies on kinds of experimental approaches, like molecular cloning [57]. RB is the first tumor

suppressor gene isolated from human genome in 1986 [57]. Therefore, it would be helpful to explore computational method to finding those proteins.

### 1.3. Introduction to microRNAs

Non-coding genes function without being translated into protein products; instead, their products function at RNA level. For many years, it was believed that there are only a few non-coding RNAs (ncRNAs), such as transfer RNA (tRNA) and ribosomal RNA (rRNA), both of which are involved in the process of translation and gene expression [58]. However, since the late 1990s, a number of new non-coding RNAs have been found to participate in various regulatory events, which open a new door to investigate gene regulatory networks.

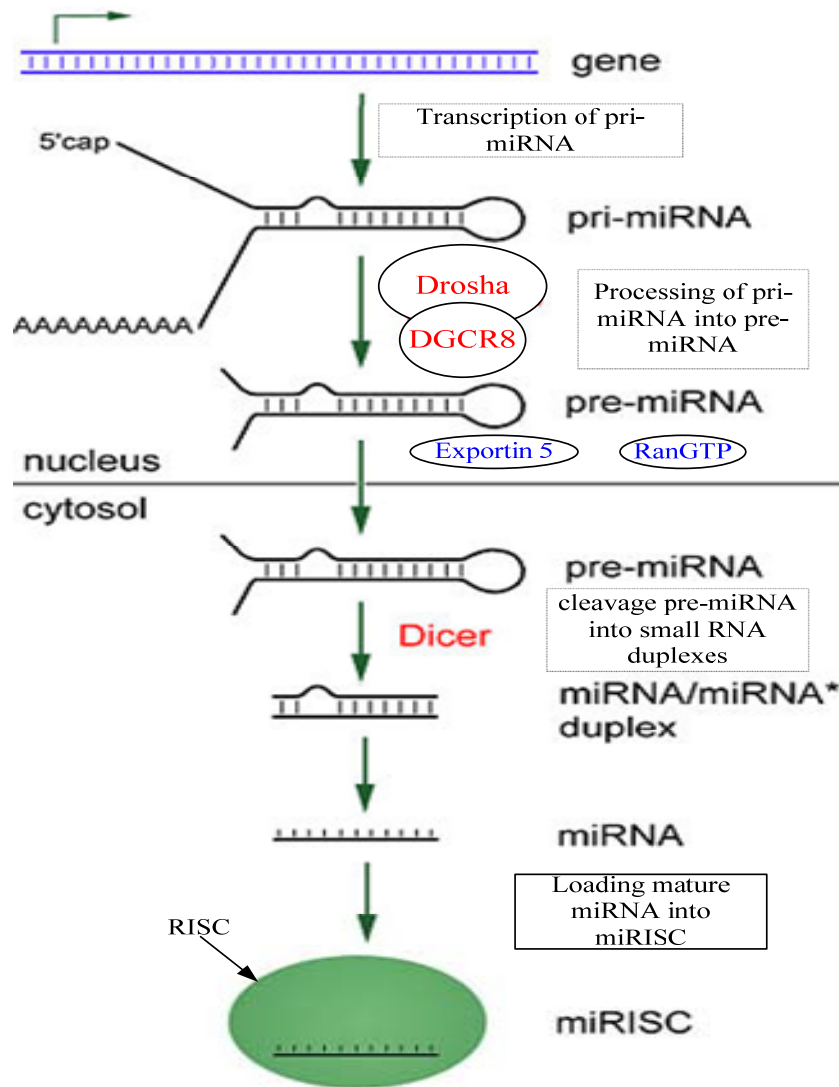
MicroRNAs (miRNAs) are a group of smallest functional ncRNAs that regulate gene expression. Since the discovery of the first miRNA in 1993 [59], miRNAs have been attracting more and more scientists' interest. MiRNA genes could be located in intergenic regions or in introns; some of them are found to be clustered [60]. Many miRNAs have heterogeneous expression profiles in different tissues, which also could be used as potential cancer markers [61-63]. The majority of miRNAs are 21 to 25 nucleotides (nt) in length [64], with 21nt long on average. Many miRNAs are both sequence and structure conserved in evolution [65]. Mature miRNAs are derived from miRNA precursors (pri-miRNAs), which are about 70-100nt long and have an imperfect stem-loop structure with one or two miRNAs in the arms [66, 67]. Figure 1-1 shows the biosynthesis of miRNAs in humans. MiRNAs are first transcribed as primary transcripts (pri-miRNAs) with a cap and poly-A tail by RNA polymerase II [68]. Pri-miRNAs are then processed into precursor miRNAs (pre-miRNAs) by



microprocessor complex, which is comprised of Drosha [69] and DGCR8 [70]. After that pre-miRNAs are transported from nucleus to cytoplasm by another complex that consists of exportin 5 and RanGTP [71]. In cytoplasm, pre-miRNAs are released and processed by Dicer into short double-stranded RNAs [72]. One segment called mature miRNA is integrated into the RNA-induced silencing complex (RISC) [73, 74]. This complex is responsible for the gene silencing observed due to miRNA expression and RNA interference [75, 76].

MiRNAs play important roles in gene regulation at post-transcription level. It is estimated that approximately one third of protein coding genes are regulated by miRNAs [77]. MiRNAs are involved in surprisingly diverse of biological processes and they are responsible for a number of human diseases [78, 79]. The exact mechanisms of gene regulation by miRNAs remain to be discovered. Evidence shows that miRNA could degrade the target transcript, or inhibit protein translation [64]. MiRNAs are able to negatively regulate their targets through sequence-specific-pairing approach [80]. MiRNAs could bind to mRNA targets at on 3'-UTRs and repress translation and mediate degradation [72]. The regulation mechanism of miRNAs in plants and animals are different. Most plants miRNAs could bind almost perfectly to their target mRNAs, and their binding sites are not limited to the 3' untranslated region (3' UTR), but could be throughout the whole genome [81]. In contrast, the pairing of animal miRNAs to their targets 3'UTR is imperfect.

**Figure 1-1 MiRNA biosynthesis.** MiRNA is produced from precursor microRNA (pre-miRNA), which in turn is formed from a miRNA primary transcript (pri-miRNA).



The number of miRNAs in a vertebrate genome is estimated to be about 800-1000 [82, 83], and approximately 0.5-1.5% of human genes are estimated to encode miRNAs [84]. Efforts have been devoted to collect and annotate miRNAs [85, 86] through various approaches. A number of experimental methods have been developed to identify and characterize miRNAs [87-90]. However, these methods are usually costly,

time consuming, and resource intensive [91, 92]. The short sequence, redundancy, and heterogeneous expression profiles make miRNA discovery even more difficult [92, 93]. Numerous computational methods are also developed to facilitate the identification of miRNAs in different genomes, including sequence alignment [94, 95], structure based approach [96] and conservation based approach [97, 98]. One statistical learning method, support vector machine (SVM), has also been applied to identify new miRNA candidates [93, 99, 100]. However, these methods usually produce too many false positives when applied to large genomes. Thus prediction of miRNAs with lower false positive rate is still a challenging task.

## 1.4. Overview of computational methods for biological function prediction

### 1.4.1. Sequence similarity method

Sequence similarity method (also named sequence alignment method) is the most popular method used in protein or RNA function prediction. The underlying assumption behind sequence similarity method is that similar sequence implies similar structure, and then similar function, which is satisfied in most of cases.

Modern sequence alignment methods begin with the global homology algorithm of Needleman-Wunsch [101], which uses an iterative matrix method for optimizing the alignment between two sequences. Since then, more rigorous methods, such as Sankoff alignment (1972) [102] and Reicher alignment (1973) [103], started to emerge, although their biological implication was difficult to formulate. Later on, Smith and Waterman developed a local sequence alignment method [104], which only

searched relatively conserved subsequence, so one single sequence may yield more than one subsequence and only these conserved sequences could contribute to the score of alignment. Although this method was more useful for searching sequences in databases, it was still quite time consuming, and had to be used in supercomputers when large databases need to be searched. In order to solve this problem, heuristic algorithms were proposed. One of first tries is FASTA program developed by Lipman and Pearson[105], which aims to identify local similar regions between two sequences using PAM matrix. The strategy significantly decreased the computation time for comparison. In 1990, a breakthrough sequence comparison method, Basic Local Alignment Search Tool (BLAST), was developed [106]. At that time, BLAST was significantly faster than any other sequence alignment tools while maintaining comparable sensitivity. It balances accuracy and computation speed. After that, Gapped BLAST [14] was developed to generate gapped alignments, with approximately three times as fast as BLAST search. Meanwhile, Position Specific Iterated BLAST (PSI-BLAST) [107] allows BLAST search to iterate, which is particular useful to identify remote homologous proteins.

Although sequence alignment methods have good performance in sequence analysis, they still have some limitations. Some proteins are so unique that it is difficult to find their “neighbors” in existing protein databases. Moreover, no all the similar proteins have analogous functions [108]. So there is a need to find other methods to assign protein function beyond sequence alignment.

#### 1.4.2. Motif based methods

Many proteins or RNAs are found to share consensus sequences or motif, which may provide important clue for their function prediction [109]. Motif based methods, such as Motifs, Prosite [109] and Sequence Clustering [110], have been developed to detect common motifs among proteins and RNAs. Motif databases, such as PROSITE, ProDom and Rfam, are also widely used in sequence analysis.

PROSITE [111] consists of a large collection of biologically significant signature patterns that were manually annotated and used to determine the function of a given protein. The first release of PROSITE was published in 1992, which contained 397 entries describing 433 different patterns. The number of patterns in the database has increased to 1318 in April 2008. The problem with PROSITE is that those patterns are usually too short, which may result in too many false positives of unrelated sequences.

In order to address this problem, structurally defined regions, called domains, are used to characterize parts of protein sequences with well defined functions. ProDom and Pfam are two examples of this kind of databases. ProDom [112] is a comprehensive database of protein domain families generated from the Swiss-Prot database by automated sequence comparisons. It can be used for analyzing domain arrangements of complex protein families and protein homology relationships. Similarly, Pfam [113, 114] database currently covers a large collection of manually curated protein domain families. Each family is represented by two multiple sequence alignments, two profile-Hidden Markov Models (profile-HMMs) and an annotation file [114]. It can automatically classify query proteins into protein domain families [115]. Pfam database current covers 9,318 families in April 2008.

Although there are so many motif databases which contain a large amount of patterns and domain information, not all newly sequenced proteins or RNAs could be covered by these databases. If a new sequence does not have any domain defined in current domain databases, its function could not be identified. So it is desirable to explore alternative methods to predict protein function besides motif based method.

#### 1.4.3. Machine learning approach

Unlike sequence similarity approach and motif based approach, machine learning methods take a different strategy to predict protein function. Machine learning methods derive rules from common characteristics within proteins, and then apply these rules to justify unseen examples. Machine learning methods have been successfully applied to the identification of novel enzymes [116], bacterial proteins [117], lipid-binding proteins [118], transporters [119] and other protein functional classes [120, 121].

A number of challenges are still waiting to be solved, such as the generation of effective negative samples, ambiguous information in biological data and data imbalance issue.

### 1.5. Scope and objective

The objective of this study is to develop computational tools to facilitate the identification of multifunctional enzymes, disease related proteins and miRNAs from their primary sequences derived physicochemical properties. Machine learning methods were employed in the study. Computational tools are expected to offer an

alternative solution to the identification of functional proteins and non-coding RNAs, and help to accelerate the pace of drug development and discover new therapeutic targets.

The objective could be divided into three parts:

1. To develop a classification system for predicting MFEs directly from their primary sequences. Further analysis of their mechanism, evolution, species distribution need to be done.
2. To develop prediction systems for disease related proteins, including antimicrobial proteins, antibiotic resistance proteins, and cancer related proteins.
3. To apply machine learning methods to predict miRNAs.

In order to achieve the 3 parts of the objective described above, a machine learning method, support vector machine (SVM), is employed to develop these prediction systems. It is particular useful for the prediction of the proteins or miRNAs that are not homologous to those with known function, where traditional sequence similarity or motif based approach are likely to fail.

This thesis includes six chapters. Chapter 1 provides the introduction to multifunctional Enzymes, disease related proteins, microRNAs and current prediction methods for protein and RNA. Chapter 2 describes algorithms of different machine learning methods, as well the construction of feature vectors. The application of machine learning methods for the prediction of multifunctional Enzymes, disease related proteins and microRNAs are described in Chapter 3,

---

Chapter 4 and Chapter 5, respectively. Chapter 6 describes conclusion and future work.



## 2. Methods

In the chapter, algorithms of four well known machine learning methods will be introduced, which will be used to develop computational methods to predict functional proteins and RNAs. Moreover, feature selection and performance evaluation will also be illustrated. As most of machine learning methods could only accept numerical values instead of protein/RNA sequences, it is essential to convert them into numerical vectors before the application of machine learning. The method of feature vector construction will be covered in the last part of this chapter.

### 2.1. Machine learning methods

The term of machine learning refers to algorithms and techniques that allow computers to extract information from past experience. Although it emerges as a separate research field in the early 1980s, the study of machine learning can be traced from the 1960s [122]. Over the past 50 years, various machine learning methods have been developed and applied in a wide spectrum of fields, such as k-nearest neighbor algorithms in text categorization, decision tree methods in pharmaceutical research, artificial neural network in stock market analysis and prediction, support vector machine in bioinformatics and cheminformatics.

Machine learning uses computational and statistical methods to build mathematical models, and make inference from training samples [123]. Machine learning is a branch of artificial intelligence (AI), and it is closely related to statistics and pattern recognition, since they all study the analysis of data. However, unlike statistics and

pattern recognition, machine learning is primarily concerned with algorithmic complexity of computational implementations [124].

In order to be learnt by computational methods, all the samples, or instances, should be represented by feature vectors, which could be categorical, binary or continuous. Machine learning could be divided into two categories: if samples are given with known classes, it is called supervised learning; otherwise, it is called unsupervised learning [125]. In supervised learning, the learning process is to optimize an objective function and predict the value of the function for any valid input object after having learnt experience from training examples. This category includes well known machine learning methods like k-nearest neighbors, support vector machines, and decision trees. On the other hand, unsupervised learning is never given the answer set, and all the answers are assumed to be latent variable. All data under investigation are allowed to speak for themselves and they are treated evenly. This category includes self organization map and clustering methods.

In the following sections, four machine learning algorithms will be introduced, including support vector machine, k-nearest neighbors, neural networks, and decision tree. Their specific properties, advantages and disadvantages in real world problems, will also be discussed.

### 2.1.1. Support Vector Machine (SVM)

Support vector machine (SVM) is one of the newest members in supervised learning family [126]. It was first officially proposed by Vladimir Vapnik in 1995[126], and

then further explained by Dr. Burges in 1998[127]. A special property of SVM is that it simultaneously minimizes the empirical classification error and maximizes the geometric margin. Over the past 20 years, SVM has been successfully applied to a wide range of real-world problems, including hand-written digit recognition [128], tone recognition [129], image classification [130-133], as well as broad fields in biology, such as protein function prediction[134, 135], protein-protein interaction prediction [136], protein remote homology detection [137, 138], and classification for discriminating coronary heart disease patients[139]. SVM is the primary method used in our study. Therefore its theory and algorithm will be discussed with more details in following sections.

#### 2.1.1.1. Linear SVM

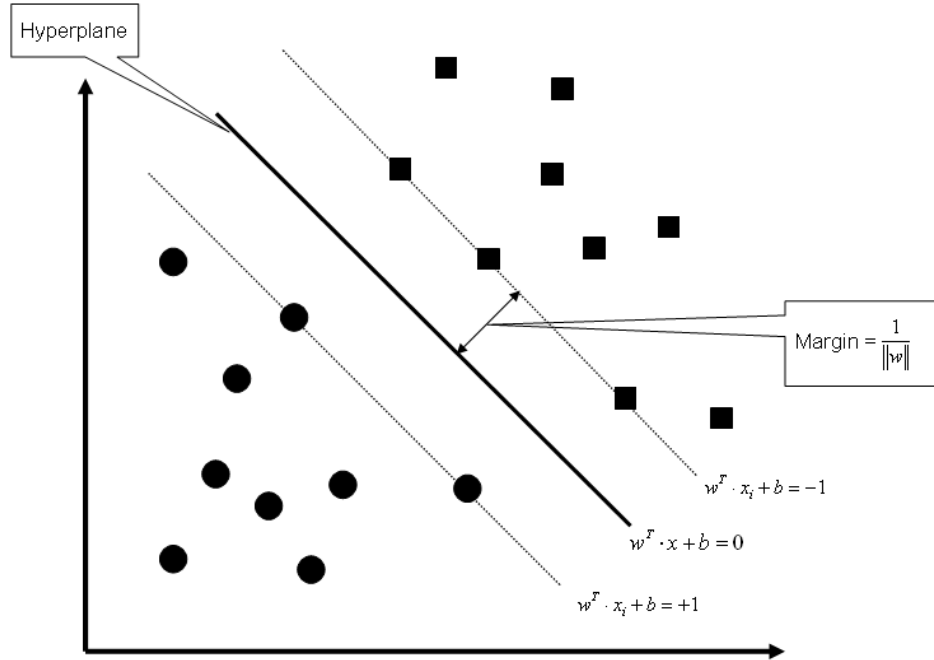
In two-class problems, SVM aims to separate examples of two classes with the maximum hyperplane (Figure 2-1). Mathematically, the data is composed of  $n$  examples of two classes, denoted as  $\mathcal{X} = \{(x_1, y_1), \dots, (x_n, y_n)\}$ , where  $x_i \in R^N$  is a vector in feature space and  $y_i \in \{-1, +1\}$  denotes its class. A hyperplane could be drawn to separate examples of one class (positive examples) from those of the other one (negative examples). The hyperplane is represented by  $w \cdot x + b = 0$ , where  $w$  is the slope and  $b$  is the bias. Thus the objective function of SVM changes to minimize Euclidean norm  $\|w\|^2$  with following limitations:

$$w \cdot x_i + b \geq +1 \quad \text{for } y_i = +1 \quad (\text{positive examples}) \quad (1)$$

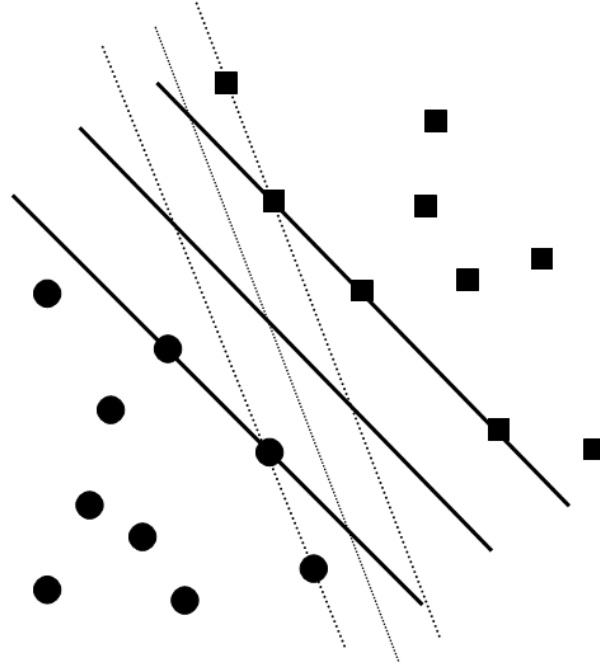
$$w \cdot x_i + b \leq -1 \quad \text{for } y_i = -1 \quad (\text{negative example}) \quad (2)$$

According which side that new instances locate, we can easily determine which class they belong to. So the decision function becomes  $f_{w,b}(x) = \text{sign}(\langle w, x \rangle + b)$ .

**Figure 2-1 Architecture of support vector machines**



Geometrically, all the points are divided into two regions by a hyperplane  $H$ . As shown in Figure 2-2, there are numerous ways through which a hyperplane can separate these examples. The objective of SVM is to choose the “optimal” hyperplane. As all new examples are supposed to be located under similar distribution as training examples, the hyperplane should be chosen such that small shifts of data do not result in fluctuations in prediction result. Therefore, the hyperplane that separates examples of two classes should have the largest margin, which is expected to possess the best generalization performance. Such hyperplane is called the Optimal Separating Hyperplane (OSH) [30].

**Figure 2-2 Different hyperplanes could be used to separate examples**

Examples locating on the margins are called support vectors, whose presentation determines the location of the hyperplane. OSH could be thus represented by a linear combination of support vectors. The margin  $\gamma_i(w, b)$  of a training point  $x_i$  is defined as the distance between  $H$  and  $x_i$  :

$$\gamma_i(w, b) = y_i(w \cdot x + b) \quad (3)$$

and the margin of a set of vectors  $S = \{x_1, \dots, x_n\}$  is defined as the minimum distance between the hyperplane  $H$  to all the vectors in  $S$  :

$$\gamma_S(w, b) = \min_{x_i \in S} \gamma_i(w, b) = \min_{\{x|y=+1\}} \frac{w \cdot x}{\|w\|} - \max_{\{x|y=-1\}} \frac{w \cdot x}{\|w\|} \quad (4)$$

So the OSH is the solution to the optimization problem:

$$\text{Maximize } \gamma_x(w, b) \quad (5)$$

$$\text{Subject to } \gamma_x(w, b) > 0 \quad (6)$$

$$\|w\|^2 = 1 \quad (7)$$

which is an equivalent statement of the problem

$$\text{Minimize } \frac{1}{2} \|w\|^2 \quad (8)$$

$$\text{Subject to } w \cdot x_i + b \geq +1 \text{ for } y_i = +1 \quad (9)$$

$$w \cdot x_i + b \leq -1 \text{ for } y_i = -1 \quad (10)$$

This optimization problem could be efficiently solved by the Lagrange method. With the introduction of Lagrangian multipliers  $\alpha_i \geq 0 (i = 1, 2, \dots, n)$ , one for each of the inequality constraints, we obtain the Lagrangian:

$$L_p(w, b, \alpha) = \frac{1}{2} w^T w - \sum_{i=1}^n \alpha_i [y_i (w \cdot x_i + b) - 1] \quad (11)$$

This is a Quadratic Programming (QP) problem. We would have to minimize  $L_p(w, b, \alpha)$  with respect to  $w$ ,  $b$  and simultaneously require that the derivatives of

$L_p(w, b, \alpha)$  with respect to the multipliers  $\alpha_i$  vanish,  $\frac{\partial}{\partial w} L_p(w, b, \alpha) = 0$  and

$$\frac{\partial}{\partial b} L_p(w, b, \alpha) = 0$$

This leads to

$$w = \sum_{i=1}^n \alpha_i y_i x_i \text{ and } \sum_{i=1}^n \alpha_i y_i = 0 \quad (12)$$

By substituting these two equations into equation (11), the QP problem becomes the Wolfe dual of the optimization problem:

$$L_D(w, b, \alpha) = \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i,j=1}^n \alpha_i \alpha_j y_i y_j (x_i \cdot x_j) \quad (13)$$

subject to constraints  $\sum_{i=1}^n \alpha_i y_i = 0$  and  $\alpha_i \geq 0, i = 1, 2, \dots, n$ .

The corresponding bias  $b_0$  can be calculated as:

$$b_0 = -\frac{1}{2} \left\{ \min_{\{x|y=+1\}} (w_0 \cdot x) - \max_{\{x|y=-1\}} (w_0 \cdot x) \right\} \quad (14)$$

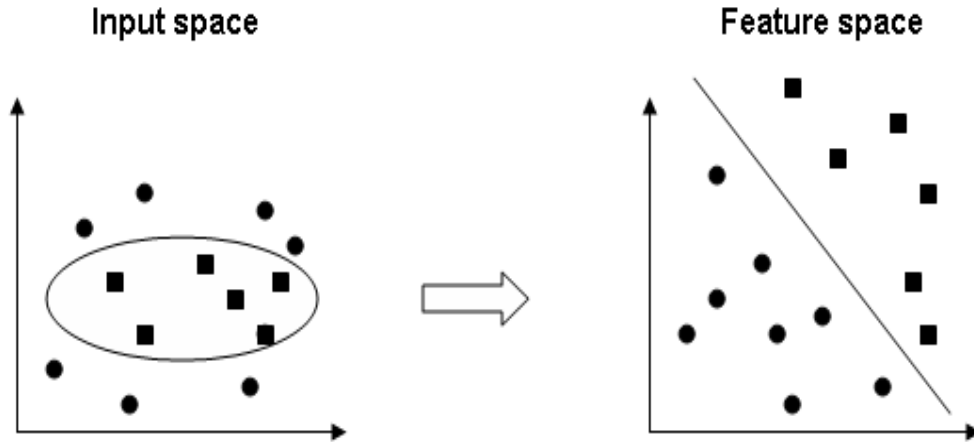
This QP problem could be efficiently solved through several standard algorithms like Sequential Minimization Optimization [140] or decomposition algorithms [141].

Once  $w_0$  and  $b_0$  are determined, the hyperplane is readily drawn. The points for which  $\alpha_i > 0$  are called support vectors, which lie on the margin [127].

### 2.1.1.2. Nonlinear SVM

Many real-world problems are usually too complicated to be solved with linear classifiers. With the introduction of kernel techniques, input data could be mapped to a higher-dimension space, where a new linear classifier can be used to classify these examples (Figure 2-3).

**Figure 2-3 Mapping input space to feature space**



Let  $\Phi$  denotes an implicit mapping function from input space to feature space  $F$ . Then all the previous equations are transformed by substituting input vector  $x_i$  and inner product  $(x_i, x)$  with  $\Phi(x_i)$  and kernel  $K(x_i, x)$  respectively, where

$$K(x_i, x) = \Phi(x_i) \cdot \Phi(x) \quad (15)$$

Equation (13) is then replaced by

$$L_D(w, b, \alpha) = \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j y_i y_j K(x_i \cdot x_j) \quad (16)$$

subject to constraints  $\sum_{i=1}^n \alpha_i y_i = 0$  and  $\alpha_i \geq 0$ , for  $i = 1, 2, \dots, n$ . The bias  $b_0$  becomes

$$b_0 = -\frac{1}{2} \left\{ \min_{\{x|y=+1\}} \left[ \sum_{SV} \alpha_i y_i K(x_i, x) \right] - \max_{\{x|y=-1\}} \left[ \sum_{SV} \alpha_i y_i K(x_i, x) \right] \right\} \quad (17)$$

and the decision function becomes

$$f(x) = \text{sign} \left[ \sum_{i=1}^n \alpha_i y_i K(x_i, x) + b_0 \right] = \text{sign} \left[ \sum_{SV} \alpha_i y_i K(x_i, x) + b_0 \right] \quad (18)$$

Note that the mapping function  $\Phi$  is never explicitly computed, which would significantly reduce the computation load. Another advantage is that the feature space may be infinitely dimensional, such as in the case of Gaussian kernel [142], where mapping function cannot be explicitly represented. A function could be used as a kernel function if and only if it satisfies Merce's conditions [143]. Followings are several well-known kernel functions:

Polynomial  $k(x, z) = (\langle x, z \rangle + 1)^p$

Sigmoid  $k(x, z) = \tanh(\kappa \langle x, z \rangle - \delta)$

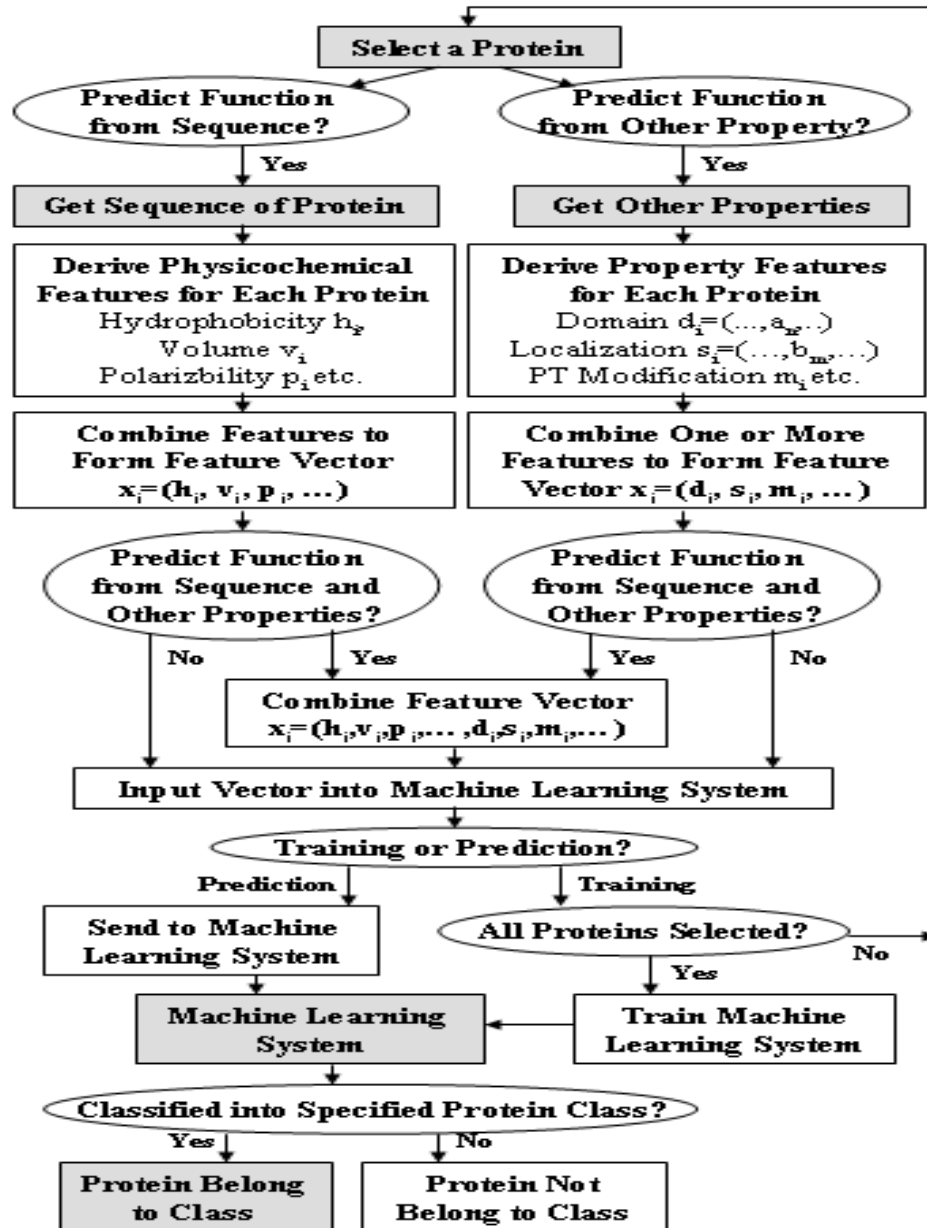
Radial basis function (RBF)  $k(x, z) = \exp(-\|x - z\|^2 / 2\sigma^2)$

In this work, RBF kernel also known as Gaussian kernel is used due to its many advantages demonstrated in previous studies [118, 144, 145]. Then SVM models developed in this study could be developed by using different  $\sigma$  values. It is thus necessary to scan a number of  $\sigma$  values to find the best model, which is evaluated by their performance on classification tasks. In our work, SVM models with  $\sigma$  value



in the range of 1 to 100 were evaluated for each classification task. On the other hand, another variable  $C$  in SVM model is assigned a value of  $10E9$ . Figure 2-4 illustrates the process of training and prediction of protein function by SVM.

**Figure 2-4** Schematic diagrams illustrating the process of the training and prediction of the functional class of proteins by using SVM. Sequence-derived feature  $h_i$ ,  $p_i$ ,  $v_i$  ... represents such structural and physicochemical properties as hydrophobicity, polarizability, and volume. Feature  $d_i$ ,  $s_i$ ,  $m_i$ , ..., represents properties such as domain information, subcellular localization, and post-translational (PT) modification profiles etc.

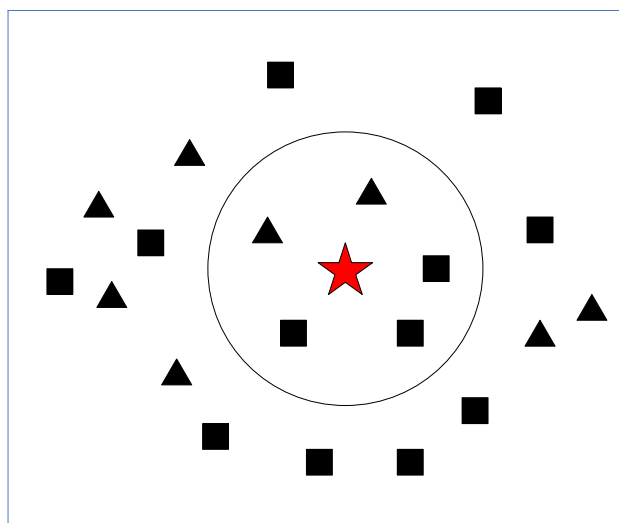


### 2.1.2. K-Nearest Neighbors (KNN)

Nearest neighbor algorithm is one of the most straightforward instance-based learning algorithms [125]. The basic idea of nearest neighbor algorithm is to assign an object to the class of its nearest neighbor. The k-nearest neighbors (KNN) is an extension of this idea, whereas an object is classified by a majority vote of its k neighbors.

In KNN classification, every example is represented by a feature vector in n-dimension feature space. Therefore, the distance between every two samples could be measured and their neighbors could be identified. According to the class of the majority of its neighbors, the class of a test example could be defined (see Figure 2-5). Thus two steps are involved in KNN classification: one is to determine the similarity between a test example and instances in training data set; the other one is to determine the class of the test example based on the classes of its k nearest neighbors.

**Figure 2-5 Example of k-nearest neighbors (squares and triangles represent training samples and the star symbol indicates an unknown sample)**



Many distance measurements could be used to determine the similarity between two examples. Following is a list of some popular ones.

$$\text{Minkowsky: } d(x_i, x_j) = \left( \sum_{l=1}^n |\alpha_l(x_i) - \alpha_l(x_j)|^r \right)^{1/r}$$

$$\text{Manhattan: } d(x_i, x_j) = \sum_{l=1}^n |\alpha_l(x_i) - \alpha_l(x_j)|$$

$$\text{Chebychev: } d(x_i, x_j) = \max_{l=1}^n |\alpha_l(x_i) - \alpha_l(x_j)|$$

$$\text{Euclidean: } d(x_i, x_j) = \left( \sum_{l=1}^n [\alpha_l(x_i) - \alpha_l(x_j)]^2 \right)^{1/2}$$

where  $x$  denotes an arbitrary instance represented by a feature vector  $\{\alpha_1(x), \alpha_2(x), \dots, \alpha_n(x)\}$ ,  $\alpha_l(x)$  denotes the value of the  $l$ th attribute of instance  $x$ ,  $l=1, 2, \dots, n$ . Based on equations above, the distances between the query instance and all the training instances could be readily calculated.

Once the distances are calculated, the  $k$  nearest neighbors of a test example could be identified, the majority of which will be used to determine the class of the test example. As shown in Figure 2-5, all training instances belong to two classes, one represented by squares and the other one represented by triangles. If five nearest neighbors are taken into consideration, the test sample (black star) should be classified into the class as squares, because there are 3 squares but only 2 triangles as the nearest neighbors. However, it is never a trivial work to select appropriate number of neighbors, namely the value of  $k$ . If  $k$  is too small, the model may not benefit from a large data set, whereas if  $k$  is too large, the larger class will overwhelm the smaller one. In theory, the value of  $k$  should be greater than or equal to one but less than  $N$ , where  $N$  is the size of the entire dataset. Dasarathy found that the ideal value of  $k$  is

usually less than  $\sqrt{N}$  [146]. In practice, the value of  $k$  is generally estimated by the cross validation and can be optimized by many trials on the training and validation sets.

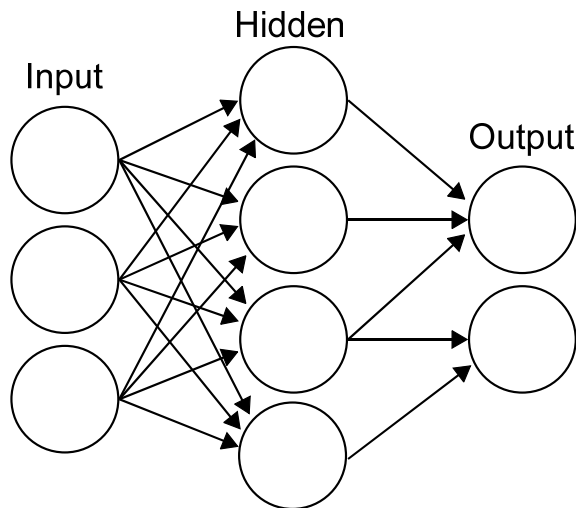
### 2.1.3. Neural Networks (NN)

Neural networks (NN), often referred to artificial neural networks, is inspired by neuroscience and designed to emulate the central nervous system. A neural network consists of an interconnected group of processing information and processing element known as artificial neurons using a connectionist approach to computation. Since an initial neural network model called perceptrons proposed by Frank Rosenblatt, neural networks have been successfully applied in pattern reorganization, drug discovery, and modeling the process of expert systems.

A neural network is a two-stage regression or classification model that trains a hidden-layer-containing network [144, 145]. A simple neural network (see Figure 2-6) could be mathematically represented as  $f(x) = g \sum_j w_{0j} h_j$ , 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  $h_j = \delta(\sum_j w_{ji} x_j + w_j)$ ,  $x_j$  represents input feature vector,  $w_{ji}$  is the input weight from an input node  $i$  to a hidden node  $j$ ,  $w_j$  is the threshold weight from an input node of value 1 to a hidden node  $j$ , and  $\delta$  is an activation function where the sigmoid function  $y = \frac{1}{1 + e^{-x}}$  is mostly used. Other alternative activation functions like Gaussian function are also widely used in neural networks, especially in probabilistic neural network (PNN) [147].

The main advantage of neural networks is that their separable structure is suitable for parallel computation, which could reach higher computation speed. However, there are still some problems during the practice of neural network. One issue is that models generated by neural networks are implicit: they work like a black box, where the relationship between input variables and output variables is difficult to formulate. Moreover, neural networks tend to overfit their models to training examples, making them difficult to be extended to unknown cases. The optimization of large number of weight parameters also presents a computationally intensive challenge.

**Figure 2-6 Architecture of a simple three-layer neural network**



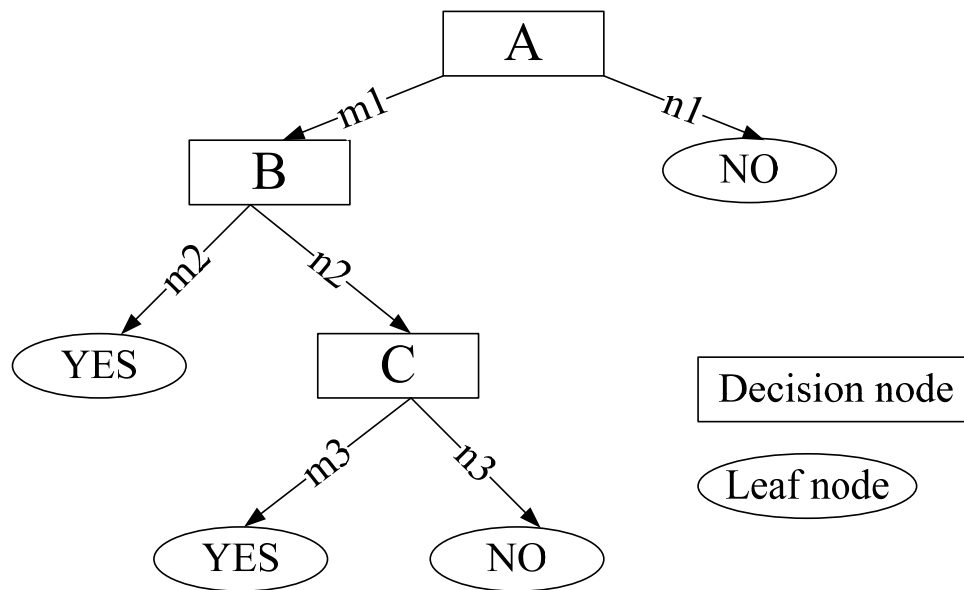
#### 2.1.4. Decision Tree (DT)

Decision tree (DT) is a powerful learning algorithm that has been applied to a variety of fields. It is simple to construct, efficient in decision making and able to generate

human readable and interpretable rules compared with other machine learning methods, such as neural networks and support vector machines.

Decision trees classify input instances through generating a series of rules on each feature property. The structure of decision tree classifier is similar to a tree, where each leaf node represents the target attribute, and each decision node represents some tests to be conducted on a specific attribute. Instances are classified starting at the root node and sorted down the tree until leaf nodes are reached. Figure 2-7 is an example of a decision tree according to the training set of Table 2-1.

**Figure 2-7 Example of a decision tree classifier**



Using the decision tree described in Figure 2-7 as an example, the instance  $A = m1$ ,  $B = n2$ ,  $C = m3$ , would sort to the nodes: A, B, and finally C, which would classify the instance as positive.

The attractiveness of decision tree is the speed and perspicuity. It could be applied to a number of different tasks such as predicting outcomes, classification or when the goal is assignment of a query to a few broad categories. One problem with decision tree is overfitting. In the original algorithm of decision tree, construction of the tree will not stop until all the training examples are classified. However, in many cases, noise or erroneous information are inevitable present in the training data, which may result in a tree too specific to be applied to unseen examples [148]. There are several approaches to alleviate this problem, such as modifying stopping criteria to stop the tree generation before leaf nodes are reached, or post-pruning the tree when it is finished. These approaches, on the other aspect, however, will inevitably come with the sacrifice of training accuracy. Therefore it is important to find a balance between the accuracy of classification and the generalization capability for unseen cases.

**Table 2-1 Example of training data for decision tree**

A	B	C	Class
m1	m2	m3	Yes
m1	m2	n3	Yes
m1	n2	m3	Yes
m1	n2	n3	No
n1	n2	n3	No

## 2.2. Feature selection

Feature selection is to select a subset of relevant features for different tasks, while removing irrelevant ones. This process is expected to help to build more robust models, and make the models more easily to understand and interpret.

A number of feature selection methods have been developed based on different strategies. Among these strategies, recursive feature elimination (RFE), has gained popularity due to its effectiveness and sensitivity. RFE has been successfully applied

in a wide range of biological tasks like cancer gene classification and drug activity analysis [149-151].

The central idea of RFE is recursively ranking features. The ranking criterion for feature selection can be based on the variation in an objective function upon removing each descriptor [152]. It iteratively ranks the contribution of each feature to the objective function, and then eliminates those features that do not reach a defined threshold. In RFE, ranking criteria is based on the change in objective function of QP problem in SVM upon removing each feature. From previous introduction of SVM, we know that the objective function is represented by a cost function for the  $i$ -th feature computed by using the training set. The cost function is minimized under the

constrain  $\sum_{i=1}^n \alpha_i y_i = 0$  where  $\alpha_i \geq 0, i = 1, 2, \dots, n$ . We then convert cost function into

following format

$$J = \frac{1}{2} \alpha^T H \alpha - \alpha^T L, \quad (19)$$

where  $H(i, j) = y_i y_j K(x_i, x_j)$ ,  $K$  is the kernel function and  $L$  is an  $l$  dimensional vector ( $l$  is the number of proteins in the training set).

For linear case, we have

$$K(x_i, x_j) = x_i \bullet x_j \quad (20)$$

When the  $i$ -th feature is removed, the effect of removing one feature could be deduced as following

$$DJ(i) = \frac{1}{2} \frac{\partial^2 J}{\partial w_i^2} (Dw_i)^2 \quad (21)$$

where the change in weight  $Dw_i = w_i - 0$  corresponds to the removal of feature  $i$ .



In nonlinear case, one assumption should be made that the values of  $\alpha$  s will not change significantly with removing one feature. Therefore, only  $H(-i)$ , which is matrix computed by using the same method as that of matrix  $H$  with its  $i$ -th component removed, needs to be re-computed. Then the resulting ranking criterion is:

$$DJ(i) = \frac{1}{2} \alpha^T H \alpha - \frac{1}{2} \alpha^T H(-i) \alpha \quad (22)$$

The features with the smallest difference  $DJ(i)$  should be removed iterated until only the features with predetermined number are obtained [153].

### 2.3. Performance evaluation

The performance evaluation aims to find out whether an algorithm is able to be applied to novel data that have not been used to develop the prediction model, or measure the generalization capacity to recognize new examples from the same data domain [154].

In this study, several statistical measurements were explored, including sensitivity (SE), specificity (SP), positive prediction value (PPV), and overall prediction accuracy (Q). The formulas to calculate these measurements are listed as following

$$SE = TP / (TP + FN)$$

$$SP = TN / (TN + FP)$$

$$PPV = TP / (TP + FP)$$

$$Q = (TP + TN) / (TP + TN + FP + FN)$$

where TP (true positive), FN (false negative), TN (true negative), and FP (false positive) represents correctly predicted positive samples, positive samples incorrectly

predicted as negative, correctly predicted negative samples, and negative samples incorrectly predicted as positive respectively. Another measurement, Matthews correlation coefficient (MCC) was also used to evaluate the randomness of the prediction. MCC is defined as following

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

where MCC is within the range of -1 to 1. Negative values of MCC indicate the disagreement between prediction and measurement, while positive values of MCC indicate the agreement between prediction and measurement. A zero value means the prediction is no better than random guess.

## 2.4. Construction of feature vectors

### 2.4.1. Protein feature vectors

As most machine learning methods could only accept numerical vectors instead of protein sequences, it is essential to convert proteins sequences into numerical vectors in order to employ the power of machine learning to classify proteins. The construction of feature vector for each protein is based on the formula used for the prediction of protein functional classes [155-157] and protein-protein interactions [136]. Each feature vector is constructed from the 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 [136, 158].

For each of these properties, amino acids are divided into three groups such that those in a particular group are regarded 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 2-2. 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 (such as hydrophobicity) 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.

Overall, there are 21 elements representing these three descriptors: 3 for C, 3 for T and 15 for D [136, 158]. 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. In this study, totally 188 elements are used as feature vector for each protein shown in Table 2-3. The following is a hypothetical protein sequence AEAAAEAEAEAAAAAEAEAEAEAEAEAEAE, as shown in Figure 2-8, which has 16 alanines ( $n_1=16$ ) and 14 glutamic acids ( $n_2=14$ ). The composition for these two amino acids are  $n_1 \times 100.00 / (n_1 + n_2) = 53.33$  and  $n_2 \times 100.00 / (n_1 + n_2) = 46.67$  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  $(15/29) \times 100.00 = 51.72$ . 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  $1/30 \times 100.00 = 3.33$ ,  $5/30 \times 100.00 = 16.67$ ,  $12/30 \times 100.00 = 40.0$ ,  $20/30 \times 100.00 = 66.67$ ,  $29/30 \times 100.00 = 96.67$ . Likewise, the D descriptor for glutamic is 6.67, 26.67, 60.0, 76.67, and 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. Table 2-4 gives the computed descriptors of the cellular tumor antigen p53 (Swiss-Prot AC P04637). The feature vector of a protein is constructed by combining all of the descriptors in sequential order.

**Table 2-2 Division of amino acids into 3 different groups by different physicochemical properties**

Property		Group 1	Group 2	Group 3
Hydrophobicity	Type	Polar	Neutral	Hydrophobic
	Amino acids	RKEDQN	GASTPHY	CVLUMFW
Van der Waals volume	Value	0~2.78	2.95~4.0	4.43~8.08
	Amino acids	GASCTPD	NVEQIL	MHKFRYW
Polarity	Value	0~0.456	0.6~0.696	0.792~1.0
	Amino acids	LIFWCMVY	PATGS	HQRKNED
Polarizability	Value	0~0.108	0.128~0.186	0.219~0.409
	Amino acids	GASDT	CPNVEQIL	KMHFRYW
Charge	Type	Positive	Neutral	Negative
	Amino acids	KR	ANCQGHILMFPSTWYV	DE
Surface tension	Value	-0.20~0.16	-0.3~ -0.52	-0.98~ -2.46
	Amino acids	GQDNAHR	KTSEC	ILMFPWYV
Secondary structure	Type	Helix	Strand	Coil
	Amino acids	EALMQKRH	VIYCWFT	GNPSD
Solvent accessibility	Type	Buried	Exposed	Intermediate
	Amino acids	ALFCGIVW	RKQEND	MPSTHY

**Table 2-3 List of features for proteins**

Feature Description	Number of Dimensions
Amino acid composition	20
Hydrophobicity	21
Van der Waals volume	21
Polarity	21
Polarizability	21
Charge	21
Surface tension	21
Secondary structure	21
Solvent accessibility	21
Total	188

**Table 2-4 Characteristic descriptors of cellular tumor antigen p53 (Swiss-Prot AC P04637). The feature vector of this protein is constructed by combining all of the descriptors in sequential order.**

Property	Elements of Descriptors									
Amino acid composition	A	C	D	E	F	G	H	I	K	L
	6.11	2.54	5.09	7.63	2.80	5.85	3.05	2.04	5.09	8.14
	M	N	P	Q	R	S	T	V	W	Y
Hydrophobicity	3.05	3.56	11.45	3.82	6.62	9.67	5.60	4.58	1.02	2.29
	31.81	44.02	24.17	26.02	16.58	19.39	0.51	33.33	58.02	81.17
	100.0	1.02	22.39	43.26	74.55	99.75	0.25	24.68	46.31	65.39
Van der waals volume	97.96									
	46.31	29.77	23.92	23.98	17.10	14.29	1.02	20.87	42.24	70.74
	100.0	0.51	24.68	51.14	72.77	98.73	0.25	34.10	59.29	81.68
Polarity	98.22									
	26.46	38.68	34.86	18.62	19.90	24.23	0.25	27.74	47.84	64.89
	97.96	1.02	21.12	39.44	74.55	99.75	0.51	34.61	58.02	81.17
Polarizability	100.0									
	32.32	43.77	23.92	28.57	13.52	17.86	1.53	22.40	46.82	76.84
	100.0	0.51	19.59	48.35	69.97	99.24	0.25	34.10	59.29	81.68
Charge	98.22									
	11.70	75.57	12.72	16.07	2.30	18.37	6.11	44.27	71.76	85.24
	98.22	0.25	23.66	45.55	69.97	99.74	0.51	12.98	52.67	74.81
Surface tension	100.0									
	34.10	30.53	35.37	19.90	25.77	20.92	1.27	27.99	52.67	75.57
	100.0	0.51	30.02	57.76	79.90	99.75	0.25	18.58	40.71	64.63
Secondary structure	99.24									
	43.51	20.87	35.62	17.86	27.30	11.22	0.25	25.70	51.40	81.17
	98.73	2.54	31.04	46.31	63.87	98.47	1.02	20.36	47.58	74.55
Solvent accessibility	100.0									
	33.08	31.81	35.11	22.45	25.0	20.15	2.54	24.68	47.84	69.72
	98.98	0.51	33.33	58.02	81.17	100.0	0.25	22.14	43.00	68.45
	99.75									

**Figure 2-8 The sequence of a hypothetical protein for illustration of derivation of the feature vector\***

Sequence	A E A A A E A E E 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																														

\*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.

### 2.4.2. MiRNA feature vectors

The construction of feature vectors for miRNAs in this work was based on different kinds of physicochemical properties of nucleotides, including molecular weight, surface area, inter-base hydrogen bonds donor (IHD) and acceptor (IHA), dipole moment, polarizability and hydrogen bonds on the side. Three descriptors, composition (C), transition (T), and distribution (D), similar to those for protein features, are used to describe global composition of each of the properties.

In order to calculate each group of properties, nucleotides are divided into distinct groups, each of which is expected to have the same property. Table 2-5 shows the grouping of nucleotides in according to their different physicochemical properties. The feature vectors of miRNAs are constructed by combining all of the descriptors in sequential order. Totally 115 elements are included in the feature vector for each miRNA as listed in Table 2-6. An example of computed descriptors of miRNA precursor (cel-mir-243) is shown in Table 2-7.

**Table 2-5 Division of nucleotides into different groups for different physicochemical properties**

Descriptors	Group 1	Group 2	Group 3
Molecular Weight*	A (491.2) G (507.2)	C (467.2) T (482.2)	
Solvent accessibility[159] *	A (213.66) T (212.67)	C (195.84)	G (236.76)
Interbase H-bonds donator (IHD), Interbase H-bonds acceptor (IHA)	A, T	C	G
Dipole moment[160]*	G (6.82), C (6.90)	A (2.68) T (4.53)	-
Polarizability*[160]	A (88.4), G (91.8)	T (75.8), C (69.5)	-
H-bonds on the side	A, C	T, G	-

\* The property value of each nucleotide is shown in parentheses.

**Table 2-6 List of features for miRNA**

Feature Description	Number of Dimensions
nucleotide composition	4
nucleotide distribution	20
nucleotide transition	16
polarizability and molecular weight	16
dipole moment	16
hydrogen bonds on the side	16
solvent accessibility and IHD and IHA	27
Total	115

**Table 2-7 Example of computed descriptors of miRNA precursor (cel-mir-243). The feature vector of this precursor is constructed by combining all the descriptors in sequential order.**

Property	Elements of Descriptors									
Nucleotide composition	A	G	C	U						
	22.45	27.55	23.47	26.53						
Nucleotide distribution	6.12	21.43	52.04	80.61	100.0	2.04	29.59	64.29	78.57	98.98
	5.10	23.47	44.90	62.24	90.82	1.02	22.45	43.88	70.41	96.94
Nucleotide transition	1.02	3.06	7.14	10.20	7.14	7.14	6.12	7.14	5.10	12.24
	2.04	4.08	9.18	5.10	8.16	4.08				
Polarizability and molecular weight	50.00	50.00	18.37	30.61	31.63	18.37	2.04	27.55	58.16	79.59
	100.00	1.02	22.45	44.90	67.35	96.94				
Dipole moment	51.02	48.98	27.55	23.47	23.47	24.49	2.04	27.55	48.98	75.51
	98.98	1.02	21.43	50.00	70.41	100.0				
Hydrogen bonds on the side	45.92	54.08	15.31	29.59	30.61	23.47	5.10	21.43	48.98	69.39
	100.00	1.02	26.53	55.10	78.57	98.98				
Solvent accessibility and IHD and IHA	23.47	27.55	48.98	2.04	12.24	9.18	6.12	7.14	14.29	15.31
	8.16	24.49	5.10	23.47	44.90	62.24	90.82	2.04	29.59	64.29
	78.57	98.98	1.02	21.43	50.00	70.41	100.0			

### **3. *In silico* search and characterization of multifunctional enzymes**

As introduced in the first chapter, multifunctional enzymes (MFEs) are enzymes that perform multiple functions. Characterization and identification of MFEs are critical for better understanding of molecular mechanisms underlying the crosstalk between different cellular processes. As an alternative approach, support vector machine (SVM) has been successfully applied for predicting different functional classes of proteins from their amino acid sequences with accuracies of 60.6%~97.8% [156, 157, 161]. It is thus expected that SVM might be equally applicable for the identification of MFEs. In this study, SVM was applied for the identification of MFEs from their primary sequences. We also analyzed the pathway, structure, and orthologs of MFEs.

#### **3.1. Selection of MFEs and non-MFEs**

A total of 3,120 MFEs were derived from a comprehensive search of Swiss-Prot database [162] using keyword “multifunctional enzyme” followed by manual check that each MFE performs at least two different kinds of catalytic activities as annotated in the database. They were further divided into two independent classes of positive datasets for model construction: MFEs with multiple catalytic domains (MCD-MFEs) (2,551 proteins) and MFEs with single multi-activity domain (SMAD-MFEs) (537 proteins). The non-MFE proteins (NMFEPs), including non-enzymatic proteins and non-MFE enzymes, were selected from seed proteins of the domain families in Pfam database [163] excluding those that contain at least one MFE. Totally, 21,833 NMFEPs were generated as the negative dataset. All these sequences were then converted into numerical vectors as described in last chapter.



These positive and negative datasets were divided into separate training, testing and independent evaluation sets by the following procedure: First, proteins were clustered into groups based on their distance in the structural and physicochemical feature-space by using the hierarchical clustering method. In the feature space, more homologous sequences will have shorter distance between them and an upper-limit of the largest separation of 20 was used for each cluster. One representative protein was randomly selected from each group to form a training set that is sufficiently diverse and broadly distributed in the feature space. One or up to 50% of the remaining proteins in each group were randomly selected to form the testing set. The selected proteins from each group were further checked to ensure that they are distinguished from the proteins in other groups. The remaining proteins were then designated as the independent evaluation set, which was also found in a reasonable level of diversity. Fragments of smaller than 50 residues were discarded. The statistics of the members and non-members in each dataset of MFE classes were given in Table 3-1.

**Table 3-1 Statistics of the datasets and prediction accuracy of individual class of MFE and that of all MFEs (6=21)**

MFE Classes	Training set		Testing set				Independent evaluation set							
	positive	negative	positive		negative		positive			negative			PPV (%)	Q(%)
			TP	FN	TN	FP	TP	FN	SE (%)	TN	FP	SP (%)		
All MFEs	1221	3897	1564	13	15435	16	303	19	94.1	2461	24	99.0	92.6	98.5
MCD-MFEs	918	2354	1342	13	16984	10	256	22	92.1	2474	11	99.6	95.9	99.8
SMAD-MFEs	263	1653	212	4	17694	1	54	4	93.1	2483	2	99.9	96.4	99.7

## 3.2. Evaluation and discussion

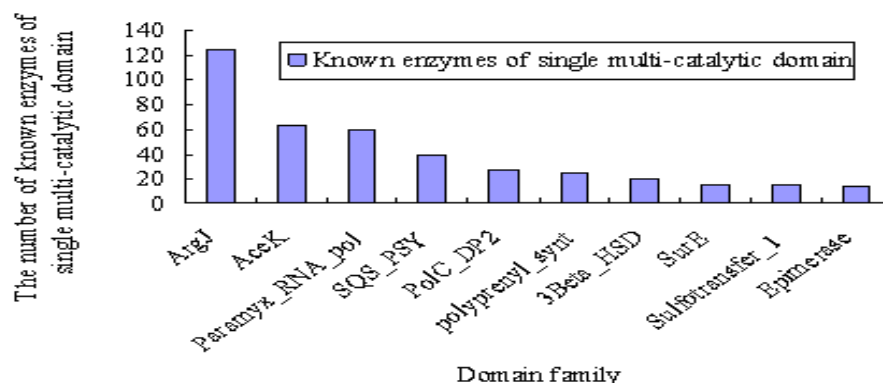
### 3.2.1. Structural preference of MFEs

Proteins perform their functions through structural and functional defined regions called domains. Knowledge of domain composition is able to provide valuable insights into the mechanism of MFEs. In this work, the domain composition of MFEs was investigated statistically against the Pfam database [163]. The distribution of top 10 Pfam domains in two classes of MFEs was shown in Figure 3-1 and Figure 3-2 respectively. The most abundant domain among SMAD-MFEs is ArgJ domain (Pfam ID: PF01960), which plays key role in both N-acetylglutamate synthase (EC 2.3.1.1) and ornithine acetyltransferase (EC 2.3.1.35) activities in the cyclic version of arginine biosynthesis [164]. Investigation of the structure of ArgJ domain indicates that the complete active-site is defined by some disconnected residues and potentially the protein C-terminus. “It is possible that the movement of the C-terminus in and out of the active site enables ArgJ to accept two different substrates by altering the substrate specificity of the binding pocket” [165]. Moreover, a number of eukaryotes enzymes contain both tetrahydrofolate dehydrogenase/cyclohydrolase NAD(P)-binding domain (PF02882) and catalytic domain (PF00763), which are also present separately in many prokaryotic single-function enzymes [166]. This indicates the emergence of gene fusion between these two ancient domains; they merged sometime during the evolution as a new protein performing double physiological functions under different conditions.

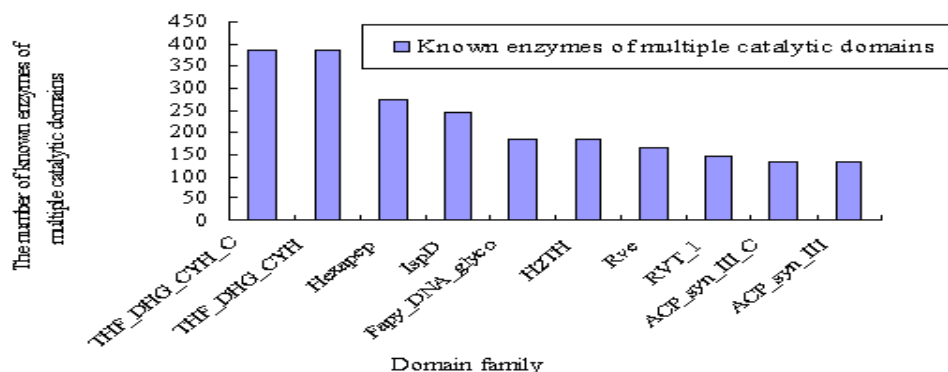
Enzymes with different functions may share the same fold. To have an overview of their structural propensities, the distribution of MFEs in SCOP fold [167] was

investigated. As illustrated in Figure 3-3, 42% of MCD-MFEs and 69% of SMAD-MFEs are found to belong to alpha and beta protein class (a/b), contrast to about 10% of enzymes contain such fold (a/b) in nature. Previous studies indicated that the alpha and beta enzymes may have diverged from a common ancestor [168, 169], which suggested that MFEs may have common evolutionary origin, while the alpha and beta fold topology is favored to preserve their multiple catalytic activities.

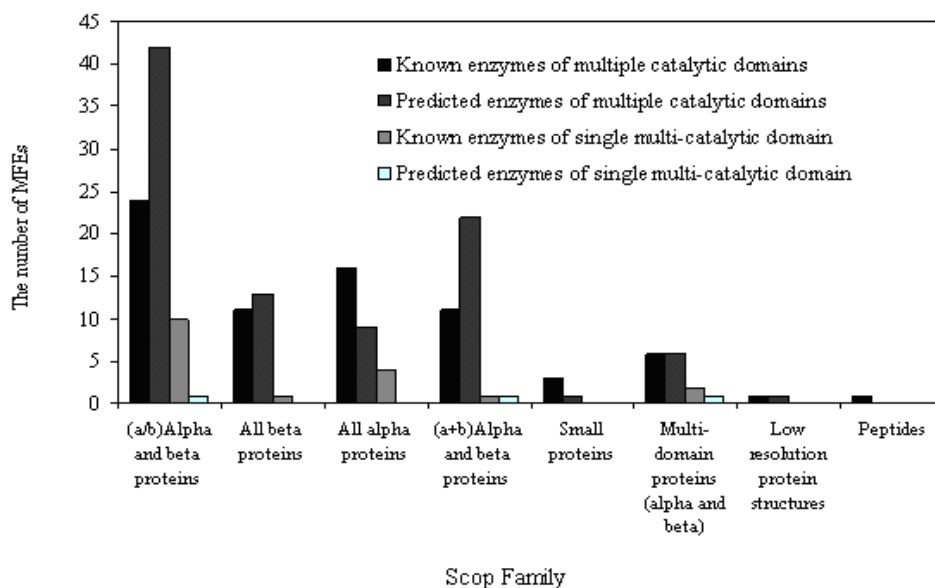
**Figure 3-1 Top 10 Pfam families for known enzymes of single multi-catalytic domain (SMAD-MFEs).** It is noted that about 38% of SMAD-MFEs contain ArgJ domain, and majority of them are involved in Urea cycle and metabolism of amino groups pathway (amino acid metabolism map00220).



**Figure 3-2 Top 10 Pfam families of known enzymes of multiple catalytic domains (MCD-MFEs)**



**Figure 3-3 Distribution of known and predicted putative MFEs (enzymes of single multi-catalytic domain SMAD-MFEs, enzymes of multiple catalytic domains MCD-MFEs) in SCOP fold families. It is noted that 42% of MCD-MFEs and 69% of SMAD-MFEs belong to the alpha and beta fold class (a/b).**



### 3.2.2. Characteristics of MFEs from pathway and evolution perspective

Biological pathways are networks of molecular interactions, which could provide valuable information of complex cellular reactions from molecular level. MFEs are believed to be involved in different biological pathways due to their multiple functionalities. To have an overview of the physiological preference of MFEs, statistics were demonstrated to investigate the distribution of MFEs in KEGG pathway database release 44.0 [170]. The physiological annotation of MFEs was acquired from the KEGG ontology (KO) annotation release 2007-08-17 by the cross-links between KEGG and the UniProt protein knowledgebase. To automate the searching and retrieving, a Perl script was programmed.

For 1,578 currently known MFEs with KEGG pathway information, nearly half of them, including 48.7% of currently known MCD-MFEs (630 out of 1,293 proteins) and 54% of SMCD-MFEs (154 out of 285 proteins), are involved in only one biological pathway (Figure 3-4). It suggests that MCD-MFEs integrate their functions together to accomplish a complicated metabolism step that might need several independent enzyme-catalyzed steps in the primitive pathway. About 38.9%, 5.34%, 1.16%, 5.88% of MCD-MFEs and about 10.9%, 13.3%, 3.51%, and 18.2% of SMAD-MFEs participate in two, three, four and five distinct pathways respectively (Figure 3-4). These enzymes with different physiological roles are difficult to characterize using traditional experimental or homology-based methods; however, they can be properly probed by some learning algorithms like support vector machines adopted in present study. It is interesting that multiple functionalities of some MFEs are not well conserved across species. For example, bifunctional protein fold in *Escherichia coli* participates in glyoxylate and dicarboxylate metabolism (KEGG: map00630) as well as one carbon pool by folate pathway (KEGG: map00670). However, bifunctional protein fold in *Mycoplasma synoviae* just participates in one carbon pool by folate pathway (KEGG: map00670). Similar phenomena can be observed in bifunctional protein glum, bifunctional aminoacyl-tRNA synthetase, and etc. The loss of multiple functionalities of MFEs in some species is either an evolutionary phenomenon, or it may suggest potential unknown mechanism of the functional regulation of MFEs.

The classification of MFEs according to their KEGG ontology (KO) annotation indicates that MCD-MFEs are involved in 4 level one, 17 level two, and 74 level three pathways (Figure 3-5); while SMAD-MFEs are involved in 3 level one, 10 level two

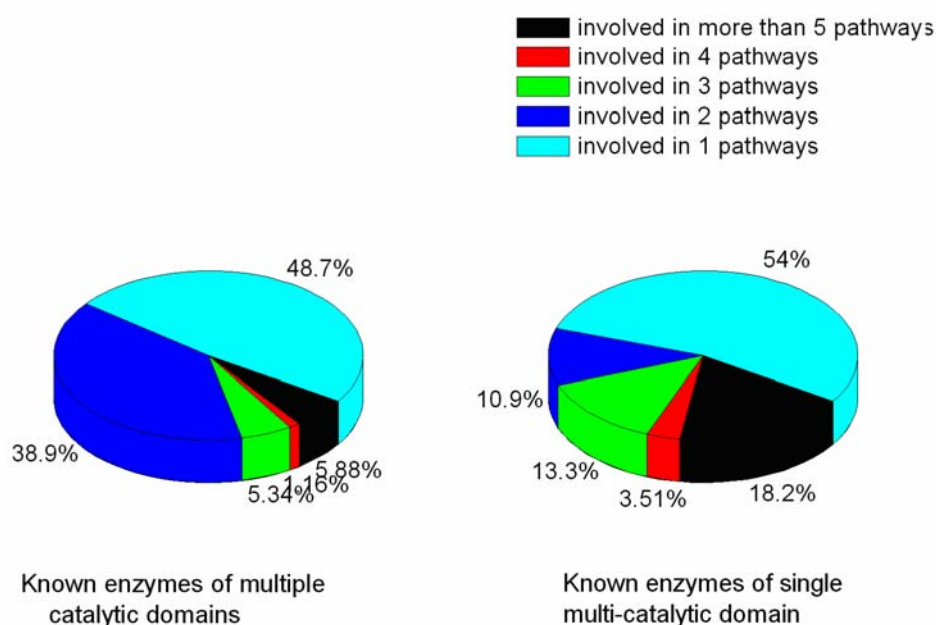
and 52 level three pathways (Figure 3-6). Almost all MFEs (99.5%) are involved in metabolism pathways, majority of which are carbohydrate metabolism (CAR, KEGG: map01110), lipid metabolism (LIP, KEGG: map01130), nucleotide metabolism (NUC, KEGG: map01140), amino acid metabolism (AAC, KEGG: map01150) and metabolism of cofactors and vitamins (COF, KEGG: map01190). Previous studies found that CAR, LIP, NUC and AAC are ancestral and part of early enzymatic burst from phylogenetic analysis of protein architecture [171]. Furthermore, about 25% of MCD-MFEs, which are involved in two pathways, glyoxylate and dicarboxylate metabolism (CAR, KEGG: map00630) and one carbon pool by folate (COF, KEGG: map00670), contain both tetrahydrofolate dehydrogenase/cyclohydrolase NAD (P)-binding domain and catalytic domain. The conservation of MFEs in those essential cellular processes like carbohydrate metabolism, nucleotide metabolism and amino acid metabolism infer the critical roles of MFEs in the origin and evolution of life forms.

In addition, MFEs are not evenly distributed in organisms. In this work, the orthologous proteins of MFEs were primarily identified by blasting the MFE protein sequences against the NCBI Clustering of Orthologous Groups (COGs) [172, 173]. A Perl script was coded to automatically BLAST [106] orthologous proteins of MFEs from local COG database downloaded from <ftp://ftp.ncbi.nih.gov/pub/COG/COG/>. A similarity E-value of  $1.0e-7$  was adopted as threshold to ensure maximum inclusion of proteins that have a homolog.

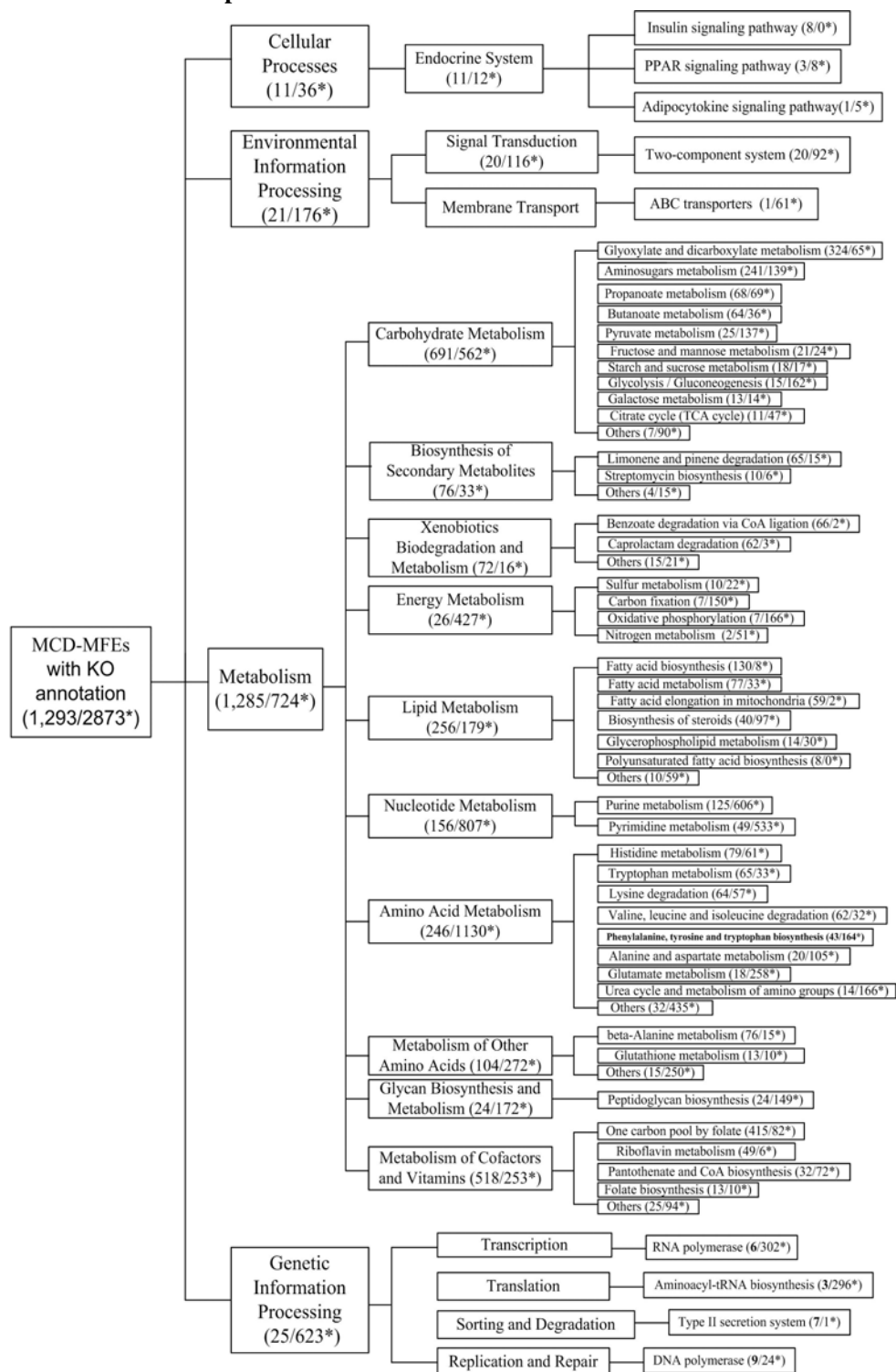
The result seems that bacteria have more MFEs than eukaryotes and archaeabacteria with current knowledge (Figure 3-7 Table 3-2, and Table 3-3). The dominance of

bacteria is significant in both known and predicted MCD-MFEs, which is not only shown in the total number of MFEs, but also the average number in each organism. For known SMAD-MFEs, eukaryotic organisms tend to possess more MFEs. However, no significant difference could be found across organisms (Table 3-2 and Table 3-3). In addition, MFEs orthologs in *S. cerevisiae* and *H. sapiens* were searched and compared. As shown in Table 3-4, 60 MFEs of *H. sapiens* and 37 MFEs of *S. cerevisiae* were found to possess COGs. Comparing these COGs, 36.7% (22 out of 60) MFEs in *H. sapiens* had their orthologs in *S. cerevisiae*, meanwhile 56.8% (21 out of 37) MFEs in *S. cerevisiae* had their orthologs in *H. sapiens*. This may implicate that MFEs are well preserved although some may be lost or gained during evolution.

**Figure 3-4 Statistics of known MFEs according to the number of biological pathways they anticipated in. Totally 1,293 known enzymes of multiple catalytic domains (MCD-MFEs) and 285 known enzymes of single multi-catalytic domain (SMAD-MFEs) were employed in this study.**

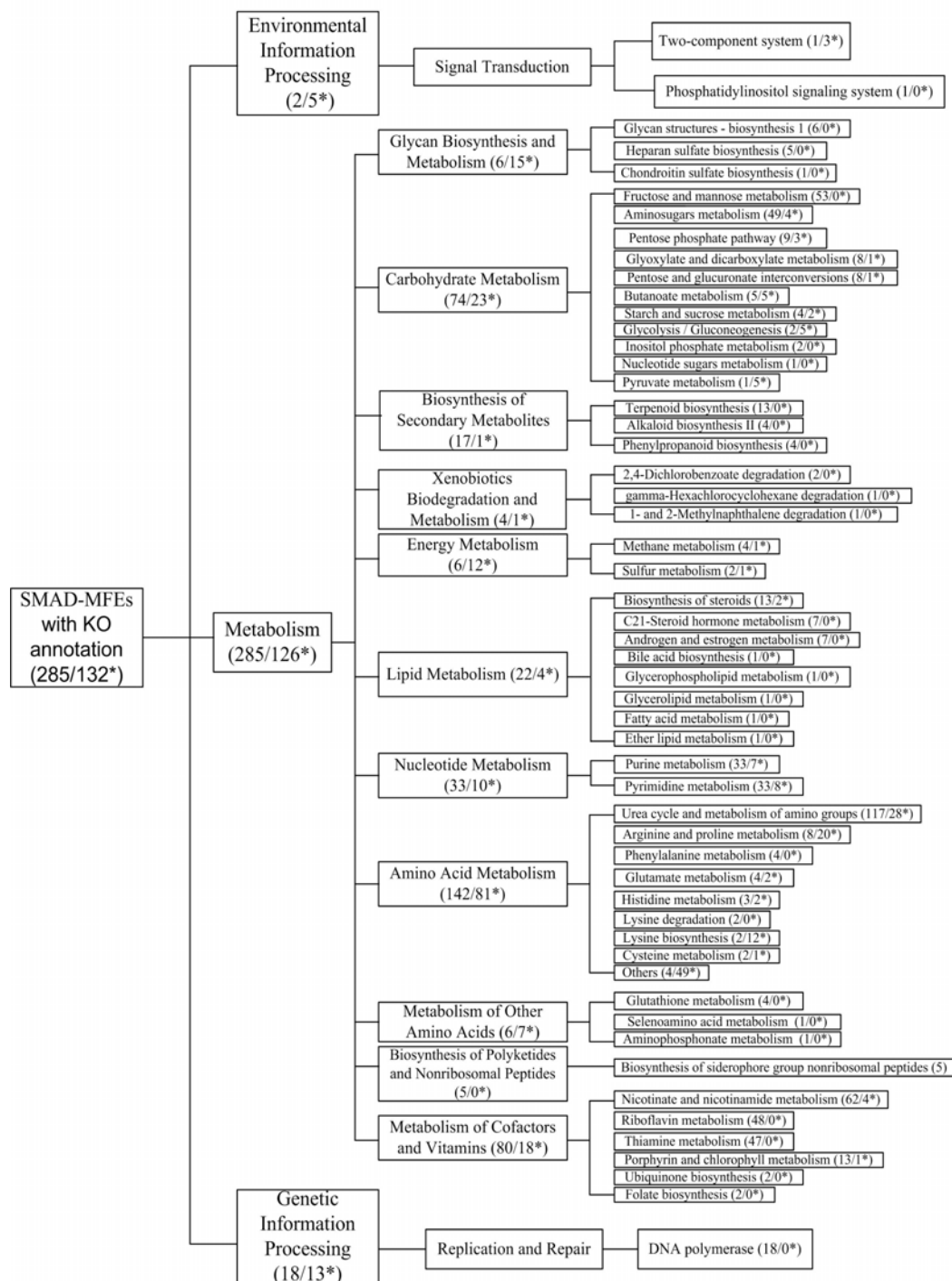


**Figure 3-5 Statistics of known and predicted enzymes of multiple catalytic domains (MCD-MFEs) with KEGG ontology (KO).** MCD-MFEs are involved in 4 level one, 17 level two, and 74 level three pathways. Majority of them anticipate in carbohydrate metabolism (CAR), lipid metabolism (LIP), nucleotide metabolism (NUC), amino acid metabolism (AAC) and metabolism of cofactors and vitamins (COF). Number with “\*” denotes the number of predicted MCD-MFEs.

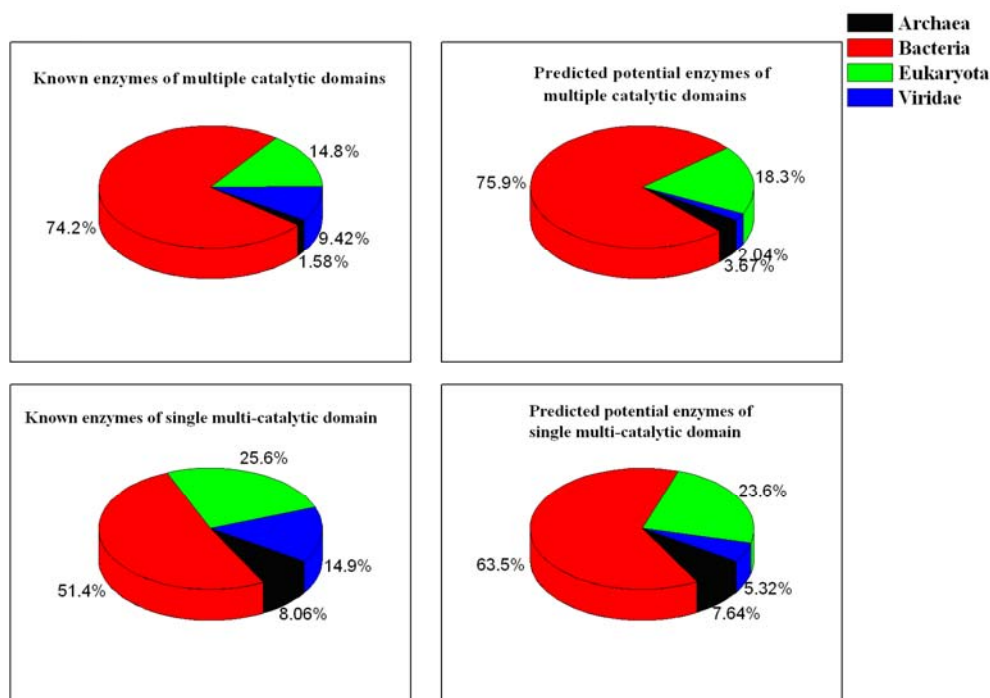




**Figure 3-6 Statistics of known enzymes of single multi-catalytic domains (SMAD-MFEs) in KEGG ontology (KO). SMAD-MFEs are involved in 3 level one, 10 level two and 52 level three pathways. Majority of them anticipate in the carbohydrate metabolism (CAR), amino acid metabolism (AAC) and metabolism of cofactors and vitamins (COF). Number with “\*” denotes the number of predicted SMAD-MFEs.**



**Figure 3-7 Distribution of MFEs in different kingdoms. Totally, 2,551 known enzymes of multiple catalytic domains (MCD-MFEs), 4,075 predicted MCD-MFEs, 537 known enzymes of single multi-catalytic domain (SMAD-MFEs), and 245 predicted SMAD-MFEs were included in the statistics. It is noted the dominance of bacteria in both known and predicted MCD-MFEs and SMAD-MFEs in total enzyme number.**



**Table 3-2 Distribution of known and predicted enzymes of multiple catalytic domains in different kingdoms and in top 20 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 databases.**

	Known enzymes of multiple catalytic domains			Predicted potential enzymes of multiple catalytic domains		
	Kingdom or species	No. of enzymes	Mean of each kingdom	Kingdom or species	No. of enzymes	Mean of each kingdom
MFes distribution in kingdom	Archaea	41	2.1	Archaea	150	4.1
	Bacteria	1894	4.5	Bacteria	3102	6.4
	Eukaryota	373	4.3	Eukaryota	739	3.4
	Viridae	243	1.0	Viridae	84	1.0
MFes distribution in top 20 species	<i>Saccharomyces cerevisiae</i>	66		<i>Homo sapiens</i>	65	
	<i>Homo sapiens</i>	53		<i>Mus musculus</i>	53	
	<i>Escherichia coli</i>	38		<i>Rattus norvegicus</i>	50	
	<i>Mus musculus</i>	37		<i>Escherichia coli</i>	45	
	<i>Bacillus subtilis</i>	26		<i>Bacillus subtilis</i>	44	
	<i>Haemophilus influenzae</i>	24		<i>Arabidopsis thaliana</i>	39	
	<i>Rattus norvegicus</i>	22		<i>Saccharomyces cerevisiae</i>	35	
	<i>Salmonella typhimurium</i>	21		<i>Mycobacterium tuberculosis</i>	35	
	<i>Escherichia coli O157:H7</i>	21		<i>Mycobacterium bovis</i>	31	
	<i>Escherichia coli O6</i>	18		<i>Bacillus halodurans</i>	31	
	<i>Pseudomonas aeruginosa</i>	18		<i>Mycobacterium leprae</i>	30	
	<i>Shigella flexneri</i>	17		<i>Haemophilus influenzae</i>	28	
	<i>Vibrio cholerae</i>	17		<i>Escherichia coli O157:H7</i>	27	
	<i>Salmonella typhi</i>	16		<i>Rhizobium meliloti</i>	27	

	<i>Schizosaccharomyces pombe</i>	15	<i>Salmonella typhimurium</i>	26
	<i>Vibrio parahaemolyticus</i>	15	<i>Vibrio parahaemolyticus</i>	23
	<i>Mycobacterium tuberculosis</i>	15	<i>Escherichia coli</i> O6	22
	<i>Vibrio vulnificus</i>	15	<i>Pasteurella multocida</i>	22
	<i>Bos taurus</i>	14	<i>Vibrio cholerae</i>	22
	<i>Synechocystis</i> sp. (strain PCC 6803)	14	<i>Pseudomonas putida</i> (strain KT2440)	22

**Table 3-3 Distribution of known and predicted enzymes with single multi-catalytic domain in different kingdoms and in top 20 host species**

	Known enzymes of single multi-catalytic domain			Predicted potential enzymes of single multi-catalytic domain		
	Kingdom or species	No. of enzymes	Mean of each kingdom	Kingdom or species	No. of enzymes	Mean of each kingdom
MFes distribution in kingdom	Archaea	43	1.7	Archaea	18	1.5
	Bacteria	274	1.4	Bacteria	145	1.3
	Eukaryota	139	2.8	Eukaryota	66	1.2
	Viridae	81	1.1	Viridae	16	1.3
Distribution of MFes in top 20 species	<i>Mus musculus</i>	17		<i>Arabidopsis thaliana</i>	5	
	<i>Homo sapiens</i>	17		<i>Haemophilus influenzae</i>	4	
	<i>Arabidopsis thaliana</i>	15		<i>Thermoplasma volcanium</i>	4	
	<i>Escherichia coli</i>	12		<i>Mycobacterium tuberculosis</i>	3	
	<i>Rattus norvegicus</i>	9		<i>Mycobacterium bovis</i>	3	
	<i>Saccharomyces cerevisiae</i>	8		<i>Saccharomyces cerevisiae</i>	3	
	<i>Bos taurus</i>	7		<i>Corynebacterium efficiens</i>	3	

	<i>Escherichia coli</i> <i>O157:H7</i>	6	<i>Bradyrhizobium</i> <i>japonicum</i>	3
	<i>Pongo pygmaeus</i>	5	<i>Chlamydia pneumoniae</i>	3
	<i>Salmonella typhimurium</i>	5	<i>Buchnera aphidicola</i> (subsp. <i>Acyrtosiphon</i> <i>pisum</i> )	2
	<i>Neurospora crassa</i>	4	<i>Escherichia coli</i>	2
	<i>Mycobacterium bovis</i>	4	<i>Brucella suis</i>	2
	<i>Pasteurella multocida</i>	4	<i>Brucella melitensis</i>	2
	<i>Mycobacterium</i> <i>tuberculosis</i>	4	<i>Buchnera aphidicola</i> (subsp. <i>Schizaphis</i> <i>graminum</i> )	2
	<i>Methanococcus</i> <i>jannaschii</i>	4	<i>Pseudomonas aeruginosa</i>	2
	<i>Thermoplasma</i> <i>volcanium</i>	3	<i>Synechocystis</i> sp. (strain <i>PCC 6803</i> )	2
	<i>Escherichia coli</i> O6	3	<i>Thermoplasma</i> <i>acidophilum</i>	2
	<i>Thermoplasma</i> <i>acidophilum</i>	3	<i>Mus musculus</i>	2
	<i>Pseudomonas</i> <i>aeruginosa</i>	3	<i>Bos taurus</i>	2
	<i>Salmonella typhi</i>	3	<i>Salmonella typhimurium</i>	2

**Table 3-4 Orthologs of multifunctional enzymes (MFEs) in *S. cerevisiae* and *H. sapiens* species. 36.7% (22 out of 60) MFEs in *H. sapiens* had their orthologs in *S. cerevisiae*, while 56.8% (21 out of 37) MFEs in *S. cerevisiae* had their orthologs in *H. sapiens*.**

orthologs of MFEs in <i>H. sapiens</i>		orthologs of MFEs in <i>S. cerevisiae</i>	
Entry Name	OCG number	Entry Name	OCG number
3BHS1_HUMAN	COG1088 COG0451	ARG56_YEAST	COG0548 COG0002
3BHS2_HUMAN	COG0451	ARGJ_YEAST	COG1364
AADAT_HUMAN	COG1167	ARO1_YEAST	COG0128 COG0337
AASS_HUMAN	COG1748	BPL1_YEAST	COG0340 COG4285
AMD_HUMAN	COG3391	C1TC_YEAST	COG2759
BLVRB_HUMAN	COG0702	C1TM_YEAST	COG0190 COG2759
BPL1_HUMAN	COG0340	COAC_YEAST	COG0439 COG4799 COG0511
C1TC_HUMAN	COG2759	DUR1_YEAST	COG0439 COG0154 COG1984
COA1_HUMAN	COG0439 COG4799	EPT1_YEAST	COG5050
COA2_HUMAN	COG0439 COG4799	FAS_YEAST	COG0294
COASY_HUMAN	COG1019 COG0237	FAS1_YEAST	COG0331 COG4981
ECHA_HUMAN	COG1250	FAS2_YEAST	COG4982 COG0304
ECHP_HUMAN	COG1250	FDFT_YEAST	COG1562
ENPP1_HUMAN	COG1524	FOX2_YEAST	COG1028
ENPP3_HUMAN	COG1524	GAL10_YEAST	COG2017 COG1087
ERN1_HUMAN	COG0515	GDE_YEAST	COG3408 COG0366
ERN2_HUMAN	COG0515	GGPPS_YEAST	COG0142
F261_HUMAN	COG0406	HFA1_YEAST	COG0439 COG4799 COG0511
F262_HUMAN	COG0406	HIS2_YEAST	COG0141
F263_HUMAN	COG0406	HIS5_YEAST	COG0107
F264_HUMAN	COG0406	IRE1_YEAST	COG0515
FAS_HUMAN	COG3321	LKHA4_YEAST	COG0308
FCL_HUMAN	COG0451	MET17_YEAST	COG2873
FDFT_HUMAN	COG1562	NPP1_YEAST	COG1524
FOLH1_HUMAN	COG2234	NPP2_YEAST	COG1524
FTCD_HUMAN	COG3643	OGG1_YEAST	COG0122
G6PE_HUMAN	COG0364	PABS_YEAST	COG0147
GDE_HUMAN	COG3408	PUR2_YEAST	COG0150 COG0151
GEPH_HUMAN	COG0303	PUR91_YEAST	COG0138
GGPPS_HUMAN	COG0142	PUR92_YEAST	COG0138
GLCNE_HUMAN	COG1940 COG0381	PYC1_YEAST	COG1038
LKHA4_HUMAN	COG0308	PYC2_YEAST	COG1038
MAAI_HUMAN	COG0625	PYR1_YEAST	COG0458 COG0505 COG0540 COG0044
MCE1_HUMAN	COG5226	THI6_YEAST	COG0352
MGA_HUMAN	COG1501	TRNL_YEAST	COG5324
MTDC_HUMAN	COG0190	TRPG_YEAST	COG0512
NALD2_HUMAN	COG2234	YL345_YEAST	COG0406
NALDL_HUMAN	COG2234	POK5_HUMAN	COG2801
NCOAT_HUMAN	COG0454	POK6_HUMAN	COG2801
NTHL1_HUMAN	COG0177	PRDX6_HUMAN	COG0450
OGG1_HUMAN	COG0122	PUR2_HUMAN	COG0150 COG0151
P5CS_HUMAN	COG0014	PUR6_HUMAN	COG0152
PAPS1_HUMAN	COG0529	PUR9_HUMAN	COG0138
PAPS2_HUMAN	COG0529	PYC_HUMAN	COG1038
PNKP_HUMAN	COG0241	PYR1_HUMAN	COG0458

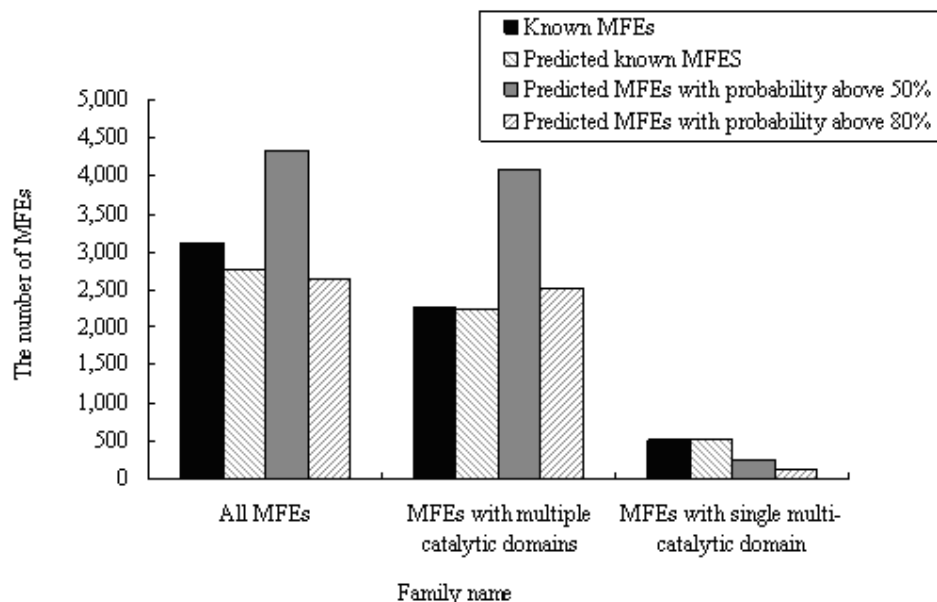
POK10_HUMAN	COG2801	PYR5_HUMAN	COG0284
POK17_HUMAN	COG2801	SUIS_HUMAN	COG1501
POK2_HUMAN	COG2801	SYEP_HUMAN	COG0008 COG0442
POK20_HUMAN	COG2801		

### 3.2.3. Identification of novel MFEs

Identification of novel MFEs is one of the best ways for understanding of multiple functionalities of enzymes; unfortunately, traditional experimental approaches or *in silico* homology-based methods have difficulties in proper and efficient identification of novel MFEs due to the variety of primary protein sequences. In present study, a SVM model was trained and optimized in an enriched way, which has been described previously [174]. The model was further evaluated independently by 322 positive and 2,481 negative data, achieving sensitivity of 94.1%, specificity of 99.0%, positive prediction accuracy of 92.6% and overall accuracy of 98.5% (Table 3-1). The model was then applied to screen the ExPASy Enzyme database for the identification of novel MFEs.

Overall, 2,641 novel MFEs with probability (the confidence of the prediction) >80% (4,320 with probability >50%) were identified from 91,140 enzymes of ExPASy Enzyme database, excluding 2,806 known MFEs. Among these novel MFEs, 126 MFEs contain single multi-activity domain, whereas 2,515 MFEs contain at least two catalytic domains (Figure 3-8). The complete list of known and predicted MFEs is searchable at <http://bioinf.xmu.edu.cn/databases/MFEs/index.htm>.

**Figure 3-8 Statistics of currently known MFEs and predicted MFEs by screening the ExPASy Enzyme database. Totally there are 3,120 currently known MFEs, including 2,279 enzymes of multiple catalytic domains (MCD-MFEs), 572 known enzymes of single multi-catalytic domain (SMAD-MFEs). Totally, 2,641 novel MFEs with prediction probability >50% (4,320 with probability >80%), including 2,515 MCD-MFEs (4,075 with probability >80%) and 126 SMAD-MFEs (245 with probability >80%) were identified from 91,140 enzymes of ExPASy Enzyme database**



### 3.2.4. Contribution of physicochemical properties in the classification of MFEs

Not all enzymes 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 sequence, mediate the same mechanistic role, due to the flexibility at the active site [175]. This plasticity is unlikely to be sufficiently described by commonly used structural and physicochemical features. Therefore, the recognition of features that properly describe this plasticity may further improve the accuracy of identification of MFEs by statistical learning methods like SVM and artificial neural networks. In this work, a total of nine feature properties were used to describe physicochemical characteristics of each protein, which have been routinely used for



the prediction of proteins of different structural and functional classes [116, 136, 158, 176-178]. It was acknowledged that not all these 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 [176]. It is thus of interest to examine which feature properties play more prominent role in the characterization of MFEs.

In an earlier study, contribution of individual feature property to protein classification was investigated by separately conducting classification using each feature property [176]. Similar approach was employed in present study. It was found that the charge, polarizability, hydrophobicity, and solvent accessibility play more prominent role than other feature properties. Previous studies found that some MFEs, i.e. ADP-ribosyl cyclase and CD38 can switch functions at different pH, indicating the importance of polarity, charge distribution and solvent accessibility in determining their multifunctionality [179]. Multiple protein-interacting modules of some MFEs, i.e. High-voltage-activated  $\text{Ca}^{2+}$  channels, involve in hydrophobic interactions [180]. Some MFEs, i.e. neuronal nitric oxide synthase, have large solvent-exposed hydrophobic surface that contains a cavity rimmed with charges [181]. Therefore, sequence features we used in this study seem to be reasonable to catch underlying common characteristics of MFEs.

### 3.3. Server for identification of multifunctional enzyme (SIME)

Server for identification of multifunctional enzyme (SIME), is a web-server that uses SVM for predicting multifunctional enzyme with multiple catalytic domain or with single catalytic domain based on sequence-derived structural and physicochemical properties. SIME could be accessed at <http://jing.cz3.nus.edu.sg/cgi-bin/sime.cgi>

(Figure 3-9). The sequence of a protein, in RAW format and containing no non-amino acid letters, can be input in a window provided. The average computing time is about 5 seconds for a typical protein and the computed result is displayed in a separate window. If the input protein is provided to SIME, the result page would show the length of the query sequence and whether it is predicted to be a multifunctional enzyme with multiple catalytic domains or with single catalytic domain (Figure 3-10a, Figure 3-10b, Figure 3-10c). If the input sequence contains invalid character or abnormal composition such as long stretch of consecutive single letters, then an error message of “Your input sequence is not a valid sequence” will be prompted.

**Figure 3-9 SIME interface.** The sequence of a protein, in RAW format and containing no non-amino acid letters, can be input in a window provided.

BioInfo & Drug Design   Databases   Softwares   Arts   Teaching   Research   Links

**SIME**   **BIDD**  
Bioinformatics and Drug Design group

Bioinformatics & Drug Design Group [BIDD]  
**SIME : Server for Identification of Multi-function Enzyme**

The sequence **MUST** be provided in **RAW** format.

SEQUENCE

MTVLKPSHWVLAELADGLPQHVSQIARMADMKPQLNGFWQQMPAHTIGLLR

Run SIME   Reset Sequence

Preliminary: Use of SIME for commercial purposes is not allowed.

**Figure 3-10a** Result page of SIME showing that a query sequence is predicted as a multifunctional enzyme with multiple catalytic domain



Bioinformatics & Drug Design Group [BIDD]  
**SIME : Server for Identification of Multi-function Enzyme**

---

Welcome to SIME network service

Query=[Sequence](#)  
Length=592 amino acids

**Classification Done.**

Based on Support Vector Machine classification,


This sequence is predicted to be	R-Value	P-Value(%)
Multi-Functional Enzyme with Multiple Catalytic Domain	0.97	99.9

Preliminary: Use of SIME for commercial purposes is not allowed.

Please click [here](#) to multi-function enzyme database.

Last update: Dec 30,2007

**Figure 3-10b** Result page of SIME showing that a query sequence is predicted as a multifunctional enzyme with single catalytic domain



Bioinformatics & Drug Design Group [BIDD]  
**SIME : Server for Identification of Multi-function Enzyme**

---

Welcome to SIME network service

Query=[Sequence](#)  
Length=346 amino acids

**Classification Done.**

Based on Support Vector Machine classification,


This sequence is predicted to be	R-Value	P-Value(%)
Multi-Functional Enzyme with Single Catalytic Domain	1.00	99.9

Preliminary: Use of SIME for commercial purposes is not allowed.

Please click [here](#) to multi-function enzyme database.

Last update: Dec 30,2007

**Figure 3-10c Result page of SIME showing that a query sequence is predicted as non multifunctional enzyme**



Bioinformatics & Drug Design Group [BIDD]  
**SIME : Server for Identification of Multi-function Enzyme**

---

**Welcome to SIME network service**

Query=[Sequence](#)  
 Length=82 amino acids

**Classification Done.**

Based on Support Vector Machine classification,

This sequence is predicted to be	R-Value	P-Value
NON-Multi-Functional Enzyme	-0.96	99.9

Preliminary: Use of SIME for commercial purposes is not allowed.  
 Please click [here](#) to multi-function enzyme database.  
 Last update: Dec 30,2007

### 3.4. MFEs database

A database was also developed to provide detailed information of known and putative MFEs, such as enzyme name, EC number, description of function, Pfam domain, prediction status and subtypes (Figure 3-11). The database can be accessed at <http://bioinf.xmu.edu.cn/databases/MFEs/index.htm>, as shown in Figure 3-12. Users can search MFEs by name, accession number, EC number and Pfam number (Figure 3-13). The database also provides the analysis of known and potential MFEs, which would help users to further investigate the underlying mechanisms of MFEs (Figure 3-14).

**Figure 3-11** Graphical searching interface of MFEs database.

The screenshot displays the MFEs database interface. At the top, there is a header with the BIRG logo (Xiamen University), the MFEs logo (bioinf.xmu.edu.cn), and the text 'Multi-Functional Enzymes'. Navigation links include HOME, KDBI, DITOP, GEPS, and CYTOSVM. Below the header is a menu bar with HOME, SEARCH, BROWSE, and STATISTICS. The main content area is titled 'DETAILED INFORMATION' and contains a table with the following data:

General Information	
Unipro AC	Q34520
Protein name	ATP phosphoribosyltransferase
EC number	EC 2.4.2.17
Organism	Bacillus subtilis. [NCBI_TaxID=1423;]
Features	
Function	Catalyzes the condensation of ATP and PRPP to form N <sup>5</sup> -5'-phosphoribosyl-ATP (PR-ATP). Has a crucial role in the pathway because the rate of histidine biosynthesis seems to be controlled primarily by regulation of hisG enzymatic activity (By similarity).
Pfam	PF01634 HisG;
MFEs Type	
MFE type	multiple catalytic domains
Status	predicted
Probability	56.52764178

**Figure 3-12** Graphical user interface of MFEs database.

The screenshot displays the MFEs database graphical user interface. At the top, there is a header with the BIRG logo (Xiamen University), the MFEs logo (bioinf.xmu.edu.cn), and the text 'Multi-Functional Enzymes'. Navigation links include HOME, KDBI, DITOP, GEPS, and CYTOSVM. Below the header is a menu bar with HOME, SEARCH, BROWSE, and STATISTICS. The main content area is titled 'Multi-Functional Enzymes' and features a 'Quicksearch' input field with a 'Submit' button. Below this is a section titled 'About MFEs' with the following text:

**Multi-Functional Enzymes (MFEs)** are a group of enzymes that perform at least two distinct enzymatic activities. They are further divided into two classes. One is MFEs with single multi-catalytic domain (SMCD), and the other is MFEs with multiple catalytic domains (MCD).



**Figure 3-13 Graphical searching interface of MFEs database**

**Multi-Functional Enzymes**

Quicksearch

ID search

- ☒ Swiss AC egp13442
- ☐ EC eg1.1.1.1
- ☐ Pfam eg PF00009

**Figure 3-14 Biological analysis results interface of MFEs**

**The supplemental materials**

Figures of MFEs

- [Figure 1 The result of screening the ExPASy Enzyme database.](#)
- [Figure 2 Distribution of MFEs in different kingdoms.](#)
- [Figure 3 Distribution of known enzymes of single multi-catalytic domain in top 10 Pfam domain families.](#)
- [Figure 4 Distribution of known enzymes of multiple catalytic domains in top 10 Pfam domain families.](#)
- [Figure 5 Distribution of different classes of known and predicted potential MFEs.](#)
- [Figure 6 The percentage of known MFEs involved in different number of biological pathways.](#)
- [Figure 7 The statistics of known enzymes of multiple catalytic domains \(MCD-MFEs\) with KEGG ontology \(KO\).](#)

### 3.5. Summary

The characterization and identification of multifunctional enzymes (MFEs) attracts recent interests of biochemical communities for better understanding of the common mechanism underlying the crosstalk of various cellular processes. In the present study, we collected and systematically analyzed MFEs by grouping them into two categories: MFEs with multiple catalytic domains (MCD-MFEs) and MFEs with single multi-activity domain (SMAD-MFEs). No obvious evidence show complex life forms like human prefer more MFEs than simple life form like yeast. Combined with pathway ontology analysis showing that the majority of MFEs are involved in several essential cellular processes, MFEs are most likely ancestor enzymes in primitive life forms. They may play key roles in catalyzing essential cellular processes so that their functions are well conserved across species. This is also supported by the evidence that almost half of MFEs participate in only one biological pathway. At the meantime, new MFEs are generated by diversification and specification in various forms of genetic variations like gene fusion and exon shuffling, since there are about half of MFEs involving in as more as five independent pathways. The alpha and beta fold topology is found to be most favored to preserve multiple functions of MFEs during evolution. The analysis of feature contribution indicates that four physiochemical properties are most important to characterize MFEs.

In this study, a sequence-based machine learning system, SVM classifier, was also constructed, which predicted 2,641 potential MFEs with statistic confidence within ExPASy enzyme database, including 2,515 MCD-MFEs and 126 SMAD-MFEs. This work introduced a new and efficient approach to identify and annotate enzymes with multiple activities in large scale. Several factors may more or less affect its

performance. One is the diversity of protein samples used for developing SVM classification system. It is likely that not all possible types of MFEs and non-MFEs are adequately represented in the training set. This can be improved with the availability of more diverse protein sequences and improved knowledge about MFEs. A broad spectrum of MFEs of diverse functions may also reduce the performance of our SVM classification system to some extent. An online classification system for novel MFEs identification and a database of known and putative MFEs were also constructed for public access, at <http://jing.cz3.nus.edu.sg/cgi-bin/sime.cgi> and <http://bioinf.xmu.edu.cn/databases/MFEs/index.htm> respectively.



## **4. Prediction of disease related proteins by support vector machine**

### **4.1. Prediction of antimicrobial proteins**

Antimicrobial peptides play important roles in innate immune defense against microbial infection. They were derived from antimicrobial proteins (AMPs) upon microbial attacks. The identification of AMPs will thus facilitate the search of therapeutic targets, which would help to design better drugs to fight against microbes. Due to their function and sequence diversity, it is desirable to develop alternative methods irrespective of sequence similarity to predict AMPs. This work explores the use of support vector machine (SVM) as such a method.

#### **4.1.1. Selection of antimicrobial proteins and non-antimicrobial proteins**

A total of 986 AMPs used in this study were collected from a comprehensive search of Swiss-Prot database at <http://us.expasy.org/sprot/> [162]. The distribution of these proteins in top 10 host species is given in Table 4-1. It could be found that these proteins are from diverse range of species. These AMPs were later divided into two subclasses, antibiotic proteins and fungicide proteins.

All distinct members in each class were used to construct a positive dataset for the corresponding SVM prediction system. A negative dataset, representing non-class members, was selected from seed proteins of the domain families in Pfam database [163] excluding those that contain at least one AMP. Members in other antimicrobial

classes were included in the negative dataset if they are not a member of the class being studied.

These proteins were further divided into separate training, testing and independent evaluation sets, which were used for developing SVM models, fine-tuning parameters and evaluating performance, respectively, as described in the previous chapter. The statistics of AMPs and subclasses is given in Table 4-2.

**Table 4-1 Distribution of AMPs in top 10 host species**

	Species
<b>List of top 10 species and number of AMPs in each species</b>	<i>Homo sapiens</i> (65)
	<i>Mus musculus</i> (59)
	<i>Bombina maxima</i> (50)
	<i>Bos Taurus</i> (44)
	<i>Pan troglodytes</i> (37)
	<i>Sus scrofa</i> (23)
	<i>Drosophila melanogaster</i> (20)
	<i>Macaca mulatta</i> (20)
	<i>Penaeus vannamei</i> (18)
	<i>Rattus norvegicus</i> (17)

**Table 4-2 Statistics of the datasets and prediction accuracy of individual class of AMPs**  
The predicted results are given in TP, FN, TN, FP, sensitivity  $SE=TP/(TP+FN)$ , specificity  $SP=TN/(TN+FP)$ , positive prediction value  $PPV=TP/(TP+FP)$  and overall accuracy  $Q=(TN+TP)/(TP+FN+TN+FP)$ . The number of members and non-members in the testing and independent evaluation sets is TP+FN or TN+FP respectively.

Classes	Training set		Testing set				Independent evaluation set							
	positive	negative	positive		negative		positive			negative			Q (%)	PPV (%)
			TP	FN	TN	FP	TP	FN	SE(%)	TN	FP	SP(%)		
Antibiotic proteins	457	3545	148	18	12185	4	50	4	92.6	7874	12	99.8	96.2	92.5
Fungicide proteins	124	3424	58	3	12408	1	9	2	81.8	7926	9	99.9	90.8	87.0
All antimicrobial proteins	631	1649	232	4	14132	4	108	11	90.8	7230	28	99.6	99.5	91.4

#### 4.1.2. Prediction performance for antimicrobial proteins

The statistics of prediction results is given in Table 4-2. The predicted sensitivity for fungicide proteins ( $\bar{n}=11$ ), antibiotic proteins ( $\bar{n}=13$ ) and all AMPs ( $\bar{n}=20$ ) is 81.1%, 92.6% and 90.8% respectively, while the corresponding predicted specificity is 99.9%, 99.8% and 99.6% respectively. The predicted PPV for fungicide proteins, antibiotic proteins and all AMPs is 87.0%, 92.5% and 91.4% respectively. These results suggest that SVM is capable to predict AMPs with reasonable high accuracy.

The performance of our prediction system was further tested by using 5-fold cross validation [182]. In 5-fold cross validation, the group of positive and negative data is each randomly divided into five subsets of approximately equal size respectively. Four of the subsets are used as the training set, and the remaining subset is used as the testing set for AMPs and non-AMPs respectively. This process is repeated five times such that every subset is used as the test set once. The result of 5-fold cross validation is given in Table 4-3. The average sensitivity for fungicide proteins, antibiotic proteins and all AMPs is 74.9%, 86.7% and 80.2% respectively, and the average specificity is 99.9%, 99.9%, and 99.6% respectively, which are comparable to those derived from the use of independent set. Therefore both validation methods give roughly similar prediction performances.

**Table 4-3 Statistics of prediction accuracy of antimicrobial proteins measured by 5-fold cross validation**

Classes	Cross validation	TP	FN	SE (%)	TN	FP	SP (%)	Q (%)
Fungicide proteins	1	26	8	76.5	4752	1	99.9	99.8
	2	23	10	69.7	4754	0	100.0	99.8
	3	29	4	87.8	4753	1	99.9	99.9
	4	23	11	67.6	4750	3	99.9	99.7
	5	24	9	72.7	4754	0	100.0	99.8
	Average			74.9			99.9	99.8
	SD			2.69			0.07	0.0
Antibiotic proteins	1	93	19	83.0	4723	2	99.9	99.6
	2	98	16	86.0	4718	5	99.9	99.6
	3	97	17	85.1	4722	1	100.0	99.6
	4	100	13	88.5	4714	10	99.8	99.5
	5	101	10	91.0	4718	7	99.8	99.6
	Average			86.7			99.9	99.6
	SD			5.65			0.07	0.0
All AMPs	1	152	46	76.8	4593	15	99.7	98.7
	2	159	38	80.7	4583	26	99.4	98.7
	3	160	37	81.2	4593	16	99.6	98.9
	4	163	35	82.3	4587	21	99.5	98.8
	5	157	39	80.1	4591	18	99.6	98.8
	Average			80.2			99.6	98.8
	SD			2.33			0.07	0.07

#### 4.1.3. Prediction of novel antimicrobial proteins

To assess the ability of our model to identify novel AMPs, especially those without homologues to known AMPs, Swiss-Prot database [162] was searched for finding those proteins having no single homologous protein in the database based on BLAST [106] results. A similarity E-value threshold of 0.1 was used for homologues search to ensure maximum exclusion of proteins that have a homologue. It is found that 69 out of 70 proteins were correctly predicted by SVM as AMPs (Table 4-4). Only one protein, acanthoscurrin-2 precursor, was incorrectly predicted as non-antimicrobial protein. Investigation of amino acid composition of acanthoscurrin-2 precursor found more than 70% amino acids of its sequence are glycines, whose composition is much higher than other AMPs, which may account for its incorrect classification. Our SVM

classification system appears to show reasonably good capability for predicting AMPs based on the set of proteins tested, especially for those without homologues in known protein databases.

**Table 4-4 Prediction results of novel antimicrobial proteins by SVM-Prot, where “+” represents proteins correctly predicted as antimicrobial proteins, and “-” represents proteins incorrectly predicted as non-antimicrobial proteins.**

Protein Name	Gene name	Swiss-Prot accession number	SVM prediction status
Acanthoscurrin-1 precursor	acantho1	Q8I948	+
Acanthoscurrin-2 precursor	acantho2	Q8I6R7	-
Antifungal protein precursor	afp	P17737	+
Antimicrobial peptide 1 precursor	none	P80915	+
Antimicrobial peptides precursor	AMP	O24006	+
Bacteriocin amylovorin-L precursor	amyL	P80696	+
Antibacterial substance A	none	P01548	+
Antifungal protein precursor	none	Q08617	+
Armadillidin precursor	none	Q64HC7	+
Beta-defensin 50 precursor	Defb50	Q6TU36	+
Colicin-M	cma	P05820	+
Colicin-V precursor	cvaC	P22522	+
Cicadin	none	P83282	+
Beta-defensin 129 precursor	DEFB129	Q9H1M3	+
Dermcidin precursor (Preproteolysin)	DCD	P81605	+
Defensin-1 precursor (Cll-dlp)	none	Q6GU94	+
Big defensin	none	P80957	+
Dermaseptin AA-2-5 precursor	none	O93222	+
Dermaseptin PD-2-2 precursor	none	O93452	+
Adenoregulin precursor	ADR	P31107	+
Dermaseptin PD-3-3 precursor	none	O93453	+
Dermaseptin PD-3-6 precursor	none	O93454	+
Dermaseptin AA-3-4 precursor	none	O93225	+
Dermaseptin PD-3-7 precursor	none	O93455	+
Dermaseptin AA-3-6 precursor	none	O93226	+
Dermaseptin DRG1 precursor	DRG1	Q90ZK3	+
Dermaseptin DRG2 precursor	DRG2	Q90ZK5	+
Dermaseptin DRG3 precursor	DRG3	P81488	+
Galensin precursor	none	Q90W78	+
Gloverin	none	P81048	+
Gomesin precursor	none	P82358	+
Nonhistone chromosomal protein H6	none	P02315	+
Halocin-H4 precursor	halH4	Q48236	+
Holotricin-3 precursor	none	Q25055	+
Hymenoptaecin precursor	none	Q10416	+
Ixosin	none	Q2LKC9	+
Bacteriocin lactocin-S precursor	lasA	P23826	+
Uncharacterized protein in cib 5'region	none	P04481	+

Bacteriocin lactacin-F subunit lafX precursor	lafX	Q48509	+
Lantibiotic lactacin 3147 A1 precursor	ltnA1	O87236	+
Lantibiotic lactacin 3147 A2 precursor	ltnA2	O87237	+
Lantibiotic epilancin precursor	elkA	Q57312	+
Lantibiotic Pep5 precursor	pepA	P19578	+
Locustin	none	P83428	+
B-enzyme (EC 3.2.1.17)	lyzB	P10773	+
Lysozyme (EC 3.2.1.17)	LYS4	Q27650	+
Bacteriocin microcin B17 precursor (MccB17)	mcbA	P05834	+
Microcin H47 precursor (MccH47)	mchB	P62531	+
Microcin H47 precursor (MccH47)	mchB	P62530	+
Microcin J25 precursor (MccJ25)	mcjA	Q9X2V7	+
Lantibiotic mersacidin precursor	mrsA	P43683	+
Metchnikowin precursor	Mtk	Q24395	+
Neuropeptide-like protein 31 precursor	nlp-31	O44662	+
Neuropeptide-like protein 33 precursor	nlp-33	Q95ZN4	+
Perinerin	none	P84117	+
Phylloseptin-12 precursor (PS-12)	ppp-12	Q17UY9	+
Phylloxin precursor	PLX	P81565	+
Propionacin-F precursor	pcfA	Q6E3K9	+
PYL <sub>a</sub> /PGL <sub>a</sub> precursor	none	Q99134	+
Scarabaecin precursor	scar	Q86SC0	+
Protein spaetzle precursor	spz	P48607	+
Stomoxyn precursor	none	Q8T9R8	+
SPBc2 prophage-derived lantibiotic sublancin-168 precursor	sunA	P68577	+
Tachystatin-A2 precursor	none	Q9U8X3	+
Temporin-B precursor	none	P79874	+
Temporin-G precursor	none	P79875	+
Tenecin-3 precursor	none	Q27270	+
Lysozyme (EC 3.2.1.17)	XV	P13559	+
Protein P35 (Holin)	XXXV	Q3T4L9	+
Bacteriocin lactacin-F subunit lafA precursor	lafA	P24022	+

With the application of AMPer tools [16], more information of antimicrobial proteins and peptides could be found. Their data were not included in our dataset and thus is useful as truly independent testing data for our model. 177 antimicrobial proteins with length above 50 amino acids were obtained from AMPer database to be predicted by our SVM model. The results showed that 105 of 177 proteins were predicted as antimicrobial proteins. Considering that AMPer also includes some predicted results, the picking-up rate of around 60% suggests that our SVM model is capable of

identifying antimicrobial proteins at a reasonably true positive rate. The result is listed in Table 4-5.

Our model was also applied to scan the whole human genome sequences downloaded from Ensembl site. 411 proteins were predicted as potential antimicrobial proteins from a total of 43,570 human proteins (less than 1%), among which 56 have been experimentally verified as antimicrobial proteins.

**Table 4-5 List of prediction results of 177 antimicrobial proteins in AMPer database (“+” represents proteins correctly predicted as antimicrobial proteins, and “-” represents proteins incorrectly predicted as non-antimicrobial proteins)**

Swiss-Prot Entry Name	Protein Name	Prediction result by our SVM model
10KD_VIGUN	10 kDa protein precursor (Clone PSAS10)	+
SRP_SOYBN	84 kDa sulfur-rich protein precursor (SE60 protein)	+
Q9NL71_CAEEL	ABF-2 precursor (Antibacterial factor related protein 2)	+
Q6KFT8_SAGLB	Alpha defensin (Fragment)	+
Q6KFT9_ATEGE	Alpha defensin (Fragment)	+
Q9TTZ8_MACMU	Alpha-defensin 2	+
CAS2_BOVIN	Alpha-S2-casein precursor	+
Q8MVY9_GALME	Antifungal peptide gallerimycin	+
Q9FPM3_MEDSA	Antifungal protein precursor	+
Q71QD7_PINSY	Antimicrobial peptide 4 (Antimicrobial peptide 2)	+
Q8WTD3_GLOMR	Antimicrobial peptide attacin AttA	+
Q9FR52_CAPBU	Antimicrobial peptide shep-GRP	+
Q71U16_AMAHP	Antimicrobial protein	+
APOA2_HORSE	Apolipoprotein A-II	+
APOA2_MACMU	Apolipoprotein A-II	+
P90683_ASCSU	ASABF precursor (ASABF-alpha)	+
O19040_SHEEP	Bactinecin 6	+
O97942_CAPHI	Beta defensin-2 precursor	+
Q865P6_HORSE	Beta-defensin-1	+
Q6TN20_CANFA	Cathelicidin	+
Q95VE8_MUSDO	Cecropin 1	+
CCKN1_XENLA	Cholecystokinin type 1 precursor	+
CCKN_RAT	Cholecystokinins precursor (CCK)	+
Q71KM5_RAT	CRAMP (Fragment)	+
Q9BK52_ACALU	Defensin 1 precursor	+
P82378_STOCA	Defensin 1a	+
P82379_STOCA	Defensin 2a	+
Q86MY3_RHOPR	Defensin A	+

Q9BLJ3_ORNMO	Defensin A	+
Q9BLJ4_ORNMO	Defensin B	+
Q8MY08_ORNMO	Defensin C	+
Q86MY1_RHOPR	Defensin C	+
O77217_AEDAL	Defensin D (Fragment)	+
Q8MY07_ORNMO	Defensin D	+
Q3L180_MOUSE	Defensin related cryptdin 26	+
Q6XL51_TRIFG	Defensin	+
Q6RSS6_PICGL	Defensin	+
D230_PEA	Disease resistance response protein 230 precursor	+
EMBP_HUMAN	Eosinophil granule major basic protein precursor (MBP)	+
EMBP_MOUSE	Eosinophil granule major basic protein precursor (MBP)	+
EMBP_RAT	Eosinophil granule major basic protein precursor (MBP)	+
EMBP_CRIGR	Eosinophil granule major basic protein precursor (MBP)	+
THGF_HELAN	Flower-specific gamma-thionin precursor (Defensin SD2)	+
THG1_NICPA	Gamma-thionin 1 precursor	+
THG_PETIN	Gamma-thionin homolog PPT precursor	+
Q71MD5_MYXGL	Hematopoietic antimicrobial peptide-29 precursor (Fragment)	+
Q71MD7_MYXGL	Hematopoietic antimicrobial peptide-37 precursor	+
Q8T3C5_CAEEL	Hypothetical protein abf-6 (ABF-6)	+
KAB7_OLDAB	Kalata-B7 precursor	+
KNL2_BOMMX	Kininogen-2 precursor (BMK-2)	+
LCR69_ARATH	Low-molecular-weight cysteine-rich protein LCR69 precursor	+
LCR72_ARATH	Low-molecular-weight cysteine-rich protein LCR72 precursor	+
P82017_CAPHI	MAP34-A protein (MAP34-B protein)	+
MEL_APIME	Melittin precursor	+
MEL_VESVN	Melittin precursor	+
MEL_VESMC	Melittin precursor	+
MEL_VESMG	Melittin precursor	+
MEL_APICC	Melittin precursor	+
MEL_POLHE	Melittin precursor	+
MEL_APICE	Melittin precursor	+
P79360_SHEEP	Myeloid antimicrobial peptide	+
O62840_HORSE	Myeloid cathelicidin 1 precursor	+
O62841_HORSE	Myeloid cathelicidin 2 precursor	+
Q9Y0B1_MYTGA	Mytilin B antimicrobial peptide	+
PFPN_ENTHI	Nonpathogenic pore-forming peptide precursor (APNP)	+
OSMO_TOBAC	Osmotin precursor	+
OSL3_ARATH	Osmotin-like protein OSM34 precursor	+
OS13_SOLCO	Osmotin-like protein OSML13 precursor (PA13)	+
OS35_SOLCO	Osmotin-like protein OSML15 precursor (PA15)	+
OS81_SOLCO	Osmotin-like protein OSML81 precursor (PA81)	+
OLPA_TOBAC	Osmotin-like protein precursor (Pathogenesis-related protein PR-5d)	+
BPT1_BOVIN	Pancreatic trypsin inhibitor precursor	+
PR5_ARATH	Pathogenesis-related protein 5 precursor (PR-5)	+
PRR3_JUNAS	Pathogenesis-related protein precursor	+
PRR1_TOBAC	Pathogenesis-related protein R major form precursor	+
PRR2_TOBAC	Pathogenesis-related protein R minor form precursor (PR-R)	+



PLF4_HUMAN	Platelet factor 4 precursor (PF-4)	+
Q90WJ0_PSEAM	Pleurocidin-like prepropolypeptide (Fragment)	+
Q90VX5_PSEAM	Pleurocidin-like prepropolypeptide (Fragment)	+
Q9U8G5_ENTDI	Pore-forming protein isoform B precursor	+
Q9U8G4_ENTDI	Pore-forming protein isoform C precursor	+
CAER5_XENLA	Preprocaerulein clone PXC202 precursor	+
CAER2_XENLA	Preprocaerulein type I' precursor (Fragment)	+
LCR68_ARATH	Probable low-molecular-weight cysteine-rich protein LCR68 precursor	+
P322_SOLTU	Probable protease inhibitor P322 precursor	+
PENK_CAVPO	Proenkephalin A precursor	+
PENK_MESAU	Proenkephalin A precursor	+
PENK_RAT	Proenkephalin A precursor	+
PENK_HUMAN	Proenkephalin A precursor	+
PENK_MOUSE	Proenkephalin A precursor	+
LEVI_XENLA	Prolevitide precursor	+
P21_SOYBN	Protein P21	+
Q9FZ31_ARATH	Putative antifungal protein	+
Q8WRP5_PENVA	Putative antimicrobial peptide (Crustin P)	+
Q9XZN6_ANOGA	Putative infection responsive short peptide	+
LCR66_ARATH	Putative low-molecular-weight cysteine-rich protein LCR66 precursor	+
Q948Z4_SOLTU	Snakin-1	+
TEMH_RANTE	Temporin-H precursor	+
THM2_THADA	Thaumatococcus precursor (Thaumatococcus II)	+
TLP1_PRUPE	Thaumatococcus-like protein 1 precursor (PpAZ44)	+
TLP1_PYRPY	Thaumatococcus-like protein 1 precursor	+
TP1A_MALDO	Thaumatococcus-like protein 1a precursor (Allergen Mal d 2)	+
TP1B_MALDO	Thaumatococcus-like protein 1b	+
TLP2_PRUPE	Thaumatococcus-like protein 2 precursor (PpAZ8)	+
TLP_PRUAV	Thaumatococcus-like protein precursor	+
Q9NL72_CAEEL	ABF-1 precursor (Antibacterial factor related protein 1)	-
O96447_LUMRU	Antimicrobial peptide lumbricin1	-
Q9U6U0_MYTGA	Antimicrobial peptide MGD2b	-
APOA2_HUMAN	Apolipoprotein A-II precursor	-
APOA2_PANTR	Apolipoprotein A-II precursor	-
APOA2_MOUSE	Apolipoprotein A-II precursor	-
APOA2_RAT	Apolipoprotein A-II precursor	-
APOA2_MACFA	Apolipoprotein A-II precursor	-
Q9FER3_MAIZE	Basal layer antifungal peptide precursor	-
Q9FER2_MAIZE	Basal layer antifungal peptide precursor	-
Q9FER1_MAIZE	Basal layer antifungal peptide precursor	-
Q9FER0_MAIZE	Basal layer antifungal peptide precursor	-
Q30KT2_CANFA	Beta-defensin 122	-
Q9Y0X4_MESMA	BmK3 (Bradykinin-potentiating peptide)	-
Q6QLQ5_CHICK	Cathelicidin (Fowlicidin-1)	-
CATG_HUMAN	Cathepsin G precursor (EC 3.4.21.20) (CG)	-
CCKN2_XENLA	Cholecystokinin type 2 precursor	-
CCKN_MOUSE	Cholecystokinin precursor (CCK)	-
CCKN_RANCA	Cholecystokinin precursor (CCK)	-
CCKN_MACFA	Cholecystokinin precursor (CCK)	-
CCKN_HUMAN	Cholecystokinin precursor (CCK)	-

CCKN_PIG	Cholecystokinins precursor (CCK)	-
CCKN_BOVIN	Cholecystokinins precursor (CCK)	-
CCKN_TRASC	Cholecystokinins precursor (CCK)	-
CCKN_CHICK	Cholecystokinins precursor (CCK)	-
CCKN_STRCA	Cholecystokinins precursor (CCK)	-
CCKN_PAROL	Cholecystokinins precursor (CCK)	-
CCKN_CARAU	Cholecystokinins precursor (CCK8)	-
CION_CIOIN	Cionin precursor	-
COLI_BOVIN	Corticotropin-lipotropin precursor	-
Q9BMA5_APIME	Defensin (Fragment)	-
Q9GYU6_AEDAL	Defensin (Fragment)	-
Q8WQZ3_MAMBR	Defensin	-
ELAF_HUMAN	Elafin precursor	-
ENV_SIVML	Envelope glycoprotein gp160 precursor (Env polyprotein)	-
ENV_HV1LW	Envelope glycoprotein gp160 precursor (Env polyprotein)	-
ENV_EIAVY	Envelope glycoprotein precursor (Env polyprotein)	-
HEMO_HYACE	Hemolin precursor	-
HEMO_MANSE	Hemolin precursor (P4 protein)	-
Q22690_CAEEL	Hypothetical protein abf-5 (ABF-5)	-
KAB2_OLDAL	Kalata-B2 precursor	-
KAB3_OLDAL	Kalata-B3/B6 precursor	-
Q8IX02_HUMAN	Lactoferrin (Fragment)	-
P91817_TACTR	Limulus factor D	-
LCR70_ARATH	Low-molecular-weight cysteine-rich protein LCR70 precursor	-
O18425_EISFO	Lysenin-related protein (Hemolysin)	-
O62842_HORSE	Myeloid cathelicidin 3 precursor	-
Q61903_MOUSE	Myeloid secondary granule protein	-
Q9BKM2_NAEFO	Naegleriapore A pore-forming peptide	-
Q9BKM1_NAEFO	Naegleriapore B pore-forming peptide	-
ANFB_HUMAN	Natriuretic peptides B precursor	-
Q91X12_CAVPO	Neutrophil cationic antibacterial polypeptide of 11 kDa	-
TRFE_CHICK	Ovotransferrin precursor	-
PRR3_JUNVI	Pathogenesis-related protein precursor	-
Q91322_RANCA	Pepsinogen precursor	-
PERF_MOUSE	Perforin-1 precursor (P1)	-
CAER1_XENLA	Preprocaerulein type-1 precursor (Preprocaerulein type I)	-
CAER3_XENLA	Preprocaerulein type-3 precursor (Preprocaerulein type III)	-
CAER4_XENLA	Preprocaerulein type-4 precursor (Preprocaerulein type IV)	-
CAER4_XENBO	Preprocaerulein type-4 precursor (Preprocaerulein type IV)	-
PENKA_XENLA	Proenkephalin A-A precursor	-
PENKB_XENLA	Proenkephalin A-B precursor (Fragment)	-
HEVE_HEVBR	Pro-hevein precursor (Major hevein)	-
RELX_HORSE	Prorelaxin precursor (RXN)	-
PSPB_HUMAN	Pulmonary surfactant-associated protein B precursor (SP-B)	-
Q6K209_ORYSA	Putative defensin (Os02g0629800 protein)	-
SECP_APIME	Secapin precursor	-
SCG1_BOVIN	Secretogranin-1 precursor	-
Q948Z5_SOLTU	Snakin2 precursor	-
P91818_TACTR	Tachycitin	-
THM1_THADA	Thaumatococin-1 (Thaumatococin I)	-

TLPH_ARATH	Thaumatococcus-like protein precursor	-
------------	---------------------------------------	---

#### 4.1.4. Contribution of feature properties

The contribution of feature properties was also studied. It was found that amino acid composition and hydrophobicity play more prominent roles than other feature properties. Previous studies suggested the binding sites of antimicrobial peptides usually appear in clusters in hydrophobic environments[183, 184]. On the other hand, specific amino acid composition and sequence motifs have been used for predicting antimicrobial peptides [185]. Previous experimental analysis revealed that fundamental composition and sequence motifs determine not only the biochemical properties of antimicrobial proteins, but also their three-dimensional configuration, which would profoundly influence their antimicrobial properties [18]. It seems that our prediction results are consistent with these experimental findings.

#### 4.1.5. Server for antimicrobial protein identification (SAPI)

A server for antimicrobial protein identification (SAPI) was also developed to facilitate the discovery of new AMPs. The server could be accessed at <http://jing.cz3.nus.edu.sg/cgi-bin/sapi.cgi>. The format of input sequences is the same as that of SIME server. The average computing time is about 3 seconds for a typical protein and the result is displayed with a separate window. Figure 4-1 shows the interface of SAPI. The result page shows that the input sequence is predicted as an antimicrobial protein (Figure 4-2).

**Figure 4-1** Graphical user interface for SAPI

Bioinformatics & Drug Design Group [BIDD]  
**SAPI: Server for Antimicrobial Protein Identification**

The sequence **MUST** be provided in [RAW](#) format.

SEQUENCE

Preliminary: Use of SAPI for commercial purposes is not allowed.  
 Last update: Jan 2, 2008

**130** visits

**Figure 4-2** Result page of SAPI showing that a query sequence is an antimicrobial protein.

Bioinformatics & Drug Design Group [BIDD]  
**SAPI: Server for Antimicrobial Protein Identification**

Welcome to SAPI network service

Query=[Sequence](#)  
 Length=78 amino acids

**Classification Done.**

Based on Support Vector Machine classification,

This sequence is predicted to be	R-Value	P-Value(%)
Antimicrobial protein - antibiotic	6	99.0

Preliminary: Use of SAPI for commercial purposes is not allowed.  
 Last update: Jan 2, 2008

[Go Back](#)

## 4.2. Prediction of antibiotic resistance proteins

Increasing antibiotic resistance has become a worldwide challenge to the clinical treatment of infectious diseases. The identification of antibiotic resistance proteins (ARPs) would be helpful in the discovery of new therapeutic targets and the design of novel drugs to control the potential spread of antibiotic resistance. In this work, a

support vector machines (SVM) based ARP prediction system was developed to facilitate the identification of proteins involved in antibiotic resistance.

#### 4.2.1. Selection of ARPs and non-ARPs

A total of 1,621 ARPs used in this study were retrieved from a comprehensive search of Swiss-Prot database at <http://us.expasy.org/sprot/> [162] using keyword “antibiotic resistance” followed by manual check that each protein is involved in antibiotic resistance. The distribution of these ARPs in top 10 bacteria species is given in Table 4-7 showing that these ARPs are from diverse sources.

All of these 1,621 ARPs were then used for constructing a positive dataset for the SVM classification system. The negative dataset, representing non-ARPs, was selected by a similar procedure as that of MFEs. In this procedure, representative proteins of curated protein families in the Pfam database [163] that contain no single known ARPs are selected as non-ARPs. These ARPs and non-ARPs are divided into separate training, testing and independent evaluation sets by the following procedure: first, proteins are clustered into groups based on their distance in the structural and physicochemical feature-space by using the hierarchical clustering method. One representative protein is randomly selected from each group to form a training set that is sufficiently diverse and broadly distributed in the feature space. One or up to 50% of the remaining proteins in each group is randomly selected to form the testing set. The selected proteins from each group are further checked to ensure that they are distinguished from the proteins in other groups. The remaining proteins are then designated as the independent evaluation set, which is also found in a reasonable level

of diversity. The statistics of ARPs and non- ARPs in each dataset is given in Table 4-8.

**Table 4-7 Distribution of ARPs in top 10 bacteria species**

Species	Number of ARPs
<i>Escherichia coli K12</i>	56
<i>Staphylococcus aureus</i>	35
<i>Pseudomonas aeruginosa</i>	33
<i>Bacillus subtilis</i>	26
<i>Salmonella typhimurium</i>	26
<i>Escherichia coli O157:H7</i>	26
<i>Staphylococcus aureus Mu50</i>	20
<i>Shigella flexneri</i>	20
<i>Staphylococcus aureus N315</i>	19
<i>Enterococcus faecalis</i>	19

#### 4.2.2. Prediction performance

The prediction result assessed by independent test (Table 4-8) shows that 277 of 313 ARPs, and 7099 of 7156 non-ARPs were successfully predicted by SVM, which means that the predicted sensitivity, specificity and overall accuracy are 88.5%, 99.2% and 98.7%, respectively.

**Table 4-8 Statistics of the datasets and prediction accuracy of ARPs ( $\sigma = 18$ )**

Data set	Antibiotic resistance proteins		Non-antibiotic resistance proteins		Prediction Accuracy				
	TP	FN	TN	FP	SE (%)	SP (%)	Q (%)	PPV (%)	MCC
Training	734	0	2372	0	100	100	100	100	1.0
Testing	572	2	13180	35	99.6	99.7	99.7	94.2	0.97
Independent evaluation	277	36	7099	57	88.5	99.2	98.7	82.9	0.85

The model was further tested by using 10-fold cross validation method [182]. As shown in Table 4-9, the average sensitivity, specificity and overall accuracy measured by 10-fold cross validation are 82.1%, 99.5%, and 98.3%, respectively, which are comparable with those derived from the use of independent test. Therefore, both validation methods give roughly similar prediction performance.

**Table 4-9 Statistics of accuracy for SVM prediction of antibiotic resistance proteins evaluated by using 10-fold cross validation**

Cross validation	TP	FN	SE (%)	TN	FP	SP (%)	Q (%)
1	127	36	78.0	2257	17	99.3	97.8
2	140	23	85.9	2264	10	99.6	98.6
3	135	28	82.8	2265	9	99.6	98.5
4	130	31	80.7	2264	12	99.5	98.2
5	132	30	81.5	2259	15	99.3	98.1
6	137	26	84.0	2262	11	99.5	98.5
7	128	33	79.5	2263	12	99.5	98.2
8	134	27	83.2	2266	9	99.6	98.5
9	136	26	84.0	2262	12	99.5	98.4
10	132	30	81.5	2264	10	99.6	98.4
Average			82.1			99.5	98.3
SD			2.47			0.21	0.42

#### 4.2.3. Prediction of novel ARPs

To assess the capability of the model to identify novel ARPs, especially those without homologous to available members, Swiss-Prot database[162] was searched to find ARPs without homologous proteins in the database based on BLAST[106]. A similarity E-value threshold of 0.1 was used for homologue search to ensure maximum exclusion of proteins that have homologues. As shown in Table 4-10, a total of 11 proteins were found from this process, all of which were correctly predicted as ARPs by our SVM classification system. Therefore, SVM appears to

show good capability for predicting ARPs based on the set of proteins tested, especially for those ARPs without homologues in known protein databases.

**Table 4-10 Prediction results of novel ARPs.**

Protein Name	Gene name	Swiss-Prot accession number
Blasticidin S-acetyltransferase	bls	P19997
Protein vanZ	vanZ	Q06242
Bacitracin transport permease protein BCRB	bcrB	P42333
Pentamidine resistance factor, mitochondrial	PNT1	P38969
Tunicamycin resistance protein	tmrB	P12921
Albicidin resistance protein	albR	P10488
Uncharacterized HTH-type transcriptional regulator in mcrB 3'region	ymcr	P43458
Mitomycin resistance protein mcrB	mcrB	P43486
Hygromycin-B kinase	hph	P00557
Acridine resistance protein	ac	P18924
Curromycin resistance protein	cre	P16961

#### 4.2.4. Scanning bacteria genomes

Our model was further tested by scanning two common bacterial genomes, *Escherichia coli K12* and *Staphylococcus aureus Mu50* to search for potential new ARPs. These two complete genomes are retrieved from NCBI website (<ftp://ftp.ncbi.nih.gov/genomes/Bacteria/>). It was found that 144 of total 4,131 proteins in *E. coli K12* were predicted as ARPs, among which 64 known ARPs, including 10 ARPs not included in our dataset, were successfully predicted. Similarly, 86 of total 2,697 proteins in *S. aureus Mu50* were predicted as potential ARPs, whilst 36 known ARPs, including 17 ARPs not included in our dataset, were successfully predicted. These results suggest that SVM is able to pick out most of known ARPs with low false positive rate. The details of full list of potential new ARPs in *E. coli K12* and *S. aureus Mu50* are provided in Table S1 and Table S2.



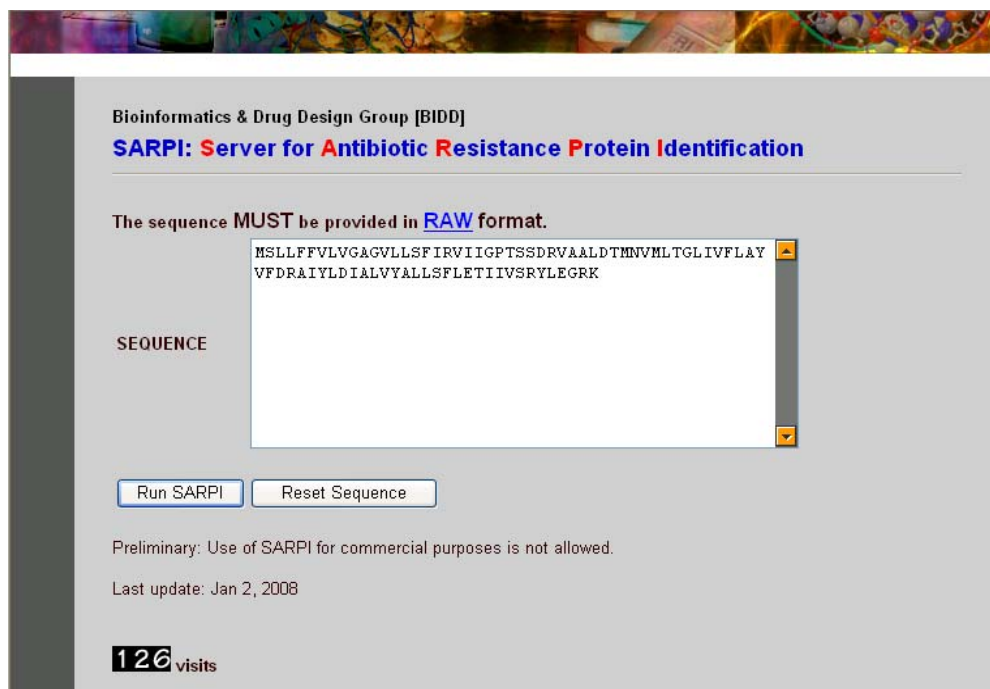
#### 4.2.5. Contribution of feature properties to the classification of ARPs

Contribution of individual feature properties to ARP classification was also investigated by separately conducting classification using each feature property[186]. Hydrophobicity and amino acid composition were found to play more prominent role than other feature properties. Previous experimental analysis revealed that many antibiotic resistance proteins contain hydrophobic pockets, which are essential to multi-drug transporters[187-189]. Some specific amino acids are essential to antibiotic resistance. For example, serine at the active site of ser- $\beta$ -lactamases performs a ring opening nucleophilic attack on the lactam ring [40]. Therefore, sequence features we used in this study seem to be reasonable to catch underlying common characters of ARPs from our results.

#### 4.2.6. Server for antibiotic resistance protein identification (SARPI)

SARPI is a prediction server for antibiotic resistance proteins, which can be accessed at <http://jing.cz3.nus.edu.sg/cgi-bin/sarpi.cgi>. The format of input query sequence is the same as that of SIME server described in the previous chapter. The average computing time is about 3 seconds for a typical protein and the computed result is also displayed in a separate window. Figure 4-3 shows the interface of SARPI. The result page would show you whether a query sequence is predicted to be an antibiotic resistance protein (Figure 4-4).

**Figure 4-3 Interface for SARPI**



Bioinformatics & Drug Design Group [BIDD]  
**SARPI: Server for Antibiotic Resistance Protein Identification**

The sequence **MUST** be provided in [RAW](#) format.

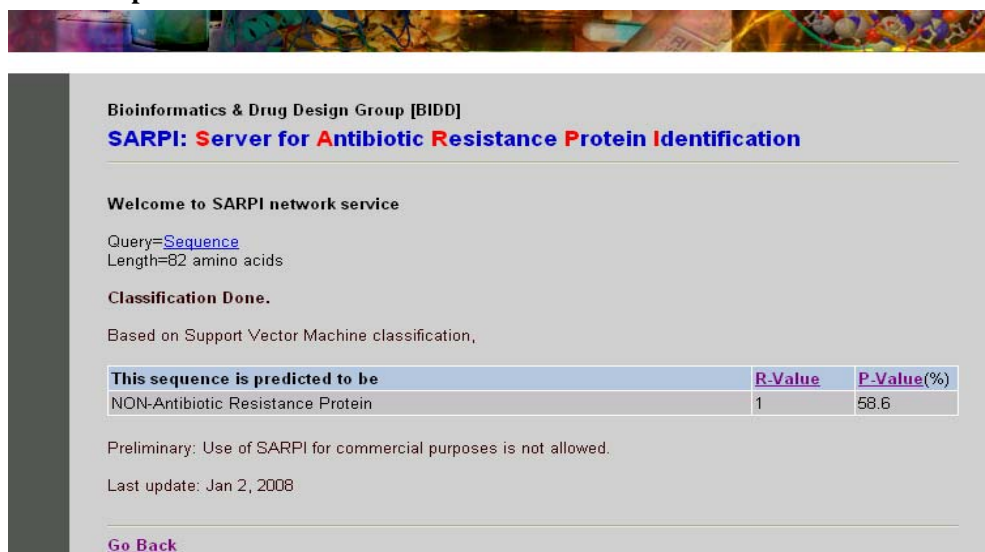
SEQUENCE

```
MSLLFFVLVGAGVLLSFIRVIIGPTSSDRVAALDTMNVMLTGLIVFLAY
VFDRAIYLDIALVYALLSFLETIIIVSRYLEGRK
```

Preliminary: Use of SARPI for commercial purposes is not allowed.  
 Last update: Jan 2, 2008

**126** visits

**Figure 4-4** Result page of SARPI showing that the query sequence is not antibiotic resistance protein



Bioinformatics & Drug Design Group [BIDD]  
**SARPI: Server for Antibiotic Resistance Protein Identification**

Welcome to SARPI network service

Query=[Sequence](#)  
 Length=82 amino acids

**Classification Done.**

Based on Support Vector Machine classification,

This sequence is predicted to be	R-Value	P-Value(%)
NON-Antibiotic Resistance Protein	1	58.6

Preliminary: Use of SARPI for commercial purposes is not allowed.  
 Last update: Jan 2, 2008

[Go Back](#)

### 4.3. Prediction of cancer associated proteins

As described in the introduction chapter, the identification of cancer associated proteins would help to understand the mechanism of cancer, and also to find better treatments for cancer therapies. In this study, a SVM based prediction system was developed to facilitate the identification of cancer associated proteins.

#### 4.3.1. Data preparation

A total of 784 cancer associated proteins (CAPs) were collected from a comprehensive search of Swiss-Prot database at <http://us.expasy.org/sprot/> [162] using keyword “tumor suppressor” and “proto-oncogene” followed by manual check that each protein is related with cancer. They were further divided into two separate classes: proto-oncogene (499 proteins) and tumor suppressors (286 proteins). The negative dataset was also selected in a similar way as that for MFEs. These positive and negative datasets were further divided into separate training set, testing set and independent evaluation set. The statistics of positives and negatives in each CAP class is given in Table 4-11, as well as the distribution of these proteins in top 10 host species is given in Table 4-12, which shows that these proteins are from diverse range of species.

**Table 4-11 Statistics of datasets and prediction accuracy of cancer associated proteins**

CAP Classes	Training set		Testing set				Independent evaluation set							
	positive	negative	positive		negative		positive			negative			PPV (%)	Q (%)
			TP	FN	TN	FP	TP	FN	SE (%)	TN	FP	SP (%)		
CAPs	403	2082	241	0	14646	3	133	7	95.0	7505	30	99.6	81.6	99.5
Tumor suppressors	142	1377	83	0	15887	0	50	11	82.0	7790	14	99.8	78.1	99.7
Proto- oncogenes	283	2107	137	0	14871	0	77	2	97.5	7626	29	99.6	72.6	99.6

**Table 4-12 Distribution of cancer associated proteins in top 10 bacteria species**

Proto-oncogene		Tumor suppressor	
Species	Number	Species	Number
<i>Homo sapiens</i>	219	<i>Homo sapiens</i>	108
<i>Mus musculus</i>	82	<i>Mus musculus</i>	65
<i>Rattus norvegicus</i>	30	<i>Rattus norvegicus</i>	36
<i>Gallus gallus</i>	26	<i>Bos taurus</i>	16
<i>Bos taurus</i>	17	<i>Canis familiaris</i>	5
<i>Sus scrofa</i>	9	<i>Pongo pygmaeus</i>	5
<i>Felis silvestris catus</i>	9	<i>Sus scrofa</i>	4
<i>Canis familiaris</i>	8	<i>Mesocricetus auratus</i>	3
<i>Pongo pygmaeus</i>	7	<i>Gallus gallus</i>	3
<i>Mesocricetus auratus</i>	5	<i>Danio rerio</i>	3

#### 4.3.2. Overall prediction accuracies and performance evaluation

The statistics of prediction results is given in Table 4-11. The predicted sensitivity for tumor suppressor ( $\sigma=26$ ), proto-oncogenes ( $\sigma=15$ ) and all CAPs ( $\sigma=23$ ) is 82.0%, 97.5% and 95.0% respectively, while the corresponding predicted specificity is 99.8%, 99.6% and 99.6% respectively. The predicted PPV for tumor suppressor, proto-oncogenes and all CAPs is 78.1%, 72.6% and 81.6% respectively. These results suggest that SVM is capable to predict CAPs at a reasonable accuracy. It was also found that the prediction accuracy of tumor suppressors is much lower than other classes. This may be due to the following two factors. One is the inadequate representatives of tumor suppressors, because of the small number in the training data. The other is the result of unbalanced dataset, which will lead to a reduced accuracy for the dataset either with a smaller number of samples or of less diversity[119]. In order to improve the accuracy of tumor suppressors, accumulation of more positive

data, and application of additional computational methods for re-adjusting biased shift of hyperplane[190] should be considered.

Our model was further tested by scanning sequences with human genome (NCBI release 36), to identify potential new CAPs. It was found that 714 of total 43,367 proteins in human genome were predicted as tumor suppressors, while 100 of 108 known tumor suppressors were successfully predicted. Similarly, 2,234 of total 43,367 proteins in human genome were predicted as potential proto-oncogene, while 204 of 219 known proto-oncogenes were successfully predicted. These results suggest that SVM is able to pick out most of known CAPs at low false positive rate.

#### 4.3.3. Contribution of feature properties to the classification of cancer associated proteins

The contribution of feature properties to the classification of cancer associated proteins was also studied. Amino acid composition and hydrophobicity were found to most important to characterize cancer associated proteins. Previous studies indicated that binding sites of proto-oncogenes usually appear in hydrophobic regions [191-193]. On the other hand, specific amino acid composition and sequence motifs are essential to for the proto-oncogene interactions [194-196]. Therefore our prediction results are some sort of consistent with these experimental findings.

#### 4.3.4. Analysis of individual feature contribution by feature selection

A more rigorous feature selection method, recursive feature elimination (RFE), as described in Method chapter, was applied to SVM classification of cancer associated proteins to select those features most relevant to the prediction of cancer associated proteins.

A total of 33 features were selected by RFE, which are given in Table 4-13. In order of prominence, hydrophobicity, secondary structure, surface tension normalized Van der Waals volume, and polarity are found to be important for predicting cancer associated proteins. This conclusion is roughly consistent with that derived from the other feature evaluation method used in this work. We further tested the usefulness of these 33 selected features by constructing a SVM classification system based solely on these features. The prediction accuracies of this new system are 96.4% and 99.9% for cancer associated proteins and non-cancer associated proteins respectively, which is slightly improved against those of 95.0% and 99.6% by using all features. This suggests that the use of selected subset of features enhances prediction performance by reducing the noise created by the redundant and irrelevant features.

**Table 4-13 Features important for characterizing cancer associated proteins as selected by recursive feature elimination method**

Feature Rank	Feature Index	Feature Description
1	F76	Polarity Group 2 2/4th Distribution
2	F91	Polarizability Group 1 1/4th Distribution
3	F23	Hydrophobicity Composition Group 3
4	F31	Hydrophobicity Group 1 4/4th Distribution
5	F72	Polarity Group 1 3/4th Distribution
6	F155	Secondary structure Group 1 2/4th Distribution
7	F134	Surface tension Group 1 2/4th Distribution

8	F141	Surface tension Group 2 4/4th Distribution
9	F49	Normalized Van der Waals volume Group 1 1/4th Distribution
10	F156	Secondary structure Group 1 3/4th Distribution
11	F183	Solvent accessibility Group 2 4/4th Distribution
12	F48	Normalized Van der Waals volume Group 1 First Distribution
13	F57	Normalized Van der Waals volume Group 2 4/4th Distribution
14	F160	Secondary structure Group 2 2/4th Distribution
15	F51	Normalized Van der Waals volume Group 1 3/4th Distribution
16	F26	Hydrophobicity Transition Group 3
17	F62	Normalized Van der Waals volume Group 3 4/4th Distribution
18	F112	Charge Group 1 1/4th Distribution
19	F41	Hydrophobicity Group 3 4/4th Distribution
20	F140	Surface tension Group 2 3/4th Distribution
21	F149	Secondary structure Composition Group 3
22	F69	Polarity Group 1 First Distribution
23	F176	Solvent accessibility Group 1 2/4th Distribution
24	F127	Surface tension Composition Group 2
25	F145	Surface tension Group 3 3/4th Distribution
26	F98	Polarizability Group 2 3/4th Distribution
27	F154	Secondary structure Group 1 1/4th Distribution
28	F146	Surface tension Group 3 4/4th Distribution
29	F40	Hydrophobicity Group 3 3/4th Distribution
30	F161	Secondary structure Group 2 3/4th Distribution
31	F153	Secondary structure Group 1 First Distribution
32	F45	Normalized Van der Waals volume Transition Group 1
33	F128	Surface tension Composition Group 3

#### 4.3.5. Cancer associated protein identification server (CAPIS)

Cancer associated protein identification server (CAPIS) was also developed to predict cancer associated proteins based on primary sequences. It is easily accessible at <http://jing.cz3.nus.edu.sg/cgi-bin/capis.cgi> as shown in Figure 4-5. The average computing time is around 3 seconds for a typical protein and the result is displayed in a separate window (Figure 4-6). Given a sequence, CAPIS will output whether this sequence is a cancer associated protein or not. If yes, it will also indicate whether it is a proto-oncogene or tumor suppressor.

**Figure 4-5 CAPIS interface.** The sequence of a protein, in RAW format and containing no non-amino acid letters, can be input in a window provided.



BioInfo & Drug Design ▾ Databases ▾ Softwares ▾ Arts ▾ Teaching ▾ Research ▾ Links ▾

**Bioinformatics & Drug Design Group [BIDD]**  
**CAPIS: Cancer Associated Protein Identification Server**

The sequence **MUST** be provided in **RAW** format.

**SEQUENCE**

```
MFGLDQFEPQVNSRNAGQGERNFNETGLSMNTHFKAPAFHTGGPPGPVD
PAMSA LGEPPI LGMNM EPGYGFHARGHSELHAGGLQAQPVHGFFGGQQPH
HGHPGSHHPHQHHPHFGGNFGGPDPGASCLHGGRL LGYGAAGGLGSP
PFAEGYEHMAESQGPEFSGPQRPGNLPDFHSSGASSHAVPAPCLPLDQS
PNRAASFHGLPSSSGSDSHSLEPRRVTNQGA VDSLEYNYNYPGEAPS GHFD
MFSPDSEGLPHYAAGROVPGGAFPGASAMPRAAGMVGLSKMHAQPPQ
QQPQQQQQPQQQQHGVFFERFSGARKMPVGLPSVGSRRPLMQPPQQA
PPPPQQPPQQPPQQPPPPGLLVQRNSCPPALPRPQQGEAGTPSGGL
QDGGPMLPSQHAQFEYPIHRLNRS MHPYSEPVFSMQHPPPPQQA PNQRL
QHFDAPPYMMNVAKRPRDFPGSAGVDRCA SWNGSMHNGALDNHLSPAY
```

Preliminary: Use of CAPIS for commercial purposes is not allowed.  
 Last update: Jan 2, 2008

**Figure 4-6 Result page of CAPIS showing that the query sequence is a proto-oncogene.**



BioInfo & Drug Design ▾ Databases ▾ Softwares ▾ Arts ▾ Teaching ▾ Research ▾ Links ▾

**Bioinformatics & Drug Design Group [BIDD]**  
**CAPIS: Cancer Associated Protein Identification Server**

**Welcome to CAPIS network service**

Query=[Sequence](#)  
 Length=531 amino acids

**Classification Done.**

Based on Support Vector Machine classification,

This sequence is predicted to be	R-Value	P-Value(%)
Cancer associated protein - Protooncogene	6	99.0

Preliminary: Use of CAPIS for commercial purposes is not allowed.  
 Last update: Jan 2, 2008

[Go Back](#)



#### 4.4. Comparison with other statistical learning methods

The performance of our SVM classification systems (Gaussian kernel SVM) for predicting disease related proteins such as AMPs, ARPs and CAPs were also compared with other statistical learning methods, including decision tree (DT), k-nearest neighbors (KNN) and probabilistic neural networks (PNN). The same datasets of these 3 classes of disease related proteins were used for constructing and testing the classifiers developed by each of these methods.

Table 4-14 shows the results of predicting antimicrobial proteins. The prediction accuracies for AMPs were in the range of 73.2%~90.8% with SVM giving the best accuracy at 90.8%. For non-AMPs, the prediction accuracies were in the range of 92.1%~99.6% with SVM giving the best accuracy at 99.6%. Table 4-15 shows the results of predicting antibiotic resistance. The accuracies of each method were in the range of 72.8%~88.5% with SVM giving the best accuracy at 88.5%. For non-ARPs, the prediction accuracies were found in the range of 91.3%~99.2% with SVM giving the best performance. Table 4-16 shows the results of predicting cancer associated proteins. The accuracies of each method were in the range of 69.3%~95.0% with SVM giving the best accuracy at 95.0%. For non-CAPs, the prediction accuracies were found in the range of 97.8%~99.8% with PNN giving the best performance. Therefore, it seems that SVM is able to predict disease related proteins with highest accuracy comparing with other machine learning methods.

**Table 4-14 Comparison of prediction performance of all AMPs and non-AMPs with different machine learning methods**

Method	Parameter	TP	FN	TN	FP	AMPs SE (%)	Non- AMPs SP (%)	Q (%)
DT		92	27	7138	120	77.3	98.3	98.0
PNN	$\bar{\sigma}=2$	87	32	6685	573	73.2	92.1	91.8
KNN	$k=17$	89	30	7216	42	74.8	99.4	99.0
SVM	$\bar{\sigma}=35$	108	11	7230	28	90.8	99.6	99.5

**Table 4-15 Comparison of prediction performance of antibiotic resistances and non-antibiotic resistances with different machine learning methods**

Method	Parameter	TP	FN	TN	FP	ARPs SE (%)	Non- ARPs SP (%)	Q (%)
DT		228	85	6933	223	72.8	96.9	95.9
PNN	$\bar{\sigma}=1.3$	233	80	6720	463	74.4	94.0	93.1
KNN	$k=9$	232	81	6535	621	74.1	91.3	90.6
SVM	$\bar{\sigma}=18$	277	36	7099	57	88.5	99.2	98.7

**Table 4-16 Comparison of prediction performance of all CAPs and non-CAPs with different machine learning methods**

Method	Parameter	TP	FN	TN	FP	CAPs SE (%)	Non- CAPs SP (%)	Q (%)
DT		97	43	7372	163	69.3	97.8	97.3
PNN	$\bar{\sigma}=0.8$	109	31	7520	15	77.9	99.8	99.4
KNN	$k=3$	124	16	6781	754	88.6	90.0	90.0
SVM	$\bar{\sigma}=35$	133	7	7505	30	95.0	99.6	99.5

## 4.5. Summary

This chapter presents my work in the prediction of disease related proteins through support vector machine approach. Three disease related protein classes, including antimicrobial proteins, antibiotic resistance proteins and cancer associated proteins were investigated in details. The accuracies for predicting members and non-members for each class were in the range of 81.8%~97.5% and 99.2%~99.9% respectively. Moreover, most of non-homologous disease related proteins were successfully predicted by our method. Genome screening was also performed to identify potential disease related proteins. Furthermore, the comparison with other machine learning

methods like KNN, DT and PNN indicates that SVM has better performance in disease related protein prediction. These results suggest the usefulness of SVM for facilitating the identification of disease related proteins. It should be noted that the performance of SVM critically depends on the diversity of training samples. The datasets used in this study are not expected to be fully representative of all the functional proteins with and without a particular functional profile. Various degrees of inadequate sampling representation are likely to affect, to a certain extent, the prediction performance.

## 5. Prediction of microRNAs by machine learning methods

MicroRNAs (miRNAs) are endogenous short non-coding RNAs of approximately 22 nucleotides long. They are able to regulate gene expression at both the transcription and translation level, targeting mRNA for degradation or translational repression. Prediction of miRNAs is important to uncover post-transcriptional gene regulatory network. Computational methods have been developed based on sequence similarity, structure properties, and probabilistic models. These methods have achieved impressive accuracies in the prediction of known miRNAs. However, their false positive rates are still too high. In this work, we applied machine learning methods to the prediction of miRNAs, which achieved relatively lower false positive rate.

### 5.1. Data preparation

#### 5.1.1. Retrieval of precursor miRNAs and non-precursor miRNAs

A total of 5,299 precursor miRNAs were collected from miRBase database (Release 10.1) [85, 197]. The distribution of these miRNAs in top 10 species is given in Table 5-1.

Non-precursor miRNAs consists of two parts. The first part contains representative examples of each non-coding RNA functional class except miRNAs extracted from Rfam database (Version 8.1) [198]. The second part came from representative cDNA sequences retrieved from an annotated human gene database, H-Invitational Database

[199]. These cDNA sequences were then cut into strands with 100 nucleotides, in order to maximize including characterized precursor miRNA sequence information. These non-precursor miRNAs were then clustered with known precursor miRNAs. Sequences which were not clustered with true precursor miRNAs were selected as non-precursor miRNAs. The final negative dataset contains 14,626 cDNA sequences and 36,431 functional RNA sequences.

**Table 5-1 Distribution of precursor miRNAs in top 10 host species**

	Species
List of top 10 species and the number of miRNAs in each of them	<i>Homo sapiens</i> (541)
	<i>Mus musculus</i> (443)
	<i>Rattus norvegicus</i> (287)
	<i>Oryza sativa</i> (243)
	<i>Physcomitrella patens</i> (220)
	<i>Populus trichocarpa</i> (215)
	<i>Arabidopsis thaliana</i> (184)
	<i>Xenopus tropicalis</i> (184)
	<i>Drosophila melanogaster</i> (152)
	<i>Caenorhabditis elegans</i> (137)

### 5.1.2. Retrieval of mature miRNAs and non-mature miRNAs

Mature miRNA sequences were also collected from miRBase database (Release 10.1) [85, 197]. 5,149 sequences were left after removing those with unknown nucleotides. The distribution of these mature miRNAs in top 10 species, as shown in Table 5-1, is similar to that of precursor miRNAs.

Non-mature miRNAs were generated by following procedures. Each mature sequence was excised from its corresponding precursor miRNA. The remaining sequences of precursor strands were processed in such a way that sequences with less than 21 nucleotides were kept aside while those with longer than 21 nucleotides were cut into

fragments with 21, 22 and 23 nucleotides long. All of these short fragments were clustered with known mature miRNAs. Those short fragments which were not clustered with true mature miRNAs were selected as non-mature miRNA. The final negative dataset contains 36,677 short sequences.

## 5.2. Evaluation and discussion

### 5.2.1. Prediction performance for precursor miRNAs and mature miRNAs

The statistics of the dataset and prediction results of miRNAs and their precursors is given in Table 5-2. The predicted sensitivity for precursor miRNAs and mature miRNAs is 92.2% and 94.8% respectively, while the corresponding predicted specificity is 98.4% and 99.5% respectively. These results suggest that SVM is capable to predict miRNAs and their precursor with reasonably high accuracy.

**Table 5-2 Statistics of the datasets and prediction accuracy for precursor miRNAs and mature miRNAs**

Classes	Training set		Testing set				Independent evaluation set							
	positive	negative	positive		negative		positive			negative			PPV (%)	Q (%)
			TP	FN	TN	FP	TP	FN	SE (%)	TN	FP	SP (%)		
Precursor miRNAs	1841	3183	2884	11	41825	12	519	44	92.2	5942	95	98.4	84.5	97.9
Mature miRNAs	2244	7354	2294	0	22603	0	579	32	94.8	6686	34	99.5	94.5	99.1

A recent work done by Yue et al. identified 245 putative precursor miRNAs in rhesus genome [200], which share high homology (>90%) with verified human precursor miRNAs. These precursor miRNAs were not included in our training dataset, thus constituted an additional independent testing data for our model. 192 (or 78%) of

them were predicted as potential precursor miRNAs by our SVM prediction system, indicating that both methods are consistent to some extent. The list of these precursor miRNAs could be found in Appendices Table S3.

Among 245 potential rhesus precursor miRNAs, eight rhesus precursor sequences were randomly selected by Yue *et al* [200]. Of these, three mature miRNAs were experimentally validated as novel mature miRNAs. All of them were also successfully predicted by our model, although the specific region is slightly different as shown in Table 5-3.

**Table 5-3 Location of predicted and validated rhesus miRNAs within putative precursor sequences. Sequences in *italic* denote those predicted by MiRDetector while those with underline denote experimentally validated miRNAs.**

miR-379	Precursor Sequences
Predicted by MiRDetector	AGAG <u>UGGUAGACUAUGGAACGUAGG</u> CGUUAUGAUUUUUGACCUA UGUAACAUGGUCCACUAACUCU
Experimentally validated	AGAGA <u>UGGUAGACUAUGGAACGUA</u> GGCGUUAUGAUUUUUGACCUA UGUAACAUGGUCCACUAACUCU
miR-422	Precursor Sequences
Predicted by MiRDetector	GAGAGAAGC <u>ACUGGACUCAGGGUCAGAAGGCC</u> CUGAGUCUCCCUG CUGCAGAUGGGCUGUGUGUCCCUGAGCCAAGCCUUGUCCUCCCUGG
Experimentally validated	GAGAGAAGCA <u>CUGGACUCAGGGUCAGAAGGCC</u> CUGAGUCUCCCUG CUGCAGAUGGGCUGUGUGUCCCUGAGCCAAGCCUUGUCCUCCCUGG
miR-648	Precursor Sequences
Predicted by MiRDetector	AGCACAGACGCCUCCA <u>AGUGUGCAGGGCACUGAUGGGG</u> GCCAGG GCAGGCCAGCCAAAGUGCAGGACCUGGCACUUAGUCGGAGGUGA GGAUG
Experimentally validated	AGCACAGACGCCUCC <u>AAGUGUGCAGGGCACUGAU</u> GGGGGCCAGG GCAGGCCAGCCAAAGUGCAGGACCUGGCACUUAGUCGGAGGUGA GGAUG

### 5.2.2. Screening non-coding RNAs within four representative genomes

In order to further test the ability of our model to identify miRNAs, non-coding RNA sequences from four representative genomes were screened by our model, including *Drosophila melanogaster*, *Homo sapiens*, *Mus musculus*, and *Saccharomyces cerevisiae*. The sequence data were downloaded from Ensembl database [201], and all known miRNAs were removed.

As shown in Table 5-5, 2.6%, 5.6%, 2.2% and 4.2% of non-coding RNAs in *Drosophila melanogaster*, *Homo sapiens*, *Mus musculus* and *Saccharomyces cerevisiae* were predicted as potential miRNA precursors respectively. Considering that some of them are un-characterized precursor miRNAs, it indicates that our model has a false positive rate less than 5%, which is relative lower than previous studies (10%)[93].

**Table 5-5 Screening results of non-coding RNAs from four representative genomes**

Species	Negative set		Percentage of positive
	No. of predicted precursors (NP)	No. of total non-coding RNA sequences (NT)	
<i>Drosophila melanogaster</i>	16	623	2.6%
<i>Homo sapiens</i>	357	6363	5.6%
<i>Mus musculus</i>	47	2089	2.2%
<i>Saccharomyces cerevisiae</i>	18	426	4.2%

### 5.2.3. Comparison with other statistical learning methods

The performance of our SVM classification systems (Gaussian kernel SVM) for predicting precursor miRNAs and mature miRNAs were also compared with other



statistical learning methods, including decision tree (DT), k-nearest neighbors (KNN) and probabilistic neural networks (PNN). The same datasets of the precursor and mature miRNAs were used for constructing and testing the classifiers developed by each of these methods.

Table 5-6 shows the results of predicting precursor miRNAs. The prediction accuracies for precursor miRNAs were in the range of 72.6%~92.2% with SVM giving the best accuracy at 92.2%. For non-precursor miRNAs, the prediction accuracies were in the range of 78.8%~98.4% with SVM giving the best accuracy at 98.4%. Table 5-7 shows the results of predicting mature miRNAs. The accuracies of each method were in the range of 78.2%~94.8% with SVM giving the best accuracy at 94.8%. For non-mature miRNAs the prediction accuracies were found in the range of 89.8%~99.5% with SVM and PNN giving the same best performance. Therefore, it seems that SVM is able to predict both precursor miRNA and mature miRNA prediction with highest accuracy comparing with other machine learning methods.

**Table 5-6 Comparison of prediction performance of precursor miRNAs and non-precursor miRNAs with different machine learning methods**

Method	Parameter	TP	FN	TN	FP	Precursor miRNA SE (%)	Non-precursor miRNA SP (%)	Q (%)
DT		409	154	5874	163	72.6%	97.3%	95.2%
PNN	$\sigma = 4$	444	119	4757	1280	78.9%	78.8%	78.8%
KNN	$k=1$	437	126	4807	1230	77.6%	79.6%	79.5%
SVM	$\sigma = 35$	519	44	5942	95	92.2%	98.4%	97.8%

**Table 5-7 Comparison of prediction performance of mature miRNAs and non-mature miRNAs with different machine learning methods**

Method	Parameter	TP	FN	TN	FP	Mature miRNAs SE (%)	Non-mature miRNA SP (%)	Q (%)
DT		549	62	6393	327	89.9%	95.1%	94.7%
PNN	$\sigma = 2$	562	49	6687	33	92.0%	99.5%	98.9%
KNN	$k=3$	478	133	6035	685	78.2%	89.8%	88.8%
SVM	$\sigma = 53$	579	32	6,686	34	94.8%	99.5%	99.1%

### 5.3. MiRNA prediction server

A web server, MicroRNA Detector (MiRDetector), was developed to facilitate the identification of novel miRNAs based on sequence derived physicochemical properties. MiRDetector could be freely accessible at <http://ang.cse.nus.edu.sg/cgi-bin/mirna/mirna.cgi>. Figure 5-1 shows its graphical user interface. The query sequence containing only the valid nucleotide characters such as ‘A’, ‘U or T’, ‘G’ and ‘C’ can be submitted to this prediction server. If a query sequence contains non-AU(T)GC characters, it would be rejected with a corresponding error message. Two steps are taken in our prediction server. First, MiRDetector will determine whether the query sequence is a precursor miRNA or not (Figure 5-2). If yes, MiRDetector will continue to identify its potential mature miRNA and the location within the precursor sequence (Figure 5-3). If not, MiRDetector will output the result without going to the second step. Figure 5-3 shows the predicted potential mature miRNA and the location within the precursor sequence. Predicted mature miRNA is highlighted in red. A predicted secondary hairpin structure is also displayed, which is generated from RNAfold [202]. The open bracket on the far left end of the sequence indicates a match with the closed bracket on the far right end. The full-stop character indicates mismatches resulting in loops or bulges.

#### 5.3.1. Comparison with other microRNA prediction servers

Many methods like MiRPred [203] and miRNA SVM [100] utilize knowledge of characteristic loops and bulges such as minimum free energy rule [204] for novel miRNA prediction [93, 205]. Recent studies also indicate that the hairpin is a good evaluating feature for miRNA prediction [206]. However, such characteristics do not

exclusively exist in miRNAs. In addition, the short length of precursors compared to that of the genome brings more complexity to this problem. MiRFinder [207] attempted to use multiple sequence alignment to remove non-hairpin-like sequences in the hope of reducing the number of false positives, but the computation cost increase significantly with little benefit. To make the problem worse, this approach may result in the loss of true precursors which are not well conserved across species. MiRDetector, on the other hand, was developed just by nucleotides derived physicochemical, without any sequence or structure conservation limitation. As it allows the characteristic features to be captured based on the input sequences alone, conserved sequence information within the samples can be extracted too. This helps to reach in higher sensitivities and specificities in the prediction of miRNAs across different species as tabulated in Table 5-2, implying that our method works well without the use of any structural information. On the other hand, it does not mean that structural features are not important. Current comparison is merely to test the possibility to exclude structural characteristics to the prediction of miRNAs. Since the use of physicochemical features produces satisfactory results, it forms an alternative solution for identifying miRNAs. It is likely that the combination of these features with the structural knowledge of miRNAs would reach even better prediction performance.

Comparative genomics based methods which apply conserved sequence and secondary structure features were found inferior to MiRDetector. miRseeker[208] achieved an accuracy of 75% in the prediction of miRNAs with *Drosophila* genome while miRAlign [209] achieved an accuracy of 89.9% in the prediction of miRNAs in animal genomes but 70% for plants. The comparatively higher sensitivity of 99.5%

in MiRDetector reassures the findings from other similar approaches in that the use of cross-species information may have a prediction advantage over species-specific models. A recent but different approach adopted by MiRPred also verified this point [203]. Although it aimed at using an *ab initio* method which relied only on precursor miRNA structural characteristics, the model was trained on human dataset, achieving 95% sensitivity over human data but later dropped to 90% for other non-human datasets. Thus MiRDetector, uses only sequence information instead, its application to cross-species data would expect to achieve better performance than MiRPred.

Comparing with other SVM based approaches, such as RNAmicro [210] and triplet-SVM [99], MiRDetector does not zoom into any specific regions of input sequences. Furthermore, triplet-SVM was trained on human dataset, resulting in an overall accuracy of 90% for human precursor miRNAs, which is comparatively lower than that of MiRDetector (99.4%). mirCoS [93] which implemented three SVM models simultaneously on human and mouse miRNA achieved a sensitivity of 85%. MiR-abela [205] is designed to predict species-specific samples. The work was built on the property that miRNA sequences are usually found in genomic clusters as some miRNAs are transcribed as polycistronic transcripts [211]. This idea is similar to the feature used for MiRDetector construction. MiRDetector was constructed based on physicochemical properties within the sample sequences, and miRNA with related sequence information are usually those which are located close to each other. Due to the similarity between the two concepts, a comparison test was performed to evaluate the specificity of each method. 40 randomly selected non-coding RNA sequences (Ensembl release 48) from *Drosophila melanogaster*, *Homo sapiens*, *Mus musculus*

and *Saccharomyces cerevisiae* were tested on both models. Both methods achieve very high accuracy indicating that they are comparable.

**Figure 5-1** Graphical user interface of MiRDetector. The sequence of a query sequence, in RAW format and containing non-AU(T)GC characters, can be input in a window provided.

Bioinformatics & Drug Design group [BIDD]  
**MiRDetector - MicroRNA Detector**  
for searching MicroRNA.


The sequence **MUST** be provided in [RAW](#) format.

SEQUENCE

UAVACCGAGAGCCCAGCUGAUUUCGUCUUGGUAUAAGCUCGUCAUUGA  
GAUUAUCACCGGGUGUAAAUCAGCUUGGCUCUGGUGUC

Submit Reset Sequence

**152** visits since Mar 6, 2008



**Bioinformatics & Drug Design group [BIDD]**  
**MiRDector - MicroRNA Detector**  
 for searching MicroRNA.

Your sequence has 87 nucleotides

**Classification Done.**

Based on Support Vector Machine classification,

This sequence is predicted to be	R-Value	P-Value(%)
MiRNA precursor ( <a href="#">Predict mature MiRNA region</a> )	5	98.7

[Go Back](#)

**152** visits since November 6, 2005

Preliminary: Use of **MiRD** for commercial purposes is not allowed.

## 5.4. Summary

The discovery of miRNAs has driven intensive interest towards better understanding of their biogenesis and functional roles. As regulatory components involved in many diverse cellular development and physiological processes, miRNAs are emerging as a promise gene regulation therapy. More and more efforts have been devoted to identify miRNAs and their targets. In this work, a SVM based predictor, MiRDetector, was developed to identify miRNAs based on primary sequences of RNAs. The overall accuracy of predicting precursor miRNAs and mature miRNAs was 97.9% and 99.1% respectively, which is slightly higher than previous studies. A further test by using 245 putative precursor miRNAs and three novel mature miRNAs from *Macaca mulatta* genome showed that 78% of precursor miRNAs and all of mature miRNA were correctly classified. In addition, genome screening found that an average of 4% of non-coding RNAs was predicted as precursor miRNAs, which indicates that our model has a relatively low false positive rate. Furthermore, the comparison with other machine learning methods like KNN, DT and PNN indicates that SVM has better performance in both precursor miRNA and mature miRNA prediction. These results suggest the usefulness of SVM for facilitating the prediction of miRNAs.

## 6. Conclusion and future work

The characterization of novel biochemical class, disease related proteins and miRNAs is essential for in-depth understanding of biological processes, and helpful to accelerate the drug target discovery. Due to the limitation of experimental approaches, various computational tools have been developed to facilitate the identification of these proteins and miRNAs. In this work, one machine learning method, support vector machine (SVM), was employed to develop prediction systems for multifunctional enzymes (MFEs), disease related proteins and miRNAs from their primary sequences derived physicochemical properties. Corresponding prediction servers were also developed to serve scientific community who are interested in further investigating these protein and RNA classes.

### 6.1. Major findings

In the study of MFEs, we collected and systematically analyzed MFEs by grouping them into two categories: MFEs with multiple catalytic domains (MCD-MFEs) and MFEs with single multi-activity domain (SMAD-MFEs). No obvious evidences were found species distribution analysis that complex life forms like human prefer more MFEs statistically than simple life form like yeast. Combined with the finding in later pathway ontology analysis that the majority of MFEs are involved in several essential cellular processes, it suggests that MFEs are most likely to be early enzymes in primitive life forms. They may play key roles in catalyzing essential cellular processes so that their functions are well conserved across species. This is also supported by the evidence that almost half of MFEs participate in only one biological pathway. At the meantime, some MFEs are generated by diversification and



specification in various forms of genetic variation like gene fusion or exon shuffling, since about half of MFEs are involving more than one pathway, as more as five independent pathways. According to current available 3D structures and orthologous analysis, the alpha and beta fold topology is the most favored to preserve multiple functions of MFEs during evolution. Later principle feature analysis found that four physiochemical properties are important for characterizing MFEs. A support vector machine based classifier was also constructed and successfully identified 2,641 novel MFEs from the ExPASy Enzyme database. In additional, from the domain information, we know that most MFEs are composed of multiple catalytic domains, which is similar with MFPs [212], so the same procedure may be extended to the identification of other types of MFPs. An online prediction system (SIME) and database, was also developed to facilitate the study of MFEs.

In the study of disease related proteins, three prediction systems were developed to identify antimicrobial proteins (including fungicide proteins, antibiotic proteins and all antimicrobial proteins), antibiotic resistance proteins and cancer associated proteins (including proto-oncogenes and tumor suppressors). Independent evaluation of these functional classes showed that the prediction accuracies for members and non-members were in the range of 81.8%~97.5% and 99.2%~99.9% respectively. The comparison with other machine learning methods like KNN, DT and PNN indicates that SVM performed better for disease related protein prediction. Moreover, a majority of novel proteins which do not have homologues in known protein databases were successfully predicted by our prediction systems. Potential disease related proteins were also identified through scanning bacterial and human genomes. These result shows that our prediction systems are potentially useful tools for the prediction

of disease related proteins and may serve as a promising complementary method to sequence similarity approach.

Three SVM prediction systems are developed to identify AMPs, ARPs and CAPs. They appear to be potentially useful tools to identify antimicrobial proteins, antibiotic resistances, proto-oncogenes and tumor suppressors with satisfying accuracy, especially for those novel disease related proteins without homologues in known protein databases. The prediction accuracy may be further enhanced with future accumulation of our knowledge about these disease related proteins particularly for those small sub classes, more refined representation of the structural and physicochemical properties of proteins, and the improvement of prediction algorithms such as the better treatment of imbalanced dataset.

In the study of miRNAs, a SVM based predictor, MiRDector, reached the accuracies of 92.2% and 94.8% for precursor miRNAs and mature miRNAs, and 98.4% and 99.5% for non-precursors and non-mature respectively. Sequences of non-coding RNAs from four representative genomes were also screened to identify potential miRNAs. An average of 4% of non-coding RNAs was predicted as precursor miRNAs, indicating our predictor is able to reach a relatively low false positive rate than previous study [213]. The comparison with other machine learning methods like KNN, DT and PNN indicates that SVM performed slightly better for both members and non-members prediction. These results suggest that our miRNA prediction system is capable of identifying miRNAs with satisfactory accuracy. Similar methodology could be ideally applied to predict other functional RNAs.

## 6.2. Limitation of methods applied in this work

It is found that the prediction accuracy of non-members appears to be better than that of members. This is because the negative data set is generally more diverse than positive data, which enables SVM to perform a better statistical learning to recognize non-members. Based on the statistics provided on the webpage of Pfam and Rfam database, there are over 9,000 families of proteins and 600 families of RNAs, from which one can generate a diverse set of non-members for each class. Because of the differences in the number of members and that of non-members in each class, there is an imbalance between each dataset. SVM based on an imbalanced datasets tends to produce feature vectors that push the hyperplane towards the side with smaller number of data, which can lead to a reduced accuracy for the dataset either with a smaller number of samples or of less diversity. This may be another reason why the prediction accuracy for members is generally lower than that for non-members. It is however inappropriate to simply reduce the size of non-members to artificially match that of members, since this compromises the diversity needed to fully represent all non-members. On the other hand, the number of actual negatives is much larger than the number of positives in reality. To solve the imbalance problem, resampling methods, such as oversampling and undersampling, are commonly used [214]. The idea of resampling methods is to either oversample the small class to make it reach to a size comparable to that of the larger class, or to undersample the larger class until it reaches to a size comparable to that of the smaller class [215].

It should be noted that the procedure for determining the contribution of feature properties in this study does not take into account of interactions between the different

feature properties. Special care should be taken if such procedures are used to analyze non-orthogonal features.

### 6.3. Future studies

The performance of our prediction system could be further improved with future accumulation of our knowledge about proteins and small RNAs. In addition, new protein and miRNA descriptors will be introduced to better represent certain types of protein functional profiles, such as structural characteristics and localization features, while folding energy can be added to small RNA features. Moreover, resampling methods would also be applied to solve the imbalance problem and further improve the prediction accuracy.

## BIBLIOGRAPHY

1. Jeffery CJ: **Multifunctional proteins: examples of gene sharing.** *Annals of medicine* 2003, **35**(1):28-35.
2. Jeffery CJ: **Moonlighting proteins: old proteins learning new tricks.** *Trends in genetics* 2003, **19**(8):415-417.
3. Copley SD: **Enzymes with extra talents: moonlighting functions and catalytic promiscuity.** *Current opinion in chemical biology* 2003, **7**(2):265-272.
4. Moore B: **Bifunctional and moonlighting enzymes: lighting the way to regulatory control.** *Trends in plant science* 2004, **9**(5):221-228.
5. Jensen RA: **Enzyme recruitment in evolution of new function.** *Annual review of microbiology* 1976, **30**:409-425.
6. Aharoni A, Gaidukov L, Khersonsky O, McQ Gould S, Roodveldt C, Tawfik DS: **The 'evolvability' of promiscuous protein functions.** *Nature genetics* 2005, **37**(1):73-76.
7. Jeffery CJ: **Molecular mechanisms for multitasking: recent crystal structures of moonlighting proteins.** *Current opinion in structural biology* 2004, **14**(6):663-668.
8. Canada KA, Iwashita S, Shim H, Wood TK: **Directed evolution of toluene ortho-monooxygenase for enhanced 1-naphthol synthesis and chlorinated ethene degradation.** *Journal of bacteriology* 2002, **184**(2):344-349.
9. Zheng W, Scheibner KA, Ho AK, Cole PA: **Mechanistic studies on the alkyltransferase activity of serotonin N-acetyltransferase.** *Chemistry & biology* 2001, **8**(4):379-389.
10. Pelegri PB, Franco OL: **Plant gamma-thionins: novel insights on the mechanism of action of a multi-functional class of defense proteins.** *Int J Biochem Cell Biol* 2005, **37**(11):2239-2253.
11. Tombes RM, Faison MO, Turbeville JM: **Organization and evolution of multifunctional Ca(2+)/CaM-dependent protein kinase genes.** *Gene* 2003, **322**:17-31.
12. Han JM, Lee MJ, Park SG, Lee SH, Razin E, Choi EC, Kim S: **Hierarchical network between the components of the multi-tRNA synthetase complex: implications for complex formation.** *J Biol Chem* 2006, **281**(50):38663-38667.
13. Gomez A, Domedel N, Cedano J, Pinol J, Querol E: **Do current sequence analysis algorithms disclose multifunctional (moonlighting) proteins?** *Bioinformatics* 2003, **19**(7):895-896.
14. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.** *Nucleic Acids Res* 1997, **25**(17):3389-3402.
15. Makwana N, Riordan FA: **Bacterial Meningitis : The Impact of Vaccination.** *CNS Drugs* 2007, **21**(5):355-366.
16. Hahn BH, Kong LI, Lee SW, Kumar P, Taylor ME, Arya SK, Shaw GM: **Relation of HTLV-4 to simian and human immunodeficiency-associated viruses.** *Nature* 1987, **330**(6144):184-186.
17. Cardona PJ: **New insights on the nature of latent tuberculosis infection and its treatment.** *Inflamm Allergy Drug Targets* 2007, **6**(1):27-39.
18. Yeaman MR, Yount NY: **Mechanisms of antimicrobial peptide action and resistance.** *Pharmacol Rev* 2003, **55**(1):27-55.

19. Leeuw. Ed, Lu W: **Human defensins: turning defense into offense?** *Infect Disord Drug Targets* 2007, **7**(1):67-70.
20. Seebah S, Suresh A, Zhuo S, Choong YH, Chua H, Chuon D, Beuerman R, Verma C: **Defensins knowledgebase: a manually curated database and information source focused on the defensins family of antimicrobial peptides.** *Nucleic Acids Res* 2007, **35**(Database issue):D265-268.
21. Chugh JK, Wallace BA: **Peptaibols: models for ion channels.** *Biochem Soc Trans* 2001, **29**(Pt 4):565-570.
22. Brahmachary M, Krishnan SP, Koh JL, Khan AM, Seah SH, Tan TW, Brusica V, Bajic VB: **ANTIMIC: a database of antimicrobial sequences.** *Nucleic Acids Res* 2004, **32**(Database issue):D586-589.
23. Fjell CD, Hancock RE, Cherkasov A: **AMPer: a database and an automated discovery tool for antimicrobial peptides.** *Bioinformatics* 2007, **23**(9):1148-1155.
24. Park IY, Park CB, Kim MS, Kim SC: **Parasin I, an antimicrobial peptide derived from histone H2A in the catfish, *Parasilurus asotus*.** *FEBS Lett* 1998, **437**(3):258-262.
25. Yamada K, Natori S: **Characterization of the antimicrobial peptide derived from sapecin B, an antibacterial protein of *Sarcophaga peregrina* (flesh fly).** *The Biochemical journal* 1994, **298 Pt 3**:623-628.
26. Buffy JJ, McCormick MJ, Wi S, Waring A, Lehrer RI, Hong M: **Solid-state NMR investigation of the selective perturbation of lipid bilayers by the cyclic antimicrobial peptide RTD-1.** *Biochemistry* 2004, **43**(30):9800-9812.
27. Kalfa VC, Jia HP, Kunkle RA, McCray PB, Jr., Tack BF, Brogden KA: **Congeners of SMAP29 kill ovine pathogens and induce ultrastructural damage in bacterial cells.** *Antimicrob Agents Chemother* 2001, **45**(11):3256-3261.
28. Matsuzaki K, Yoneyama S, Miyajima K: **Pore formation and translocation of melittin.** *Biophys J* 1997, **73**(2):831-838.
29. Gabay JE, Scott RW, Campanelli D, Griffith J, Wilde C, Marra MN, Seeger M, Nathan CF: **Antibiotic proteins of human polymorphonuclear leukocytes.** *Proc Natl Acad Sci U S A* 1989, **86**(14):5610-5614.
30. Beovic B: **The issue of antimicrobial resistance in human medicine.** *Int J Food Microbiol* 2006, **112**(3):280-287.
31. McDermott W, Rogers DE: **Social ramifications of control of microbial disease.** *Johns Hopkins Med J* 1982, **151**(6):302-312.
32. Neuhooff V, Schill WB, Sternbach H: **Micro-analysis of pure deoxyribonucleic acid-dependent ribonucleic acid polymerase from *Escherichia coli*. Action of heparin and rifampicin on structure and function.** *Biochem J* 1970, **117**(3):623-631.
33. Lohner K, Blondelle SE: **Molecular mechanisms of membrane perturbation by antimicrobial peptides and the use of biophysical studies in the design of novel peptide antibiotics.** *Comb Chem High Throughput Screen* 2005, **8**(3):241-256.
34. Tenover FC: **Mechanisms of antimicrobial resistance in bacteria.** *Am J Infect Control* 2006, **34**(5 Suppl 1):S3-10; discussion S64-73.
35. Amin AN, Rehm SJ: **Infections in hospitalized patients: what is happening and who can help?** *Cleve Clin J Med* 2007, **74 Suppl 4**:S2-5.
36. Alekshun MN, Levy SB: **Molecular mechanisms of antibacterial multidrug resistance.** *Cell* 2007, **128**(6):1037-1050.

37. Zgurskaya HI: **Molecular analysis of efflux pump-based antibiotic resistance.** *Int J Med Microbiol* 2002, **292**(2):95-105.
38. Heym B, Cole ST: **Multidrug resistance in Mycobacterium tuberculosis** *International Journal of Antimicrobial Agents* 1997, **8**(1):61-70.
39. Tanaka M, Nakayama H, Huruya K, Konomi I, Irie S, Kanayama A, Saika T, Kobayashi I: **Analysis of mutations within multiple genes associated with resistance in a clinical isolate of Neisseria gonorrhoeae with reduced ceftriaxone susceptibility that shows a multidrug-resistant phenotype.** *Int J Antimicrob Agents* 2006, **27**(1):20-26.
40. Wright GD: **Bacterial resistance to antibiotics: enzymatic degradation and modification.** *Adv Drug Deliv Rev* 2005, **57**(10):1451-1470.
41. Scaria J, Chandramouli U, Verma SK: **Antibiotic Resistance Genes Online (ARGO): a Database on vancomycin and beta-lactam resistance genes.** *Bioinformation* 2005, **1**(1):5-7.
42. Zhou CE, Smith J, Lam M, Zemla A, Dyer MD, Slezak T: **MvirDB--a microbial database of protein toxins, virulence factors and antibiotic resistance genes for bio-defence applications.** *Nucleic Acids Res* 2007, **35**(Database issue):D391-394.
43. Chaieb K, Zmantar T, Chehab O, Boucham O, Ben Hasen A, Mahdouani K, Bakhrouf A: **Antibiotic resistance genes detected by multiplex PCR assays in Staphylococcus epidermidis strains isolated from dialysis fluid and needles in a dialysis service.** *Jpn J Infect Dis* 2007, **60**(4):183-187.
44. Perez A, Canle D, Latasa C, Poza M, Beceiro A, Del Mar Tomas M, Fernandez A, Mallo S, Perez S, Molina F *et al*: **Cloning, Nucleotide Sequencing and Analysis of the AcrAB-TolC Efflux Pump of Enterobacter cloacae and its Involvement in Antibiotic Resistance in a Clinical Isolate.** *Antimicrob Agents Chemother* 2007.
45. Davies C, Bussiere DE, Golden BL, Porter SJ, Ramakrishnan V, White SW: **Ribosomal proteins S5 and L6: high-resolution crystal structures and roles in protein synthesis and antibiotic resistance.** *J Mol Biol* 1998, **279**(4):873-888.
46. Chen CY, Nace GW, Solow B, Fratamico P: **Complete nucleotide sequences of 84.5- and 3.2-kb plasmids in the multi-antibiotic resistant Salmonella enterica serovar Typhimurium U302 strain G8430.** *Plasmid* 2007, **57**(1):29-43.
47. de Gruijl FR, van Kranen HJ, Mullenders LH: **UV-induced DNA damage, repair, mutations and oncogenic pathways in skin cancer.** *J Photochem Photobiol B* 2001, **63**(1-3):19-27.
48. Benjamin CL, Ullrich SE, Kripke ML, Ananthaswamy HN: **p53 tumor suppressor gene: a critical molecular target for UV induction and prevention of skin cancer.** *Photochem Photobiol* 2008, **84**(1):55-62.
49. Ariumi Y, Serhan F, Turelli P, Telenti A, Trono D: **The integrase interactor 1 (INI1) proteins facilitate Tat-mediated human immunodeficiency virus type 1 transcription.** *Retrovirology* 2006, **3**:47.
50. Yokota J: **Tumor progression and metastasis.** *Carcinogenesis* 2000, **21**(3):497-503.
51. Alberti L, Carniti C, Miranda C, Roccato E, Pierotti MA: **RET and NTRK1 proto-oncogenes in human diseases.** *J Cell Physiol* 2003, **195**(2):168-186.
52. Gotoh A, Broxmeyer HE: **The function of BCR/ABL and related proto-oncogenes.** *Curr Opin Hematol* 1997, **4**(1):3-11.

53. Torrey DS: **Proto-oncogenes and germ-cell differentiation.** *Am J Reprod Immunol* 1992, **27**(3-4):167-170.
54. Weinberg RA: **How cancer arises.** *Scientific American* 1996, **275**(3):62-70.
55. Vattermi E, Claudio PP: **Tumor suppressor genes as cancer therapeutics.** *Drug News Perspect* 2007, **20**(8):511-520.
56. Finlan LE, Hupp TR: **p63: the phantom of the tumor suppressor.** *Cell Cycle* 2007, **6**(9):1062-1071.
57. Friend SH, Bernards R, Rogelj S, Weinberg RA, Rapaport JM, Albert DM, Dryja TP: **A human DNA segment with properties of the gene that predisposes to retinoblastoma and osteosarcoma.** *Nature* 1986, **323**(6089):643-646.
58. Huttenhofer A, Schattner P, Polacek N: **Non-coding RNAs: hope or hype?** *Trends Genet* 2005, **21**(5):289-297.
59. Wightman B, Ha I, Ruvkun G: **Posttranscriptional regulation of the heterochronic gene *lin-14* by *lin-4* mediates temporal pattern formation in *C. elegans*.** *Cell* 1993, **75**(5):855-862.
60. Lagos-Quintana M, Rauhut R, Meyer J, Borkhardt A, Tuschl T: **New microRNAs from mouse and human.** *Rna* 2003, **9**(2):175-179.
61. Lu J, Getz G, Miska EA, Alvarez-Saavedra E, Lamb J, Peck D, Sweet-Cordero A, Ebert BL, Mak RH, Ferrando AA *et al*: **MicroRNA expression profiles classify human cancers.** *Nature* 2005, **435**(7043):834-838.
62. Slack FJ, Weidhaas JB: **MicroRNA in cancer prognosis.** *The New England journal of medicine* 2008, **359**(25):2720-2722.
63. Roccaro AM, Sacco A, Chen C, Runnels J, Leleu X, Azab F, Azab AK, Jia X, Ngo HT, Melhem MR *et al*: **microRNA expression in the biology, prognosis and therapy of Waldenstrom macroglobulinemia.** *Blood* 2008.
64. He L, Hannon GJ: **MicroRNAs: small RNAs with a big role in gene regulation.** *Nature reviews* 2004, **5**(7):522-531.
65. Pasquinelli AE, McCoy A, Jimenez E, Salo E, Ruvkun G, Martindale MQ, Baguna J: **Expression of the 22 nucleotide *let-7* heterochronic RNA throughout the Metazoa: a role in life history evolution?** *Evolution & development* 2003, **5**(4):372-378.
66. Washietl S, Hofacker IL, Lukasser M, Huttenhofer A, Stadler PF: **Mapping of conserved RNA secondary structures predicts thousands of functional noncoding RNAs in the human genome.** *Nat Biotechnol* 2005, **23**(11):1383-1390.
67. **Finishing the euchromatic sequence of the human genome.** *Nature* 2004, **431**(7011):931-945.
68. Lee Y, Kim M, Han J, Yeom KH, Lee S, Baek SH, Kim VN: **MicroRNA genes are transcribed by RNA polymerase II.** *The EMBO journal* 2004, **23**(20):4051-4060.
69. Lee Y, Ahn C, Han J, Choi H, Kim J, Yim J, Lee J, Provost P, Radmark O, Kim S *et al*: **The nuclear RNase III Drosha initiates microRNA processing.** *Nature* 2003, **425**(6956):415-419.
70. Gregory RI, Yan KP, Amuthan G, Chendrimada T, Doratotaj B, Cooch N, Shiekhattar R: **The Microprocessor complex mediates the genesis of microRNAs.** *Nature* 2004, **432**(7014):235-240.



71. Bohnsack MT, Czaplinski K, Gorlich D: **Exportin 5 is a RanGTP-dependent dsRNA-binding protein that mediates nuclear export of pre-miRNAs.** *Rna* 2004, **10**(2):185-191.
72. Bartel DP: **MicroRNAs: genomics, biogenesis, mechanism, and function.** *Cell* 2004, **116**(2):281-297.
73. Schwarz DS, Hutvagner G, Du T, Xu Z, Aronin N, Zamore PD: **Asymmetry in the assembly of the RNAi enzyme complex.** *Cell* 2003, **115**(2):199-208.
74. Hammond SM, Boettcher S, Caudy AA, Kobayashi R, Hannon GJ: **Argonaute2, a link between genetic and biochemical analyses of RNAi.** *Science* 2001, **293**(5532):1146-1150.
75. Jackson AL, Bartz SR, Schelter J, Kobayashi SV, Burchard J, Mao M, Li B, Cavet G, Linsley PS: **Expression profiling reveals off-target gene regulation by RNAi.** *Nat Biotechnol* 2003, **21**(6):635-637.
76. Brennecke J, Stark A, Russell RB, Cohen SM: **Principles of microRNA-target recognition.** *PLoS Biol* 2005, **3**(3):e85.
77. Huttenhofer A, Cavaille J, Bachellerie JP: **Experimental RNomics: a global approach to identifying small nuclear RNAs and their targets in different model organisms.** *Methods Mol Biol* 2004, **265**:409-428.
78. Argaman L, Hershberg R, Vogel J, Bejerano G, Wagner EG, Margalit H, Altuvia S: **Novel small RNA-encoding genes in the intergenic regions of Escherichia coli.** *Curr Biol* 2001, **11**(12):941-950.
79. Wassarman KM, Repoila F, Rosenow C, Storz G, Gottesman S: **Identification of novel small RNAs using comparative genomics and microarrays.** *Genes Dev* 2001, **15**(13):1637-1651.
80. Miska EA: **How microRNAs control cell division, differentiation and death.** *Current opinion in genetics & development* 2005, **15**(5):563-568.
81. Hake S: **MicroRNAs: a role in plant development.** *Curr Biol* 2003, **13**(21):R851-852.
82. Bentwich I, Avniel A, Karov Y, Aharonov R, Gilad S, Barad O, Barzilai A, Einat P, Einav U, Meiri E *et al*: **Identification of hundreds of conserved and nonconserved human microRNAs.** *Nat Genet* 2005, **37**(7):766-770.
83. Berezikov E, Guryev V, van de Belt J, Wienholds E, Plasterk RH, Cuppen E: **Phylogenetic shadowing and computational identification of human microRNA genes.** *Cell* 2005, **120**(1):21-24.
84. Carthew RW: **Gene regulation by microRNAs.** *Current opinion in genetics & development* 2006, **16**(2):203-208.
85. Griffiths-Jones S, Grocock RJ, van Dongen S, Bateman A, Enright AJ: **miRBase: microRNA sequences, targets and gene nomenclature.** *Nucleic Acids Res* 2006, **34**(Database issue):D140-144.
86. Hsu SD, Chu CH, Tsou AP, Chen SJ, Chen HC, Hsu PW, Wong YH, Chen YH, Chen GH, Huang HD: **miRNAMap 2.0: genomic maps of microRNAs in metazoan genomes.** *Nucleic Acids Res* 2008, **36**(Database issue):D165-169.
87. Lau NC, Lim LP, Weinstein EG, Bartel DP: **An abundant class of tiny RNAs with probable regulatory roles in Caenorhabditis elegans.** *Science* 2001, **294**(5543):858-862.
88. Lee RC, Ambros V: **An extensive class of small RNAs in Caenorhabditis elegans.** *Science* 2001, **294**(5543):862-864.

89. Lagos-Quintana M, Rauhut R, Yalcin A, Meyer J, Lendeckel W, Tuschl T: **Identification of tissue-specific microRNAs from mouse.** *Curr Biol* 2002, **12**(9):735-739.
90. Watanabe T, Takeda A, Tsukiyama T, Mise K, Okuno T, Sasaki H, Minami N, Imai H: **Identification and characterization of two novel classes of small RNAs in the mouse germline: retrotransposon-derived siRNAs in oocytes and germline small RNAs in testes.** *Genes Dev* 2006, **20**(13):1732-1743.
91. Abbott AL, Alvarez-Saavedra E, Miska EA, Lau NC, Bartel DP, Horvitz HR, Ambros V: **The let-7 MicroRNA family members mir-48, mir-84, and mir-241 function together to regulate developmental timing in *Caenorhabditis elegans*.** *Dev Cell* 2005, **9**(3):403-414.
92. Berezikov E, Cuppen E, Plasterk RH: **Approaches to microRNA discovery.** *Nat Genet* 2006, **38 Suppl**:S2-7.
93. Sheng Y, Engstrom PG, Lenhard B: **Mammalian MicroRNA Prediction through a Support Vector Machine Model of Sequence and Structure.** *PLoS ONE* 2007, **2**(9):e946.
94. Legendre M, Lambert A, Gautheret D: **Profile-based detection of microRNA precursors in animal genomes.** *Bioinformatics* 2005, **21**(7):841-845.
95. Nam JW, Shin KR, Han J, Lee Y, Kim VN, Zhang BT: **Human microRNA prediction through a probabilistic co-learning model of sequence and structure.** *Nucleic Acids Res* 2005, **33**(11):3570-3581.
96. Missal K, Zhu X, Rose D, Deng W, Skogerbo G, Chen R, Stadler PF: **Prediction of structured non-coding RNAs in the genomes of the nematodes *Caenorhabditis elegans* and *Caenorhabditis briggsae*.** *J Exp Zool B Mol Dev Evol* 2006, **306**(4):379-392.
97. Ohler U, Yekta S, Lim LP, Bartel DP, Burge CB: **Patterns of flanking sequence conservation and a characteristic upstream motif for microRNA gene identification.** *Rna* 2004, **10**(9):1309-1322.
98. Xie X, Lu J, Kulbokas EJ, Golub TR, Mootha V, Lindblad-Toh K, Lander ES, Kellis M: **Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals.** *Nature* 2005, **434**(7031):338-345.
99. Xue C, Li F, He T, Liu GP, Li Y, Zhang X: **Classification of real and pseudo microRNA precursors using local structure-sequence features and support vector machine.** *BMC Bioinformatics* 2005, **6**:310.
100. Helvik SA, Snove O, Jr., Saetrom P: **Reliable prediction of Drosha processing sites improves microRNA gene prediction.** *Bioinformatics* 2007, **23**(2):142-149.
101. Needleman SB, Wunsch CD: **A general method applicable to the search for similarities in the amino acid sequence of two proteins.** *J Mol Biol* 1970, **48**(3):443-453.
102. Sankoff D: **Matching sequences under deletion-insertion constraints.** *Proc Natl Acad Sci U S A* 1972, **69**(1):4-6.
103. Reichert TA, Cohen DN, Wong AK: **An application of information theory to genetic mutations and the matching of polypeptide sequences.** *J Theor Biol* 1973, **42**(2):245-261.
104. Smith TF, Waterman MS: **Identification of common molecular subsequences.** *J Mol Biol* 1981, **147**(1):195-197.
105. Lipman DJ, Pearson WR: **Rapid and sensitive protein similarity searches.** *Science* 1985, **227**(4693):1435-1441.

106. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool.** *J Mol Biol* 1990, **215**(3):403-410.
107. Altschul SF, Koonin EV: **Iterated profile searches with PSI-BLAST--a tool for discovery in protein databases.** *Trends Biochem Sci* 1998, **23**(11):444-447.
108. Benner SA, Chamberlin SG, Liberles DA, Govindarajan S, Knecht L: **Functional inferences from reconstructed evolutionary biology involving rectified databases--an evolutionarily grounded approach to functional genomics.** *Res Microbiol* 2000, **151**(2):97-106.
109. Gattiker A, Gasteiger E, Bairoch A: **ScanProsite: a reference implementation of a PROSITE scanning tool.** *Appl Bioinformatics* 2002, **1**(2):107-108.
110. Enright AJ, Van Dongen S, Ouzounis CA: **An efficient algorithm for large-scale detection of protein families.** *Nucleic Acids Res* 2002, **30**(7):1575-1584.
111. Bairoch A, Bucher P, Hofmann K: **The PROSITE database, its status in 1995.** *Nucleic Acids Res* 1996, **24**(1):189-196.
112. Bru C, Courcelle E, Carrere S, Beausse Y, Dalmar S, Kahn D: **The ProDom database of protein domain families: more emphasis on 3D.** *Nucleic Acids Res* 2005, **33**(Database issue):D212-215.
113. Sonnhammer EL, Eddy SR, Durbin R: **Pfam: a comprehensive database of protein domain families based on seed alignments.** *Proteins* 1997, **28**(3):405-420.
114. Finn RD, Mistry J, Schuster-Bockler B, Griffiths-Jones S, Hollich V, Lassmann T, Moxon S, Marshall M, Khanna A, Durbin R *et al*: **Pfam: clans, web tools and services.** *Nucleic Acids Res* 2006, **34**(Database issue):D247-251.
115. Ribeiro Ede O, Zerlotini GG, Lopes IR, Ribeiro VB, Melo AC, Walter ME, Costa MM: **A distributed computation of Interpro Pfam, PROSITE and ProDom for protein annotation.** *Genet Mol Res* 2005, **4**(3):590-598.
116. Han LY, Cai CZ, Ji ZL, Cao ZW, Cui J, Chen YZ: **Predicting functional family of novel enzymes irrespective of sequence similarity: a statistical learning approach.** *Nucleic Acids Res* 2004, **32**(21):6437-6444.
117. Cui J, Han LY, Cai CZ, Zheng CJ, Ji ZL, Chen YZ: **Prediction of functional class of novel bacterial proteins without the use of sequence similarity by a statistical learning method.** *J Mol Microbiol Biotechnol* 2005, **9**(2):86-100.
118. Lin HH, Han LY, Zhang HL, Zheng CJ, Xie B, Chen YZ: **Prediction of the functional class of lipid binding proteins from sequence-derived properties irrespective of sequence similarity.** *J Lipid Res* 2006, **47**(4):824-831.
119. Lin HH, Han LY, Cai CZ, Ji ZL, Chen YZ: **Prediction of transporter family from protein sequence by support vector machine approach.** *Proteins* 2006, **62**(1):218-231.
120. Kumar KK, Shelokar PS: **An SVM method using evolutionary information for the identification of allergenic proteins.** *Bioinformation* 2008, **2**(6):253-256.
121. Mishra NK, Kumar M, Raghava GP: **Support vector machine based prediction of glutathione S-transferase proteins.** *Protein Pept Lett* 2007, **14**(6):575-580.

122. Briscoe G, Caelli T: **A compendium of machine learning** vol. 1. Norwood: Ablex; 1996.
123. Alpaydm E: **Introduction to Machine learning** Cambridge: The MIT Press; 2004.
124. Dietterich TG: **Machine Learning** In: *Nature Encyclopedia of Cognitive Science*. London: Macmillan; 2003.
125. Kotsiantis SB: **Supervised Machine Learning: A Review of Classification Techniques**. *Informatica* 2007, **31**:249-268.
126. Vapnik VN: **The Nature of Statistical Learning Theory**. New York: Springer-Verlag New York Inc; 1995.
127. BURGES CJC: **A Tutorial on Support Vector Machines for Pattern Recognition**. *Data Mining and Knowledge Discovery* 1988, **2**:121-167.
128. Vapnik V: **The nature of statistical learning theory**. New York: Springer-Verlag New York, Inc.; 1995.
129. Nuttakorn Thubthong, Boonserm Kijirikul: **Support vector machines for Thai phoneme recognition**. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 2001, **9**(6).
130. Ben-Yacoub S, Abdeljaoued, Y. & Mayoraz E: **Fusion Face and Speech Data for Person Identity Verification**. *IEEE Transactions on Neural Networks* 1999, **10**:1065-1074.
131. Karlsen RE, Gorsich DJ, Gerhart GR: **Target classification via support vector machines**. *Optical Engineering* 2000, **39**:704-711.
132. Papageorgiou C, Poggio T: **A trainable system for object detection**. *International Journal of Computer Vision* 2000, **38**:15-33.
133. Huang C, Davis LS, Townshend JRG: **An assessment of support vector machines for land cover classification**. *International Journal of Remote Sensing* 2002, **23**:725-749.
134. Karchin R, Karplus K, Haussler D: **Classifying G-protein coupled receptors with support vector machines**. *Bioinformatics* 2002, **18**(1):147-159.
135. 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* 2005, **33**(Web Server issue):W148-153.
136. Bock JR, Gough DA: **Predicting protein--protein interactions from primary structure**. *Bioinformatics* 2001, **17**(5):455-460.
137. Busuttill S, Abela J, Pace GJ: **Support vector machines with profile-based kernels for remote protein homology detection**. *Genome Inform* 2004, **15**(2):191-200.
138. Webb-Robertson BJ, Oehmen C, Matzke M: **SVM-BALSA: remote homology detection based on Bayesian sequence alignment**. *Comput Biol Chem* 2005, **29**(6):440-443.
139. Hongzong S, Tao W, Xiaojun Y, Huanxiang L, Zhide H, Mancang L, BoTao F: **Support vector machines classification for discriminating coronary heart disease patients from non-coronary heart disease**. *West Indian Med J* 2007, **56**(5):451-457.
140. Platt JC: **Sequential Minimal Optimization: A fast algorithm for training support vector machines**. *Microsoft Research Technical Report MSR-TR-98-14* 1998.

141. Osuna E, Freund, R. and Girosi, F.: **An improved training algorithm for support vector machines.** *Neural Networks for Signal Processing VII-Proceedings of the 1997 IEEE Workshop* 1997:276-285.
142. Aizerman MA, Braverman EM, er LIR: **Theoretical foundations of the potential function method in pattern recognition and learning.** *Automation and Remote Control* 1964, **25**:821--837.
143. Courant R, Hilbert D: **Methods of Mathematical Physics:** John Wiley & Sons; 1989.
144. Widjaja E, Zheng W, Huang Z: **Classification of colonic tissues using near-infrared Raman spectroscopy and support vector machines.** *International journal of oncology* 2008, **32**(3):653-662.
145. De Smet F, De Brabanter J, Van den Bosch T, Pochet N, Amant F, Van Holsbeke C, Moerman P, De Moor B, Vergote I, Timmerman D: **New models to predict depth of infiltration in endometrial carcinoma based on transvaginal sonography.** *Ultrasound Obstet Gynecol* 2006, **27**(6):664-671.
146. Dasarathy BV: **Nearest neighbor (NN) norms: NN pattern classification techniques.** In: 1990: IEEE Computer Society Press; 1990: 477.
147. Specht DF: **Probabilistic neural networks.** *Neural Networks* 1990, **3**(1):109-118.
148. Mitchell TM: **Machine Learning.** New York: McGraw-Hill; 1997.
149. Guyon I, Weston J, Barnhill, , S. and Vapnik: **Gene selection for cancer classification using support vector machines.** *Machine Learning* 2002, **46**:389-422.
150. Mao Y, Zhou X, Pi D, Sun Y, Wong ST: **Multiclass cancer classification by using fuzzy support vector machine and binary decision tree with gene selection.** *J Biomed Biotechnol* 2005, **2005**(2):160-171.
151. Xue Y, Yap CW, Sun LZ, Cao ZW, Wang JF, Chen YZ: **Prediction of P-glycoprotein substrates by a support vector machine approach.** *J Chem Inf Comput Sci* 2004, **44**(4):1497-1505.
152. Kohavi R, GH J: **Wrappers for feature subset selection.** *Artificial Intelligence* 1997, **97**(1-2):273-324.
153. Guyon I, Weston, J., Barnhill, S. and Vapnik, V.: **Gene selection for cancer classification using support vector machines.** *Machine Learning* 2002, **46**:389-422.
154. Baldi P, Brunak S, Chauvin Y, Andersen CA, Nielsen H: **Assessing the accuracy of prediction algorithms for classification: an overview.** *Bioinformatics* 2000, **16**(5):412-424.
155. Cai CZ, Han LY, Ji ZL, Chen X, Chen YZ: **SVM-Prot: Web-based support vector machine software for functional classification of a protein from its primary sequence.** *Nucleic acids research* 2003, **31**(13):3692-3697.
156. Han LY, Cai CZ, Ji ZL, Cao ZW, Cui J, Chen YZ: **Predicting functional family of novel enzymes irrespective of sequence similarity: a statistical learning approach.** *Nucleic acids research* 2004, **32**(21):6437-6444.
157. Han LY, Cai CZ, Lo SL, Chung MC, Chen YZ: **Prediction of RNA-binding proteins from primary sequence by a support vector machine approach.** *Rna* 2004, **10**(3):355-368.
158. Cai CZ, Han LY, Ji ZL, Chen X, Chen YZ: **SVM-Prot: Web-based support vector machine software for functional classification of a protein from its primary sequence.** *Nucleic Acids Res* 2003, **31**(13):3692-3697.

159. Alden CJ, Kim SH: **Solvent-accessible surfaces of nucleic acids.** *J Mol Biol* 1979, **132**(3):411-434.
160. Nakagawa S: **Polarizable model potential function for nucleic acid bases.** *J Comput Chem* 2007, **28**(9):1538-1550.
161. Karchin R, Karplus K, Haussler D: **Classifying G-protein coupled receptors with support vector machines.** *Bioinformatics* 2002, **18**(1):147-159.
162. Bairoch A, Apweiler R: **The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000.** *Nucleic Acids Res* 2000, **28**(1):45-48.
163. Bateman A, Birney E, Cerruti L, Durbin R, Eddy SR, Griffiths-Jones S, Howe KL, Marshall M, Sonnhammer EL: **The Pfam protein families database.** *Nucleic acids research* 2002, **30**(1):276-280.
164. Sakanyan V, Charlier D, Legrain C, Kochikyan A, Mett I, Pierard A, Glansdorff N: **Primary structure, partial purification and regulation of key enzymes of the acetyl cycle of arginine biosynthesis in *Bacillus stearothermophilus*: dual function of ornithine acetyltransferase.** *J Gen Microbiol* 1993, **139**(3):393-402.
165. Elkins JM, Kershaw NJ, Schofield CJ: **X-ray crystal structure of ornithine acetyltransferase from the clavulanic acid biosynthesis gene cluster.** *Biochem J* 2005, **385**(Pt 2):565-573.
166. Hum DW, Bell AW, Rozen R, MacKenzie RE: **Primary structure of a human trifunctional enzyme. Isolation of a cDNA encoding methylenetetrahydrofolate dehydrogenase-methenyltetrahydrofolate cyclohydrolase-formyltetrahydrofolate synthetase.** *J Biol Chem* 1988, **263**(31):15946-15950.
167. Andreeva A, Howorth D, Brenner SE, Hubbard TJ, Chothia C, Murzin AG: **SCOP database in 2004: refinements integrate structure and sequence family data.** *Nucleic Acids Res* 2004, **32**(Database issue):D226-229.
168. Farber GK, Petsko GA: **The evolution of alpha/beta barrel enzymes.** *Trends Biochem Sci* 1990, **15**(6):228-234.
169. Reardon D, Farber GK: **The structure and evolution of alpha/beta barrel proteins.** *Faseb J* 1995, **9**(7):497-503.
170. Kanehisa M, Goto S, Hattori M, Aoki-Kinoshita KF, Itoh M, Kawashima S, Katayama T, Araki M, Hirakawa M: **From genomics to chemical genomics: new developments in KEGG.** *Nucleic Acids Res* 2006, **34**(Database issue):D354-357.
171. Caetano-Anolles G, Kim HS, Mitternthal JE: **The origin of modern metabolic networks inferred from phylogenomic analysis of protein architecture.** *Proc Natl Acad Sci U S A* 2007, **104**(22):9358-9363.
172. Tatusov RL, Koonin EV, Lipman DJ: **A genomic perspective on protein families.** *Science* 1997, **278**(5338):631-637.
173. Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B, Koonin EV, Krylov DM, Mazumder R, Mekhedov SL, Nikolskaya AN *et al*: **The COG database: an updated version includes eukaryotes.** *BMC Bioinformatics* 2003, **4**:41.
174. Xu JR, Zhang JX, Han BC, Liang L, Ji ZL: **CytoSVM: an advanced server for identification of cytokine-receptor interactions.** *Nucleic Acids Res* 2007, **35**(Web Server issue):W538-542.
175. Todd AE, Orengo CA, Thornton JM: **Plasticity of enzyme active sites.** *Trends in biochemical sciences* 2002, **27**(8):419-426.

176. Ding CH, Dubchak I: **Multi-class protein fold recognition using support vector machines and neural networks.** *Bioinformatics* 2001, **17**(4):349-358.
177. Cai YD, Liu XJ, Xu XB, Chou KC: **Prediction of protein structural classes by support vector machines.** *Comput Chem* 2002, **26**(3):293-296.
178. Cai YD, Liu XJ, Xu XB, Chou KC: **Support Vector Machines for predicting HIV protease cleavage sites in protein.** *J Comput Chem* 2002, **23**(2):267-274.
179. Lee HC, Graeff RM, Walseth TF: **ADP-ribosyl cyclase and CD38. Multi-functional enzymes in Ca<sup>2+</sup> signaling.** *Adv Exp Med Biol* 1997, **419**:411-419.
180. Chen YH, Li MH, Zhang Y, He LL, Yamada Y, Fitzmaurice A, Shen Y, Zhang H, Tong L, Yang J: **Structural basis of the alpha1-beta subunit interaction of voltage-gated Ca<sup>2+</sup> channels.** *Nature* 2004, **429**(6992):675-680.
181. Tochio H, Ohki S, Zhang Q, Li M, Zhang M: **Solution structure of a protein inhibitor of neuronal nitric oxide synthase.** *Nat Struct Biol* 1998, **5**(11):965-969.
182. Bhasin M, Reinherz EL, Reche PA: **Recognition and classification of histones using support vector machine.** *J Comput Biol* 2006, **13**(1):102-112.
183. Zasloff M: **Antimicrobial peptides of multicellular organisms.** *Nature* 2002, **415**(6870):389-395.
184. Matsuzaki K: **Why and how are peptide-lipid interactions utilized for self-defense? Magainins and tachyplesins as archetypes.** *Biochim Biophys Acta* 1999, **1462**(1-2):1-10.
185. Lata S, Sharma BK, Raghava GP: **Analysis and prediction of antibacterial peptides.** *BMC Bioinformatics* 2007, **8**(1):263.
186. Fierro-Monti I, Mathews MB: **Proteins binding to duplexed RNA: one motif, multiple functions.** *Trends Biochem Sci* 2000, **25**(5):241-246.
187. Paulsen IT: **Multidrug efflux pumps and resistance: regulation and evolution.** *Curr Opin Microbiol* 2003, **6**(5):446-451.
188. Gatzeva-Topalova PZ, May AP, Sousa MC: **Structure and mechanism of ArnA: conformational change implies ordered dehydrogenase mechanism in key enzyme for polymyxin resistance.** *Structure* 2005, **13**(6):929-942.
189. Sugantino M, Roderick SL: **Crystal structure of Vat(D): an acetyltransferase that inactivates streptogramin group A antibiotics.** *Biochemistry* 2002, **41**(7):2209-2216.
190. Brown MP, Grundy WN, Lin D, Cristianini N, Sugnet CW, Furey TS, Ares M, Jr., Haussler D: **Knowledge-based analysis of microarray gene expression data by using support vector machines.** *Proc Natl Acad Sci U S A* 2000, **97**(1):262-267.
191. Frazier DP, Wilson A, Graham RM, Thompson JW, Bishopric NH, Webster KA: **Acidosis regulates the stability, hydrophobicity, and activity of the BH3-only protein Bnip3.** *Antioxid Redox Signal* 2006, **8**(9-10):1625-1634.
192. Lewis C, Baro MF, Marques M, Gruner M, Alonso A, Bravo IG: **The first hydrophobic region of the HPV16 E5 protein determines protein cellular location and facilitates anchorage-independent growth.** *Virol J* 2008, **5**:30.
193. Campbell-Valois FX, Michnick SW: **The transition state of the ras binding domain of Raf is structurally polarized based on Phi-values but is energetically diffuse.** *J Mol Biol* 2007, **365**(5):1559-1577.

194. Ptacek JB, Edwards AP, Freeman-Cook LL, DiMaio D: **Packing contacts can mediate highly specific interactions between artificial transmembrane proteins and the PDGFBeta receptor.** *Proc Natl Acad Sci U S A* 2007, **104**(29):11945-11950.
195. Ma Y, Cunningham ME, Wang X, Ghosh I, Regan L, Longley BJ: **Inhibition of spontaneous receptor phosphorylation by residues in a putative alpha-helix in the KIT intracellular juxtamembrane region.** *J Biol Chem* 1999, **274**(19):13399-13402.
196. Espinoza-Fonseca LM, Trujillo-Ferrara JG: **Transient stability of the helical pattern of region F19-L22 of the N-terminal domain of p53: a molecular dynamics simulation study.** *Biochem Biophys Res Commun* 2006, **343**(1):110-116.
197. Griffiths-Jones S: **The microRNA Registry.** *Nucleic Acids Res* 2004, **32**(Database issue):D109-111.
198. Griffiths-Jones S, Moxon S, Marshall M, Khanna A, Eddy SR, Bateman A: **Rfam: annotating non-coding RNAs in complete genomes.** *Nucleic Acids Res* 2005, **33**(Database issue):D121-124.
199. Imanishi T, Itoh T, Suzuki Y, O'Donovan C, Fukuchi S, Koyanagi KO, Barrero RA, Tamura T, Yamaguchi-Kabata Y, Tanino M *et al*: **Integrative annotation of 21,037 human genes validated by full-length cDNA clones.** *PLoS Biol* 2004, **2**(6):e162.
200. Yue J, Sheng Y, Orwig KE: **Identification of novel homologous microRNA genes in the rhesus macaque genome.** *BMC Genomics* 2008, **9**:8.
201. Hubbard TJ, Aken BL, Beal K, Ballester B, Caccamo M, Chen Y, Clarke L, Coates G, Cunningham F, Cutts T *et al*: **Ensembl 2007.** *Nucleic Acids Res* 2007, **35**(Database issue):D610-617.
202. Hofacker IL: **Vienna RNA secondary structure server.** *Nucleic Acids Res* 2003, **31**(13):3429-3431.
203. Brameier M, Wiuf C: **Ab initio identification of human microRNAs based on structure motifs.** *BMC Bioinformatics* 2007, **8**:478.
204. Zuker M, Stiegler P: **Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information.** *Nucleic Acids Res* 1981, **9**(1):133-148.
205. Sewer A, Paul N, Landgraf P, Aravin A, Pfeffer S, Brownstein MJ, Tuschl T, van Nimwegen E, Zavolan M: **Identification of clustered microRNAs using an ab initio prediction method.** *BMC Bioinformatics* 2005, **6**:267.
206. Ng Kwang Loong S, Mishra SK: **Unique folding of precursor microRNAs: quantitative evidence and implications for de novo identification.** *Rna* 2007, **13**(2):170-187.
207. Huang TH, Fan B, Rothschild MF, Hu ZL, Li K, Zhao SH: **MiRFinder: an improved approach and software implementation for genome-wide fast microRNA precursor scans.** *BMC Bioinformatics* 2007, **8**:341.
208. Lai EC: **microRNAs: runts of the genome assert themselves.** *Curr Biol* 2003, **13**(23):R925-936.
209. Wang X, Zhang J, Li F, Gu J, He T, Zhang X, Li Y: **MicroRNA identification based on sequence and structure alignment.** *Bioinformatics* 2005, **21**(18):3610-3614.
210. Hertel J, Stadler PF: **Hairpins in a Haystack: recognizing microRNA precursors in comparative genomics data.** *Bioinformatics* 2006, **22**(14):e197-202.



211. Baskerville S, Bartel DP: **Microarray profiling of microRNAs reveals frequent coexpression with neighboring miRNAs and host genes.** *Rna* 2005, **11**(3):241-247.
212. Chothia C: **Proteins. One thousand families for the molecular biologist.** *Nature* 1992, **357**(6379):543-544.
213. Jiang P, Wu H, Wang W, Ma W, Sun X, Lu Z: **MiPred: classification of real and pseudo microRNA precursors using random forest prediction model with combined features.** *Nucleic Acids Res* 2007, **35**(Web Server issue):W339-344.
214. TAEHO JO HJ: **A multiple resampling method for learning from imbalanced data sets.** *Computational Intelligence* 2004, **20**:19.
215. Tang Y, Zhang YQ, Chawla NV, Krasser S: **SVMs Modeling for Highly Imbalanced Classification.** *IEEE Trans Syst Man Cybern B Cybern* 2008.

## APPENDICES

**S1 Scanning results of *E. coli* K12 genome (# indicates that data were not included in our model development)**

Access number in NCBI	Gene Name	Entry Name in Swiss-Prot	Protein Name	Function described Swiss-Prot	Status
AAC73137.1	ileS	SYI_ECOLI	Isoleucyl-tRNA synthetase (EC 6.1.1.5) (Isoleucine--tRNA ligase) (IleRS)	Aminoacyl-tRNA synthetase; Antibiotic resistance; ATP-binding; Complete proteome; Cytoplasm; Direct protein sequencing; Ligase; Metal-binding; Nucleotide-binding; Protein biosynthesis; Zinc.	known
AAC73159.1	folA	DYR_ECOLI	Dihydrofolate reductase (EC 1.5.1.3)	3D-structure; Antibiotic resistance; Complete proteome; Direct protein sequencing; Methotrexate resistance; NADP; One-carbon metabolism; Oxidoreductase; Trimethoprim resistance.	known
AAC73162.1	ksgA	KSGA_ECOLI	Dimethyladenosine transferase (EC 2.1.1.-) (S-adenosylmethionine-6-N', N'-adenosyl(rRNA) dimethyltransferase) (16S rRNA dimethylase) (High level kasugamycin resistance protein ksgA) (Kasugamycin dimethyltransferase)	3D-structure; Antibiotic resistance; Complete proteome; Methyltransferase; RNA-binding; rRNA processing; S-adenosyl-L-methionine; Transferase.	known
AAC73195.1	ftsI	FTSI_ECOLI	Peptidoglycan synthetase ftsI precursor (EC 2.4.1.129) (Peptidoglycan glycosyltransferase 3) (Penicillin-binding protein 3) (PBP-3)	Antibiotic resistance; Cell cycle; Cell division; Cell shape; Cell wall biogenesis/degradation; Complete proteome; Glycosyltransferase; Inner membrane; Membrane; Multifunctional enzyme; Peptidoglycan synthesis; Transferase; Transmembrane.	known
AAC73260.1	mrcB	PBPB_ECOLI	Penicillin-binding protein 1B (PBP-1b) (PBP1b) (Murein polymerase)	Alternative initiation; Antibiotic resistance; Cell shape; Cell wall biogenesis/degradation; Complete proteome; Direct protein sequencing; Glycosyltransferase; Hydrolase; Inner membrane; Membrane; Multifunctional enzyme; Peptidoglycan synthesis; Signal-anchor; Transferase; Transmembrane.	known
AAC73290.1	lpxD	LPXD_ECOLI	UDP-3-O-[3-hydroxymyristoyl] glucosamine N-acyltransferase (EC 2.3.1.-)	Acyltransferase; Antibiotic resistance; Complete proteome; Direct protein sequencing; Lipid A biosynthesis; Lipid synthesis; Repeat; Transferase.	known

AAC73581.1	fsr	FSR_ECOLI	Fosmidomycin resistance protein	Antibiotic resistance; Complete proteome; Inner membrane; Membrane; Transmembrane.	known
AAC73928.1	ybjG	YBJG_ECOLI	Putative undecaprenyl-diphosphatase ybjG (EC 3.6.1.27) (Undecaprenyl pyrophosphate phosphatase).	Antibiotic resistance; Cell shape; Cell wall biogenesis/degradation; Complete proteome; Hydrolase; Inner membrane; Membrane; Peptidoglycan synthesis; Transmembrane.	known
AAC74137.1	mdtG	MDTG_ECOLI	Multidrug resistance protein mdtG	Antibiotic resistance; Complete proteome; Inner membrane; Membrane; Transmembrane; Transport.	known
AAC74149.2	mdtH	MDTH_ECOLI	Multidrug resistance protein mdtH	Antibiotic resistance; Complete proteome; Inner membrane; Membrane; Transmembrane; Transport.	known
AAC74370.1	fabI	FABI_ECOLI	Enoyl-[acyl-carrier-protein] reductase [NADH]	3D-structure; Antibiotic resistance; Complete proteome; Direct protein sequencing; Fatty acid biosynthesis; Inner membrane; Lipid synthesis; Membrane; NAD; Oxidoreductase.	known
AAC74671.1	mdtI	MDTI_ECOLI	Multidrug resistance protein mdtI	Complete proteome; Inner membrane; Membrane; Transmembrane; Transport.	#known
AAT48136.1	mdtK	MDTK_ECOLI	Multidrug resistance protein mdtK (Multidrug-efflux transporter)	Antibiotic resistance; Antiport; Complete proteome; Inner membrane; Ion transport; Membrane; Sodium; Sodium transport; Transmembrane; Transport.	known
AAC75243.1	bcr	BCR_ECOLI	Bicyclomycin resistance protein (Sulfonamide resistance protein)	Antibiotic resistance; Complete proteome; Inner membrane; Membrane; Transmembrane; Transport.	known
AAC75271.1	yojI	YOJI_ECOLI	ABC transporter ATP-binding protein yojI	Antibiotic resistance; ATP-binding; Complete proteome; Inner membrane; Membrane; Nucleotide-binding; Transmembrane; Transport.	known
AAC75291.1	gyrA	GYRA_ECOLI	DNA gyrase subunit A (EC 5.99.1.3)	3D-structure; Antibiotic resistance; Complete proteome; DNA-binding; Isomerase; Topoisomerase.	known
AAC75319.2	pmrD	PMRD_ECOLI	Signal transduction protein pmrD	Antibiotic resistance; Complete proteome.	known
AAC75523.1	acrD	ACRD_ECOLI	Probable aminoglycoside efflux pump	Complete proteome; Inner membrane; Membrane; Transmembrane; Transport.	#known
AAC76066.1	parE	PARE_ECOLI	DNA topoisomerase 4 subunit B (EC 5.99.1.-)	3D-structure; Antibiotic resistance; ATP-binding; Complete proteome; Isomerase; Nucleotide-binding; Topoisomerase.	known

AAC76093.1	bacA	UPPP_ECOLI	Undecaprenyl-diphosphatase (EC 3.6.1.27)	Antibiotic resistance; Cell shape; Cell wall biogenesis/degradation; Complete proteome; Hydrolase; Inner membrane; Membrane; Peptidoglycan synthesis; Transmembrane.	known
AAC76209.2	folP	DHPS_ECOLI	Dihydropteroate synthase (EC 2.5.1.15)	3D-structure; Antibiotic resistance; Complete proteome; Direct protein sequencing; Folate biosynthesis; Transferase.	known
AAC76214.1	dacB	PBP4_ECOLI	Penicillin-binding protein 4 precursor (PBP-4)	3D-structure; Antibiotic resistance; Cell cycle; Cell division; Cell shape; Cell wall biogenesis/degradation; Complete proteome; Direct protein sequencing; Hydrolase; Peptidoglycan synthesis; Periplasm; Signal.	known
AAC76321.1	rpsD	RS4_ECOLI	30S ribosomal protein S4	3D-structure; Antibiotic resistance; Complete proteome; Direct protein sequencing; Repressor; Ribonucleoprotein; Ribosomal protein; RNA-binding; rRNA-binding; Transcription; Transcription regulation; Transcription termination; Translation regulation.	known
AAC76328.1	rpsE	RS5_ECOLI	30S ribosomal protein S5	3D-structure; Acetylation; Antibiotic resistance; Complete proteome; Direct protein sequencing; Ribonucleoprotein; Ribosomal protein; RNA-binding; rRNA-binding.	known
AAC76330.1	rplF	RL6_ECOLI	50S ribosomal protein L6	3D-structure; Antibiotic resistance; Complete proteome; Direct protein sequencing; Ribonucleoprotein; Ribosomal protein; RNA-binding; rRNA-binding.	known
AAC76336.1	rpsQ	RS17_ECOLI	30S ribosomal protein S17	3D-structure; Antibiotic resistance; Complete proteome; Direct protein sequencing; Ribonucleoprotein; Ribosomal protein; RNA-binding; rRNA-binding.	known
AAC76344.1	rplD	RL4_ECOLI	50S ribosomal protein L4	3D-structure; Antibiotic resistance; Complete proteome; Direct protein sequencing; Repressor; Ribonucleoprotein; Ribosomal protein; RNA-binding; rRNA-binding; Transcription; Transcription regulation; Transcription termination; Translation regulation.	known
AAC76364.1	tufA	EFTU_ECOLI	Elongation factor Tu (EF-Tu) (P-43)	3D-structure; Acetylation; Antibiotic resistance; Complete proteome; Cytoplasm; Direct protein	#known

				sequencing; Elongation factor; GTP-binding; Membrane; Methylation; Nucleotide-binding; Phosphorylation; Protein biosynthesis.	
AAC76367.1	rpsL	RS12_ECOLI	30S ribosomal protein S12	3D-structure; Antibiotic resistance; Complete proteome; Direct protein sequencing; Ribonucleoprotein; Ribosomal protein; RNA-binding; rRNA-binding; tRNA-binding.	known
AAT48201.1	gyrB	GYRB_ECOLI	DNA gyrase subunit B (EC 5.99.1.3)	3D-structure; Antibiotic resistance; ATP-binding; Complete proteome; Direct protein sequencing; Isomerase; Nucleotide-binding; Topoisomerase.	known
AAC76733.1	mdtL	MDTL_ECOLI	Multidrug resistance protein mdtL	Antibiotic resistance; Complete proteome; Inner membrane; Membrane; Transmembrane; Transport.	known
AAC76961.1	rpoB	RPOB_ECOLI	DNA-directed RNA polymerase subunit beta (EC 2.7.7.6) (RNAP subunit beta) (Transcriptase subunit beta) (RNA polymerase subunit beta)	Complete proteome; DNA-directed RNA polymerase; Nucleotidyltransferase; Transcription; Transferase.	#known
AAC77074.1	basR	BASR_ECOLI	Transcriptional regulatory protein basR/pmrA.	Antibiotic resistance; Complete proteome; Cytoplasm; DNA-binding; Phosphorylation; Transcription; Transcription regulation; Two-component regulatory system.	known
AAC77110.1	ampC	AMPC_ECOLI	Beta-lactamase precursor (EC 3.5.2.6)	3D-structure; Antibiotic resistance; Complete proteome; Direct protein sequencing; Hydrolase; Periplasm; Signal.	known
AAC77293.1	mdtM	MDTM_ECOLI	Multidrug resistance protein mdtM	Antibiotic resistance; Complete proteome; Inner membrane; Membrane; Transmembrane; Transport.	known
AAC73564.1	acrB	ACRB_ECOLI	Acriflavine resistance protein B	3D-structure; Complete proteome; Inner membrane; Membrane; Transmembrane; Transport.	#known
AAC73565.1	acrA	ACRA_ECOLI	Acriflavine resistance protein A precursor	3D-structure; Antibiotic resistance; Complete proteome; Inner membrane; Lipoprotein; Membrane; Palmitate; Signal; Transport.	known
AAC73736.1	mrda	PBP2_ECOLI	Penicillin-binding protein 2 (PBP-2)	Antibiotic resistance; Cell shape; Cell wall biogenesis/degradation; Complete proteome; Inner membrane; Membrane; Multifunctional enzyme; Peptidoglycan synthesis.	known
AAC73965.2	macA	MACA_ECOLI	Macrolide-specific efflux protein macA precursor	Antibiotic resistance; Complete proteome; Inner membrane; Membrane; Signal; Transport.	known

AAC73966.1	macB	MACB_ECOLI	Macrolide export ATP-binding/permease protein macB (EC 3.6.3.-)	Antibiotic resistance; ATP-binding; Complete proteome; Hydrolase; Inner membrane; Membrane; Nucleotide-binding; Transmembrane; Transport.	known
AAC74511.1	tehA	TEHA_ECOLI	Tellurite resistance protein tehA	Antibiotic resistance; Complete proteome; Inner membrane; Membrane; Tellurium resistance; Transmembrane.	known
AAC74512.1	tehB	TEHB_ECOLI	Tellurite resistance protein tehB	Antibiotic resistance; Complete proteome; Cytoplasm; Tellurium resistance.	known
AAC74602.1	marC	MARC_ECOLI	Multiple antibiotic resistance protein marC	Antibiotic resistance; Complete proteome; Inner membrane; Membrane; Transmembrane.	known
AAC74603.2	marR	MARR_ECOLI	Multiple antibiotic resistance protein marR	3D-structure; Antibiotic resistance; Complete proteome; DNA-binding; Repressor; Transcription; Transcription regulation.	known
AAC74604.2	marA	MARA_ECOLI	Multiple antibiotic resistance protein marA	3D-structure; Activator; Antibiotic resistance; Complete proteome; DNA-binding; Repeat; Transcription; Transcription regulation.	known
AAC74605.1	marB	MARB_ECOLI	Multiple antibiotic resistance protein marB	Antibiotic resistance; Complete proteome.	known
AAC75136.1	mdtB	MDTB_ECOLI	Multidrug resistance protein mdtB	Complete proteome; Inner membrane; Membrane; Transmembrane; Transport.	#known
AAC75137.1	mdtC	MDTC_ECOLI	Multidrug resistance protein mdtC	Complete proteome; Inner membrane; Membrane; Transmembrane; Transport.	#known
AAC75138.1	mdtD	MDTD_ECOLI	Putative multidrug resistance protein mdtD	Complete proteome; Inner membrane; Membrane; Transmembrane; Transport.	#known
AAC75313.3	arnB	ARNB_ECOLI	UDP-4-amino-4-deoxy-L-arabinose--oxoglutarate aminotransferase (EC 2.6.1.-)	Aminotransferase; Antibiotic resistance; Complete proteome; Lipid A biosynthesis; Lipid synthesis; Lipopolysaccharide biosynthesis; Pyridoxal phosphate; Transferase.	known
AAC75314.1	arnC	ARNC_ECOLI	Undecaprenyl-phosphate 4-deoxy-4-formamido-L-arabinose transferase (EC 2.7.8.-)	Antibiotic resistance; Complete proteome; Glycosyltransferase; Inner membrane; Lipid A biosynthesis; Lipid synthesis; Lipopolysaccharide biosynthesis; Membrane; Transferase; Transmembrane.	known
AAC75315.1	arnA	ARNA_ECOLI	Bifunctional polymyxin resistance protein arnA (Polymyxin resistance protein pmrI)	3D-structure; Antibiotic resistance; Complete proteome; Lipid A biosynthesis; Lipid synthesis; Lipopolysaccharide biosynthesis; Methyltransferase; Multifunctional enzyme; NAD; Oxidoreductase; Transferase.	known

AAC75732.1	emrA	EMRA_ECOLI	Multidrug resistance protein A	Antibiotic resistance; Complete proteome; Inner membrane; Membrane; Transmembrane; Transport.	known
AAC75733.1	emrB	EMRB_ECOLI	Multidrug resistance protein B	Antibiotic resistance; Complete proteome; Inner membrane; Membrane; Transmembrane; Transport.	known
AAC76297.1	acrE	ACRE_ECOLI	Acriflavine resistance protein E precursor (Protein envC)	Cell cycle; Cell division; Complete proteome; Inner membrane; Lipoprotein; Membrane; Palmitate; Signal.	#known
AAC76298.1	acrF	ACRF_ECOLI	Acriflavine resistance protein F (Protein envD)	Cell cycle; Cell division; Complete proteome; Inner membrane; Membrane; Transmembrane; Transport.	#known
AAC76340.1	rplV	RL22_ECOLI	50S ribosomal protein L22	3D-structure; Antibiotic resistance; Complete proteome; Direct protein sequencing; Ribonucleoprotein; Ribosomal protein; RNA-binding; rRNA-binding.	known
AAC76421.2	mrcA	PBPA_ECOLI	Penicillin-binding protein 1A (PBP-1a) (PBP1a)	Antibiotic resistance; Cell shape; Cell wall biogenesis/degradation; Complete proteome; Direct protein sequencing; Glycosyltransferase; Hydrolase; Inner membrane; Membrane; Multifunctional enzyme; Peptidoglycan synthesis; Signal-anchor; Transferase; Transmembrane.	known
AAC76538.1	mdtE	MDTE_ECOLI	Multidrug resistance protein mdtE precursor	Antibiotic resistance; Complete proteome; Inner membrane; Lipoprotein; Membrane; Palmitate; Signal; Transport.	known
AAC76539.1	mdtF	MDTF_ECOLI	Multidrug resistance protein mdtF	Antibiotic resistance; Complete proteome; Inner membrane; Membrane; Transmembrane; Transport.	known
AAC76954.1	tufB	EFTU_ECOLI	Elongation factor Tu (EF-Tu) (P-43)	3D-structure; Acetylation; Antibiotic resistance; Complete proteome; Cytoplasm; Direct protein sequencing; Elongation factor; GTP-binding; Membrane; Methylation; Nucleotide-binding; Phosphorylation; Protein biosynthesis.	known
AAD13463.1	mdtP	MDTP_ECOLI	Multidrug resistance outer membrane protein mdtP precursor	Antibiotic resistance; Complete proteome; Lipoprotein; Membrane; Outer membrane; Palmitate; Signal.	known
AAD13464.2	mdtO	MDTO_ECOLI	Multidrug resistance protein mdtO	Antibiotic resistance; Complete proteome; Inner membrane; Membrane; Transmembrane;	known

				Transport.	
AAD13465.1	mdtN	MDTN_ECOLI	Multidrug resistance protein mdtN	Antibiotic resistance; Complete proteome; Inner membrane; Membrane; Transmembrane; Transport.	known
AAC73176.1	yabI	YABI_ECOLI	Inner membrane protein yabI	Complete proteome; Inner membrane; Membrane; Transmembrane.	predicted
AAC73309.1	metI	METI_ECOLI	D-methionine transport system permease protein metI	Amino-acid transport; Complete proteome; Inner membrane; Membrane; Transmembrane; Transport.	predicted
AAC73356.2	ykfA	YKFA_ECOLI	Uncharacterized protein ykfA	Complete proteome; GTP-binding; Nucleotide-binding.	predicted
AAC73388.1	yagS	YAGS_ECOLI	Putative xanthine dehydrogenase yagS FAD-binding subunit (EC 1.17.1.4)	Complete proteome; FAD; Flavoprotein; NAD; Oxidoreductase; Purine metabolism; Purine salvage.	predicted
AAC73424.1	yahG	YAHG_ECOLI	Uncharacterized protein yahG.	Complete proteome; Membrane; Transmembrane.	predicted
AAC73443.1	cynS	CYNS_ECOLI	Cyanate hydratase (EC 4.2.1.104)	3D-structure; Complete proteome; Direct protein sequencing; Lyase.	predicted
AAC73456.2	mhpT	MHPT_ECOLI	Putative 3-hydroxyphenylpropionic acid transporter	Complete proteome; Inner membrane; Membrane; Symport; Transmembrane; Transport.	predicted
AAC73499.1	araJ	ARAJ_ECOLI	Protein araJ	Complete proteome; Inner membrane; Membrane; Transmembrane.	predicted
AAC73511.1	secD	SECD_ECOLI	Protein-export membrane protein secD	Complete proteome; Inner membrane; Membrane; Protein transport; Translocation; Transmembrane; Transport.	predicted
AAC73543.1	hupB	DBHB_ECOLI	DNA-binding protein HU-beta	Complete proteome; Direct protein sequencing; DNA condensation; DNA-binding.	predicted
AAC73676.1	cusA	CUSA_ECOLI	Cation efflux system protein cusA	Complete proteome; Copper; Copper transport; Inner membrane; Ion transport; Membrane; Transmembrane; Transport.	predicted
AAC73692.1	entS	ENTS_ECOLI	Enterobactin exporter entS	Complete proteome; Inner membrane; Membrane; Transmembrane; Transport.	predicted
AAC73735.1	mrdb	RODA_ECOLI	Rod shape-determining protein rodA	Cell shape; Complete proteome; Inner membrane; Membrane; Transmembrane.	predicted
AAC75729.1	ygaZ	YGAZ_ECOLI	Inner membrane protein ygaZ	Complete proteome; Inner membrane; Membrane; Transmembrane; Transport.	predicted
AAC75730.1	ygaH	YGAH_ECOLI	predicted inner membrane protein	Complete proteome.	predicted



AAC76341.1	rpsS	RS19_ECOLI	30S ribosomal protein S19	3D-structure; Complete proteome; Direct protein sequencing; Ribonucleoprotein; Ribosomal protein; RNA-binding; rRNA-binding; tRNA-binding.	predicted
AAC76420.2	hofM	YRFD_ECOLI	predicted pilus assembly protein	Complete proteome.	predicted
AAT48187.1	rbbA	YHIH_ECOLI	Uncharacterized ABC transporter ATP-binding protein yhiH	ATP-binding; Complete proteome; Inner membrane; Membrane; Nucleotide-binding; Repeat; Transmembrane; Transport.	predicted
AAC76512.1	yhiI	YHII_ECOLI	Uncharacterized protein yhiI precursor	Complete proteome; Signal.	predicted
AAC76955.1	secE	SECE_ECOLI	Preprotein translocase subunit secE	3D-structure; Complete proteome; Inner membrane; Membrane; Protein transport; Translocation; Transmembrane; Transport.	predicted
AAC75282.1	atoA	ATOA_ECOLI	Acetate CoA-transferase subunit beta (EC 2.8.3.8)	Complete proteome; Fatty acid metabolism; Lipid metabolism; Transferase.	predicted
AAC75284.1	atoB	ATOB_ECOLI	Acetyl-CoA acetyltransferase (EC 2.3.1.9)	Acyltransferase; Complete proteome; Cytoplasm; Fatty acid metabolism; Lipid metabolism; Transferase.	predicted
AAC73796.1	ybfB	YBFB_ECOLI	predicted inner membrane protein	Complete proteome; Membrane; Transmembrane.	predicted
AAC73831.1	tolQ	TOLQ_ECOLI	membrane spanning protein in TolA-TolQ-TolR complex	Bacteriocin transport; Complete proteome; Inner membrane; Membrane; Protein transport; Transmembrane; Transport.	predicted
AAC73878.1	ybhQ	YBHQ_ECOLI	Inner membrane protein ybhQ	Complete proteome; Inner membrane; Membrane; Transmembrane.	predicted
AAC73882.1	ybhG	YBHG_ECOLI	UPF0194 membrane protein ybhG precursor	Coiled coil; Complete proteome; Periplasm; Signal.	predicted
AAC73915.1	iaaA	ASGX_ECOLI	Putative L-asparaginase precursor (EC 3.5.1.1)	3D-structure; Complete proteome; Hydrolase; Signal.	predicted
AAC73917.1	gsiB	GSIB_ECOLI	Glutathione-binding protein gsiB precursor	3D-structure; Complete proteome; Periplasm; Signal; Transport.	predicted
AAC73923.1	bssR	BSSR_ECOLI	Biofilm regulator bssR	Complete proteome.	predicted
AAC73991.2	ycaO	YCAO_ECOLI	UPF0142 protein ycaO	Complete proteome.	predicted
AAC74110.1	ycdT	YCDT_ECOLI	Inner membrane protein ycdT	Complete proteome; Inner membrane; Membrane; Transmembrane.	predicted
AAC74274.1	dadX	ALR2_ECOLI	Alanine racemase, catabolic	Complete proteome; Isomerase; Pyridoxal phosphate.	predicted

AAC74388.1	pspC	PSPC_ECOLI	Phage shock protein C	Activator; Complete proteome; Stress response; Transcription; Transcription regulation.	predicted
ABD18660.1	rzoR	RZOR_ECOLI	Outer membrane lipoprotein RzI from lambdoid prophage Rac precursor	Complete proteome; Lipoprotein; Membrane; Outer membrane; Palmitate; Phage lysis protein; Signal.	predicted
AAC74534.1	yncE	YNCE_ECOLI	Uncharacterized protein yncE precursor	ATP-binding; Complete proteome; Nucleotide-binding; Signal.	predicted
AAD13437.3	yddG	YDDG_ECOLI	Inner membrane protein yddG	Complete proteome; Inner membrane; Membrane; Transmembrane.	predicted
AAC74648.1	ydfD	YDFD_ECOLI	Uncharacterized protein ydfD	Complete proteome.	predicted
AAC74668.1	ynfM	YNFM_ECOLI	Inner membrane transport protein ynfM	Complete proteome; Inner membrane; Membrane; Transmembrane; Transport.	predicted
AAC74675.1	pntA	PNTA_ECOLI	NAD(P) transhydrogenase subunit alpha (EC 1.6.1.2)	3D-structure; Complete proteome; Direct protein sequencing; Inner membrane; Membrane; NAD; NADP; Oxidoreductase; Transmembrane.	predicted
AAC74703.1	rsxG	RNFG_ECOLI	Electron transport complex protein rnfG	Complete proteome; Electron transport; Inner membrane; Membrane; Transmembrane; Transport.	predicted
AAC74712.1	anmK	ANMK_ECOLI	Anhydro-N-acetylmuramic acid kinase (EC 2.7.1.-) (AnhMurNAc kinase)	ATP-binding; Carbohydrate metabolism; Complete proteome; Kinase; Nucleotide-binding; Transferase.	predicted
AAC74716.2	ydhJ	YDHJ_ECOLI	Uncharacterized protein ydhJ	Complete proteome; Membrane; Transmembrane.	predicted
AAC74729.1	ydhP	YDHP_ECOLI	Inner membrane transport protein ydhP	Complete proteome; Inner membrane; Membrane; Transmembrane; Transport.	predicted
AAC74740.1	ydhU	PHSC_ECOLI	Protein phsC homolog	Complete proteome; Inner membrane; Membrane; Transmembrane.	predicted
AAC74828.2	ynjF	YNJF_ECOLI	Inner membrane protein ynjF	Complete proteome; Inner membrane; Membrane; Transferase; Transmembrane.	predicted
AAC74858.2	yoaI	YOAI_ECOLI	Uncharacterized protein yoaI	Complete proteome; Membrane; Transmembrane.	predicted
AAC74868.1	leuE	YEAS_ECOLI	neutral amino-acid efflux system	Complete proteome; Membrane; Transmembrane.	predicted
AAC74896.1	mgrB	YOBG_ECOLI	Uncharacterized protein yobG	Complete proteome.	predicted
AAC74898.2	yebQ	YEBQ_ECOLI	Uncharacterized transporter yebQ	Complete proteome; Inner membrane; Membrane; Transmembrane; Transport.	predicted
AAC75121.2	wzc	WZC_ECOLI	Tyrosine-protein kinase wzc (EC 2.7.10.2)	Complete proteome; Exopolysaccharide synthesis; Inner membrane; Kinase; Membrane; Phosphorylation; Transferase; Transmembrane;	predicted

				Tyrosine-protein kinase.	
AAC75153.1	gatC	PTKC_ECOLI	Galactitol permease IIC component	Complete proteome; Galactitol metabolism; Inner membrane; Membrane; Phosphotransferase system; Sugar transport; Transmembrane; Transport.	predicted
AAC75206.1	yeiS	YEIS_ECOLI	predicted inner membrane protein	Complete proteome; Membrane; Transmembrane.	predicted
AAC75218.1	yeiE	YEIE_ECOLI	Uncharacterized HTH-type transcriptional regulator yeiE	Complete proteome; DNA-binding; Transcription; Transcription regulation.	predicted
AAC75372.1	purF	PUR1_ECOLI	Amidophosphoribosyltransferase (EC 2.4.2.14)	3D-structure; Complete proteome; Direct protein sequencing; Glutamine amidotransferase; Glycosyltransferase; Magnesium; Metal-binding; Purine biosynthesis; Transferase.	predicted
AAC75382.1	yfcJ	YFCJ_ECOLI	UPF0226 protein yfcJ	Complete proteome; Inner membrane; Membrane; Transmembrane.	predicted
AAC75387.1	yfcA	YFCA_ECOLI	Inner membrane protein yfcA	Complete proteome; Inner membrane; Membrane; Transmembrane.	predicted
AAC75468.1	ptsH	PTHP_ECOLI	Phosphocarrier protein HPr	3D-structure; Complete proteome; Cytoplasm; Direct protein sequencing; Phosphorylation; Phosphotransferase system; Sugar transport; Transport.	predicted
AAC75556.1	yfgF	YFGF_ECOLI	Inner membrane protein yfgF	Complete proteome; Inner membrane; Membrane; Transmembrane.	predicted
AAC75589.1	hcaT	HCAT_ECOLI	Probable 3-phenylpropionic acid transporter	Complete proteome; Inner membrane; Membrane; Symport; Transmembrane; Transport.	predicted
AAC75599.1	yphD	YPHD_ECOLI	Probable ABC transporter permease protein yphD	Complete proteome; Inner membrane; Membrane; Transmembrane; Transport.	predicted
AAC75736.1	yqaA	YQAA_ECOLI	Inner membrane protein yqaA	Complete proteome; Inner membrane; Membrane; Transmembrane.	predicted
AAC76003.1	yqgA	YQGA_ECOLI	predicted inner membrane protein	Complete proteome; Membrane; Transmembrane.	predicted
AAC76008.2	pppA	PPPA_ECOLI	bifunctional prepilin leader peptidase/ methylase	Complete proteome; Inner membrane; Membrane; Transmembrane.	predicted
AAC76134.1	yqjE	YQJE_ECOLI	Inner membrane protein yqjE	Complete proteome; Inner membrane; Membrane; Transmembrane.	predicted
AAC76167.2	agaV	PTPB2_ECOLI	N-acetylgalactosamine-specific phosphotransferase enzyme IIB component 2 (EC 2.7.1.69)	Complete proteome; Cytoplasm; Phosphorylation; Phosphotransferase system; Sugar transport; Transferase; Transport.	predicted

AAC76273.1	aaeA	AAEA_ECOLI	p-hydroxybenzoic acid efflux pump subunit aaeA	Complete proteome; Inner membrane; Membrane; Transmembrane; Transport.	predicted
AAC76353.1	gspG	GSPG_ECOLI	Putative general secretion pathway protein G precursor	Complete proteome; Methylation; Transport.	predicted
AAC76610.1	yiaV	YIAV_ECOLI	Inner membrane protein yiaV precursor	Complete proteome; Inner membrane; Membrane; Signal; Transmembrane.	predicted
AAC76645.1	rfaC	RFAC_ECOLI	Lipopolysaccharide heptosyltransferase 1	Complete proteome; Direct protein sequencing; Glycosyltransferase; Lipopolysaccharide biosynthesis; Transferase.	predicted
AAC76700.1	yidI	YIDI_ECOLI	Inner membrane protein yidI	Complete proteome; Inner membrane; Membrane; Transmembrane.	predicted
AAT48222.1	rhtC	RHTC_ECOLI	Threonine efflux protein	Complete proteome; Inner membrane; Membrane; Transmembrane; Transport.	predicted
AAT48224.1	pldB	PLDB_ECOLI	Lysophospholipase L2 (EC 3.1.1.5)	Complete proteome; Hydrolase; Inner membrane; Lipid synthesis; Membrane.	predicted
AAC76860.1	yihG	YIHG_ECOLI	Probable acyltransferase yihG (EC 2.3.-.-)	Acyltransferase; Complete proteome; Inner membrane; Membrane; Transferase; Transmembrane.	predicted
AAC77004.1	malE	MALE_ECOLI	Maltose-binding periplasmic protein precursor	3D-structure; Complete proteome; Direct protein sequencing; Periplasm; Signal; Sugar transport; Transport.	predicted
AAC77023.1	alr	ALR1_ECOLI	Alanine racemase, biosynthetic (EC 5.1.1.1)	Cell shape; Cell wall biogenesis/degradation; Complete proteome; Isomerase; Peptidoglycan synthesis; Pyridoxal phosphate.	predicted
AAC77043.1	nrfD	NRFD_ECOLI	Protein nrfD	Complete proteome; Inner membrane; Membrane; Transmembrane.	predicted
AAC77089.1	yjdK	YJDK_ECOLI	Uncharacterized protein yjdK	Complete proteome.	predicted
AAC77118.1	yjeO	YJEO_ECOLI	Inner membrane protein yjeO	Complete proteome; Inner membrane; Membrane; Transmembrane.	predicted
AAC77182.1	chpB	CHPB_ECOLI	PemK-like protein 2	Complete proteome; DNA-binding.	predicted
AAC77288.1	yjiJ	YJIJ_ECOLI	predicted inner membrane protein	Complete proteome; Membrane; Transmembrane.	predicted

**S2 Scanning results of *S. aureus* Mu50 genome (\*indicates functional classification by SVMProt followed by probability of correct characterization P-value, while # indicates the data are not included in our model data set)**

Access number in NCBI	Gene Name	Entry Name in Swiss-Prot	Protein Name	Protein Function	Status
NP_370529.1	gyrB	GYRB_STAAM	DNA gyrase subunit B (EC 5.99.1.3)	ATP-binding; Complete proteome; Isomerase; Nucleotide-binding; Topoisomerase.	#known
NP_370530.1	gyrA	GYRA_STAAM	DNA gyrase subunit A (EC 5.99.1.3)	Antibiotic resistance; Complete proteome; DNA-binding; Isomerase; Topoisomerase.	known
NP_370558.1	bleO	BLE_STAAM	Bleomycin resistance protein (Bleomycin-binding protein) (BRP)	Antibiotic resistance; Complete proteome.	#known
NP_370565.1	mecA	Q54113_STAAM	penicillin binding protein 2 prime	Complete proteome.	#known
NP_370566.1	mecR1	MECR_STAAM	Methicillin resistance mecR1 protein	Antibiotic resistance; Complete proteome.	known
NP_370567.1	mecI	MECI_STAAM	Methicillin resistance regulatory protein mecI	Antibiotic resistance; Complete proteome; DNA-binding; Repressor; Transcription; Transcription regulation.	known
NP_371066.1	rpoB	RPOB_STAAM	DNA-directed RNA polymerase beta chain (EC 2.7.7.6) (RNAP betasubunit) (Transcriptase beta chain) (RNA polymerase subunit beta)	Complete proteome; DNA-directed RNA polymerase; Nucleotidyltransferase; Transcription; Transferase.	#known
NP_371067.1	rpoC	RPOC_STAAM	DNA-directed RNA polymerase beta' chain (EC 2.7.7.6) (RNAP beta'subunit) (Transcriptase beta' chain) (RNA polymerase beta' subunit)	Complete proteome; DNA-directed RNA polymerase; Nucleotidyltransferase; Transcription; Transferase.	#known
NP_371898.1	femA	FEMA_STAAM	Aminoacyltransferase femA (EC 2.3.2.-) (Factor essential forexpression of methicillin resistance A)	Acytransferase; Antibiotic resistance; Cell shape; Cell wall biogenesis/degradation; Complete proteome; Cytoplasm; Peptidoglycan synthesis; Transferase.	known
NP_371899.1	femB	FEMB_STAAM	Aminoacyltransferase femB (EC 2.3.2.-) (Factor essential forexpression of methicillin resistance B)	Acytransferase; Antibiotic resistance; Cell shape; Cell wall biogenesis/degradation; Complete proteome; Cytoplasm; Peptidoglycan synthesis; Transferase.	known
NP_372179.1	ermA	ERMA_STAAM	rRNA adenine N-6-methyltransferase (EC 2.1.1.48) (Macrolide-lincosamide-streptogramin B resistance protein) (Erythromycinresistance protein)	Antibiotic resistance; Complete proteome; Methyltransferase; RNA-binding; S-adenosyl-L-methionine; Transferase.	known

NP_372180.1	ant(9)	S3AD_STAAM	Streptomycin 3"-adenylyltransferase (EC 2.7.7.47) (AAD(9))	Antibiotic resistance; Complete proteome; Nucleotidyltransferase; Transferase; Transposable element.	known
NP_372408.1	vraR	VRAR_STAAM	Response regulator protein vraR	Antibiotic resistance; Activator; Complete proteome; Cytoplasm; DNA-binding; Phosphorylation; Transcription; Transcription regulation; Two-component regulatory system.	known
NP_372409.1	vraS	VRAS_STAAM	Sensor protein vraS (EC 2.7.13.3)	Complete proteome; Kinase; Membrane; Phosphorylation; Transferase; Transmembrane; Two-component regulatory system.	#known
NP_372786.1	fmhB	FEMX_STAAM	Aminoacyltransferase femX (EC 2.3.2.-) (Factor essential forexpression of methicillin resistance X)	Acytransferase; Antibiotic resistance; Cell shape; Cell wall biogenesis/degradation; Complete proteome; Cytoplasm; Peptidoglycan synthesis; Transferase.	known
NP_370858.1	mepA	MEPA_STAAM	Multidrug export protein mepA	Antibiotic resistance; Complete proteome; Membrane; Transmembrane; Transport.	known
NP_370922.1	tetM	TETM_STAAM	Tetracycline resistance protein tetM (TetA(M))	Antibiotic resistance; Tetracycline resistance protein tetM (TetA(M)).	known
NP_371017.2	ksgA	KSGA_STAAM	Dimethyladenosine transferase (EC 2.1.1.-) (S-adenosylmethionine-6-N',N'-adenosyl(rRNA) dimethyltransferase) (16S rRNA dimethylase) (Highlevel kasugamycin resistance protein ksgA) (Kasugamycindimethyltransferase)	Antibiotic resistance; Complete proteome; Methyltransferase; RNA-binding; rRNA processing; S-adenosyl-L-methionine; Transferase.	known
NP_371038.1	folP	DHPS_STAAM	Dihydropteroate synthase (EC 2.5.1.15) (Dihydropteroatepyrophosphorylase) (DHPS)	Antibiotic resistance; Complete proteome; Folate biosynthesis; Transferase.	known
NP_371069.1	rpsL	RS12_STAAM	30S ribosomal protein S12	Complete proteome; Ribonucleoprotein; Ribosomal protein; RNA-binding; rRNA-binding; tRNA-binding.	#known
NP_371207.1	uppP	UPPP_STAAM	Undecaprenyl-diphosphatase (EC 3.6.1.27) (Undecaprenyl pyrophosphatephosphatase) (Bacitracin resistance protein)	Antibiotic resistance; Cell shape; Cell wall biogenesis/degradation; Complete proteome; Hydrolase; Membrane; Peptidoglycan synthesis; Transmembrane.	known
NP_371581.1	fmt	FMTA_STAAM	Protein fmtA precursor	Antibiotic resistance; Cell wall biogenesis/degradation; Complete proteome; Membrane; Secreted; Signal.	known

NP_371705.1	pbpA	Q99UT1_STAAM	Penicillin-binding protein 1	Complete proteome.	#known
NP_371756.1	hmrB	ACP_STAAM	Acyl carrier protein (ACP)	Antibiotic resistance; Complete proteome; Cytoplasm; Fatty acid biosynthesis; Lipid synthesis; Phosphopantetheine.	known
NP_371788.1	polC	DPO3_STAAM	DNA polymerase III polC-type (EC 2.7.7.7) (PolIII)	Complete proteome; Cytoplasm; DNA replication; DNA-directed DNA polymerase; Exonuclease; Hydrolase; Nuclease; Nucleotidyltransferase; Transferase.	#known
NP_371879.1	parC	PARC_STAAM	DNA topoisomerase 4 subunit A (EC 5.99.1.-) (Topoisomerase IV subunitA)	Complete proteome; Lipoprotein; Membrane; Palmitate; Signal; Transmembrane.	#known
NP_371884.1	fmtC	MPRF_STAAM	Probable lysylphosphatidylglycerol synthetase (LPG synthetase)(Multiple peptide resistance factor)	Antibiotic resistance; Complete proteome; Lipid metabolism; Membrane; Transmembrane; Virulence.	known
NP_371950.1	dfrA	DYR_STAAM	Dihydrofolate reductase (EC 1.5.1.3) (DHFR)	Complete proteome; NADP; One-carbon metabolism; Oxidoreductase.	#known
NP_372243.1	rpsD	RS4_STAAM	30S ribosomal protein S4	Complete proteome; Ribonucleoprotein; Ribosomal protein; RNA-binding; rRNA-binding.	#known
NP_372757.1	rpsE	RS5_STAAM	30S ribosomal protein S5	Complete proteome; Ribonucleoprotein; Ribosomal protein; RNA-binding; rRNA-binding.	#known
NP_372765.1	rpsQ	RS17_STAAM	30S ribosomal protein S17	Complete proteome; Ribonucleoprotein; Ribosomal protein; RNA-binding; rRNA-binding.	#known
NP_372773.1	rplD	RL4_STAAM	50S ribosomal protein L4	Complete proteome; Ribonucleoprotein; Ribosomal protein; RNA-binding; rRNA-binding.	#known
NP_372857.1	fosB	FOSB_STAAM	Metallothiol transferase fosB (EC 2.5.1.-) (Fosfomycin resistanceprotein).	Antibiotic resistance; Complete proteome; Cytoplasm; Magnesium; Metal-binding; Transferase.	known
NP_373212.1	drp35	DRP35_STAAM	Lactonase drp35	3D-structure; Calcium; Complete proteome; Cytoplasm; Hydrolase; Metal-binding.	#known
NP_370559.1	aadD	O87369_STAAM	Kanamycin nucleotidyltransferase.	Complete proteome; Transferase.	#known
NP_372784.1	-		hypothetical protein	*Transmembrane (65.4%); EC 1.9:	predicted

				Oxidoreductases - Acting on a heme group of donors (62.2%); Iron-binding (58.6%); Calcium-binding (58.6%); EC 3.6: Hydrolases - Acting on Acid Anhydrides (58.6%); Virulence (58.6%); Copper-binding (58.6%); Magnesium-binding (58.6%);	
NP_372785.1	-	-	similar to acriflavin resistance protein	*Transmembrane (99.2%); EC 3.6: Hydrolases - Acting on Acid Anhydrides (92.9%); All lipid-binding proteins (92.1%); Metal-binding (85.4%); TC 3.A.3 P-type ATPase (P-ATPase) family (58.6%); Copper-binding (58.6%); Calcium-binding (58.6%); TC 1.C. Channels/Pores - Pore-forming toxins (proteins and peptides) (58.6%);	predicted
NP_373226.1	vraE	Q7A2K2_STAAM	VraE protein.	Complete proteome; Membrane; Transmembrane.	predicted
NP_373227.1	-	-	hypothetical protein	*TC 1.E. Channels/Pores - Holins (58.6%); Zinc-binding (58.6%); Metal-binding (58.6%); Transmembrane (58.6%); Photoreceptor (58.6%); TC 3.A.5 Type II (general) secretory pathway (IISP) family (58.6%); Magnesium-binding (58.6%);	predicted
NP_370597.1	kdpB(SCCmec)	ATKB1_STAAM	Potassium-transporting ATPase B chain 1 (EC 3.6.3.12) (Potassium-translocating ATPase B chain 1) (ATP phosphohydrolase [potassium-transporting] B chain 1) (Potassium-binding and translocating subunitB 1).	ATP-binding; Complete proteome; Hydrolase; Ion transport; Magnesium; Membrane; Metal-binding; Nucleotide-binding; Phosphorylation; Potassium; Potassium transport; Transmembrane; Transport.	predicted
NP_370797.1	-	-	hypothetical protein	*EC 2.7: Transferases - Transferring Phosphorus-Containing Groups (78.4%); All lipid-binding proteins (71.3%); Nickel-binding (58.6%); DNA repair (58.6%); Calcium-binding (58.6%); Magnesium-binding (58.6%);	predicted
NP_370820.1	-	-	hypothetical protein	*Transmembrane (93.6%); Virulence	predicted



				(80.4%); All lipid-binding proteins (62.2%);	
NP_370855.1	-	-	hypothetical protein	*EC 2.5: Transferases - Transferring Alkyl or Aryl Groups, Other than Methyl Groups (80.4%); All lipid-binding proteins (68.5%); EC 3.1: Hydrolases - Acting on Ester Bonds (62.2%); Chlorophyll biosynthesis (58.6%); EC 3.4: Hydrolases - Acting on peptide bonds (Peptidases) (58.6%); TC 3.A.1 ATP-binding cassette (ABC) family (58.6%); Outer membrane (58.6%); Magnesium-binding (58.6%);	predicted
NP_370876.1	-	-	hypothetical protein	*Transmembrane (78.4%); EC 1.9: Oxidoreductases - Acting on a heme group of donors (76.2%); Iron-binding (58.6%); EC 3.6: Hydrolases - Acting on Acid Anhydrides (58.6%); TC 3.A.5 Type II (general) secretory pathway (IISP) family (58.6%); TC 3.A.1 ATP-binding cassette (ABC) family (58.6%); Magnesium-binding (58.6%);	predicted
NP_370942.1	-	-	probable transposase	*All DNA-binding (58.6%) Zinc-binding (58.6%) DNA recombination (58.6%) Repressor (58.6%) DNA repair (58.6%) DNA-directed RNA polymerase (58.6%)	predicted
NP_370973.1	-	-	hypothetical protein	*Photosynthesis (58.6%); All lipid-binding proteins (58.6%); Transmembrane (58.6%); Photosystem I (58.6%); Photosystem II (58.6%); TC 1.C. Channels/Pores - Pore-forming toxins (proteins and peptides) (58.6%);	predicted
NP_370980.1	-	-	hypothetical protein	*Transmembrane (98.7%); Calcium-binding (58.6%); TC 3.A.1 ATP-binding cassette (ABC) family (58.6%);	predicted
NP_370987.1	-	-	ABC transporter permease protein	-	predicted

NP_371057.1	-	-	hypothetical protein	*Zinc-binding (73.8%); EC 3.6: Hydrolases - Acting on Acid Anhydrides (65.4%); Nickel-binding (62.2%); All DNA-binding (62.2%); EC 2.3: Transferases - Acyltransferases (58.6%); EC 4.2: Lyases - Carbon-Oxygen Lyases (58.6%); RNA-binding Proteins (58.6%); DNA repair (58.6%); Magnesium-binding (58.6%);	predicted
NP_371172.1	fhuB	Q99VX2_STAAM	Ferrichrome transport permease	Complete proteome.	predicted
NP_371185.1	vraF	Q99VW0_STAAM	ABC transporter ATP-binding protein	ATP-binding; Complete proteome; Nucleotide-binding.	predicted
NP_371188.1	-	-	low-affinity inorganic phosphate transporter	-	predicted
NP_371233.1	-	-	hypothetical protein	*Transmembrane (88.1%); EC 2.7: Transferases - Transferring Phosphorus-Containing Groups (65.4%); Lipoprotein (65.4%); Virulence (62.2%); Iron-binding (58.6%); TC 3.D. Primary Active Transporters - Oxidoreduction-driven transporters (58.6%); Lipopolysaccharide biosynthesis (58.6%); TC 2.A.3 Amino acid-polyamine-organocation (APC) family (58.6%);	predicted
NP_371251.1	-	-	di-tripeptide ABC transporter	-	predicted
NP_371279.1	-	-	hypothetical protein	*EC 3.5: Hydrolases - Acting on Carbon-Nitrogen Bonds, other than Peptide Bonds (78.4%); All lipid-binding proteins (73.8%); EC 3.4: Hydrolases - Acting on peptide bonds (Peptidases) (62.2%); Plant defense (58.6%); Metal-binding (58.6%); Lipopolysaccharide biosynthesis (58.6%); Photosystem I (58.6%);	predicted
NP_371302.1	secG	SECG_STAAM	Probable protein-export membrane protein secG	Complete proteome; Membrane; Protein transport; Translocation; Transmembrane; Transport.	predicted

NP_371328.1	-	-	hypothetical protein	*EC 3.1: Hydrolases - Acting on Ester Bonds (71.3%); Lipoprotein (68.5%); TC 1.C. Channels/Pores - Pore-forming toxins (proteins and peptides) (58.6%);	predicted
NP_371335.1	fmb	CLFA_STAAM	Clumping factor A precursor (Fibrinogen-binding protein A) (Fibrinogenreceptor A)	Cell wall; Complete proteome; Peptidoglycan-anchor; Secreted; Signal; Virulence.	predicted
NP_371446.1	-	-	hypothetical protein	*Transmembrane (95.7%); All lipid-binding proteins (71.3%); Metal-binding (58.6%); Calcium-binding (58.6%);	predicted
NP_371534.1	-	-	Na <sup>+</sup> /H <sup>+</sup> antiporter homolog		predicted
NP_371556.1	-	-	hypothetical protein	*Antimicrobial (58.6%); Photosynthesis (58.6%); Fungicide (58.6%); Innate immunity (58.6%); Growth factor (58.6%); TC 1.C. Channels/Pores - Pore-forming toxins (proteins and peptides) (58.6%); Immune response (58.6%); All lipid-binding proteins (58.6%); Hormone (58.6%); Inflammatory response (58.6%); Lipid-binding (58.6%); Antibiotic (58.6%); rRNA-binding Proteins (58.6%);	predicted
NP_371641.1	-	-	hypothetical protein	*Transmembrane (95.7%); Iron-binding (62.2%); TC 3.A.3 P-type ATPase (P-ATPase) family (58.6%); TC 3.A.1 ATP-binding cassette (ABC) family (58.6%);	predicted
NP_371845.1	-	-	similar to two-component sensor histidine kinase	*EC 2.7: Transferases - Transferring Phosphorus-Containing Groups (94.7%); Transmembrane (92.1%); All lipid-binding proteins (76.2%); Lipid transport (62.2%); Metal-binding (58.6%); TC 3.A.1 ATP-binding cassette (ABC) family (58.6%); EC 1.16: Oxidoreductases - Oxidising metal ions (58.6%); Copper-binding (58.6%);	predicted

				Magnesium-binding (58.6%);	
NP_371924.1	lysA	Q7A2R9_STAAM	Diaminopimelate decarboxylase (EC 4.1.1.20)	Amino-acid biosynthesis; Complete proteome; Decarboxylase; Lyase; Lysine biosynthesis.	predicted
NP_371975.1	-	-	hypothetical protein	*EC 2.4: Transferases - Glycosyltransferases (73.8%); Zinc-binding (62.2%); EC 3.6: Hydrolases - Acting on Acid Anhydrides (58.6%); Manganese-binding (58.6%); RNA-binding Proteins (58.6%); EC 4.2: Lyases - Carbon-Oxygen Lyases (58.6%);	predicted
NP_372076.1	pbp3	Q99TU2_STAAM	Penicillin-binding protein 3	Complete proteome.	predicted
NP_372140.1	-	-	putative Holliday junction resolvase	*Cobalt-binding(62.2%) All DNA-binding (62.2%) EC 3.6.-.-: Hydrolases - Acting on Acid Anhydrides (62.2%) Iron-binding (58.6%) Metal-binding (58.6%) DNA condensation (58.6%) Outer membrane (58.6%) DNA repair (58.6%)	predicted
NP_372144.1	-	-	hypothetical protein	*EC 3.1: Hydrolases - Acting on Ester Bonds (76.2%); Zinc-binding (71.3%); EC 2.4: Transferases - Glycosyltransferases (62.2%); mRNA-binding Proteins (58.6%); Proto oncogene (58.6%);	predicted
NP_372203.1	rpmI	RL35_STAAM	50S ribosomal protein L35.	Complete proteome; Ribonucleoprotein; Ribosomal protein.	predicted
NP_372214.1	polA	Q99TH2_STAAM	DNA polymerase I.	Complete proteome; Hydrolase; Nuclease.	predicted
NP_372238.1	-	-	hypothetical protein	*Transmembrane (98.0%); TC 2.A.1 Major facilitator family (MFS) (58.6%); TC 3.A.15 The Outer Membrane Protein Secreting Main Terminal Branch (MTB) family (58.6%); TC 3.A.1 ATP-binding cassette (ABC) family (58.6%);	predicted
NP_372299.1	-	-	similar to glucosaminidase	*All lipid-binding proteins (97.3%); Zinc-binding (73.8%); Metal-binding (58.6%); Copper-binding (58.6%);	predicted

				Calcium-binding (58.6%); TC 1.C. Channels/Pores - Pore-forming toxins (proteins and peptides) (58.6%);	
NP_372324.1	-	-	hypothetical protein	*Zinc-binding (88.1%); Immune response (62.2%); EC 3.1: Hydrolases - Acting on Ester Bonds (62.2%); Antigen (62.2%); All DNA-binding (58.6%); Metal-binding (58.6%); DNA repair (58.6%); Calcium-binding (58.6%); Magnesium-binding (58.6%);	predicted
NP_372450.1	-	-	hypothetical protein	*All lipid-binding proteins (76.2%); Virulence (62.2%); Outer membrane (58.6%);	predicted
NP_372549.2	-	-	hypothetical protein	*All lipid-binding proteins (65.4%); DNA repair (58.6%); Magnesium-binding (58.6%);	predicted
NP_372573.1	-	-	similar to O-sialoglycoprotein endopeptidase	*Zinc-binding (98.8%); EC 2.3: Transferases - Acyltransferases (96.4%); EC 3.4: Hydrolases - Acting on peptide bonds (Peptidases) (96.1%); EC 2.7: Transferases - Transferring Phosphorus-Containing Groups (94.7%); EC 4.2: Lyases - Carbon-Oxygen Lyases (94.2%); All lipid-binding proteins (92.9%); Manganese-binding (92.9%); Lipid synthesis (85.4%); Metal-binding (76.2%); EC 4.1: Lyases - Carbon-Carbon Lyases (73.8%); Lipid-binding (71.3%); EC 2.5: Transferases - Transferring Alkyl or Aryl Groups, Other than Methyl Groups (62.2%); Chlorophyll biosynthesis (58.6%); Photosystem I (58.6%);	predicted
NP_372600.1	kdpB	ATKB2_STAAM	Potassium-transporting ATPase B chain 2 (EC 3.6.3.12) (Potassium-translocating ATPase B chain 2) (ATP phosphohydrolase [potassium-transporting] B chain 2) (Potassium-binding and translocating subunitB	ATP-binding; Complete proteome; Hydrolase; Ion transport; Magnesium; Membrane; Metal-binding; Nucleotide-binding; Phosphorylation; Potassium; Potassium transport;	predicted

			2)	Transmembrane; Transport.	
NP_372614.1	-	OXAA_STAAM	Membrane protein oxaA precursor	Complete proteome; Lipoprotein; Membrane; Palmitate; Signal; Transmembrane.	predicted
NP_372632.1	atpE	Q99SF0_STAAM	ATP synthase C chain (EC 3.6.3.-)	CF(0); Complete proteome; Hydrogen ion transport; Ion transport; Lipid-binding; Membrane; Transmembrane; Transport.	predicted
NP_372744.1	-	-	similar to cobalt transport protein	*Transmembrane (98.5%); Zinc-binding (71.3%); Copper-binding (58.6%); Lipid transport (58.6%); TC 3.A.1 ATP-binding cassette (ABC) family (58.6%)	predicted
NP_372907.1	-	-	hypothetical protein	*EC 3.4: Hydrolases - Acting on peptide bonds (Peptidases) (58.6%); TC 1.C. Channels/Pores - Pore-forming toxins (proteins and peptides) (58.6%);	predicted
NP_372934.1	-	-	hypothetical protein	*Hormone (58.6%); Lipid-binding (58.6%); Lipid synthesis (58.6%); Magnesium-binding (58.6%);	predicted
NP_373037.1	-	-	alkaline phosphatase	-	predicted
NP_373154.1	clfB	CLFB_STAAM	Clumping factor B precursor (Fibrinogen-binding protein B) (Fibrinogenreceptor B)	Cell wall; Complete proteome; Peptidoglycan-anchor; Secreted; Signal; Virulence.	predicted
NP_373199.1	hisH	HIS5_STAAM	Imidazole glycerol phosphate synthase subunit hisH (EC 2.4.2.-) (IGP synthase glutamine amidotransferase subunit) (IGP synthase subunit hisH) (ImGP synthase subunit hisH) (IGPS subunit hisH)	Amino-acid biosynthesis; Complete proteome; Cytoplasm; Glutamine amidotransferase; Histidine biosynthesis; Transferase.	predicted
NP_373219.1	-	-	2-oxoglutarate/malate translocator	-	predicted

**S3 Prediction result of potential precursor miRNAs (“+” and “-” indicates that the RNA is predicted as precursor miRNA and non-precursor miRNA, respectively)**

miRNA Id	Sequence information	SVM prediction status
mml-miR-10b	CCAGAGGTTGTAACGTTGTCTATATATACCCTGTAGAAC CGAATTTGTGTGGTATCCATATAGTCACAGATTCGATTC TAGGGGAATATATGGTCGATGCAAAAACTTCA	+
mml-miR-122a	CCTTAGCAGAGCTGTGGAGTGTGACAATGGTGTGTTGTG TCTAAACTATCAAACGCCATTATCACACTAAATAGCTA CTACTAGGC	+
mml-miR-125a	TGCCAGTCTCTGGGTCCCTGAGACCCTTTAACCTGTGAG GACATCCAGGGTCACAGGTGAGGTTCTTGGGAGCCTGG CGTCTGGCC	+
mml-miR-126	CGCTGGTGATGGGACATTATTACTTTTGGTACGCGCTGT GACACTTCAAACCTCGTACCGTGAGTAATAATGCGCTGT CCACAGCA	+
mml-miR-130b	GGCCTGCCCCGACACTCTTTCCTGTTGCACTACTGTGGG CCACTGGGAAGCAGTGCAATGATGAAAGGGCATCGGTC AGGTC	+
mml-miR-134	CAGGGTGTGTGACTGGTTGACCAGAGGGGCGTGCAGTGT TGTTACCCCTGTGGGCCACCTAGTCACCAACCCTC	+
mml-miR-143	GCGCAGCGCCGTGTCTCCAGCCTGAGGTGCAGTGTGTG CATCTCTGGTCAGTTGGGAGTCTGAGATGAAGCACTGT AGCTCAGGAAGAGAGAAGTTGTTCTGCAGC	+
mml-miR-144	TGGGGCCCTGGCTGGGATATCATCATATACTGTAAGTTT GTGATGAGACACTACAGTATAGATGATGTACTAGTCCG GGCACCCCC	+
mml-miR-146a	CCTATGTGTATCCTCAGCTTTGAGAACTGAATTCCATGG GTTGTGTGTCAGTGTGACACCTGTGAAATTCAGTTCTTCAG CTGGGATATCTCTGTCGTCGT	+
mml-miR-147	AATCTAAAGAAAACATTTCTGCACACACACCAGACTAT TGAAGCCAGTGTGTGGAATGCTTCTGCTACATT	+
mml-miR-147b	TATAAATCTAGTGGAACATTTCCGCACAACTAGATT CTGGACACCAGTGTGCGGAAGTGCTTCTGCTGCATTTTT AGG	+
mml-miR-149	GCCGGCGCCCAAGCTCTGGCTCCGTGTCTTCACTCCCGT GTTTGTCCGAGGAGGAGGGAGGGACGGGGGCTGTGCT GGGGCAGCCGGA	+
mml-miR-154	GAGGTACTTGAAGATAGGTTATCCGTGTTGCCTTCGCTT TATTTGTGACGAATCATACACGGTTGACCTATTTTTCAG TACCAA	+
mml-miR-155	CTGTTAATGCTAATCGTGATAGGGGTTTTTACCTCCAAC TGAATCCTACATGTTAGCATTAACAG	+
mml-miR-16-2	GTTCCACTCTAGCAGCACGTAAATATTGGCGTAGTGAA ATATGTATTAAACACCAATATTACTGTGCTGCTTCAGTG TGAC	+
mml-miR-181b-2	CTGATGGCTGCACTCAACATTCATTGCTGTGCGGTGGGTT TGAGTCTGAATCAACTCACTGATCGATGAATGCAAACT GCGGACCAAACA	+
mml-miR-184	TCAGTCACGTCCCCTTATCACTTTTCCAGCCCAGCTTTA TGAATGTAAGTGTGGACGGAGAACTGATAAGGGTAGG TGATTGA	+
mml-miR-185	AGGGCGCGAGGGATTGGAGAGAAAGGCAGTTCCTGAT GGTCCCCCTCCTCAGGGGCTGGCTTTTCTCTGGTCCTTCC	+

	CTCCCA	
mml-miR-186	TGCTTGTAACTTTCCAAAGAATTCTCCTTTTGGGCTTTCT GGTTTTATTTTAAGCCCAAAGGTGAATTTCTTGGGAAGT TTGAGCT	+
mml-miR-187	GGTCAGGCTCACTATGACACAGTGTGAGACCTCGGGCT ACAACACAGGACCCGGGTGCTGCTCTGACCCCTCGTGT CTTGTGTTGCAGCCGGAGGGACGCAGGTCCGCA	+
mml-miR-190b	TGCTTCTGTGTGATATGTTTGATATTGGGTGTTTAATT AGGAACCAACTAAATGTCAAACATATTCTTACAGCAGC TG	+
mml-miR-192	GCTGAGACCGAGTGCACAGGGCTCTGACCTATGAATTG ACAGCCAGTGCTCTCGTCTCCCCTCTGGCTGCCAATTCC ATAGGTCACAGGTATGTTTCGCCTCAATGCCAGC	+
mml-miR-193a	GGATGGGAGCTGAGGGCTGGGTCTTTGCGGGCGAGATG AGGGTGTGCGATCAACTGGCCTACAAAGTCCCAGTCCT CGGCCCCCG	+
mml-miR-194-2	TGGCTCCCGCCCCCTGTAACAGCAACTCCATGTGGAAG TGTCCTACTGATTCCAGTGGGGCTGCTGTTATCTGGGGCG AGGGCCCG	+
mml-miR-195	AGCTTCCCTGGCTCTAGCAGCACAGAAATATTGGCACA GGGAAGCAAGTCTGCCAATATTGGCTGTGCTGCTCCAG GCAGGGTGGTG	+
mml-miR-199a-2	GCCAACCCAGTGTTTCAGACTACCTGTTTCAGGAGGCTCT CAACGTGTACAGTAGTCTGCACATTGGTTAGGC	+
mml-miR-203	GTGCTGGGGACTCGCGCGCTGGGTCCAGTGGTTCTTAA CAGTTCAACAGTTCTGTAGCGCAATTGTGAAATGTTTAG GACCACTAGACCCGCGGGGCACGGCGACAGCGA	+
mml-miR-208	TGACAGGCGAGCTTTTGGCCCGGGTTATACCTGATGCTC ACGTATAAGACGAGCAAAAAGCTTGTTGGTCA	+
mml-miR-210	ACCCGGCAGTCCCTCCAGGCGCAGGGCAGCCCCCTGCCC ACCGCACACTGCGCTGCCCCAGACCCACTGTGCGTGTG ACAGCGGCTGATCTGTGCCTGGGCAGCGCGACCC	+
mml-miR-212	CGGGGCACCCCGCCCGGACAGCGCGCCGGCACCTTGGC TCTAGACTGCTTACTGCCCCGGGCCGCCCTCAGTAACAGT CTCCAGTCAGGGCCACCGACGCCTGGCCCCGCC	+
mml-miR-216a	GATGGCTGTGAGTTGGCTTAATCTCAGCTGGCAACTGT GAGATGTTTCATACAATCCCTCACAGTGGTCTCTGGGATT ACGCTAAACAGAGCAATTTCTTGGCCCTCGCGA	+
mml-miR-216b	GCAGACTGGAAAATCTCTGCAGGCAAATGTGATGTCAC TGAAGAAATCACACACTTACCCGTAGAGATTCTACAGT CTGACA	+
mml-miR-220c	GACAGCGTGGCATTGTAGGGCTCCACCACTGTGTCTGA CACCTTGGGCGAGGGCAGCAGCTGAAGGTGTTTCATGA TGCGGTCCGGATACTCCTCAG	+
mml-miR-220d	GTGGCGTTGTAGGGCTCCACCACCGTGTCTGACACCTTG GGTGAGGGCATGACGCTGAAGGTGTTTCATGATGCGGTC TGGGTACTCTTCCCGGATCTTGCTGATG	+
mml-miR-222	GCTGCTGGAAGGTATAGGTACCCTCAATGGCTCAGTAG CCAGTGTAGATCCTGTCTTTCGTAATCAGCAGCTACATC TGGCTACTGGGTCTCTGATGGCATCTTCTAGCT	+
mml-miR-298	TCAGGTCTTCAGCAGAAGCCGGGTGGTTCTCCCAGTGG TTTTCTTGACTGTGAGGAAGTACCTGCTGTTTTGCTC AGGAATGAGCT	+
mml-miR-299-5p	AAGAAATGGTTTACCGTCCCACATACATTTTCAATATGT ATGTGGGACGGTAAACCGCTTCTT	+
mml-miR-302d	CCTCTACTTTAACATGGAGGCACTTGCTGTGGTATGACA	+



	AAAATAAGTGCTTCCATGTTTGAGTGTGG	
mml-miR-30c-2	AGATACTGTAAACATCCTACACTCTCAGCTGTGGAAAG TAAGAAAGCTGGGAGAAGGCTGTTTACTCTCTCT	+
mml-miR-325	ATGCAGTGCTTGGTTCTAGTAGGTGTCCAGTAAGTGTT TGTTACATAATTTGTTTATTGAGGACCTCCTATCAATCA AGCACTGTGCTAGGCTCTGG	+
mml-miR-329-1	GTGGTACCTGAAGGGAGGTTTTCTGGGTTTTCTGTTTCTT TAATGAGGATGAAACACACCTGGTTAACCTCTTTTCCA GTATCAA	+
mml-miR-329-2	GGTACCTGAAGGGAGGTTTTCTGGGTCTCTGTTTCTTTA CTGAGGATGAAACACACCTGGTTAACCTCTTTTCCAGTA TC	+
mml-miR-331	GAGTTTGTTTTGTTTGGGTTTGTCTAGGTATGGTCCC AGGGATCCCAGATCAAACAGGCCCTGGGCCTATCCT AGAACCAACCTAAACTC	+
mml-miR-338	TCTCCAACAATATCCTGGTGCTGAGTGATGACTCAGGT GACTCCAGCATCAGTGATTTTGTTGAAGA	+
mml-miR-339	CGGGGCGGCCGCTCTCCCTGTCCTCCAGGAGCTCACGT GTGCCTGCCTGTGAGCGCCTCGACGACAGAGCCGGCGC CCGCCCCAGTGTCTGCGC	+
mml-miR-33b	GCGGGCGGCCCCGCGGTGCATTGCTGTTGCATTGCACG TGTGTGAGGCGGGTGCAGTGCCTCGGCAGTGCAGCCCG GAGCCGGCCCCCTGGCACCGC	+
mml-miR-34c	AGTCTAGTTACCAGGCAGTGTAGTTAGCTGATTGCTGAT AGTACCAATCACTAACCACACGGCCAGGTAAAAAGATT	+
mml-miR-365-2	AGAGTGTTCAAGGACAGCAAGAAAAATGAGGGACTTTT AGGGGCAGCTGTGTTTTCTGACTCAGTCATAATGCCCT AAAAATCCTTATTGTTCTTGCACTGTGCATCAGG	+
mml-miR-367	CCACTACTGTTGCTAATATGCAACTCTGTTGAACACAAA TTGGAATTGCACCTTAGCAATGGTGATGG	+
mml-miR-370	AGACAGAGAAGCCAGGTCACGTCTCTGCAGTTACACAG CTCATGAGTGCCTGCTGGGGTGGAACCTGGTCTGTCT	+
mml-miR-371	GTGGCACTCAAACGTGGGGGCACCTTCTGCTCTCTGGT GAAAAAAGTGCCGCCATGTTTTGAGTGTTAC	+
mml-miR-372	GTGATCCTCAAATGTGGAGCACTATTCTGATGTCCAAGT GGAAAGTGCTGCGACATTTGAGCGTCAC	+
mml-miR-374a	TACATCGGCCATTATAATAACAACCTGATAAGTGTTACA GCACTTATCAGATTGTATTGTAATTGTCTGTGTA	+
mml-miR-380-5p	AAGATGGTTGACCATAGAACATGCGCTATCTCTGTGTC GTATGTAATATGGTCCACGTCTT	+
mml-miR-410	GGTACCTGAGGAGAGGTTGTCTGTGATGAGTTTCGCTTTT ATTAATGACGAATATAACACAGATGGCCTGTTTTCACT ACC	+
mml-miR-422a	GAGAGAAGCACTGGACTCAGGGTCAGAAGGCCTGAGT CTCCCTGCTGCAGATGGGCTGTGTGTCCCTGAGCCAAG CCTTGTCCTCCCTGG	+
mml-miR-425-5p	GAAAGCGCTTTGGAATGACACGATCACTCCCGTTGAGT GGGCCCCCGAGAAGCCATCGGGAATGTCGTGTCCGCCC AGTGCTCTTTC	+
mml-miR-429	CGCCGGCCGATGAGCGTCTTACCAGACACGGTTAGACC TGGCTCTCTGTCTAATACTGTCTGGTAAAACCGTCCATC CGCGGC	+
mml-miR-432	TGACTCCTCCATGTCTTGAGTAGGTCATTGGGTGGATC CTCTATTTCTTATGTGGGCCACTGGATGGCTCCTCCAT GTCTTGAGTAGATCA	+
mml-miR-433	CCAGGGAGAAGTACGGTGAGCCTGTCATTATTCAGAGA	+

	GGCTAGATCCTCTGTGTTGAGAAGGATCATGATGGGCT CCTCGGTGTTCTCCAGG	
mml-miR-448	GCCGGGAGGTTGAACATCCTGCATAGTGCTGCCAGGAA ATCCCTATTTCACTAAGAGGGGCTGGCTGGTTGCATA TGTAGGATGTCCCATCTCCCAGCCTACTTCGTCA	+
mml-miR-449a	CTGTGTGTGATGAGCTGGCAGTGTATTGTTAGCTGGTTG AATATGTGAATGGCATCAGCTAACATGCAACTGCTGTC TTATTGCATATACA	+
mml-miR-449b	TGACCTGAATCAGGTAGGCAGTGTATTGTTAGCTGGCT GCTTGAGTCAAGTCAGCAGCCACAACCTACCCTGCCACT TGCTTCTGGATAAAATTCTTCT	+
mml-miR-450a-1	AAATGATACTAAACTGTTTTTGCATGTGTTCTAATAT GTACTATAAATATATTGGGAACATTTTGCATGTGTAGTT TTGTATCAATATA	+
mml-miR-451	CTTGGGAATGGCAAGGAAACCGTTACCATTACTGAGTT TAGTAATGGTAAGGGTCTCTTGCTATATCCAGA	+
mml-miR-454	TCTGTTTATCACCAGATCCTAGAACCCTATCAATATTGT CTCTGCTGTGTAATAGTTCTGAGTAGTGCAATATTGCT TATAGGGTTTTGGTGTGGGAAGACAATGGGCAGG	+
mml-miR-487a	GGTACTTGGAGAGTGGTCATCCCTGCTGTGTTTCGCTTTG TTTATGACGAATCATAACAGGGACATCCAGTTTTTCAGTA TC	+
mml-miR-487b	TTGGTACTTGGAGAGTGGTTATCCCTGTCCTGTTTCGTTT TGCTCGTGTGCAATCGTACAGGGTCATCCACTTTTTTCAG TATCAA	+
mml-miR-488	GAGAATCATCTCTCCAGATAATGGCACTCTCAAACAA GTTTCCAAGTTGTTTGAAAGGCTATTTCTTGGTCAGATG ACTCTC	+
mml-miR-489	GTGGCAGCTTGGTGGTCGTATGTGTGGCGCCATTTACTT GAACCTTTAGGAGTGACATCACATATACGGCAGCTAAA CTGTTAC	+
mml-miR-494	GATACTCGAAGGAGAGGTTGTCCGTGTTGTCTTCTCTTT ATTTATGATGAAACATACACGGGAAACCTCTTCTTTAGT ATC	+
mml-miR-496	CCCGAGTCAGGTACTCGAATGGAGGTTGTCCATGGTGT GTTCATTTTATTTATGATGAGTATTACATGGCCAATCTC CTTTCGGTACTCAATTCTTCTTGGG	+
mml-miR-499-5p	GCCCTGTCCCGTGTCTTGGGCGGGCAGCTGTAAAGACT TGCAGTGATGTTTAACTCCTCTCCACGTGAACATCACAG CAAGTCTGTGCTGCTTCCCGTCCCTACGCTGCCTGGGCA GGGT	+
mml-miR-500	GCTCCCCCTCTCTAATCCTTGCTACCTGGGTGAGAGTGC TATCTGAATGCAATGCACCTGGGCAAGGATTCTGAGAG CGAGAGC	+
mml-miR-501-5p	GCTCTTCCTCTCTAATCCTTTGTCCCTGGGTGAGAGTGC TTTCTGAATGCAGTGCACCCAGGCAAGGATTCTGAGAG GGTGAGC	+
mml-miR-502-5p	CCCTCTCTAATCCTTGCTATCTGGGTGCTAGTGCTGTCT CAATGCAATGCACCTGGGCAAGGATTCTGAGAGGGGG AGCT	+
mml-miR-503	TGCCCTAGCAGCGGGAACAGTTCTGCAGTGAGTGATCA GTACTCTGGAGTATTGTTTCCGCTGCCAGGGTA	+
mml-miR-504	GCTGCTGTTGGGAGACCCTGGTCTGCACTCTATCTGTAT TCTTACTGAAGGGAGCGCAGGGCAGGGTTTCCCATACA GAGGGC	+
mml-miR-506	GCCACCACCATCAGCCATGCTATGTGTAGTGCCTTATTC	+

	AGGAAGGTGTTACTTAATATATTAATATTTGTAAGGCA CCCTTCTGAGTAGAGTAATGTGCAACATGGACATCATTT GTGGTGGC	
mml-miR-507	GTGCTGTGTGTAGTGCTTCACTTCAATAAGTGCCATTCA TGTGTCTAGAAATATGTTTTGCACCTTTTGGAGTGAAAT AATGCACAACAGGTAC	+
mml-miR-508	CCATCTTCAGCTGAGTGTCGTGCTCTACTCCAGAGGGCG TCACTCACATAAACTAAACATGATTGTCGCCTTTTGA GTAGAGTAATACACATCACGTAAGGCATATTTGGTGG	+
mml-miR-509-1	CATGCTGTGTGTGGTACCCTACTACAGGCAGTGGAAT CATGTATAGTTAAAAATGATTGGTATGTCTGTGGGTAG AGTAATGCATGACACATG	+
mml-miR-509-2	CATGTTGTGTGTGGTACCCTACTGCAGGCAGTGGAAT CATGTATAGTTAAAAATGATTGGTATGTCTGTGGGTAG AGTAATGCATGACACATG	+
mml-miR-510	GTGGTATCCTACTCCGGAGAGTGGAATCACATATAAT TAAGTGTGATTGAAACCTCTAAGAGTGGAAGTAACAC	+
mml-miR-511-1	CAATAGACACCCAcCtTGTCTTTTGTCTGTCAGTCAGTAA ATATTTTTTTGTGAATGTGTAGCAAAAGACAGAATGGgG GTCCATTG	+
mml-miR-511-2	CAATAGACACCCACCTTGTCTTTTGTCTGTCAGTCAGTA AATATTTTTTTGTGAATGTGTAGCAAAAGACAGAATGG GGGTCCATTG	+
mml-miR-513-1	GGGATGCCACATTCAGCCATTCAAGTGTACAGTGCCTTTC ACAGGGAGGTGTCATTTATGTGAACTAAATATAAATT TCACCTTTCTGAGAAGAGTAATGTACAGCATGCACTGC ATATGTGGTGTCCC	+
mml-miR-513-2	GGGATGCCGCATTCAGCCATTCAAGTGGTGTACAGTGCC TTTACAGGGAGGTGTCATTTATGTGAACTAACTATA AATGTCACCTTTCTGCGAAGGGTAATGTACATCATGCA CTGCATATGTGGTGTCCC	+
mml-miR-513-3	GGGATGCCACATTCACCCATTTACTGTACATTGCCTTTC ACAGGGAGGTGTCATTTATGTGAACTAACTATAAATG TCACTTTTCTGAGAAGAGTAATGTACAGCATGCACTGC ATATGTGGTGTCCC	+
mml-miR-513b-1	GGGATGCCACATTCAGCCATTCAAGTGTGTCAGTGCCTTTC ACAAGGAGGTGTCATTTATGTGAACTAACTATAAATG TCACCTTTTGGGAAGAGTAATGTACAACATGCACTGC ATATGTGGTGTCCCT	+
mml-miR-513b-2	GGGATGCCACATTCAGCCATTCCGTTTACAGTGCCTTTC ACAAGGAGGTGTCATTTATGTGAACTAACTATAAATG TCACCTTTTGGGAAGAGTAATGTACAACATGCACTGC AAATGTGGTGTCCC	+
mml-miR-514-1	AACATGTTGTCTGTGGTACCCTACTCTGGAGAGTGACA ATCATGTATAATTAAATTTGATTGACACTTCTGTGAGTA GAGTAATGCATGACACGTGCG	+
mml-miR-514-2	GTTGTCTGTGGTACCCTACTCTGGAGAGTGACAATCATG TATAATTAAATTTGATTGACACTTCTGTGAGTAGAGTAA TGCATGACAC	+
mml-miR-516a-1-5p	TCTCAGGCTGTGACCgTCTCGAGGAAAGAAGCACTTTCT GTTGTCTAAAGAAAAGgAAGTGtTTCCTTcCcGAGGGTTA CGGTTTGAGA	+
mml-miR-516a-2-5p	TCTCAGGCTGTGACCGTCTCGAGGAAAGAAGCACTTTC TGTTGTCTAAAGAAAAGGAAGTGTTTCCTTCCCGAGGG TTACGGTTTGAGA	+
mml-miR-517a	CTCATGCAGTGACCCTCTAGATGGAAGCACTGTCTGTG	+

	GTCTAAAAGAAAAGATCGTGCATCCTTTTAGAGTGTTA CCGTTTGAGA	
mml-miR-517b	GTGACCCTCTAGATGGAAGCACTGTCTGTGGTCTAAAA GAAAAGATCGTGCATCCTTTTAGAGTGTTAC	+
mml-miR-518a-1	TCTCA <sub>t</sub> GCTGTGAC <sub>cc</sub> TaCAAAGGGAAGCCCTTTCTGTTG TCTaAAcGAaAAGAAAGtGCTTcCTTTGCTGGgTTACGGT TTGAGA	+
mml-miR-518b	TCAGGCTGTGACCCTCCAGAGGGAAGCACTTTCTGTTGT CTGAAAGAAAAGCAAAGCGCTCCCCTTTAGAGGATTACG GTTTGA	+
mml-miR-518d	CATGCTGTGACTCTCTGGAGGGAAGCGCTTTCTGTTGTC TGAAAGAAAACAAAGCGCTTCTCTTTAGAGAGTTACGG TTTGAGA	+
mml-miR-518e	TCTCAGGCTGTGACCCTCTAGAGGGAAGCGaTTTCTGTga tCTgAAAGAAAAGAAAatGgTTCCCTTtAGAGTGTTActgTT TGAGA	+
mml-miR-519a-1	CTCAGGCTGTGACCCTCTAGAGGGAAGCGCTTTCTGTG GTCTGAAAGAAAAGAAAGTGCTTCCTTTTAGAGGGTTA CCGTTTGAG	+
mml-miR-519b	CATGCTGTGACCCTCTGGAGGGAAGCGCTTTCTGTTGTC TGAAAGAAAAGAACGTGCATCCCTTTAGAGGGTTACTC TTTG	+
mml-miR-519c	TCTCAGTCTGTGACCCTCTAGAAGGAAGCACTTTCTGTT GTTTGAAAGAAAAGAAAGTGCAATTTTAGAGGATTA CAGTTTGAGA	+
mml-miR-519d	TCCCAAGCTGTGACCCTCCAAAGGGAAGCACTTTCTGTT TGTTGTCTGAGAGAAAACAAAGTGCTTCCTTTTAGAGT GTGACCGCTTGGGA	+
mml-miR-520a	CTCAGGCTGTGACCCTCCAGAGGGAAGTATTTCTGTTG TCTGAAGGAAAAGAAAGTGCTTCCTTTGGACTGTTTC GGTTTGAG	+
mml-miR-520b	CCCTCTAGAGGGAAGCGCTTTCTGTGGTCTGAAAGAAA AGAAAGTGCTTCCTTTTAGAGGG	+
mml-miR-520c	TCTCAGGCTGTgacCCTCTAGAGGGAAGCgCTTTCTGTgG TCTGAAAGAAAAGAAAGTGCTTCCTTTTAGAGGGTTAC CGTTTGAGA	+
mml-miR-520d	TCTCATGCTGTGACCCTACAAAGGGAAGCCCTTTCTGTT GTCTAAACGAAAAGAAAGTGCTTCTCTTTGCTGGGTTA CGGTTTGAGA	+
mml-miR-520e	GCTGTGACCCTCTAGAGGGAAGCGCTTTCTGTGGTCTG AAAGAAAAGAAAGTGCTTCCTTTTAGAGGGTTACCGTT TGAGA	+
mml-miR-520f	TCTCAGGCTGTGACCCTCTAGAGGGAAGCGCTTTCTGTG GTCTGAAAGAAAAGAAAGTGCTTCCTTTTAGAGGGTTA CCGTTTGAGA	+
mml-miR-520g	TCCCATGCTGTGGCCCTCTAGAGAAAGCACTTTCTGTTT GTTGTCTGAGGAAAAACAAAGTGCTTCCTTCAGAGTG TGGCTGTTTGGGA	+
mml-miR-520h	TCCCAAGCTGTGACCCTCCAAAGGGAAGCACTTTCTGTT TGTTGTCTGAGAGAAAACAAAGTGCTTCCTTTTAGAGT GTG	+
mml-miR-521	TCTCATGCTGTGACCCTCCAAAGGGAAGTACTTTCTGTT GTCTAAAAGAAAAGAACGCACTTCCTTTGGAGTGTTA CCGTTTGAGA	+
mml-miR-522	TCTCAGGCTGTGACCCTCTAGAGGGAAGCGATTTCTGT GATCTGAAAGAAAAGAAAATGGTTCCCTTTAGAGTGTT	+

	ACTGTTTGAGA	
mml-miR-523a	TCTCAGGCTGTGACCCTCTAGAGGGAAGCACTTTCTGTT GTCTGGAAGAAAAGAATGCGCTTCCCTTTAGAGGGTTA CTCTCTGAGA	+
mml-miR-523c-1	CATGCTGTGACCCTCTGGAGGGAAGCGCTTTCTGTTGTC TGAAAGAAAAGAACGTGCATCCCTTTAGAGGGTACTC TTTGAGA	+
mml-miR-523c-2	TCCCATGCTGTGACCCTCTGGAGGGAAGCGCTTTCTGTT GTCTGAAAGAAAAGAACGTGCATCCCTTTAGAGGGTTA CTCTTTGAGAAGA	+
mml-miR-542-5p	CAGACCTCAGACATCTCGGGGATCATCATGTCACGAGA TACCACTGTGCACTTGTGACAGATTGATAACTGAAAGG TCTGGGAGCCATTCACTTCA	+
mml-miR-548a	TCCAGGGAGGTATTAAGTTGGTGCAAAAGTAATTGTGG TTTTTGGCATTAAAAAGTAATGACAATACTGGCAATTAC TTTCCCTCCAAACCTGATATT	+
mml-miR-548b	CAGGCTATGTATTTAGGTTGGTGCAAAAGTAATTGGGG CTTGGGCCCTTTATTTTCAATGGCAAAAACCTCAATTGCT TTTGTGCCAACCTAATACTT	+
mml-miR-548c	TGTGATGTATTAGGTTGATGCAAAAGTAATTGGGGTTTT TTGTCATTAAAAGTAGTGACAAAACCGGCAATTACTTC TGCACCAAACTAATATAA	+
mml-miR-548d	AAACAAGTTGTATTAGGTTGGTGCAAAAGTAATTGTGG TTCTTGCCTATAAAAGTAATGGCAAAAACCACAATTTCT TTTGCACCAAACTAATAAAG	+
mml-miR-548f	ATTTAGGTTGGTGCAAAAGTAATTGCGGATTTTGCCATT GAAAGTAATGGCCAAAACCACAGTTCCTTTTGACCAA TCTATAGA	+
mml-miR-549	AGACATGCAACTCAAGAATATATTGAGAGCTCATCCAT AGTTGTCACTGTCTCAGATCATGACAATTATGGATGAG CTCTTAATATATCCCAGGC	+
mml-miR-550-1	TGATGCTTTGCTGGCTGGTGCAAGTGCCTGAGGGAGTAA GAGCCCTGTTGTTGTAAGATAGTGTCTACTCCCTCAGG CACATCTCCAGCAAGT	+
mml-miR-551a	GGGGACTGCCGGGTGACCCTGGAAATCCAGAGTGGGTG GGGCCTGTCTGACCATTTCTAGGCGACCCACTCTTGTT TCCAGGGTTGCCCTGGAAA	+
mml-miR-552	ACCATTCAAATATACCACAGTTTGTGTTGACCATTAACCT GTTTGTGTAAGATGCCTTTCAACGGGTGACTGGTTAGAC AAACTGTGGTATATTCA	+
mml-miR-557	AGAATGGGCAAATGAATAGTAAATTTGGAGGCCTGGGG CCCTCCCTGCTGCTGGACAAGTGTCTGCATGGGTGAGC CTTATCTTTGAAAGGAGGTGGA	+
mml-miR-558	GTGTGTGTGTGTTTGTGTTTATTTGGCATAGTAGCT CTAGACTCTATTATAGTTTCTGAGCTGCTGTACCAAAA TACCACAACTGCCTG	+
mml-miR-56	GGTATTGTAGATTAATTTTGTGGGACATTAACAACAGC ATCAGCAGCAACATCAGCTTTAGTTAATGAATCCTGGA AAGTTAAGTGACTTTATTT	+
mml-miR-562	AGTGAAATTGCTGGGTCATATGGTCAGTCTACTTTTCTCAGA GTAATTGTGAAAGTATTTTCAAAGTAGCTGTACCATTT GCATTCCCTGTGGCAAT	+
mml-miR-570	TATTAGGTTGGTGCAAAACGTAATTGCAGTTTTTGGCATT ACTTTTAAAGGCAAAAGTAGCAATTACCTTTGCACCAA CCT	+
mml-miR-576	TACAATCCAGTGAGGATTCTAATTTCTCCACATCTTTGG	+

	TAATAAGTTTTGGCAAAGATGTGGAAAAATTGGAATCC TCATTGGATTGGTTATAA	
mml-miR-578	GATAAATATATAGACAAAATACAATCCTGGACTATAAG AAGCTCCTATAGCTCCTGTAGCTTCTTGTGCTCTGGGAT TGTATTTTGTATATAT	+
mml-miR-579	CATATTAGGTAAATGCAAAAGTAATCGCGGTTTGTGCC AAATGGCGATTTGAATTAATAAATTCATTTGGTACAAA CCGCGATTACTTTTGCATCAGC	+
mml-miR-580	ATAAAAATTTCCAGTTGGAACCTAATGATTCATCAGACTC AGATATTTAAGTTAACAGTATTTGAGTCTGATGAATCAT TAGGTTCCAGTCAGAAATT	+
mml-miR-581	GTTCTGTGAACGTATTCTTGTGTTCTGTAGATCAGTGCT TTTAGAAAATTTGTGTGATCTAGAGAACACAAAGAATA CCTACACAGAACCATCTGC	+
mml-miR-582	ATCTGTGCTCTTTGATTACAGTTGTTCAACCAGTTACTA ATCTACCTAATTGTAACCTGGTTGAACAACCTGAACCCAA AGGGTGCAAAGTAGAAACATT	+
mml-miR-584	TAGGGTGACCAGCCATTATGGTTTGCCTGGGACTGAGG AATTTGCTGGGATATGTCAGTTCCAGGCCAACAGGCT GGTTGGTTCCCTGAAGCAAC	+
mml-miR-586	ATGGGGTAAAACCATTATGCATATTGTATTTTAGGTCC CAATACGTGTGGACCCTAAAAATGCAATGCATAATGGT TTTATACTCTTATCTTCTTAT	+
mml-miR-593	CCCCCAGAGTGTGTCAGGCATCAGCCAGGCATCGCTCA GCCCCCTTCCCTCTGGGGGAGCAAGGAGTGGTGCTGGG TTTGTCTCTGCTGGGGTTTCTCCT	+
mml-miR-597	TACTTACTCTACATGTGTGTCAGTTGACGACCACTGTGA AGAGAGTAAAATGTACAGTGGTTCTCTTGGGGCTCAAG CGTAACGTAGAGTGCTGGTC	+
mml-miR-601	TGCATGAGTTCATCTTGGTCTAGGATTGTTGGAGGAGTC AGAAAAATTACCCAGGGATCCTGAAGTCATTGGGGTG GA	+
mml-miR-609	TGCTCTGCTTTTCTAGGGTGTTGCTCTCATCTCTGGTCT ATAATGGGGTAAATGTAGAGATGAGGGCAACAGCCTA GGAACAGCAGAGGAACC	+
mml-miR-611	AAAATGGTGAGAGGGTTAAGGGGAGTTCCCGACGGAG ATGCGAGGACCCCTCGGGGTCTGACCCACA	+
mml-miR-615	CTCGGGAGGGGCGGAAGGGGGGTCCCGGTGCTCGGAT CTCGAGGGTGCTTATTGTTCCGTCCGAGCCTGGGTCTCC CTCTTCCCCCAACCCCCC	+
mml-miR-616	TTAGGTAATTCCTCTCTCAAAACCCTCCAATGACTTCC CTGACATGACATAGGAAGTCACTGGAGAGTTTGTAGCA GAGGAATGACCTGTTTTAAAA	+
mml-miR-619	CGCCACCTCAGCCTCCCAAAATGCTGGGATTACAGGC ATGAGCCACCGCAGTCGACCATGATCTGGACATGTTTG TGCCTGGGATTGTCAGTTTGAG	+
mml-miR-625	AGGGTAGAGGTATAAGGGGGGAAAGTTCTGCAGGCCT GTAATTAGATCTCAGGACTGTAGAACTTTCTCCCTCACC TCTGCCCT	+
mml-miR-626	ACCGATATCTTTGTCTTATTTCTGAGCTGAGGGGTATT TTTATGCAGTCTAAATGATCTCAGCTGTCCGAAAATGTC TTCAAGTTTAAAGGCTT	+
mml-miR-628	ATAGCTGTTGTGTCAGTTTCTCATGCTGACATATTTACT AGAGGGTAAAATTAATAACCTTCTAGTAAGAGTGGCAG TCGAAGGGAAGGACTCAT	+
mml-miR-632	CGCCTCCTGCCGCAGTGCCTGACGGGAGGCGGAGCGGC	+

	GAACGAGGCCGTCGGCCATTTTGTGTCTGCTTCCTGTGG GACGCGGTTCGTAGCCGT	
mml-miR-636	TGGCGGCCTGGGCGGGAGCGCGCGGGCGGGGCCGGCC CCGCTGCCTGGAATTAACCCCGCTGTGCTTGCTCGTCCC GCCTGCAGCCCTAGGCGGCGTCG	+
mml-miR-638	GTAAGCGGGCGCGGCAGGGATCGCGGGCGGGCGGCGG CCTAGGGTGCAGAGGGCGGACCGGAATGGCGCTCCCT GCGCCGCCGGCGTAACCTGCGGCGCT	+
mml-miR-639	TGGCCGACGGGGCGCGCGCGGCCGGGAGGGGCGGGGC GGACGCACAGCCGCGTTTAGTCTAGCGCAGCGGTGCGG AGCGCTCTGGGTATCCTGTCCTG	+
mml-miR-640	GTGACCCTGGGCAAGTTTCCTGAAGATCAAACACATCAG ATCCCTTATCTGTAAATGGGCATGATCCAGGAACCTG CCTCTATGGTTGCCTTGGAG	+
mml-miR-650a-2	CAGTGCTGGGATCTCAGGAGGCAGCGCTCTCAGGACTT CTCCACCATGGTCTGGGCTCTGCTCCTCCTCACCCTCCT CACTCAGGGCACAGGTGA	+
mml-miR-650c	CAGTGCTGGGGTGTGAGGAGGCAGCGCTCTCAGTCTCC ACCATGGCCTGGGCTCTGCTCCTCCTCACTCTCCTCACT CATGGCACGGGTGA	+
mml-miR-650d	CAGTGCTGGGGTCTCAGGAGACAGTGCTGTGCGGGACGT CTCCACCATGGCCTGGGCTCTGCTCCTCCTCACCCTTCT CACTCAAGGCACAGG	+
mml-miR-652	ACGAATGGCTATGCACTGCACAACCCTAGGAGAGGGTG CCATTACATAGACTATAATTGAATGGCGCCACTAGGG TTGTGCAGTGCACAACCTGCAC	+
mml-miR-653	TTCATTCTTCAGTGTTGAAACAATCTCTACTGAACCAG CTTCAAACAAATTCAGTGGAGTTTGTTCATATTGCAA GAATGATAAGATGGAAGC	+
mml-miR-656	CTGAAATAGGTTGTCTGTGAGGTGTTCACTTTCTATATG ATGAATATTATACAGTCAACCTCTTTCCGATATCGAATC	+
mml-miR-657	GGAGGAGAGGGTCTGGAGAAGCGTGGACGGCTCCAG GTGGGTTCTGGCAGGTCTCACCCTCTCTAGGCCCCATT CTC	+
mml-miR-660	CTGCTCCTTCTCCCATACCCATTGCATATCGGAGTTGTA AATTCTCAAAACACCTCCTGTGTGCATGGATTACAGGA GGGTGAGCCTTGTATCGTG	+
mml-miR-662	GCTGTTGAGGCTGTACAGCCAGGACCTGACGGTGGGGT GGCTTCGGGCCTTCTGCAGGTCTCCACGTTGTGGCCCA GCAGCGCAGTCACGTTGC	+
mml-miR-663	CCGTTGCGCGTCCCAGGCGGGGCGCTGCGGGACCGCCC TCGTGTCTGTGGCGGTGGGATCCCGTGGCCGTGTTTCC TGGTGGCCCGGCC	+
mml-miR-664	CTGGCTAGGGAAAATGATTGGATAGAAAATGTTATTCT ATTCATTTATCCCCAGCCTA	+
mml-miR-675	CCCAGGGTCTGGTGCGGAGAGGGCCACAGTGGACTTG GTGACACTGTATGCCCTCACCCTCAGCCCCTGGG	+
mml-miR-7-1	TTGGATGTTGGCCTAGTTCTGTGTGGAAGACTAGTGATT TTGTTGTTTTTAGATAACTAAATTGACAACAAATCACAG TCTGCCATATGGCACAGGCCATGCCTCTACAG	+
mml-miR-7-2	CTGGATACAGAGTGAAGTGGCTGGCCCCGTCTGGAAGA CTAGTGATTTTGTGTTGTCTTACTGCGCTCAACAACAA ATCCAGTCTGCCGAATGGTGCCAGCCATTGCA	+
mml-miR-7-3	AGATTAGAGTGGCTATGGTCTAGTGCTGTGTGGAAGAC TAGTGATTTTGTGTTCTGATGTGCTACGACAACAAATC ACAGCCGGCCTCATAGCGCAGACTCCCTTCGAC	+

mml-miR-758	GCCTGGATACGTGAGATGGTTGACCAGAGAGCACACGC TTTATATGTGCCGTTTGTGACCTGGTCCACTACCCCTCA GTATCTAATGC	+
mml-miR-767	GCTTTTATATTGTAGGTTTTTGCTCATGCACCATGGTTG TCTGAGCATGCAGCATGCTTGTCTGCTCATACCCCATGG TTTCTGAGCAGGAATCTTCATTGTCTACTGCT	+
mml-miR-768	CTGTGCTTTGTGTGTTGGAGGATGAAAGTACGGAGTGA TCCATCGGCTAAGTGTCTTATCACAATGCTGACACTCAA ACTGCTGACAGCACACGTTTTTCACAG	+
mml-miR-874	TTAGCCCTGCGGCCCCACGCACCGGGTAAGAGAGAGT CTCGCTTCCTGCCCTGGCCCGAGGGACCGACTGGCTGG GC	+
mml-miR-875-5p	TTAGTGGTACTATACCTCAGTTTTATCAGGTGTTCTTAA AATCACCTGGAAATACTGAGGTTGTGTCTCACTGAAC	+
mml-miR-877	GCTAGAGAAGGTAGAGGAGATGGCGCAGGGGACACGG GCTAAGACTCGGGGGTTCTGGGACCCTCAGACATGTG TCCTCTTCTCCCTCCTCCCAGGTGT	+
mml-miR-886-5p	CCGGGTCGGAGTTAGCTCAAGCGGTTACCTCCTCATGC CGCACTTTCTAACTGTCCATCTCTGTGCTGGGGTTCGAG ACCCGCGGGTGCTTACTGACCCTTTTATGCACTAA	+
mml-miR-888	GGCAGTGCCCTACTCAAAAAGCTGTCAGTCACTTATGTT ACATGTGACTGACACCTCTTTAGATGAAGGAAGGCTCA	+
mml-miR-892	GCAGTGCTCTACTTAGAAAGGTGCCAGTCACTTACATT ACATGTCACTGTGTCTTTCTGCGTAGAGTAAGGCTC	+
mml-miR-920	GTAGTTGTCTCTgCAGAAGACCTGGATGTGgAaGAGCTAA GACACACTCCAGGGGAGCTGTaGAAGCgGTAACACG	+
mml-miR-922	TGGCGTTCTCTCTCTCCCTGTCTGGACTGGGGTCAGAC CGTGCCCCGAGGAGAAGCAGCAGAGAATGAGACTACG TCGT	+
mml-miR-924	AATAGAGTCTTGTGTTGTCTTGCTTAAAGGCCATCCAAC CTAGAGTCTA	+
mml-miR-92b	CGGGCCCCGGGCGGGCGGGAGGGACGGGACGCGGTGC AGTGTGTTGTTCTTTCCCCCGCCAATATTGCACTCGTCCCG GCCTCCGGCCCCCCCCGGCCC	+
mml-miR-9-3	GGAGGCCCCGTTTCTCTCTTTGGTTATCTAGCTGTATGAG TGCCACAGAGCCGCTCTCAAGCTAGATAACCGAAAGTA GAAATGACTCTCA	+
mml-miR-937	AGCACTGCCCCCGGTGAGTCAGGGTGGGGCTGGCCCCC TGCTTCGCGCCCATCCGCACTCTGACTCTCCACCTGCCT GCAGGAGCT	+
mml-miR-939	TGTGGGCAGGGCCCTGGGGAGCTGAGGCTCTGGGGGTG GCCGGGGCTGACCCCTGGGCCTCTGCTCCCCAGTGTCTG ACCGTG	+
mml-miR-940	GTGGGGTGTGGGCCCGGCCCCAGGAGCGGGGCCTGGGC AGCCCCGTGTGTTGAGGAAGGAAGGCAGGGCCCCCGCT CCCCGGGCCTGACCCAC	+
mml-miR-942	ATTAAGAGAGTACCTTCTCTGTTTTGGCCATGTGTGTAC TCACAGCCCCCTCACACGTGGCCGAAACAGAGAAGGTAC TTTCCTAAT	+
mml-miR-944	GTTCCAGACACATCTCATCTGATATACAATATTTCTTA AATTGTAAAAAGAGAAATTATTGTATATCAGATGAGAT GTGTCTGGGGT	+
mml-miR-let-7a-2	AGGCTGAGGTAGTAGGTTGTATAGTTTAGAATTACATC AAGGGAGATAACTGTACAGCCTCCTAGCTTTCCT	+
mml-miR-133b	CCTCAGAAGAAAGATGCCCCCTGCTCTGGCTGGTCAAA CGGAACCAAGTCCGTCTTCCTGAGAGGTTTGGTCCCCTT	-



	CAACCAGCTACAGCAGGGCTGGCAATTCCCAGTCCTTG GAGA	
mml-miR-181d	GTCCCCTCCCCTAGGCCACAGCCAAGGTCACAATCAAC ATTCATTGTTGTCGGTGGGTTGTGAGGACCGAGGCCAG ACCCACCGGGGGATGAATGTCAGTGTGGCTGGGCCAGA CACGGCTTAAGGGGAATGGGGAC	-
mml-miR-217	AATATAATTATTACATAGTTTTTGTATGTCGCAGATTCTG CATCAGGAACTGATTGGATAAGAATCAGTCACCATCAG TTCCTAATGCATTGCCTTCAGCATCTAAACAAG	-
mml-miR-220b	GACAGCGTGGCGTTGTAGGGCTCCACCACCGTGTCCGA CACCTTGGGCGAGGGCATGACGCTGAAGGTGTTTCATGA TGCGGTCCGGGAACCTCCTCGCGGATCTTGCTGATG	-
mml-miR-297	TGTATGTATGTGTGCATGTGCATATATGTGTGTATAT ATATATATGTATTATGTACTCATATATCA	-
mml-miR-30e	GGGCAGTCTTCGCTACTGTAAACATCCTTGACTGGAAG CTGTAAGGTGTTTCAGAGGAGCTTTCAGTCGGATGTTTAC AGCGGCAGGCTGCCA	-
mml-miR-340	TTGTACCTGGTGTGATTATAAAGCAATGAGACTGATTGT CATATGTTGTTTGTGGGATCCGTCTCAGTTACTTTATAG CCATACCTGGTATCTTA	-
mml-miR-345	AAACCCTAGGTCGGCTGACTCCTAGTCAAGGGCTCGTG GTGGCTGGTGGGCCCTGAACGAGGGTTCTGGAGGCCTG GGTTTGAATATC	-
mml-miR-362	CTCGAATCCTTGGAACCTAGGTGTGAGTGCTATTTTCAGT GCAACACACCTATTCAAGGATTCAAA	-
mml-miR-378	AGGGCTCCTGACTCCAGGTCCTGTGTGTTACCTCGAAAT AGCACTGGACTTGGAGTCAGAAGGCCT	-
mml-miR-379	AGAGATGGTAGACTATGGAACGTAGGCGTTATGATTTT TGACCTATGTAACATGGTCCACTAACTCT	-
mml-miR-384	TGTTAAATTAGGAATTGTAAACAATTCCTAGGCAATAT GTATAATGTTTCATAAGACATTCTAGAAATTGTTTCATAA TGCCTGTAACA	-
mml-miR-450b-5p	GCAGAATTATTTTTGCAATATGTTCTGAATATGTAGTA TAAGCGTATTGGGATCATTTTGCATCCATAGTTTTGTAT	-
mml-miR-492	ACTACAGCCACTACTACAAGACCTTCGAGGACCTGCGG GACAAGATTCTTGGTGCCGTCAATGAGAACTCCAGGAT TGTCTTGACAGATCAACAATGCCTGTCTGGCTGCAGATG	-
mml-miR-498	AATCCTCCTTGGGAAGTGAAGCTCAGGCTGTGATTTCA AGCCAGGGGGCGTTTTTCTGTGACTGGATGAAAAGCAC CTCCGGGGCTTGAAGCTCACAGTTTGAGAGCAATCATC TAAGGAAGTT	-
mml-miR-512-1-5p	TCTCACTCTGTGGCACTCAGCCTCGGGGGCACTTTCTGG TGTCAGAATGAAAGTGCTGTCATTGCTGAGATCCAATG ACTGAGG	-
mml-miR-512-2-5p	GGTACTTCTCACTCTGTGGCACTCAGCCTCGGGGGCACT TTCTGGTGTGAGAATGAAAGTGCTGTCATTGCTGAGATC CAATGACTGAGGCGAGCACC	-
mml-miR-523b	TCTCATGATGTGACCCTCTAGAGCGAAGCGCTTTCTGTT GGCTAGAAAAGAATAGGAAGCGCTTCCCTTTAGAGTGT TACGCTTTGAGA	-
mml-miR-548e	CCTAGAATGTTACTAGGTTGGTGCAAAAGTAATTGCGA GTTTTACCATTACTTTCAATGGCAAAACCGGCAGTTACT TTTGCACCAACGTAATACTT	-
mml-miR-551b	AGATGTGCTCTCCTGGCCCATGAAATCAAGCGTGGGTG AGACCTGGTGCAGAACAGGAAGGCGACCCATACTGGT TTCAGAGGCTGCGAGAATA	-

mml-miR-553	CTTCAATTTTATTTGAAAAAGGTGAGGTTTGTGTTTGTGCTGAGAAAATCTCACTGTTTTAGACTGAGG	-
mml-miR-554	ACCTGAGTAACCTTTGCTAGTCCTGACTCAGCCAGTACTGATCTTACACTGGCAGTGGGTCAGGGTTCATATTTGGCATCTCTCTCTGGGCATCT	-
mml-miR-556	GATAGTAATGAGAAAGATGAACTCATTGTAATATGAGCTTCATTTATGCATTTTCATATTACAATTAGCTGATCTTTTTTTT	-
mml-miR-563	AGCAAAGAAGTGTGTTGCCCTCCAGGAAATGTGTGTTGCTCTGATGTAATTAGGCTGACATACATTCCCTGGTAGCCA	-
mml-miR-567	GGATTCTTACAGGACACTATGTTCTTCCAGGACAGAACATTCTTTGCTATTTTGTACTGGAAGAACATGCAAACTTTAAAAAAGTTATTGCT	-
mml-miR-572	GTCGAGGCCGTGGCCCGGAAGTGATCGGGGCCGCCGCGGACGGAAGGGCGCCTCTGCTTCGTCGCTCGGCGGTGGCCCAGCCAGGCCCGCGGGA	-
mml-miR-577	TGGGGGAATGAAGAGTAGATAAAAATATTGGTACCTGATGAGTGTGAGGCCAGGTTTCAATACTTTATCTGCTCTTCATTTTCCCATATCTACTTAC	-
mml-miR-583	AACTCGCACATTTACCAAAGAGGAAGGTCCCAGTACTGCAGGGATCTTAGCAGTACTGGGACCTACCTCTTTGGT	-
mml-miR-587	CTCCTAGGCACCCTCTTTCCACAGGTGATGAGTTACAGGGCCCAGGGAATGTGTCTGCACCTGTGACTCATCACTGGTGAAGCCCATAC	-
mml-miR-589	TCCAGCCTGTGCCCAGCAGCCCCTGAGAACCACGTCTGCTCTGAGCTGGGTACTGCCTGTTTCAGAACAGACGCTGCTTCCCAGACGCTGCCAGCTGGCC	-
mml-miR-590	TAGCCAGTCAGAAATGAGCTTATTCATAAAAGTGCAGTATGGTGGAGTCAGTCTGTAATTTTATGTATAAGCTGGTCTCTAACTGAAACGTGCAGCA	-
mml-miR-600	AAGTCACTTACTGTGTCTCCAGCTTACAGGAAGGCTCTGTGTCTGTCAGGCAGTGGAGTTACAGACAAGAGCCTTGC TCAGGCCAGCCCTGCCC	-
mml-miR-604	AGAGCATCGTGCTTGACCTTCCACGCTCCCGTGTCCACTAGCAGGCAGGTTTTCTGACACGGGCTGCGGGATTTCAGGACAGCGCATCACGGAGA	-
mml-miR-605	CCCTAGCTTGTTCTAAATCCCACGGTGCCTTCTCCTTGGGAAAAACAGAGAAGGCACTGTGGGATTTAGAACCAAGTTAGG	-
mml-miR-607	TCGCCCCAAAGTCACACAGGTTATAGATCTGGATTGGAA CCCAGGTAGCCAGACTGCCTGGGTTTGAATCCAGATCTGTAACCTGTGTGACTTTGG	-
mml-miR-612	TCTCATCTGGACCCCACTGGGGAGGGCTTCTGAGCTCCTCAGCACTGGCAGGAGGGGCTCCAGGGGCCCTCCCTCCATGGCAGCCAGGACAGGACTCTCA	-
mml-miR-618	TCTTGTTTACAACCAAACTCTACTTGTCTTCTGAGTGTGATTACGCCCATGGAGTAGCTCAGGAGGCAAACAGGGTTACCCTGTGGATAGGTCTGAAAA	-
mml-miR-624	AATGCTGTTTCAAGGTAGTACCAGTATCTTGTGTTCACTGGAACCAAGGTAAACACAAGATACTGGTATTACCTTGATAGCATTAACACCTAAGTG	-
mml-miR-633	AACCTCTCTTAGCCTCTGTTTCTTTACTGTGGTAGATAC TATTAGCCTAAAATAAGAAGGCTAATAGTATCTACCACAATAAAATTGTTGTGATGATA	-
mml-miR-643	ACCAACTGATACGCATTATCTACGTGAGCTAGAATACA	-

	AGTAGTTGGTGTCTTCAGAGACACTTGTATTCTAGCTCA GGTAGATACTGAATGGAAAA	
mml-miR-644	TTTTATTTAGTATTCTTCCATCAGTGTTTCATAAGGGATG TTGGTCTGTAGTTTTCTTATAGTGTGGCTTGCTTAGAGC AAAGGTGGTTCCCT	-
mml-miR-648	AGCACAGACGCCTCCAAGTGTGCAGGGCACTGATGGGG GCCAGGGCAGGCCAGCCAAAGTGCAGGACCTGGCACT TAGTCGGAGGTGAGGATG	-
mml-miR-649	GCCCTAGCCAAATACTGTATTTTTATCAACATTTGGTT GAAAAACATCTGTGTATTAGTAAACCTGTGTTGTTCAA GAGTCCGCTGTGCTTTGCTG	-
mml-miR-650a-1	CAGTGCTGGGATCTCAGGAGGCAGCGCTCTCAGGACGT CTCCACCATGGTCTGGGCTCTGCTCCTCCTCACCTCCT CACTCAGGGCACAGGTGA	-
mml-miR-650b	CAGTGCTGGGGTCTCAGGAGGCAGCGCTCTCGGGACAT CTCCACCATGGCCTGGGATCTGCTCCTCTTCACCCTCCT CACTCAGGGCACAGGTGA	-
mml-miR-651	AAGCTATCACTGCTTTTTAGAATAAGCTTGACTTTTGTT CAAATAAAAACGCAAAAGGAAAGTGTATCTTAAAAGG CAATGACAGTTTAATATGTTT	-
mml-miR-661	GGAGAGGCTGTGCTGTGGGGCAGGCGCTGGCCTGGGTG GCCTGAGCCCTGATTTTGGGCTGCCTGGGTATCTGGCCC GTGCGTGACCTTGGGGCGGCT	-
mml-miR-765	TTTAGGGGCTGATGAAAGTGGAGTTCAGTAGACAACCC TTTTCAAGCCCTGCAAGAACTGGGGTTTCTGGAGGAG AGGGAAGGTGCTGAAGGGGCTGCTCTCGTGAGCCTGAA	-
mml-miR-802	GTTCTGTTATTTGCAATCAGTAACAAAGATTCATCCTTG TGTCCATCATGCAGCAAGGAGAATCTTTGTCACTTAGTG TAATTAATAGCTGGAC	-
mml-miR-934	AGgAATAAGGCTTCTGTCTACTACTGGAGACACTGaTAG TgTAAAACCCAGAGTCTtCgGTAATGGACGGGAGCCTTA TTTCT	-
mml-miR-936	AGGAATAAGGCTTCTGTCTACTACTGGAGACACTGATA GTGTA AAAACCCAGAGTCTTCGGTAATGGACGGGAGCCT TATTCT	-
mml-miR-936	AGGAATAAGGCTTCTGTCTACTACTGGAGACACTGATA GTGTA AAAACCCAGAGTCTTCGGTAATGGACGGGAGCCT TATTCT	-
mml-miR-938	GAAAGTGTACCATGTGCACTTAAAGATGAAGCCGGTGC ACCTTCATGAACTGTGGTACACCTTAAGA ACTTGGT	-

## LIST OF PUBLICATIONS

- 1) Zhang HL, Lin HH, Chen Xin, Chen YZ. MiRDetector: A web server for predicting microRNAs from sequence derived physicochemical properties by support vector machine approach (manuscript in preparation)
- 2) Zhang HL, Jia J, Ma XH, Lin HH, Chen YZ. Prediction of cancer associated proteins from sequence derived physicochemical descriptors (manuscript in preparation)
- 3) Zhang HL, Han LY, Cai CZ, Lin HH, Zheng CJ, Chen YZ. Prediction of antimicrobial proteins from sequence derived physicochemical descriptors (under review)
- 4) Zhang HL, Huang WJ, Lin HH, Han LY, Cui J, and Ji ZL. *In silico* search and characterization of multifunctional enzymes (under review)
- 5) Li HL, Zhang HL, Kang L, Luo XM, Zhu WL, Chen KX, Wang XC, Jiang HL. An Effective Docking Method for Virtual Screening Developed Based on Multi-objective Optimization Algorithm (under review)
- 6) Zhang HL, Lin HH, Tao L, Ma XH, Dai JL, Jia J and Cao ZW. Prediction of Antibiotic Resistance Proteins from Sequence Derived Properties Irrespective of Sequence Similarity. **International Journal of Antimicrobial Agents**, 2008, 32(3):221-6.
- 7) Gao ZT, Li HL, Zhang HL, Liu XF, Kang L, Yang K, Luo XM, Zhu WL, Chen KX, Wang XC and Jiang HL. PDTD: a web-accessible protein database for drug target identification. **BMC Bioinformatics**, 2008, 19(9):104.
- 8) Tang ZQ, Lin HH, Zhang HL, Han LY, Chen X, Chen YZ. Prediction of Functional Class of Proteins and Peptides Irrespective of Sequence Homology by Support Vector Machines. **Bioinformatics and Biology Insights**, 2007, 1: 19-47
- 9) Cui J, Han LY, Lin HH, Zhang HL, Tang ZQ, Zheng CJ, Cao ZW, and Chen YZ. Prediction of MHC-Binding Peptides of Flexible Lengths from Sequence-Derived Structural and Physicochemical Properties. **Mol. Immunol.** 2007, 44(5): 866-877
- 10) Zheng CJ, Han LY, Xie B, Liew CY, Ong S, Cui J, Zhang HL, Z.Q.Tang, S.H. Gan, L. Jiang and Chen YZ. PharmGED: Pharmacogenetic Effect Database. **Nucleic Acids Res.** 2007, 35:D794-D799
- 11) Li HL, Gao ZT, Kang L, Zhang HL, Yang K, Luo XM, Chen KX, J. H. Shen, Wang XC and Jiang HL. TarFisDock: a web server for identifying drug targets with docking approach. **Nucl. Acids Res.** 2006, 34:W219-W224
- 12) Lin HH, Han LY, Zhang HL, Zheng CJ, Xie B, and Chen YZ. Prediction of the Functional Class of Metal-Binding Proteins from Sequence Derived Physicochemical Properties by Support Vector Machine Approach. **BMC Bioinformatics**, 2006, 7(S5), S13
- 13) Zheng CJ, Han LY, Chen X, Cao ZW, Cui J, Lin HH, Zhang HL, Li H and Chen YZ. Information of ADME-associated proteins and potential application for pharmacogenetic prediction of drug responses. **Curr. Pharmacogenomics**, 2006, 4(2): 87-103

- 14) Lin HH, Han LY, Zhang HL, Zheng CJ, Xie B, and Chen YZ. Prediction of the Functional Class of Lipid-Binding Proteins from Sequence Derived Properties Irrespective of Sequence Similarity. **J. Lipid Res.**, 2006, 47(4):824-31
- 15) Han LY, Zheng CJ, Lin HH, Cui J, Li H, Zhang HL, Tang ZQ, and Chen YZ. Prediction of Functional Class of Novel Plant Proteins by a Statistical Learning Method. **New Phytologist**, 2005, 168:109-121