

**INVESTIGATION INTO THE USE OF SUPPORT VECTOR
MACHINE FOR –OMICS APPLICATIONS**

GUO YANGFAN
(B.Sc, DUT, China)

**A THESIS SUBMITTED
FOR THE DEGREE OF MASTERS IN SCIENCE
DEPARTMENT OF PHARMACY
NATIONAL UNIVERSITY OF SINGAPORE**

2011

ACKNOWLEDGMENT

First and foremost, I would like to express my sincere and deepest gratitude to my supervisors, Assistant Professor Yap Chun Wei and Professor Chen Yu Zong. Their excellent guidance and invaluable advices and suggestions helped and enlightened me in last two years studies in National University of Singapore.

I am grateful to my labmates and friends for their insight suggestions and collaborations in my research work: Ms Liew Chin Yee, Ms He Yuye, Mr Woo Sze Kwang, Mr Bhaskaran David Prakash, and Mr Nitin Sharma from PaDEL group, Dr Zhu Feng, Dr Jia Jia, Ms Liu Xin and Mr Zhang Jingxian from BIDD group and Dr. Pasikanti Kishore Kumar from MPRG group.

Lastly, I would like to thank my parents and friends for their encouragement and understanding. It would have been impossible for me to finish this work without them.

The financial support from NUS research scholarship is gratefully acknowledged.

TABLE OF CONTENTS

ACKNOWLEDGMENT.....	II
TABLE OF CONTENTS.....	III
ABSTRACT.....	V
LIST OF TABLES.....	VI
LIST OF FIGURES.....	VII
LIST OF ABBREVIATIONS.....	VIII
1 INTRODUCTION.....	1
1.1 Applications of SVM in bioinformatics.....	1
1.1.1 Applications of SVM in genomics.....	1
1.1.2 Applications of SVM in proteomics.....	3
1.1.3 Applications of SVM in metabonomics.....	6
1.2 Underlying difficulties in using SVM.....	7
1.3 Objectives and organization of this thesis.....	9
1.3.1 Objectives of this thesis.....	9
1.3.2 Organization of this thesis.....	13
2 METHODOLOGY.....	14
2.1 Support vector machines (SVMs) method.....	14
2.1.1 Linear SVM.....	14
2.1.2 Nonlinear SVM.....	20
2.2 Performance evaluation.....	22
3 MHC BINDING PREDCITION.....	24
3.1. Data Preparation.....	24
3.2. Descriptor Generation.....	27
3.3. Overview of SVM modeling procedure.....	31
3.4. Results and Performance evaluation.....	32
3.4.1. Self consistency testing accuracy of dataset without generated non-binders..	32
3.4.2. Self consistency testing accuracy of dataset with generated non-binders.....	32
3.5. Summary and Discussion.....	36
4 METABOLITES SELECTION IN METABONOMICS.....	37

4.1. Data collection and normalization.....	37
4.2. Overview of SVM-RFE selection procedure	38
4.3. Results and Discussion.....	42
4.3.1. Comparison of prediction performance of multiple machine learning methods.....	42
4.3.2 The predictive performance of identified metabolites biomarkers.	44
4.3.3. The list of selected metabolite biomarkers	49
4.3.4. Performance evaluation with multiple classifiers	58
5. CONCLUSION AND FUTURE WORK	60
BIBLIOGRAPHY	63

ABSTRACT

Machine learning methods have frequently been used in early stage diagnosis at the proteomic level, such as the MHC binding peptides prediction and biomarkers selection for metabonomics. Although many computational methods have been designed for such studies, it is necessary to develop more stable and smart system to improve predictive performance. Support vector machine, an artificial intelligence technique, demonstrates remarkable generalization performance. Two groups of MHC binding peptides and two bladder cancer metabonomics datasets with different number of metabolites has been investigated by support vector machine and other machine learning methods. Recursive feature elimination, an effective feature selection algorithm, has also been applied to investigate the metabonomics data. The results of MHC binding peptide study showed that the prediction system can achieve satisfactory performance by constructing the model with sufficient generated non-binding peptides. The second study on metabonomics prediction suggested that metabolites biomarkers can be effectively selected from the metabonomics dataset by support vector machine-recursive feature elimination method.

LIST OF TABLES

Table 1	Division of amino acids for different physicochemical properties.	29
Table 2	Prediction performance of MHC binding peptides without generated non-binders.	33
Table 3	Datasets and the binder and non-binder prediction accuracies for HLA alleles I.	34
Table 4	Prediction performance with metabolites selection for 75 BC samples with 189 metabolites by multiple machine learning methods.....	43
Table 5	Overall prediction accuracies of 20 times SVM-RFE selection for 75 BC samples with 189 metabolites.	45
Table 6	Selected metabolites list for 75 BC samples with 189 metabolites.	46
Table 7	Overall prediction accuracies of 20 times SVM-RFE selection for 75 BC samples with 398 metabolites.	47
Table 8	Selected metabolites list for 75 BC samples with 398 metabolites.	48
Table 9	List of 31 Selected metabolites (repeated rate > 80%) for 75 BC samples with 398 metabolites.....	50
Table 10	List of structures of the 31 Selected metabolites (repeated rate > 80%).....	52
Table 11	List of evaluation performance of the 31 Selected metabolites (repeated rate > 80%)	59

LIST OF FIGURES

Figure 1	General pipeline of data mining and knowledge discovery in metabonomics analysis	12
Figure 2	Diagrams of the process for training and predicting targets.....	15
Figure 3	Architecture of support vector machines.....	16
Figure 4	Different hyper planes could be used to separate examples.....	16
Figure 5	Mapping input space to feature space	20
Figure 6	Workflow of SVM-RFE metabolites selection procedure.....	40

LIST OF ABBREVIATIONS

ANN	Artificial Neural Networks
BC	Bladder Cancer
CE	Capillary Electrophoresis
GC-MS	Gas Chromatography-Mass Spectrometry
kNN	K Nearest Neighbor
LC-MS	Liquid Chromatography-Mass Spectrometry
NMR	Nuclear Magnetic Resonance
PCA	Principle Component Analysis
PLS	Partial Least Square
PNN	Probabilistic Neural Network
PQN	Probabilistic Quotient Normalization
RFE	Recursive Feature Elimination
SVM	Support Vector Machine

1 INTRODUCTION

Support vector machines (SVMs) are a group of supervised learning methods that can be applied to classification or regression problems. The support vector (SV) algorithm is a nonlinear generalization of the *Generalized Portrait* algorithm developed in the early 60's.^{1,2} In the past few decades, SVM showed excellent performance in many real-world applications such text categorization, hand-written character recognition, image classification and etc. With the advent of the genomic, proteomic and metabonomics era, the availability of human genome provides an opportunity to elucidate the genetic basis of biological processes and human diseases. However, the huge amount of data requires the development of high-throughput analysis tools and powerful computational capacity to facilitate the data analysis. Facing these challenges, bioinformatics has created many techniques, of which SVM as one of them. In the following sections, the increasing applications of SVM in bioinformatics, specifically genomics, proteomics and metabonomics, are reviewed.

1.1 Applications of SVM in bioinformatics

1.1.1 Applications of SVM in genomics

The Human Genome Project (HGP) was launched in 1989 with the initial goal of producing a draft sequence of the human genome. A working draft of genome was announced in 2000 and completed version in 2003. But knowledge of the genomic sequence is just the first step towards the understanding of the development and functions of organisms. The next key landmark will be an overview of the characteristics and

activities of the proteins encoded in the genes. Since not all genes are expressed at the same time, a further question is which genes are active under which circumstances. One of the immediate goals of comparative genomics is the understanding of the evolutionary trajectories of genes and integrating them into plausible evolutionary scenarios for entire genomes. A prerequisite for this process is a phylogenetic classification of genes.

The fast progress in genome sequencing projects calls for rapid, reliable and accurate functional assignments of gene products. Genome annotation³ enables the structural and functional understanding of genome. Computational analysis has been extensively explored to perform automatic annotation to co-exist with and complement mutual annotation. The basic level of annotation is annotating genomes based on BLAST based similarities. Nowadays a lot more additional information is added to the annotation platform including genome context information, similarity scores, experimental data and integrations of other resources and a variety of software tools have been developed to annotate sequences on a large scale. In recent years, the application of SVMs in genome annotation was aroused.⁴⁻⁸ These automated annotation systems develop binary classifiers based on sequence data and assign these sequences to certain Gene Ontology (GO) terms.⁴⁻⁸ Compared to other existing genome annotation systems, these SVMs based annotation tools outperform to some extent with more stable prediction results and better generalization capacity.⁵

With the accomplishment of HGP, genome-wide association studies (GWAS) are largely launched to derive gene signatures to determine common and complex diseases such as age-related macular degeneration (ARMD)⁹ and diabetes.¹⁰ In 2005, a GWAS found an association between ARMD and a variation in the gene of complement factor H (CFH).

Together with four other variants, these genes can predict half the risk of ARMD between siblings and make it the earliest and most successful example of GWAS.⁹ In 2007, a GWAS found an association between type 2 diabetes (T2B) and a variation in several single nucleotide polymorphisms (SNPs) in the genes TCF7L2, SLC30A8 and others.¹⁰ In recent years, SVMs have been applied to detect the variations associated with various diseases. Listgarten *et al.* explored combinations of SNPs from 45 genes and detected their potential relevance to breast cancer etiology in 174 patients and accuracy of 69% was obtained by using SVMs as the learning algorithm.¹¹ They concluded that multiple SNPs from different genes over distant parts of the genome are better at identifying breast cancer patients than any single SNP alone. Waddell *et al.* have applied SVMs to predict the susceptibility to multiple myeloma.¹² Their work had 71% accuracy on a dataset containing 40 cases and 40 controls.¹² In 2009, by using several machine learning techniques including SVM, Uhm *et al.* predicted patients' susceptibility to chronic hepatitis from SNPs.¹³ More recently, Ban *et al.* investigated 408 SNPs in 87 genes involved in major T2D related pathways in 462 T2D patients and 456 healthy controls using SVM and achieved a 65.3% prediction rate with a combination of 14 SNPs in 12 genes.¹⁴ As the high-throughput technology for genome-wide SNPs improves, it is likely that a much higher prediction rate with biologically more interesting combination of SNPs can be acquired and this will further benefit future drug discovery efforts and choosing of proper treatment strategies.

1.1.2 Applications of SVM in proteomics

After genomics, proteomics is considered the next step in the study of biological systems. It is much more complicated than genomics mostly because while an organism's genome

is more or less constant, the proteome differs from cell to cell and from time to time. This is because distinct genes are expressed in distinct cell types. This means that even the basic set of proteins which are produced in a cell needs to be determined. In the past, this was done by mRNA analysis but it was found not to correlate with protein content.^{15,16} It is now known that mRNA is not always translated into protein, and the amount of protein produced for a given amount of mRNA depends on the gene it is transcribed from and on the current physiological state of the cell. Besides, not only does the translation from mRNA cause differences, many proteins are also subjected to a wide variety of chemical modifications after translation. Many of these post-translational modifications, such as phosphorylation, ubiquitination, methylation, acetylation, glycosylation, oxidation, nitrosylation and etc., are critical to the protein's function.

Despite the difficulties in proteomic studies, scientists are still interested in proteomics because it gives a much better understanding of the functions of an organism than genomics. Functional clues contained in the amino acid sequence of proteins and peptides¹⁷⁻²⁰ have been extensively explored for computer prediction of protein function and functional peptides. A particular challenge is to derive functional properties from sequences that show low or no homology to proteins of known function.

Recently, SVMs have been explored for functional study of proteins and peptides by determining whether their amino acid sequence derived properties conform to those of known proteins of a specific functional class²¹⁻²⁵. The advantage of this approach is that more generalized sequence-independent characteristics can be extracted from the sequence derived structural and physicochemical properties of the multiple samples that share common functional profiles irrespective of sequence similarity. These properties

can be used to derive classifiers¹⁹⁻³⁰ for predicting other proteins that have the same functional or interaction profiles.

The task of predicting the functional class of a protein or peptide can be considered as a two-class (positive class and negative class) classification problem for separating members (positive class) and non-members (negative class) of a functional or interaction class. SVM and other well established two-class classification-based machine learning methods can then be applied for developing an artificial intelligence system to classify a new protein or peptide into the member or non-member class, which is predicted to have a functional or interaction profile if it is classified as a member.

The reported prediction accuracies for class members (P+) and non-members (P-) of SVM for predicting protein functional classes are in the range of 25.0%~100.0% and 69.0%~100.0%, with the majority concentrated in the range of 75%~95% and 80%~99.9% respectively^{21-24,31-45}. Based on these reported results, SVM generally shows a certain level of capability for predicting the functional class of proteins and protein-protein interactions. In many of these reported studies, the prediction accuracy for the non-members appears to be better than that for the members. The higher prediction accuracy for non-members likely results from the availability of more diverse set of non-members than that of members, which enables SVM to perform a better statistical learning for recognition of non-members.

Prediction of protein-binding peptides have primarily been focused on MHC-binding peptides,²⁷ the reported P+ and P- values for MHC binding peptides are in the range of 75.0%~99.2% and 97.5%~99.9%, with the majority concentrated in the range of

93.3%~95.0% and 99.7%~99.9% respectively.⁴⁶⁻⁴⁸ These studies have demonstrated that, apart from the prediction of protein functional classes, SVM is equally useful for predicting protein-binding peptides and small molecules.

From the above reported results, it can be easily concluded that SVM shows promising potential for a wide spectrum of protein and peptide classes including some of the low- and non-homologous proteins. This method can thus be explored as a potential tool to complement alignment-based, clustering-based, and structure-based methods for predicting protein function and interactions.

1.1.3 Applications of SVM in metabonomics

Metabonomics is the comprehensive and quantitative assessment of low molecular weight analytes (<1500Da) that define the metabolic status of an organism under a given condition.⁴⁹ In complementation with genomics and proteomics, the direct measurement of metabolite expression is essential in the systematic understanding of biological process. Metabolomics is increasingly enjoying widespread applications in areas such as functional genomics, identification of the onset and progression of disease, pharmacogenomics, nutrigenomics, and system biology.⁵⁰⁻⁵³

Because of its sensitivity and coverage, mass spectrometry (MS) is a favorable technology for metabolomics study. One major bottleneck for current MS-based metabolomics is the identification of metabolites. To identify the correct metabolite from a large volume of MS/MS spectra, a proper comparison or scoring scheme is needed. In machine learning, SVMs are widely considered to represent the state of the art in classification accuracy. Recently, SVMs have been applied to the supervised

classification of cancer versus control sample sets from MS data.⁵⁴⁻⁶³ Xue *et al.* investigated the serum metabolic difference between hepatocellular carcinoma (HCC) male patients and normal male subjects by stepwise discriminant analysis (SDA) and SVM based on gas chromatography (GC)/MS data.⁶¹ The resultant diagnostic model could discriminate between HCC patients and normal subjects with 20-fold cross validation classifying accuracy of 75% and error count estimate for each group of 0%.⁶¹ Hennes *et al.* constructed breast cancer predictive models by profiling of urinary RNA metabolites using SVM-based feature selection from data obtained from liquid chromatography ion trap (LC-IT) MS, and had classification sensitivity and specificity of 83.5% and 90.6% respectively.⁶³ The performance of SVM for the classification of liquid chromatography/time-of-flight (LC/TOF) MS metabolomics data focusing on recognizing combinations of potential metabolic ovarian cancer diagnostic biomarkers was evaluated by Guan *et al.*⁵⁴ The classification of the serum sample test set was 90% accurate, which suggests that the developed approach might lead to the development of an accurate and reliable metabolomics-based approach for detecting ovarian cancer.⁵⁴ More recently, Zhou *et al.* collected MS/MS spectra for 21 metabolites from both in-house data and publicly available data from the Human Metabolite Database (HMDB) and utilized SVM to incorporate both peak and profile similarity measures for spectral matching. The models had accuracies and F-measure ranging from 94.6%~96.3% and 80.7%~85.1% respectively.⁶⁴ By comparing the identification performance with other algorithms (NIST, MassBank and SpectraST) and the correlation method, it was observed that SVM can achieve 7% to 10% improvement on identification performance.⁶⁴

1.2 Underlying difficulties in using SVM

The performance of SVM critically depends on the diversity of samples in a training dataset and the appropriate representation of these samples. The datasets used in many of the reported studies are not expected to be fully representative of all of the proteins, peptides and small molecules with and without a particular functional and interaction profile. Various degrees of inadequate sampling representation likely affect, to a certain extent, the prediction accuracy of the developed statistical learning models. SVM is not applicable for proteins, peptides and small molecules with insufficient knowledge about their specific functional and interaction profile. Searching of the information about proteins, peptides and small molecules known to possess a particular profile and those that do not possess the profile is key to more extensive exploration of statistical learning methods for facilitating the study of functional and interaction profiles.

In the datasets of some of the reported studies, there appears to be an imbalance between the number of samples having a profile and those without the profile. SVM method tends to produce feature vectors that push the hyper-plane towards the side with smaller number of data,⁶⁵ which often lead to a reduced prediction accuracy for the class with a smaller number of samples or less diversity (usually members) than those of the other class (usually 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. Computational methods for re-adjusting biased shift of hyper-plane are being explored.⁶⁶ Application of these methods may help improving the prediction accuracy of SVM in the cases involving imbalanced data.

While a number of descriptors have been introduced for representing proteins and peptides,^{19,31,67,68} most reported studies typically use only a portion of these descriptors. It

has been found that, in some cases, selection of a proper subset of descriptors is useful for improving the performance of SVM.⁶⁹⁻⁷¹ Therefore, there is a need to explore different combination of descriptors and to select an optimum set of descriptors using feature selection methods.⁶⁹⁻⁷¹ Efforts have also been directed at the improvement of the efficiency and speed of feature selection methods,⁷² which will enable a more extensive application of feature selection methods. Moreover, indiscriminate use of the existing descriptors, particularly those of overlapping and redundant descriptors, may introduce noise as well as extending the coverage of some aspects of these special features. Thus, it may be necessary to introduce new descriptors for the systems that have been described by overlapping and redundant descriptors. Investigations of cases of incorrectly predicted samples have also suggested that the currently-used descriptors may not always be sufficient for fully representing the structural and physicochemical properties of proteins, peptides and small molecules.^{30,55,73} These have prompted works for developing new descriptors.⁴²

1.3 Objectives and organization of this thesis

1.3.1 Objectives of this thesis

The main objective of this thesis is to investigate and develop novel systems of support vector machine for –omics application. Two types of studies were included in this investigation. These are MHC binding prediction for proteomics level, and metabolites selection for metabonomics level.

The first study is to explore an improved flexible prediction system for MHC binding prediction. Generally, there are several inevitable limitations of the current prediction

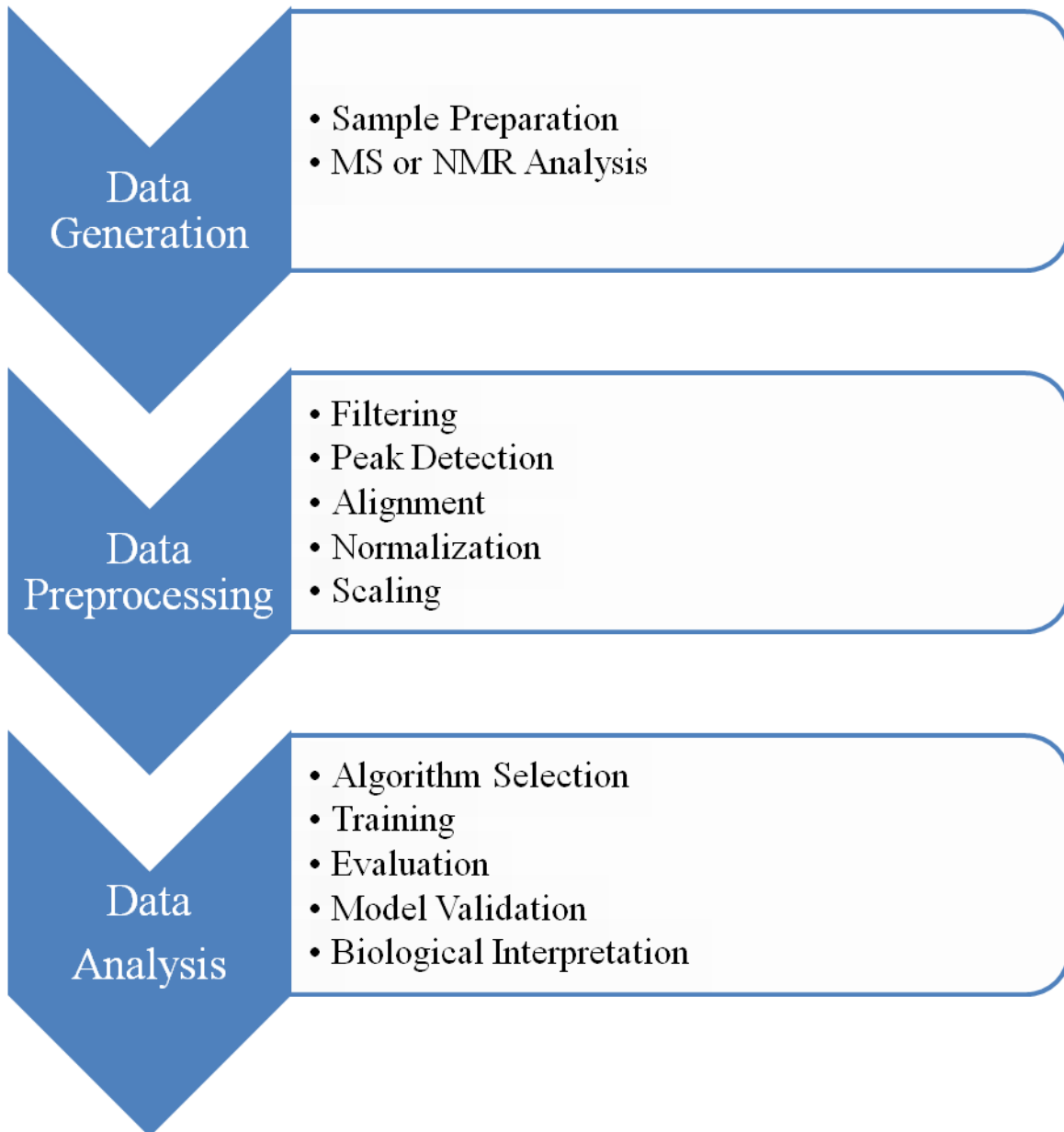
systems. First of all, most prediction systems were particularly designed for peptides with fixed lengths. Secondly, the dataset size of the existing systems, especially the training dataset of non-binders are not adequate for building a reliable prediction model. Thirdly, some of the prediction systems represented peptides not by the structural and physicochemical properties, but by sequence of peptides directly. Last but not least, most MHC binding prediction systems only cover a limited number of MHC alleles, which leads to a lack of statistically significant number of known peptides in the commonly studied length ranges.

There are several feasible ways to alleviate the above problems. These include choosing a prediction algorithm which works for peptides with flexible lengths; representing the peptides with sequence-derived structural and physicochemical properties; and conducting the training data with sufficiently diverse set of non-binders. All of these improvements can be achieved in the studies by using support vector machine. According to previous studies, SVM has shown promising capability for prediction of specific functional group of flexible lengths with sequence-derived structural and physicochemical properties. Moreover, peptides in same specific functional group are generally diverse but share similar structural and physicochemical features. To some extents, the MHC binding peptides in specific alleles share similar characteristics, which mean they have similar structural and physicochemical features. Therefore, SVM is expected to be a potential eligible algorithm to be applied for predicting MHC binding and non-binding peptides.

The second part of this thesis is to investigate a new approach of metabolites selection by using support vector machine feature selection system. The development of a new

approach of metabolites selection is one of the major topics in the area of data mining in metabonomics studies. It is important to find the marker metabolites responsible for disease reaction. This may help in early diagnosis and correct prediction of disease. The general workflow of data mining in metabonomics analysis can be found in **Figure 1**. There are two major sub-objectives for the second part of study. (1) Discovery of marker metabolites responsible for the distinction between groups of samples related to the specific interests. (2) Development the better metabolites selection methods by advanced machine learning algorithm. Compared with the traditional methods of metabolites selection, the new approach will be derived from the strategies of gene selection in microarray data. Several feature selection methods and algorithms (e.g.: SVM recursive feature elimination, forward/backward weighting methods based on Decision tree, Naïve Bayes kernel function and other traditional weighting methods) will be compared to determine their performance and usability for metabolite selection.

Figure 1 General pipeline of data mining and knowledge discovery in metabonomics analysis



1.3.2 Organization of this thesis

Chapter 1 introduces the history of SVMs and reviews their increasing applications in bioinformatics especially in genomics, proteomics and metabolomics.

Chapter 2 describes in detail the mathematical theory of SVM as a combination of two main concepts: *Maximal Margin Hyperplanes* (also called *Optimal Separating Hyperplanes*) and *kernel functions*. The general criteria for evaluating the classifying performance are also introduced.

Chapter 3 elucidated the real application of SVM in MHC binding prediction. Several SVM prediction systems were developed and evaluated for the multiple MHC alleles. The accuracies of these prediction systems were validated using fivefold cross validation.

Chapter 4 elaborated the application of SVM for metabolites selection in metabonomics. Urine samples of 75 subjects of bladder cancers were investigated with the methods of metabonomics. The advances of SVM system in metabolites selection were demonstrated by comparison with several feature selection algorithms.

Chapter 5 concludes the achievement and limitation of current work. Future works are also introduced in this chapter.

2 METHODOLOGY

2.1 Support vector machines (SVMs) method

The process of training and using a SVM model for screening peptides based on their physicochemical property descriptors is schematically illustrated in **Figure 2**. SVM is based on the structural risk minimization principle of statistical learning theory,⁷⁴⁻⁷⁹ which consistently shows outstanding classification performance, is less penalized by sample redundancy, and has lower risk for over-fitting.⁸⁰⁻⁸²

2.1.1 Linear SVM

In two-class problems, SVM aims to separate examples of two classes with the maximum hyper plane (**Figure 3**). 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 hyper plane could be drawn to separate examples of one class (positive examples) from those of the other one (negative examples). The hyper plane is represented by $w \cdot x + b = 0$, where w is slope and b is bias. Thus the objective function of SVM changes to minimize Euclidean norm $\|w\|^2$ with following limitations:

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

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

Figure 2 Diagrams of the process for training and predicting targets

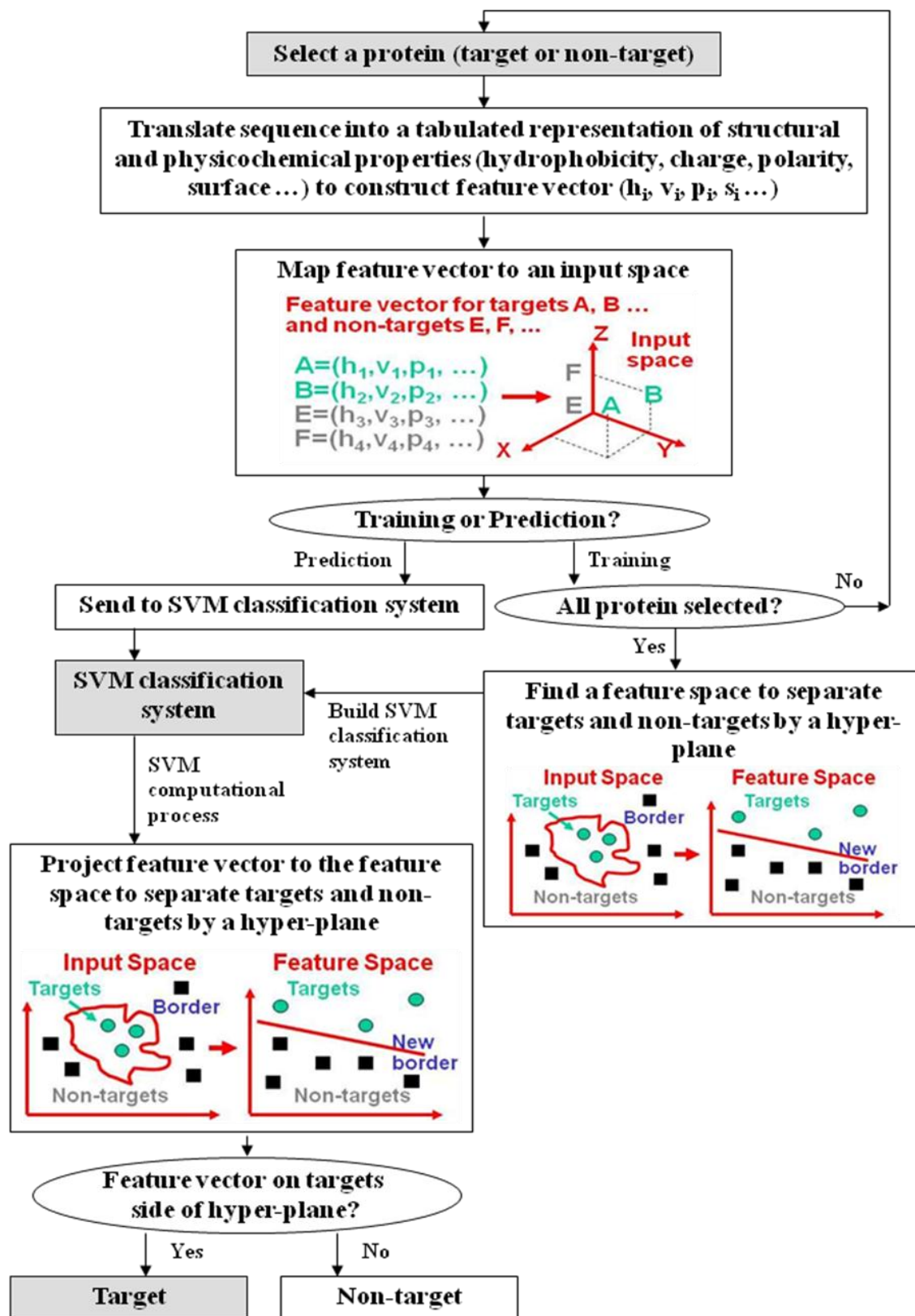


Figure 3 Architecture of support vector machines

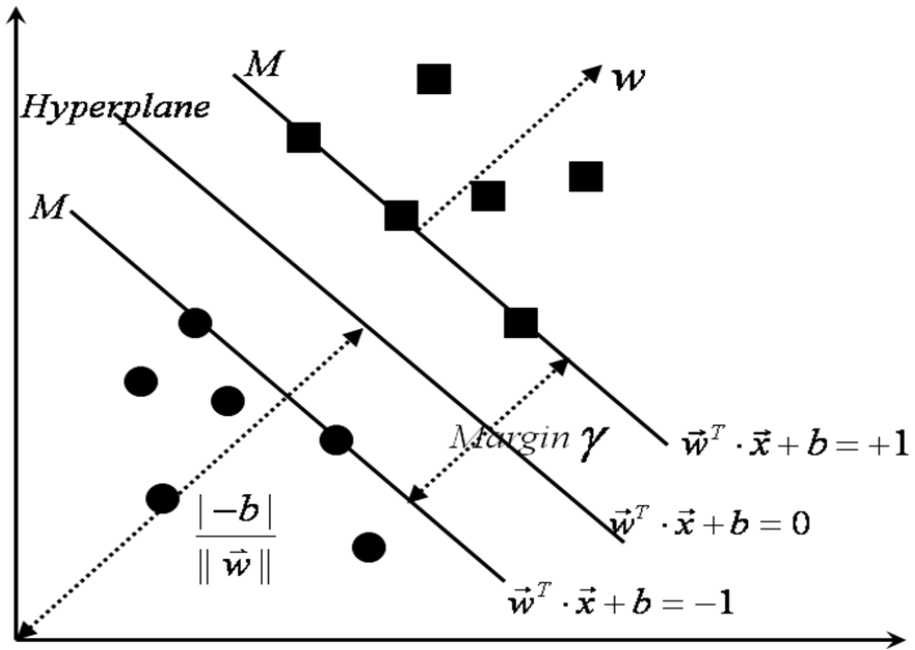
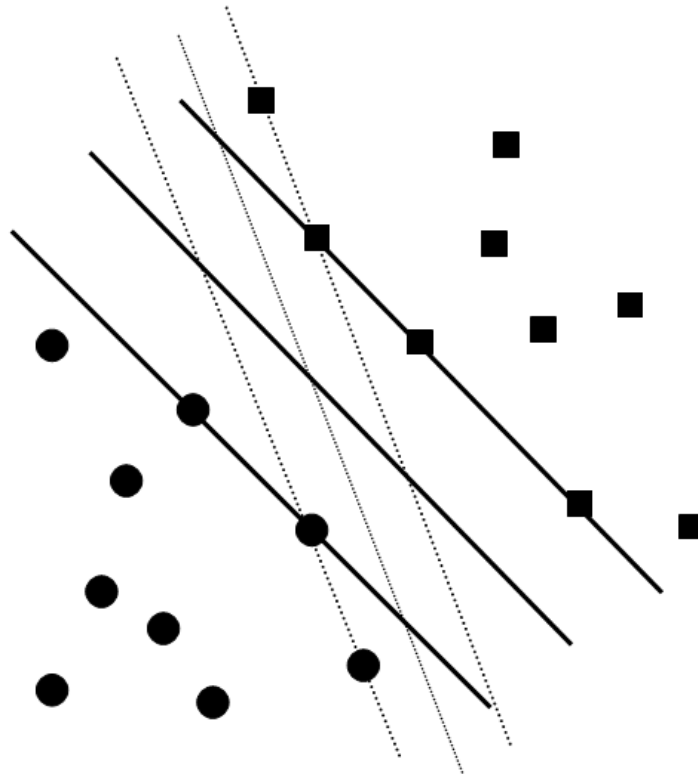


Figure 4 Different hyper planes could be used to separate examples



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

Geometrically, all the points are divided into two regions by a hyper plane H . As shown in **Figure 4**, there are numerous ways through which a hyper plane can separate these examples. The objective of SVM is to choose the “optimal” hyper plane. As all new examples are supposed to be located under similar distribution as training examples, the hyper plane should be chosen such that small shifts of data do not result in fluctuations in prediction result. Therefore, the hyper plane that separates examples of two classes should have the largest margin, which is expected to possess the best generalization performance. Such hyper plane is called the Optimal Separating Hyper plane (OSH).⁸³

Examples locating on the margins are called support vectors, whose presentation determines the location of the hyper plane. 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 hyper plane 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:^{84,85}

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

Subject to:

$$\gamma_x(w, b) > 0 \quad (6)$$

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

which is an equivalent statement of the problem

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

Subject to:

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

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

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 α_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:

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

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:

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

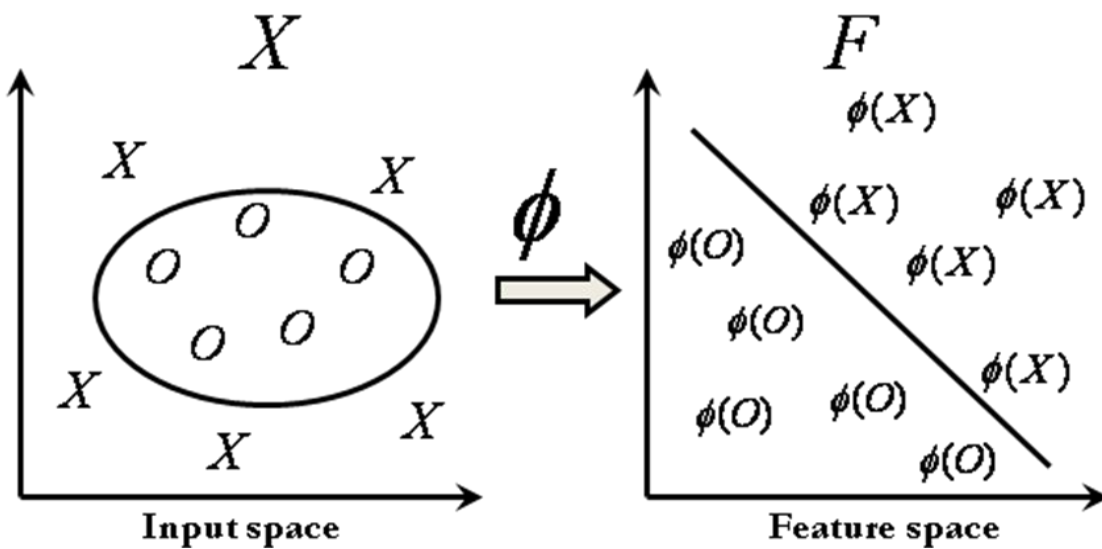
This QP problem could be efficiently solved through several standard algorithms like Sequential Minimization Optimization⁸⁶ or decomposition algorithms.⁸⁷

Once w_0 and b_0 are determined, the hyper plane is readily drawn. The points for which $\alpha_i > 0$ are called support vectors, which lie on the margin⁸⁸.

2.1.2 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 5).

Figure 5 Mapping input space to feature space



Let Φ 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 Φ 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,⁸⁹ where mapping function cannot be explicitly represented. A function could be used as a kernel function if and only if it satisfies Mercer's condition.⁹⁰ Followings are 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 is used due to its many advantages demonstrated in previous studies. Different SVM models could be developed by using different σ values. It is thus necessary to scan a number of σ values to find the best model, which is evaluated

by their performance on classification tasks. **Figure 1** illustrates the schematic diagrams of the process of training and prediction of drug targets by SVM. Sequence-derived feature h_i , p_i , v_i ... represents such structural and physicochemical properties as hydrophobicity, polarizability, and volume. The calculation of the structural and physicochemical properties used for representing MHC binding peptides is described in **Chapter 3** and the Recursive Feature Elimination (RFE) method used for metabolites prediction is introduced in **Chapter 4**.

2.2 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.⁹¹

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 follows:

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

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

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

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

where TP, FN, TN, and FP represent correctly predicted positive data, positive data incorrectly predicted as negative, correctly predicted negative data, and negative data

incorrectly predicted as positive respectively. Another measurement, Matthews correlation coefficient (MCC), was also used to evaluate the randomness of the prediction.

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

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

3 MHC BINDING PREDICTION

This work developed several prediction systems for 22 MHC Class I and 17 MHC Class II alleles by SVM. An original dataset without the pseudo non-binding peptides has been tested. All peptide of this dataset were collected from the database. The 29520 binder peptides and 24848 non-binder peptides were collected from IEDB have been tested with the five-fold cross validation. As a comparison, serial tests were conducted based on each allele. The pseudo non-binding peptides generated from the splitting proteins have been included in these tests. Fivefold cross validation has been applied to evaluate the performance of these prediction systems.

3.1. Data Preparation

Data collection from databases

Binding peptides and non-binding peptides of 22 MHC class I and 17 MHC class II alleles were collected from 2 databases: IEDB (Immune Epitope Database www.immuneepitope.org/) and SYFPEITHI (www.syfpeithi.de). A total of 70692 MHC binding peptides were collected from these two databases. After removing the duplicated binders, there were 29520 peptides left. 93734 MHC non-binding peptides were collected from these two databases. After removing the duplicated non-binders, there were 24848 peptides left.

It had been discovered that the number of tested peptides can severely affected the model's prediction performance, especially when the number is less than 150⁹². Thus,

only alleles with more than 150 binding peptides had been chosen to be studied in this project, to ensure a good performance of the prediction model.

There are 452, 5015, 856, 882, 796, 1176, 1134, 65, 308, 324, 226, 547, 209, 609, 517, 488, 335, 526, 454, 252, 209, 1274, 339, 288, 254, 1993, 370, 874, 270, 238, 373, 240, 221, 498, 236, 379, 150,254, 374 binders for class I and class II allele HLA-A*0101, HLA-A*0201, HLA-A*0202, HLA-A*0203, HLA-A*0206, HLA-A*0301, HLA-A*1101, HLA-A*2601, HLA-A*2902, HLA-A*3001, HLA-A*3002, HLA-A*3101, HLA-A*330, HLA-A*6801, HLA-A*6802, HLA-B*0702, HLA-B*0801, HLA-B*1501, HLA-B*3501, HLA-B*4402, HLA-A*11, HLA-A*2, HLA-DR*1, HLA-DR*4, HLA-DR*7, HLA-DRB1*0101, HLA-DRB1*0301, HLA-DRB1*0401, HLA-DRB1*0404, HLA-DRB1*0405, HLA-DRB1*0701, HLA-DRB1*0802, HLA-DRB1*0901, HLA-DRB1*1101, HLA-DRB1*1302, HLA-DRB1*1501, HLA-DRB3*0301, HLA-DRB4*0101, HLA-DRB5*0101 respectively. The detail information of datasets can be found in Table 3.

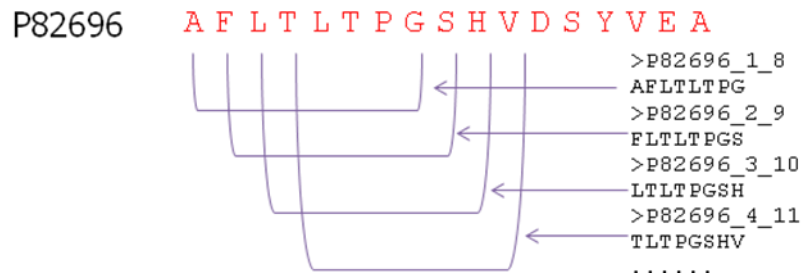
MHC Non-binders generation

Theoretically, an n-mer peptides can lead to $20n$ possible combinations. Compared to these enormous combinations, the limited number of known non-binding peptides is much smaller than the total number of the possible combinations, which cannot sufficiently represent the entire sequence space. A similar situation happened in proteins functional families^{24,92}. According to other researchers' works^{24,92,93}, additional numbers of proteins without the specific functions can be created by grouping these pseudo proteins into specific domain families and populating the whole protein space by

selecting representative proteins from each group of these un-functional families. Such kinds of efforts are expected to be applicable for MHC non-binders generation.

In this work, the additional non-binder peptides were generated from splitting the representative protein from each protein family. The steps are outlined as below:

- 1) 10082 representative proteins were selected from the 10000+ protein families respectively.
- 2) Each selected protein has been split into small peptides with different lengths from 8 amino acids to 25 amino acids. The splitting procedure is shown as below.



- 3) The peptides were removed from the generated peptides if they were identical to the binder peptides from the database. The purpose of this step is to ensure the binding peptides were not included in the generated dataset. 472,118 peptides were removed from the generated peptides. 78,000,000 peptides were left and can be treated as the negative dataset.
- 4) Because the generated non-binder dataset is too large to be used in further modeling steps, an eligible selection procedure is necessary to be applied to select the representative negative dataset from the entire negative dataset. Peptides should be

clustered into groups based on their structural and physicochemical feature space.

Then the representative peptides were randomly selected from each group to form a training set that is sufficiently diverse and broadly distributed in the feature space.

However, due to the large number of generated non-binding peptides in this work, a very long time would be needed to cluster 78,000,000 peptides into specific groups, especially when each peptide is described using hundreds of descriptors. A classical K-means clustering method would take several months to complete the entire clustering process. Therefore, as a more simplified clustering method, randomly selection algorithm has been applied to select specific number of peptides from each group. Representative peptide is randomly selected from each group to form the dataset which is sufficiently diverse and equally distributed in the feature space. The representative non-binders have been equally selected from different lengths of peptides, from 8-mer to 25-mer, and distributed into each allele group, according to a certain ratio of binders to non-binders.

3.2. Descriptor Generation

Several descriptors development methods have been designed to construct the feature space for peptides^{94,95}. For instance, the peptide sequence can be straightforwardly represented by direct sequence of amino acids.

In this study, as the binders and non-binders datasets were combined by flexible lengths of peptides, the straightforward vector representation method would create different number of descriptors for each peptide, which is not suitable for following modeling procedures. Therefore, a feature representation method with the structural and physicochemical properties of a peptide has been developed with a well-formulated

procedure. The same number of descriptors can be developed for different lengths of peptides by this method. Given the sequence of a peptide, the physical and chemical properties, as well as the composition of every constituent amino acid can be computed with certain formulas and then generated to be vectors. These computed amino acid properties include hydrophobicity, normalized van der Waals volume, polarity, polarizability, charge, surface tension, secondary structure, solvent accessibility⁹² and three global composition descriptors: composition, transition and distribution.

For each of the properties, amino acids can be divided into three or six groups such that those in a particular group are regarded to have approximately the same property. For instance, charge of amino acid can be divided into three groups: positive (KR), Neutral (ANCQGHILMFPSTWYV), and Negative (DE). Secondary structure of amino acid can be divided into three groups: Helix (EALMQKRH), Strand (VIYCWFT), and Coil (GNPSD). The detailed division of amino acids can be found in **Table 1**.

The global composition of amino acids includes three descriptors: composition (C), transition (T), and distribution (D), C represents the number of amino acids of a specific property divided by the number of total number of amino acids in an entire peptide. T is the percent frequency of amino acids with a particular property followed by amino acid with different properties. D characters the distribution of the properties along the sequence within which the first, 25%, 50%, 75% and 100% of the amino acids of a particular property are located respectively.

Table 1 Division of amino acids for different physicochemical properties.

6 Dimensions Property	Divisions					
	Group 1	Group 2	Group 3	Group 4	Group 5	Group 6
Hydrophobicity	0~0.043	0.165~0.359	0.45~0.501	0.616~0.738	0.825~0.888	0.943~1
	RDE	HNQKS	TG	ACPM	VWY	ILF
Van der Waals volume	0~1.6	2.43~2.78	2.95~3	3.78~4.0	4.43~4.77	5.89~8.08
	GAS	CTPD	NV	EQIL	MHK	FRYW
Polarity	0	0.352~0.456	0.6~0.608	0.648~0.696	0.792~0.888	0.864~1.0
	VI	LFWCMY	PA	TGS	HQR	KNED
Polarizability	0~0.062	0.105~0.108	0.128~0.15	0.18~0.186	0.219~0.23	0.29~0.409
	GAS	DT	CPNVE	QIL	KMH	FRYW
3 Dimensions Property	Group 1		Group 2		Group 3	
Charge	Positive		Neutral		Negative	
	KR		ANCQGHILMFPSTWYV		DE	
Surface tension	-0.20~0.16		-0.3~ -0.52		-0.98~ -2.46	
	GQDNAHR		KTSEC		ILMFPWYV	
Secondary structure	Helix		Strand		Coil	
	EALMQKRH		VIYCWFT		GNPSD	
Solvent accessibility	Buried		Exposed		Intermediate	
	ALFCGIVW		RKQEND		MPSTHY	

For instance, consider a sequence KRACQTDKDLERWTS. According to the charge division in **Table 1**, the charge descriptor of this peptide is encoded as 112222313231222.

Its composition descriptor can be calculated as

$$C = \left(\frac{n_1 * 100}{N}, \frac{n_2 * 100}{N}, \dots, \frac{n_m * 100}{N} \right) \quad (19)$$

n_m is the number of m in the encoded sequence and N is the length of this sequence. According to the example, the number of encoded class “1” is 4, “2” is 8, “3” is 3. The composition are $4/15=26.7\%$, $8/15=53.4\%$ and $3/15=20\%$ respectively.

Its transition descriptor can be calculated as

$$T = \left(\frac{T_{G_1G_2} \times 100}{N-1}, \frac{T_{G_1G_3} \times 100}{N-1}, \frac{T_{G_2G_3} \times 100}{N-1} \right), \quad (20)$$

where G_1G_2 , G_1G_3 , and G_2G_3 are 12, 13, 23 respectively. $T_{G_1G_2}$, $T_{G_1G_3}$, $T_{G_2G_3}$ are the numbers of dipeptide encoded as 12, 13, 23 respectively in the sequence, T is the length of the sequence.

Its distribution descriptors can be calculated as

$$D = \left(\frac{P_{i0} \times 100}{N}, \frac{P_{i25} \times 100}{N}, \frac{P_{i50} \times 100}{N}, \frac{P_{i75} \times 100}{N}, \frac{P_{i100} \times 100}{N} \right) \quad (21)$$

There are five distribution descriptors for each encoded number and they describe the position percents in the whole sequence for the 0%, 25%, 50%, 75% and 100% residues respectively.

To sum up, there are 20 dimensions of composition descriptors, 51 dimensions of hydrophobicity, Van der Waals volume, Polarity, and Polarizability, 21 dimensions of Charge, Surface tension, Secondary structure and Solvent accessibility respectively. The total number of descriptors is 308.

3.3. Overview of SVM modeling procedure.

- (1) Import the original pre-processed dataset into a matrix.
- (2) Derive physical and chemical features from sequence for each peptide. i.e. Hydrophobicity, Volume, and Polarizability etc. 308 descriptors were generated for each peptide.
- (3) Normalized descriptors to the same scale using the formula $\frac{X - \min}{\max - \min}$, the range of descriptors is from 0 to 1.
- (4) Randomized the dataset into five subgroups. Held one as the testing set, and rest are training sets. Created five training sets and 5 testing sets by this step, as the fivefold cross validation.

- (5) Chosen the appropriate SVM parameters to identify the most suitable model for each dataset.

3.4. Results and Performance evaluation

3.4.1. Self consistency testing accuracy of dataset without generated non-binders

The 29520 binder peptides and 24848 non-binder peptides downloaded from IEDB have been used to run the whole procedure as the sample tests. The average accuracy of the test was around 40%, which is shown in **Table 2**.

The main reason for this poor result is due to the lack of the clustering. The negative data cannot be selected from the entire feature space without the effective clustering. Without these negative datasets, the prediction algorithm cannot create an effective model to properly distinguish the positive data and negative data.

3.4.2. Self consistency testing accuracy of dataset with generated non-binders

Table 3 gives the results of the SVM prediction systems based on the fivefold cross validation sets. As shown in the table, the overall accuracies were in the range of 90% to 99% for all alleles, except the HLA-A*0201 and HLA-DRB1*0101, which were 86.97% and 89.24% respectively. The overall accuracies of 30 alleles were above 96%, 7 alleles were above 90% and the other 2 alleles were above 85%. These results demonstrated the manifest improvement of the prediction accuracies due to the application of generated negative datasets.

Table 2 Prediction performance of MHC binding peptides without generated non-binders.

SVM parameters		Fivefold cross validation performance		
C	gamma	Sensitivity	Specificity	Testing Accuracy
1000	0.1	68.1%	0.4%	37.3%
1000	0.6	66.3%	4.1%	38.4%
10000	1.1	52.7%	32.8%	42.8%
10000	1.6	49.4%	37.9%	44.3%
100000	2.1	47.8%	37.0%	42.9%

Table 3 Datasets and the binder and non-binder prediction accuracies for HLA alleles I.

HLA Allele	Training set		Accuracies of 5-fold cross validation
	Binders	Non-Binders	
HLA-A*0101	452	14225	96.9%
HLA-A*0201	5015	15091	87.0%
HLA-A*0202	856	14731	94.5%
HLA-A*0203	882	14736	94.4%
HLA-A*0206	796	14919	94.9%
HLA-A*0301	1176	16300	93.3%
HLA-A*11	209	6625	96.9%
HLA-A*1101	1134	15560	91.7%
HLA-A*2	1274	16993	93.0%
HLA-A*2601	65	2069	97.0%
HLA-A*2902	308	9721	96.9%
HLA-A*3001	324	10241	97.9%
HLA-A*3002	226	7148	96.9%
HLA-A*3101	547	17168	96.9%
HLA-A*3301	209	6624	96.9%
HLA-A*6801	609	19072	95.7%
HLA-A*6802	517	16224	93.3%
HLA-B*0702	488	15350	96.3%
HLA-B*0801	335	10580	94.3%
HLA-B*1501	526	16530	96.9%
HLA-B*3501	454	14307	92.4%
HLA-B*4402	252	7983	96.3%

Continued Table 3: Datasets and the binder and non-binder prediction accuracies for HLA alleles II.

HLA Allele	Training set		Accuracies of 5-fold cross validation
	Binders	Non-Binders	
HLA-DR*1	339	10693	96.7%
HLA-DR*4	288	9115	96.3%
HLA-DR*7	254	8047	96.4%
HLA-DRB1*0101	1993	16527	89.2%
HLA-DRB1*0301	370	11683	96.9%
HLA-DRB1*0401	874	16975	95.1%
HLA-DRB1*0404	270	8542	93.9%
HLA-DRB1*0405	238	7545	96.9%
HLA-DRB1*0701	373	11781	96.3%
HLA-DRB1*0802	240	7604	96.0%
HLA-DRB1*0901	221	7013	94.5%
HLA-DRB1*1101	498	15650	96.1%
HLA-DRB1*1302	236	7479	94.1%
HLA-DRB1*1501	379	11962	92.9%
HLA-DRB3*0301	150	4771	95.2%
HLA-DRB4*0101	254	8052	94.6%
HLA-DRB5*0101	374	11800	96.9%

3.5. Summary and Discussion

The prediction accuracies of binding peptides by the SVM systems were 90%-96% for 37 alleles and 86%-89% for 2 alleles, which were much better than the previous model which was built using the original datasets. Thus, we can conclude the false binder prediction rate is significantly reduced by adding the generated negative datasets.

It should be noted that the performance of MHC binding prediction might be affected by several factors. The first one is the diversity of binding peptide samples. A good prediction system cannot be established without adequate samples. Thus higher accuracies will be achieved with more MHC binder information. Secondly, the imbalanced dataset should be created to represent the entire feature space. A smaller number of negative data can lead to reduced accuracy or less diversity.

4 METABOLITES SELECTION IN METABONOMICS

4.1. Data collection and normalization

The aim of this study was to investigate the role of urinary metabonomics in the diagnosis of human bladder cancer and determine the stage of tumor growth. There were 75 subjects, which included 24 bladder cancer (BC) patients and 51 non-bladder cancer (non-BC) subjects in the study. All the urine samples were collected from the 75 subjects and stored at -80 °C for further processing. Gas chromatography (GC)/time-of-flight (TOF) mass spectrometry has been applied for these urine samples after the serial processing of urine preparation. Data acquisition was performed in the full scan mode from m/z 40 to 600 with an acquisition rate of 20 spectra/sec. Baseline correction, noise reduction, smoothing, library matching and area calculation had been applied for data pre-processing of each chromatogram obtained from GC/TOF analysis. Two sets of data were produced after data pre-processing: (1) 75 urines samples (24 BC and 51 non-BC) with 189 metabolites for each sample. (2) 75 urines samples (24 BC and 51 non-BC) with 398 metabolites for each sample.

Normalization is a systematic way of ensuring that a dataset structure is suitable for general-purpose querying and free of certain undesirable characteristics. After redundancy elimination, data organization and potential data anomalies reduction, the biological difference among different samples can be determined and compared using machine learning methods. In this study, all the values were derived from the GC/TOF chromatogram and processed using the same data pre-processing procedure. Thus normalization can be performed for all the samples.

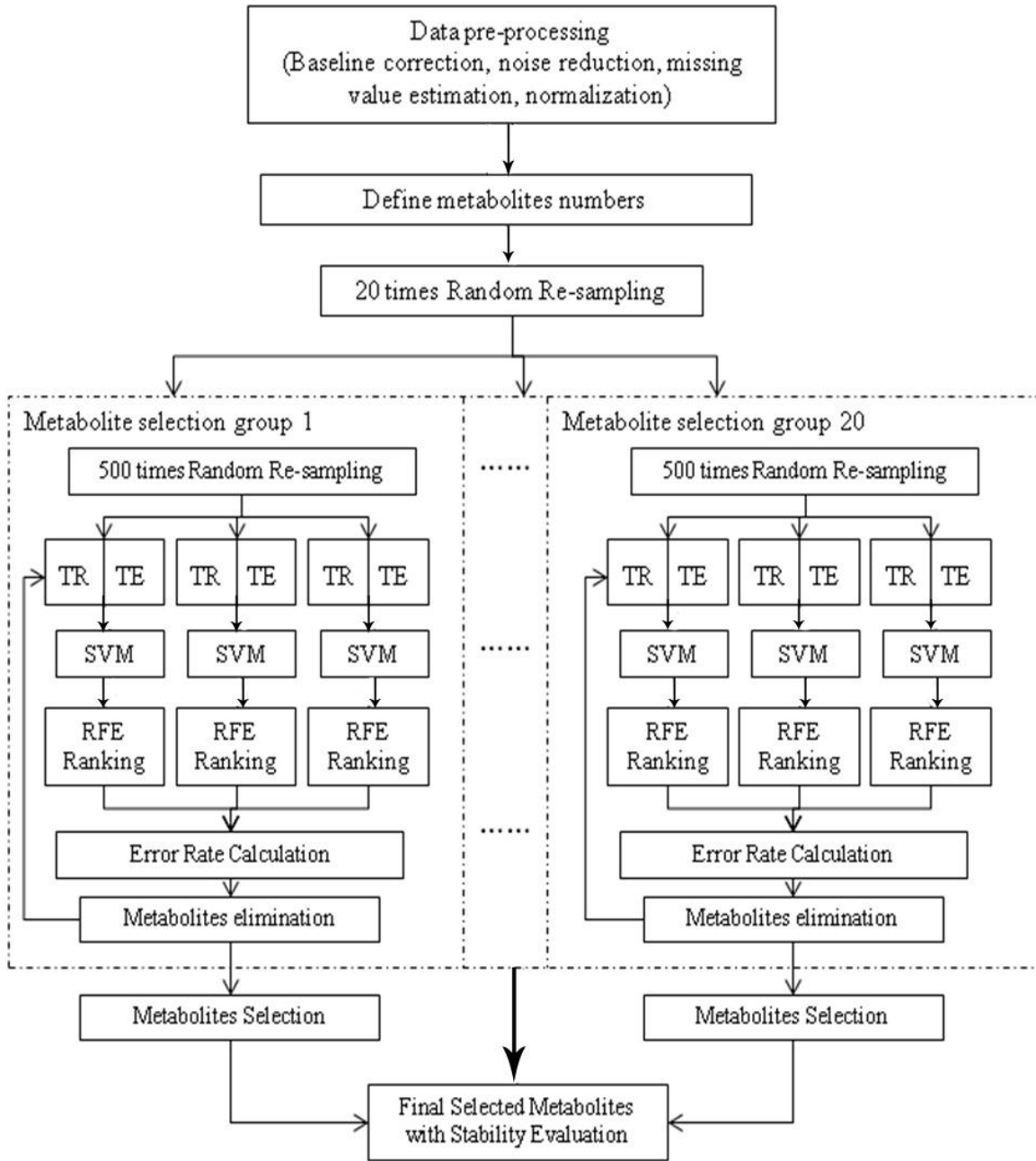
Similar to microarray experiments, the major normalization methods for metabonomic studies include global normalization performed for all metabolites on the array, and housekeeping metabolites normalization using constantly expressed housekeeping/invariant metabolites. The housekeeping normalization method might introduce extra potential errors since the metabolites were found to be not expressed constantly. Thus, we used global normalization in this study with following procedure. (1) Transforming the raw dataset into a two-dimensional matrix. The rows represent different patient samples and columns indicate different metabolites detected in patients' samples. (2) Summing up all values for each column. The sum of peak value of each metabolite can be expressed by $M_{(i)}$, i represents the number of column. (3) Dividing every row's value of each column by the absolute value of $M_{(i)}$. Then the values of each metabolite can be ranged from -1 to +1. Normalization is a key step in the pre-processing of metabonomics data and can have a large impact in identifying differential metabolites marks and classification for diagnosis. By taking normalization, random or systemic variations, such as the influence of detecting efficiencies for each patient's sample can be well identified and removed. Thus the data to be analysed are independent of particular experiment and technology used. This can help to avoid the bias caused by variations in sample preparation and GC/TOF analysis. The above metabonomics dataset were kindly provided by Metabolic Profiling Research Group at NUS Pharmacy.

4.2. Overview of SVM-RFE selection procedure

We developed a metabolites selection procedure based on algorithm of support vector machines (SVM) and the procedure of recursive feature elimination (RFE). An overview

of the procedure is shown in **Figure 6** and the steps are outlined as below:

Figure 6 Workflow of SVM-RFE metabolites selection procedure



- (1) Import the original pre-processed dataset into a matrix where each row represents a sample and each column represents a metabolite.
- (2) Use the random re-sampling method on the dataset matrix for 20 times to generate 20 groups of training-testing datasets for further analysis.
- (3) Divide each of the 20 groups of training-testing dataset into 500 subgroups, which means 500 different sample combinations.
- (4) Use SVM classifier on each training-testing sample combination to determine the class of samples (BC or non-BC).
- (5) Use RFE ranking criteria to sequentially rank the importance of each metabolite for the SVM classifier. For each group, 500 lists of ranked metabolites will be generated.
- (6) Perform the consistency evaluation based on the sequence of metabolites to determine the value of contribution. Remove the metabolite with the least contribution.
- (7) Iteratively repeat steps (5) and (6) until the SVM classifier achieves highest accuracy and no more metabolites can be removed.
- (8) Generate a metabolites list with highest accuracy from the 500 lists of metabolites. These metabolites are the biomarkers for the original dataset.

(9) Repeat steps (3) to (8) for the rest of 19 groups for the stability evaluation. The final selected metabolites will be determined after comparing the results from all the 20 groups.

4.3. Results and Discussion

4.3.1. Comparison of prediction performance of multiple machine learning methods.

Five machine learning methods, decision tree, Naïve Bayes with kernel function, k-nearest neighbor algorithm, neural network and SVM, were used to develop models using the dataset with 189 metabolites. The results, given in **Table 4**, showed that the overall accuracies of all the classifiers were in the range of 66.68%-72.76%. According to Table 4, the accuracies of KNN and Neural Network are over 75%, which are higher than other 3 methods; but the specificity of Neural Network is much lower than KNN's, which only reached 48%. On the other hand, both sensitivity and specificity of KNN are higher than 60%, and its AUC is the highest among the five methods. Therefore, KNN can be recommended as the best algorithm to build the predict model. However, such performance was also much weaker than the model developed by using the metabolites selection procedure, which will be further introduced in following sections. The differences in performance between the five models are slight and may be due to the fact that the choice of metabolites in a model has a stronger influence on the performance than the modeling algorithm.

Table 4 Prediction performance with metabolites selection for 75 BC samples with 189 metabolites by multiple machine learning methods.

Classifier	Analysis platform	Sensitivity	Specificity	Accuracy	AUC (area under curve)
Decision tree	Rapid miner version 5.0	80.00% +/- 10.95%	40.00% +/- 17.89%	72.76% +/- 7.87%	0.848 +/- 0.099
Naïve Bayes (kernel)	Rapid miner version 5.0	84.00% +/- 10.20%	28.00% +/- 16.00%	70.04% +/- 5.97%	0.736 +/- 0.221
KNN	Rapid miner version 5.0	68.00% +/- 17.20%	60.00% +/- 12.65%	76.94% +/- 6.70%	0.880 +/- 0.067
Neural Network	Rapid miner version 5.0	86.00% +/- 13.56%	48.00% +/- 24.00%	77.11% +/- 9.24%	0.688 +/- 0.181
SVM	LibSVM Version 3.0	15.32% +/- 1.59%	92.49% +/- 5.83%	66.68% +/- 2.43%	N.A.

4.3.2 The predictive performance of identified metabolites biomarkers.

The predictive performance of models developed using the identified biomarkers are given in **Tables 5** and **7**. For the bladder cancer dataset with 189 metabolites, the performance accuracies (Q) were in range of 81.98% - 83.92% and the numbers of selected metabolites were in the range of 27 - 35. For the dataset with 398 metabolites, the corresponding values are 97.12% - 99.20% and 31 - 55 respectively. The prediction performance of the dataset with 398 metabolites outperformed the dataset with 189 metabolites. Furthermore, analysis of sensitivity (how well cancer patients can be detected) and specificity (how well controls can be detected) suggested the dataset with 398 metabolites had a better balance between sensitivity and specificity.

The results also show a good stability in the overall accuracy. For example, in **Table 7**, the difference between the different trials is less than 2.1%. This is mainly due to two reasons. Firstly, the globally optimized parameters were determined using grid search and thus the best classification accuracy over multi-time modeling and testing steps can be found. Secondly, the additional metabolites ranking evaluation on top of the normal RFE procedure prevents the selection of less important metabolites.

Table 5 Overall prediction accuracies of 20 times SVM-RFE selection for 75 BC samples with 189 metabolites.

Sampling set	Selected metabolite number	Overall performance in 500 testing datasets		
		Sensitivity	Specificity	Q
1	32	58.61%	95.43%	83.11%
2	35	57.94%	96.13%	83.32%
3	34	59.21%	95.68%	83.41%
4	28	58.01%	96.34%	83.50%
5	33	56.92%	95.67%	82.45%
6	27	57.39%	96.07%	83.12%
7	27	58.97%	96.12%	83.68%
8	29	57.08%	95.43%	82.50%
9	27	56.97%	96.39%	83.16%
10	30	56.23%	95.71%	82.37%
11	33	57.55%	95.42%	82.70%
12	28	58.25%	96.07%	83.40%
13	32	58.50%	95.85%	83.24%
14	34	56.59%	94.81%	81.98%
15	28	57.58%	96.42%	83.42%
16	33	60.17%	95.52%	83.67%
17	29	60.67%	95.69%	83.92%
18	31	57.73%	95.95%	83.11%
19	28	57.65%	96.02%	83.15%
20	27	57.12%	96.47%	83.17%

Table 6 Selected metabolites list for 75 BC samples with 189 metabolites.

Sampling set	Selected metabolite number	Metabolite ID lists
1	32	3, 13, 15, 17, 23, 38, 41, 47, 53, 54, 66, 81, 84, 85, 86, 90, 94, 102, 108, 120, 125, 144, 156, 165, 166, 172, 175, 176, 181, 182, 183, 186
2	35	3, 13, 14, 15, 17, 23, 25, 26, 38, 41, 47, 53, 54, 66, 81, 84, 85, 86, 90, 94, 102, 120, 125, 126, 144, 156, 165, 166, 172, 175, 176, 181, 182, 183, 186
3	34	3, 13, 14, 15, 17, 23, 25, 26, 38, 41, 47, 53, 54, 66, 81, 84, 85, 86, 90, 94, 102, 120, 125, 126, 144, 156, 165, 166, 175, 176, 181, 182, 183, 186
4	28	15, 17, 23, 26, 38, 41, 47, 53, 54, 66, 81, 84, 85, 86, 94, 102, 120, 125, 144, 165, 166, 172, 175, 176, 181, 182, 183, 186
5	33	3, 15, 17, 23, 25, 26, 38, 41, 47, 53, 54, 66, 81, 84, 85, 86, 90, 94, 102, 108, 120, 125, 144, 156, 165, 166, 172, 175, 176, 181, 182, 183, 186
6	27	3, 14, 17, 23, 26, 38, 41, 47, 53, 54, 66, 81, 85, 86, 94, 102, 108, 120, 125, 144, 165, 166, 175, 176, 181, 182, 183
7	27	13, 15, 17, 23, 38, 41, 47, 54, 57, 66, 81, 85, 86, 90, 94, 102, 120, 125, 126, 144, 166, 175, 176, 181, 182, 183, 186
8	29	13, 17, 23, 36, 38, 47, 53, 54, 57, 66, 81, 85, 86, 90, 94, 102, 119, 120, 125, 126, 144, 166, 172, 175, 176, 181, 182, 183, 186
9	27	10, 13, 17, 23, 38, 41, 47, 53, 54, 66, 81, 85, 86, 90, 94, 102, 120, 125, 126, 144, 165, 166, 175, 176, 181, 182, 186
10	30	13, 17, 23, 36, 38, 41, 47, 53, 54, 57, 66, 81, 85, 86, 90, 94, 102, 120, 125, 126, 144, 165, 166, 172, 175, 176, 181, 182, 183, 186
11	33	3, 15, 17, 23, 25, 26, 38, 41, 47, 53, 54, 66, 81, 84, 85, 86, 90, 94, 102, 108, 120, 125, 144, 156, 165, 166, 172, 175, 176, 181, 182, 183, 186
12	28	13, 15, 17, 23, 26, 38, 41, 47, 53, 54, 66, 81, 85, 86, 90, 94, 102, 120, 125, 126, 144, 166, 175, 176, 181, 182, 183, 186
13	32	3, 14, 15, 17, 23, 25, 26, 38, 41, 47, 66, 76, 81, 84, 85, 86, 90, 94, 102, 120, 125, 126, 144, 156, 165, 166, 175, 176, 181, 182, 183, 186
14	34	3, 13, 15, 17, 23, 25, 38, 41, 47, 53, 54, 66, 71, 81, 84, 85, 86, 90, 94, 102, 108, 119, 120, 125, 144, 156, 166, 172, 175, 176, 181, 182, 183, 186
15	28	3, 13, 17, 23, 38, 41, 47, 53, 54, 66, 81, 85, 86, 94, 102, 120, 125, 126, 144, 165, 166, 172, 175, 176, 181, 182, 183, 186
16	33	3, 13, 15, 17, 23, 25, 26, 38, 41, 47, 54, 66, 81, 84, 85, 86, 90, 94, 102, 120, 125, 126, 144, 156, 165, 166, 172, 175, 176, 181, 182, 183, 186
17	29	3, 4, 17, 23, 26, 36, 38, 47, 54, 66, 81, 84, 85, 86, 90, 94, 102, 120, 125, 126, 144, 165, 166, 175, 176, 181, 182, 183, 186
18	31	3, 13, 14, 17, 23, 25, 26, 38, 41, 47, 53, 54, 66, 81, 85, 86, 94, 102, 120, 125, 126, 144, 165, 166, 172, 175, 176, 181, 182, 183, 186
19	28	10, 13, 15, 17, 23, 25, 38, 47, 54, 66, 81, 84, 85, 86, 90, 102, 119, 120, 125, 126, 144, 165, 166, 175, 176, 181, 182, 186
20	27	3, 17, 23, 26, 38, 41, 47, 53, 54, 66, 81, 85, 86, 94, 102, 120, 125, 126, 144, 165, 166, 175, 176, 181, 182, 183, 186

Table 7 Overall prediction accuracies of 20 times SVM-RFE selection for 75 BC samples with 398 metabolites.

Sampling set	Selected metabolite number	Overall performance in 500 testing datasets		
		Sensitivity	Specificity	Q
1	36	98.61%	99.42%	98.88%
2	34	98.88%	99.86%	99.20%
3	35	97.60%	99.60%	98.26%
4	55	98.26%	99.76%	98.76%
5	34	98.33%	99.75%	98.80%
6	37	98.03%	99.83%	98.62%
7	33	97.53%	99.30%	98.11%
8	47	96.67%	99.53%	97.62%
9	38	97.08%	99.88%	97.99%
10	36	95.85%	99.68%	97.12%
11	54	98.14%	99.83%	98.70%
12	36	97.87%	99.75%	98.48%
13	39	96.30%	99.56%	97.37%
14	43	98.48%	99.63%	98.86%
15	31	97.87%	99.71%	98.47%
16	46	97.23%	99.36%	97.92%
17	52	97.01%	99.86%	97.94%
18	55	98.34%	99.97%	98.87%
19	32	98.74%	99.90%	99.12%
20	37	97.46%	98.87%	97.92%

Table 8 Selected metabolites list for 75 BC samples with 398 metabolites.

Sampling set	Selected metabolite number	Metabolite ID lists
1	36	68, 72, 104, 105, 106, 107, 108, 116, 127, 149, 150, 152, 163, 180, 188, 193, 217, 230, 249, 250, 256, 262, 266, 284, 287, 288, 299, 302, 304, 316, 350, 352, 354, 365, 371, 382
2	34	61, 68, 72, 104, 105, 106, 107, 108, 115, 116, 127, 149, 150, 152, 163, 180, 188, 217, 230, 249, 250, 256, 266, 284, 287, 288, 302, 304, 316, 350, 352, 365, 371, 382
3	35	46, 61, 68, 72, 104, 105, 106, 107, 108, 116, 127, 149, 150, 152, 179, 180, 188, 217, 230, 249, 250, 256, 266, 284, 287, 288, 302, 304, 316, 350, 352, 365, 371, 382, 388
4	55	46, 61, 68, 72, 75, 97, 104, 105, 106, 107, 108, 116, 124, 127, 132, 133, 135, 149, 150, 152, 163, 179, 180, 184, 188, 210, 217, 218, 230, 234, 249, 250, 252, 256, 262, 266, 284, 287, 288, 289, 291, 299, 302, 304, 316, 350, 352, 354, 360, 363, 365, 368, 371, 382, 388
5	34	24, 61, 72, 104, 105, 106, 107, 108, 116, 127, 149, 150, 152, 179, 180, 188, 217, 230, 249, 250, 256, 266, 287, 288, 291, 302, 304, 316, 350, 352, 354, 365, 371, 382
6	37	46, 61, 68, 72, 97, 104, 105, 106, 107, 108, 115, 116, 127, 149, 150, 152, 163, 180, 188, 217, 230, 249, 250, 256, 266, 284, 287, 288, 299, 302, 304, 316, 350, 352, 365, 371, 382
7	33	46, 61, 68, 72, 105, 106, 107, 108, 116, 127, 149, 150, 152, 179, 180, 188, 217, 230, 249, 250, 256, 266, 284, 287, 288, 302, 304, 316, 350, 352, 365, 371, 382
8	47	3, 46, 61, 68, 72, 75, 97, 104, 105, 106, 107, 108, 116, 124, 127, 133, 149, 150, 152, 163, 179, 188, 217, 230, 249, 250, 252, 256, 262, 266, 284, 287, 288, 291, 299, 302, 304, 316, 350, 352, 354, 360, 363, 365, 371, 382, 388
9	38	24, 46, 61, 68, 72, 97, 104, 105, 106, 107, 108, 116, 127, 149, 150, 152, 163, 180, 188, 217, 230, 249, 250, 256, 262, 266, 284, 287, 288, 299, 302, 304, 316, 350, 352, 365, 371, 382
10	36	46, 61, 68, 72, 97, 104, 105, 106, 107, 108, 116, 127, 149, 152, 180, 188, 210, 217, 230, 249, 250, 256, 266, 284, 287, 288, 291, 299, 302, 304, 316, 350, 352, 365, 371, 382
11	54	46, 61, 68, 72, 75, 89, 97, 104, 105, 106, 107, 108, 115, 116, 124, 127, 132, 133, 149, 150, 152, 179, 180, 184, 188, 202, 217, 218, 230, 234, 249, 250, 252, 256, 262, 266, 284, 287, 288, 292, 294, 299, 302, 304, 316, 350, 352, 360, 363, 365, 368, 371, 382, 388
12	36	61, 72, 75, 104, 105, 106, 107, 108, 115, 116, 127, 149, 150, 152, 179, 180, 188, 217, 230, 249, 250, 256, 266, 287, 288, 291, 302, 316, 350, 352, 360, 363, 365, 371, 382, 388
13	39	46, 61, 68, 72, 75, 97, 104, 105, 106, 107, 108, 116, 124, 127, 149, 152, 179, 180, 188, 217, 230, 249, 250, 256, 262, 266, 287, 288, 302, 316, 350, 352, 360, 363, 365, 368, 371, 382, 388
14	43	46, 61, 68, 72, 75, 97, 104, 105, 106, 107, 108, 116, 124, 127, 149, 150, 152, 163, 179, 180, 188, 217, 230, 249, 250, 256, 262, 266, 287, 288, 291, 299, 302, 304, 316, 350, 352, 354, 360, 363, 365, 371, 382
15	31	24, 61, 72, 104, 105, 106, 107, 116, 127, 149, 150, 152, 179, 180, 188, 217, 230, 249, 250, 256, 266, 287, 288, 302, 316, 350, 352, 354, 363, 371, 382
16	46	46, 61, 68, 72, 75, 89, 97, 104, 105, 106, 107, 108, 115, 116, 127, 133, 149, 150, 152, 179, 184, 188, 217, 230, 234, 249, 250, 256, 262, 266, 284, 287, 288, 294, 299, 302, 304, 316, 350, 352, 360, 363, 365, 368, 371, 382
17	52	46, 61, 68, 72, 84, 89, 97, 104, 105, 106, 107, 108, 115, 116, 124, 127, 133, 148, 149, 150, 152, 163, 179, 180, 188, 202, 217, 230, 249, 250, 252, 256, 262, 266, 284, 287, 288, 291, 292, 299, 302, 304, 316, 350, 352, 354, 360, 363, 365, 371, 382, 388
18	55	3, 46, 61, 68, 72, 75, 84, 89, 97, 104, 105, 106, 107, 108, 115, 116, 127, 132, 133, 149, 150, 152, 163, 179, 180, 188, 193, 202, 217, 218, 230, 234, 249, 250, 256, 262, 266, 284, 287, 288, 291, 294, 299, 302, 304, 316, 350, 352, 360, 363, 365, 371, 378, 382, 388
19	32	61, 72, 104, 105, 106, 107, 108, 115, 116, 127, 149, 150, 152, 180, 188, 217, 228, 230, 249, 250, 256, 266, 284, 287, 288, 302, 304, 316, 350, 352, 371, 382
20	37	46, 61, 68, 72, 104, 105, 106, 107, 108, 116, 124, 127, 149, 150, 152, 179, 180, 188, 217, 230, 249, 250, 256, 262, 266, 284, 287, 288, 302, 304, 316, 350, 352, 365, 371, 382, 388

4.3.3. The list of selected metabolite biomarkers

Tables 6 and **8** show the ID list of the selected metabolite biomarkers for two datasets. For the dataset with 189 metabolites, 27-35 metabolites were identified as the biomarkers for the bladder cancer. The median number of chosen biomarkers was 29 and the stability was also adequate enough. For the dataset with 398 metabolites, 31-55 biomarkers were chosen. Furthermore, the IDs of metabolites chosen by each time are similar. 31 metabolites were identified in at least 16 out of the 20 experiments. The ID and name of these metabolites are listed in **Table 9**.

To further analyze the biological meaning of these selected biomarkers, it is necessary to understand their functions in metabolic pathway network and the relationship between these metabolites and the mechanism of bladder cancer. Several steps need to be performed for such purpose. Firstly, the structures of the selected metabolites should be derived from their chemical compound names, which are illustrated in Table 10. Secondly, determine the chemical and biological information about this compound. It can be achieved by querying online chemical compounds resources such as PubChem and ChEMBL, as well as analyzing designed experiments. Once the chemical and biological properties of these compounds are clear, the next steps is to identify the roles of these marker metabolites in related metabolic pathways by building the model of pathway networks for them. These steps will be gradually accomplished in further studies.

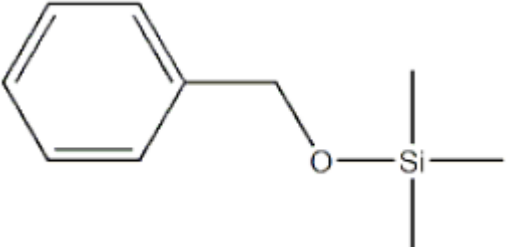
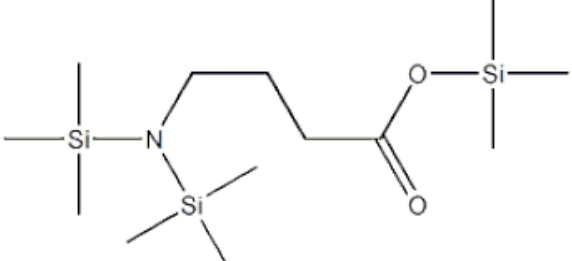
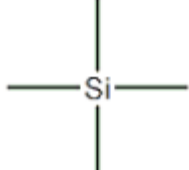
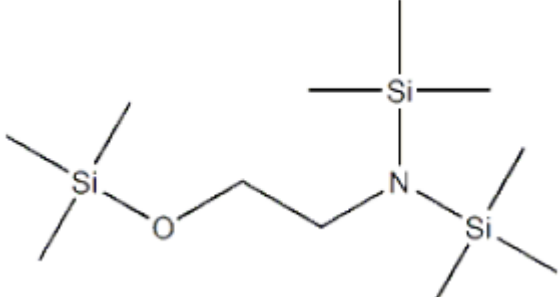
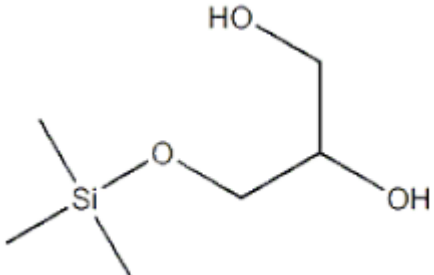

Table 9 List of 31 Selected metabolites (repeated rate > 80%) for 75 BC samples with 398 metabolites

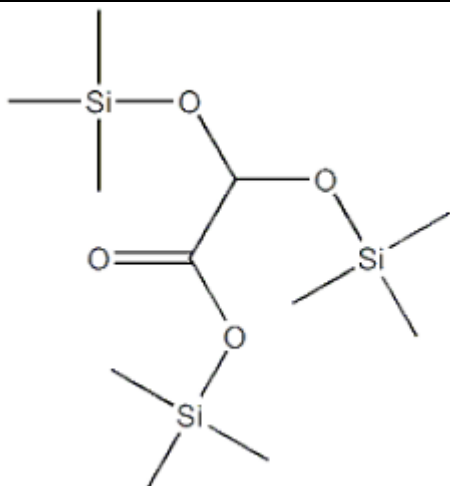
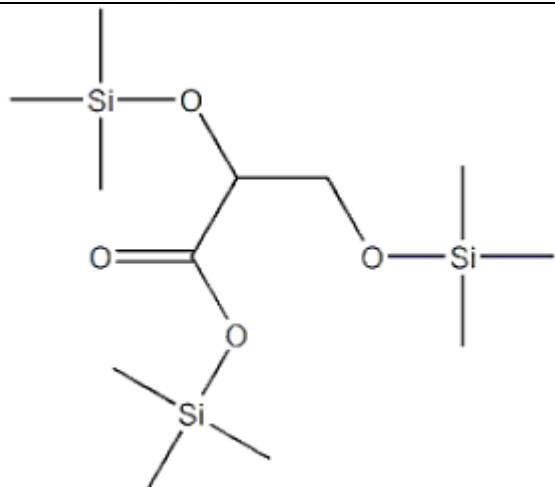
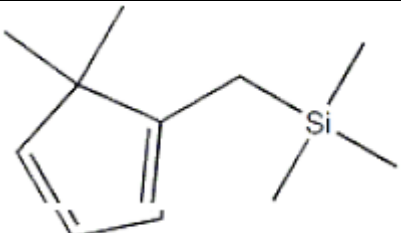
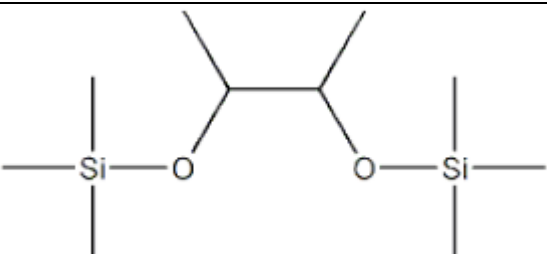
ID of selected metabolite biomarker	Name of selected metabolite biomarker
61	Silane, trimethyl(phenylmethoxy)
68	Butanoic acid, 4-[bis(trimethylsilyl)amino]-, trimethylsilyl ester
72	Silane, tetramethyl-
104	Silanamine, 1,1,1-trimethyl-N-(trimethylsilyl)-N-[2-[(trimethylsilyl)oxy]ethyl]-
105	Trimethylsilyl ether of glycerol
106	Tetradecane
107	Ethyl aminomalonate bis-(trimethylsilyl)- deriv.
116	Acetic acid, bis[(trimethylsilyl)oxyl]-, trimethylsilyl ester
127	Propanoic acid, 2,3-bis[(trimethylsilyl)oxy]-, trimethylsilyl ester
149	1,3-Cyclopentadiene, 5,5-dimethyl-1-(trimethylsilylmethyl)-
150	Butane, 2,3-bis(trimethylsiloxy)-
152	N,O,O-Tris(trimethylsilyl)-L-threonine
179	Glycine, N-formyl-N-(trimethylsilyl)-, trimethylsilyl ester
180	Propanoic acid, 3-[bis(trimethylsilyl)amino]-2-methyl-, trimethylsilyl ester
188	cis-4-Trimethylsilyloxy-cyclohexyl(trimethylsilyl)carboxylate
217	Pentanedioic acid, 3-methyl-3-[(trimethylsilyl)oxy]-, bis(trimethylsilyl) ester
230	3-Ketovaleric acid, bis(trimethylsilyl)-
249	Analyte 473 (1)
250	Analyte 473 (2)
256	Mannose, 6-deoxy-2,3,4,5-tetrakis-O-(trimethylsilyl)-, L-

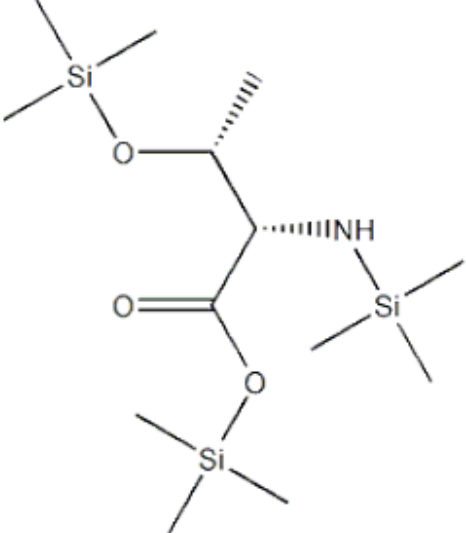
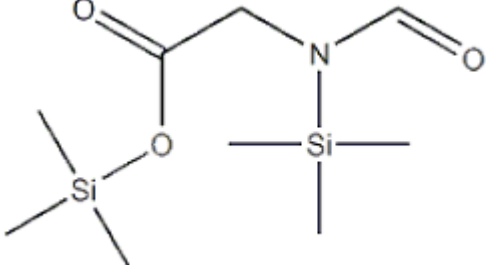
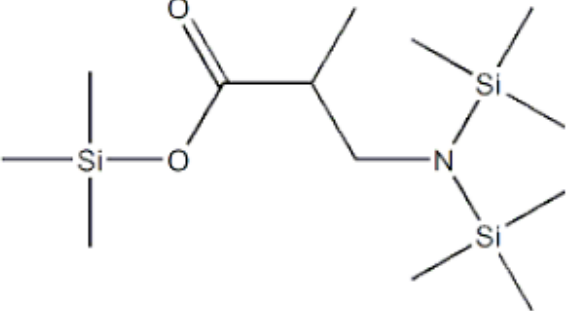
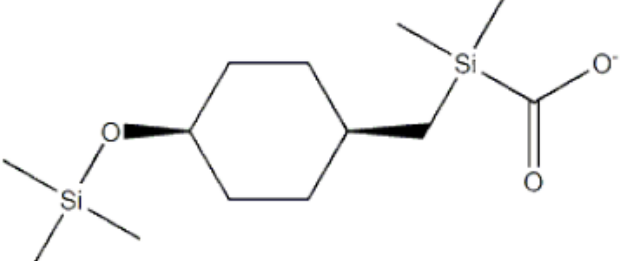
Continued Table 9

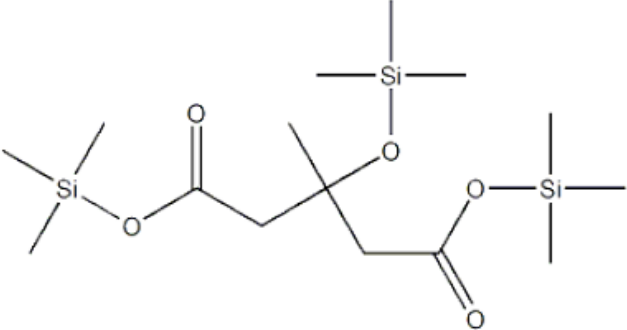
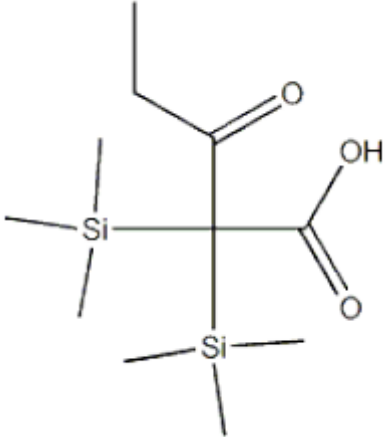
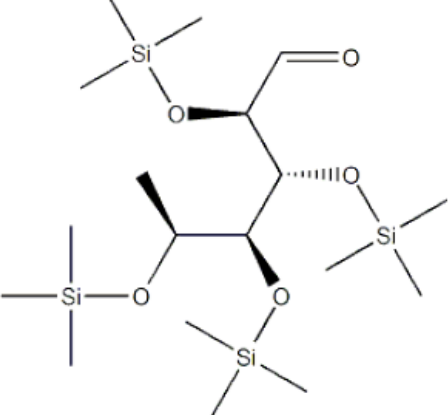
ID of selected metabolite biomarker	Name of selected metabolite biomarker
266	Ribitol, 1,2,3,4,5-pentakis-O-(trimethylsilyl)-
284	Heptasiloxane, 1,1,3,3,5,5,7,7,9,9,11,11,13,13-tetradecamethyl-
287	Tyrosine, O-trimethylsilyl-, trimethylsilyl ester
288	Glycine, N-benzoyl-, trimethylsilyl ester
302	D-Galactose-MOX-TMS-peak2
304	Acrylic acid, 2,3-bis[(trimethylsilyl)oxy]-, trimethylsilyl ester
316	D-Gluconic acid, 2,3,4,5,6-pentakis-O-(trimethylsilyl)-, trimethylsilyl ester
350	Mercaptoacetic acid, bis(trimethylsilyl)-
352	Analyte 1023
371	Analyte 799
382	2-Furanacetaldehyde, tetrahydro- α 3,4,5-tetrakis[(trimethylsilyl)oxy]-

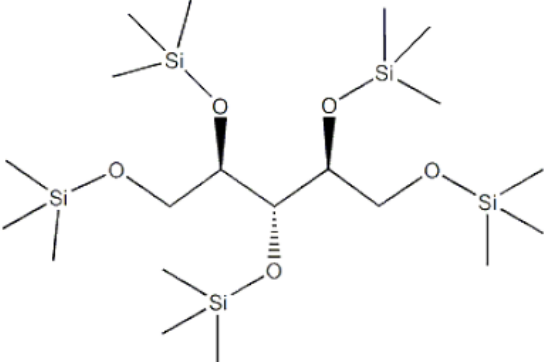
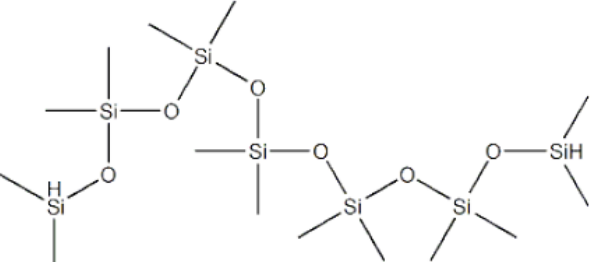
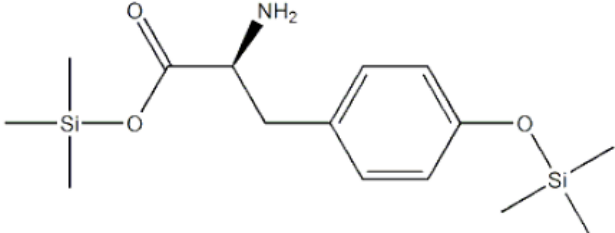
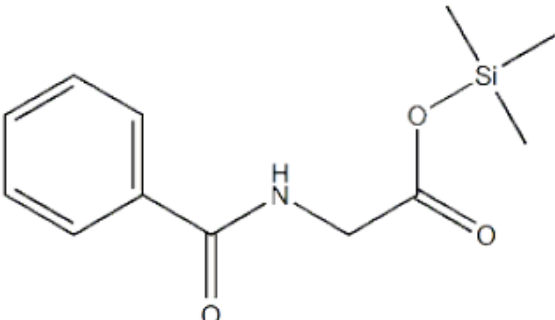
Table 10 List of structures of the 31 Selected metabolites (repeated rate > 80%)

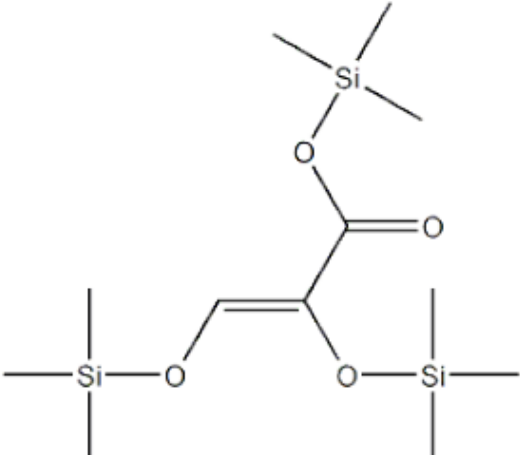
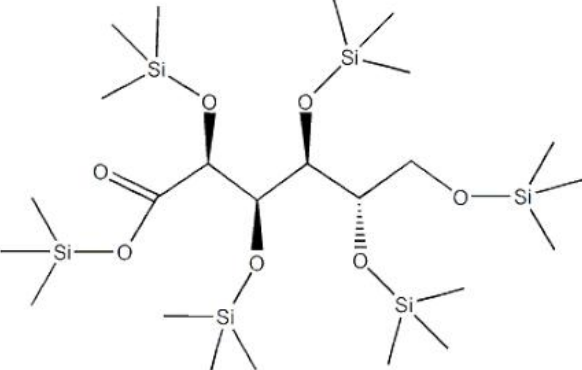
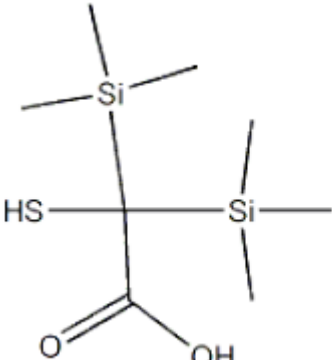
ID of selected metabolite biomarker	Name of selected metabolite biomarker	Structure of selected metabolites biomarker
61	Silane, trimethyl(phenylmethoxy)	
68	Butanoic acid, 4-[bis(trimethylsilyl)amino]-, trimethylsilyl ester	
72	Silane, tetramethyl-	
104	Silanamine, 1,1,1-trimethyl-N-(trimethylsilyl)-N-[2-[(trimethylsilyl)oxy]ethyl]-	
105	Trimethylsilyl ether of glycerol	
106	Tetradecane	

107	Ethyl aminomalonate bis-(trimethylsilyl)- deriv.	N.A.
116	Acetic acid, bis[(trimethylsilyl)oxyl]-, trimethylsilyl ester	
127	Propanoic acid, 2,3-bis[(trimethylsilyl)oxy]-, trimethylsilyl ester	
149	1,3-Cyclopentadiene, 5,5-dimethyl-1-(trimethylsilylmethyl)-	
150	Butane, 2,3-bis(trimethylsiloxy)-	

152	N,O,O-Tris(trimethylsilyl)-L-threonine	
179	Glycine, N-formyl-N-(trimethylsilyl)-, trimethylsilyl ester	
180	Propanoic acid, 3-[bis(trimethylsilyl)amino]-2-methyl-, trimethylsilyl ester	
188	cis-4-Trimethylsilyloxy-cyclohexyl(trimethylsilyl)carboxylate	

217	Pentanedioic acid, 3-methyl-3-[(trimethylsilyl)oxy]-, bis(trimethylsilyl) ester	
230	3-Ketovaleric acid, bis(trimethylsilyl)-	
249	Analyte 473	N.A.
250	Analyte 473	N.A.
256	Mannose, 6-deoxy-2,3,4,5-tetrakis-O-(trimethylsilyl)-, L-	

266	Ribitol, 1,2,3,4,5-pentakis-O-(trimethylsilyl)-	
284	Heptasiloxane, 1,1,3,3,5,5,7,7,9,9,11,11,13, 13-tetradecamethyl-	
287	Tyrosine, O-trimethylsilyl-, trimethylsilyl ester	
288	Glycine, N-benzoyl-, trimethylsilyl ester	
302	D-Galactose-MOX-TMS-peak2	N.A.

304	Acrylic acid, 2,3-bis[(trimethylsilyl)oxy]-, trimethylsilyl ester	 <p>The structure shows an acrylic acid derivative where the hydroxyl group is replaced by a trimethylsilyloxy group. The 2-position of the acrylic acid backbone is also substituted with a trimethylsilyloxy group. The silicon atoms are represented as 'Si' with three generic bonds.</p>
316	D-Gluconic acid, 2,3,4,5,6-pentakis-O-(trimethylsilyl)-, trimethylsilyl ester	 <p>The structure shows a gluconic acid derivative where all five hydroxyl groups (at positions 2, 3, 4, 5, and 6) are replaced by trimethylsilyloxy groups. The silicon atoms are represented as 'Si' with three generic bonds.</p>
350	Mercaptoacetic acid, bis(trimethylsilyl)-	 <p>The structure shows a mercaptoacetic acid derivative where both the hydroxyl group and the thiol group are replaced by trimethylsilyloxy groups. The silicon atoms are represented as 'Si' with three generic bonds.</p>
352	Analyte 1023	N.A.
371	Analyte 799	N.A.
382	2-Furanacetaldehyde, tetrahydro- 3,4,5-tetrakis[(trimethylsilyl)oxy]-	N.A.

4.3.4. Performance evaluation with multiple classifiers

In order to evaluate the performance of the selected biomarkers, multiple classification models had been built to re-train the datasets with the selected metabolites. The performance of these models can be found from the **Table 11**. As shown in **Table 11**, overall accuracies of all classifiers were above 79%, in particular, the accuracy of Naïve Bayes (kernel) and the accuracy of SVM were above 90%. Sensitivity values of all classifiers were above 92%, except for decision tree classifier. Specificity values of these classifiers were not as high as the sensitivity values. However, all of them were above 75%, except for KNN classifier. The performance of these classifiers suggests that the selected metabolites were representative of the original data. Moreover, these selected metabolites can be used as the biomarkers of the original dataset for further analysis.

Table 11 List of evaluation performance of the 31 Selected metabolites (repeated rate > 80%)

Classifier	Analysis Platform	Sensitivity	Specificity	Accuracy	AUC (area under curve)
Decision Tree	Rapid miner version 5.0	75.00% +/- 19.49%	81.47% +/- 4.52%	79.33% +/-8.02%	0.952 +/-0.046
Naïve Bayes (kernel)	Rapid miner version 5.0	96.00% +/- 8.00%	87.96% +/- 9.81%	90.57% +/-6.76%	0.964 +/-0.037
KNN	Rapid miner version 5.0	100.00% +/- 0.00%	71.47% +/- 11.40%	80.95% +/-7.52%	0.983 +/-0.012
Neural Network	Rapid miner version 5.0	92.00% +/- 9.080%	75.07% +/- 8.72%	80.76% +/-6.68%	0.912 +/-0.055
SVM	LibSVM	100.00% +/- 0.00%	98.00% +/- 4.00%	98.67% +/-2.67%	0.996 +/-0.008

5. CONCLUSION AND FUTURE WORK

Accurate identification of peptides binding to specific MHC molecules is fundamental for understanding the mechanisms of both humoral and adaptive immunity, and important for developing effective epitope-based vaccines for immunotherapy of infectious, autoimmune, and cancer diseases. Experimental methods for identifying MHC binding peptides are costly and time-consuming. In-silico methods have thus been explored for facilitating epitope screening to complement laboratory experiments in reducing the cost and time for vaccine design. In this study, we showed that MHC binding prediction methods were able to predict MHC binding peptides with high accuracy. The method developed here can be used to identify promising candidate epitopes for further experimental verification.

In the MHC binding peptide prediction study, the performances of prediction systems were compared between the original datasets and datasets with the generated non-binding peptides. It was found that the separated datasets by alleles with the generated non-binding peptides works much more effectively than the original dataset. The positive accuracies showing the percentage of the correctly predicted known binding peptides have a high level of precision. Based on the principle of the SVM algorithm, SVM shows good performance when the samples could sufficiently represent the whole space. Therefore, the diversity and representative ability of datasets are the major concerns of SVM prediction system. Although certain extent of evaluation have been made for the SVM prediction system, further validation is still necessary. Independent evaluations by new experimental samples and screening with specific genome could be appropriate ways

to validate this MHC-binding prediction system.

Metabonomics investigation on urine samples of bladder cancer patients could lead to an overview of the metabolic disturbances taking place in the patients, which is essential for the understanding of physiological progress of bladder cancer. This study demonstrates a feasible way of metabonomics research by selecting metabolites markers for specific disease. GC/TOF mass spectrometry is the major analytical techniques, which played important role in deriving data from biological sample, the feature selection algorithm; SVM-RFE has been applied to select the discriminative and meaningful metabolites from the metabolic profiling data. The result of feature selection achieved an average classification accuracy rate of 98.35%, which indicated the metabolites selection by SVM-RFE could discriminate well among and are biologically meaningful for metabonomics studies.

To further evaluate the identified metabolite biomarkers of bladder cancer diagnosis, several steps should be performed. Firstly, because the significant improvement of performance accuracy was achieved when SVM-RFE metabolites selection procedure was applied, and when comparing with other machine learning algorithms without metabolites selection, SVM did not show obvious advantage, we believe that as an effective way to select the appropriated feature, recursive feature elimination can be combined with the other machine learning methods, such as neural network, genetic algorithm and k nearest neighbor, to develop several new RFE procedures.

Secondly, we can further analysis the selected 31 metabolite biomarkers for bladder cancer by unsupervised algorithms, such as PCA. Since these biomarkers showed high

accuracies when tested by SVM classifier, they should show good distinction abilities when analyzed using PCA. The PCA score plot and loading plot can be drawn to determine how well these biomarkers can separate the bladder cancer samples and non-bladder cancer controls.

Thirdly, we can further interpret the biological relations of identified biomarkers with bladder cancer. The metabolite pathway of bladder cancer could be complicated and related to the physiological and biochemical properties of certain cells, organs and entire human system. Thus, it is necessary to investigate roles of biomarkers and highlighted metabolites in whole metabolic pathway networks, for better understanding of the pathway network profile and even improving the network modeling. Currently, there are several metabolic pathway resources for further investigation of metabonomics studies and reconstructing metabolic models, such as Kyoto Encyclopedia of Genes and Genomes (KEGG), BioCyc, EcoCyc, and MetaCyc

Fourthly, since our SVM-RFE method exhibited good performances for metabolites selection of bladder cancer, we can investigate the metabonomics dataset of other types of cancers, such as the breast cancer, colon cancer and lung cancer, with our metabolites selection methods.

BIBLIOGRAPHY

1. Vapnik V and Chervonenkis A, *A note on one class of perceptrons*. Automation and Remote Control, 1964. **25**.
2. Vapnik V and Lerner A, *Pattern recognition using generalized portrait method*. Automation and Remote Control, 1963. **24**.
3. Kawaji H and Hayashizaki Y, *Genome annotation*. Methods Mol Biol, 2008. **452**: p. 125-39.
4. Theodosiou T, Angelis L, Vakali A, et al., *Gene functional annotation by statistical analysis of biomedical articles*. Int J Med Inform, 2007. **76**(8): p. 601-13.
5. Vinayagam A, Konig R, Moormann J, et al., *Applying Support Vector Machines for Gene Ontology based gene function prediction*. BMC Bioinformatics, 2004. **5**: p. 116.
6. Schweikert G, Zien A, Zeller G, et al., *mGene: accurate SVM-based gene finding with an application to nematode genomes*. Genome Res, 2009. **19**(11): p. 2133-43.
7. Chen Y, Li Z, Wang X, et al., *Predicting gene function using few positive examples and unlabeled ones*. BMC Genomics, 2010. **11 Suppl 2**: p. S11.
8. Vinayagam A, del Val C, Schubert F, et al., *GOPET: a tool for automated predictions of Gene Ontology terms*. BMC Bioinformatics, 2006. **7**: p. 161.
9. Manolio TA, *Genomewide association studies and assessment of the risk of disease*. N Engl J Med, 2010. **363**(2): p. 166-76.
10. Sladek R, Rocheleau G, Rung J, et al., *A genome-wide association study identifies novel risk loci for type 2 diabetes*. Nature, 2007. **445**(7130): p. 881-5.
11. Listgarten J, Damaraju S, Poulin B, et al., *Predictive models for breast cancer susceptibility from multiple single nucleotide polymorphisms*. Clinical Cancer Research, 2004. **10**(8): p. 2725-2737.
12. Waddell M, Page D, Zhan F, et al. *Predicting Cancer Susceptibility from Single-Nucleotide Polymorphism Data: A Case Study in Multiple Myeloma*. in *BIOKDD '05*. 2005. Chicago, IL, USA.
13. Uhm S, Kim DH, Ko YW, et al., *A study on application of single nucleotide polymorphism and machine learning techniques to diagnosis of chronic hepatitis*. Expert Systems, 2009. **26**: p. 60-69.
14. Ban HJ, Heo JY, Oh KS, et al., *Identification of Type 2 Diabetes-associated combination of SNPs using Support Vector Machine*. BMC Genetics, 2010. **11**: p. -.
15. Rogers S, Girolami M, Kolch W, et al., *Investigating the correspondence between transcriptomic and proteomic expression profiles using coupled cluster models*. Bioinformatics, 2008. **24**(24): p. 2894-900.
16. Dhingra V, Gupta M, Andacht T, et al., *New frontiers in proteomics research: a perspective*. Int J Pharm, 2005. **299**(1-2): p. 1-18.

17. Bork P, Dandekar T, Diaz-Lazcoz Y, et al., *Predicting function: from genes to genomes and back*. J Mol Biol, 1998. **283**(4): p. 707-25.
18. Eisenberg D, Marcotte EM, Xenarios I, et al., *Protein function in the post-genomic era*. Nature, 2000. **405**(6788): p. 823-6.
19. Bock JR and Gough DA, *Predicting protein--protein interactions from primary structure*. Bioinformatics, 2001. **17**(5): p. 455-60.
20. Lo SL, Cai CZ, Chen YZ, et al., *Effect of training datasets on support vector machine prediction of protein-protein interactions*. Proteomics, 2005. **5**(4): p. 876-84.
21. Cai YD and Lin SL, *Support vector machines for predicting rRNA-, RNA-, and DNA-binding proteins from amino acid sequence*. Biochim Biophys Acta, 2003. **1648**(1-2): p. 127-33.
22. Cai CZ, Han LY, Ji ZL, et al., *Enzyme family classification by support vector machines*. Proteins, 2004. **55**(1): p. 66-76.
23. Cai YD and Doig AJ, *Prediction of Saccharomyces cerevisiae protein functional class from functional domain composition*. Bioinformatics, 2004. **20**(8): p. 1292-300.
24. Han LY, Cai CZ, Lo SL, et al., *Prediction of RNA-binding proteins from primary sequence by a support vector machine approach*. RNA, 2004. **10**(3): p. 355-68.
25. Dobson PD and Doig AJ, *Predicting enzyme class from protein structure without alignments*. J Mol Biol, 2005. **345**(1): p. 187-99.
26. Ben-Hur A and Noble WS, *Kernel methods for predicting protein-protein interactions*. Bioinformatics, 2005. **21 Suppl 1**: p. i38-46.
27. Bhasin M and Raghava GP, *Prediction of CTL epitopes using QM, SVM and ANN techniques*. Vaccine, 2004. **22**(23-24): p. 3195-204.
28. Bock JR and Gough DA, *Whole-proteome interaction mining*. Bioinformatics, 2003. **19**(1): p. 125-34.
29. Martin S, Roe D, and Faulon JL, *Predicting protein-protein interactions using signature products*. Bioinformatics, 2005. **21**(2): p. 218-26.
30. Xue Y, Yap CW, Sun LZ, et al., *Prediction of P-glycoprotein substrates by a support vector machine approach*. J Chem Inf Comput Sci, 2004. **44**(4): p. 1497-505.
31. Cai CZ, Han LY, Ji ZL, et al., *SVM-Prot: Web-based support vector machine software for functional classification of a protein from its primary sequence*. Nucleic Acids Res, 2003. **31**(13): p. 3692-7.
32. Cai YD and Chou KC, *Predicting enzyme subclass by functional domain composition and pseudo amino acid composition*. J Proteome Res, 2005. **4**(3): p. 967-71.
33. Lin HH, Han LY, Cai CZ, et al., *Prediction of transporter family from protein sequence by support vector machine approach*. Proteins, 2006. **62**(1): p. 218-31.
34. Saha S and Raghava GP, *AlgPred: prediction of allergenic proteins and mapping of IgE epitopes*. Nucleic Acids Res, 2006. **34**(Web Server issue): p. W202-9.
35. Cui J, Han LY, Li H, et al., *Computer prediction of allergen proteins from sequence-derived protein structural and physicochemical properties*. Mol Immunol, 2007. **44**(4): p. 514-20.

36. Smialowski P, Schmidt T, Cox J, et al., *Will my protein crystallize? A sequence-based predictor*. Proteins, 2006. **62**(2): p. 343-55.
37. Kumar M, Verma R, and Raghava GP, *Prediction of mitochondrial proteins using support vector machine and hidden Markov model*. J Biol Chem, 2006. **281**(9): p. 5357-63.
38. Bhasin M and Raghava GP, *GPCRpred: an SVM-based method for prediction of families and subfamilies of G-protein coupled receptors*. Nucleic Acids Res, 2004. **32**(Web Server issue): p. W383-9.
39. Guo YZ, Li M, Lu M, et al., *Classifying G protein-coupled receptors and nuclear receptors on the basis of protein power spectrum from fast Fourier transform*. Amino Acids, 2006. **30**(4): p. 397-402.
40. Yabuki Y, Muramatsu T, Hirokawa T, et al., *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): p. W148-53.
41. Bhasin M and Raghava GP, *Classification of nuclear receptors based on amino acid composition and dipeptide composition*. J Biol Chem, 2004. **279**(22): p. 23262-6.
42. Bhardwaj N, Langlois RE, Zhao G, et al., *Kernel-based machine learning protocol for predicting DNA-binding proteins*. Nucleic Acids Res, 2005. **33**(20): p. 6486-93.
43. Lin HH, Han LY, Zhang HL, et al., *Prediction of the functional class of lipid binding proteins from sequence-derived properties irrespective of sequence similarity*. J Lipid Res, 2006. **47**(4): p. 824-31.
44. Wang M, Yang J, Liu GP, et al., *Weighted-support vector machines for predicting membrane protein types based on pseudo-amino acid composition*. Protein Eng Des Sel, 2004. **17**(6): p. 509-16.
45. Huang N, Chen H, and Sun Z, *CTKPred: an SVM-based method for the prediction and classification of the cytokine superfamily*. Protein Eng Des Sel, 2005. **18**(8): p. 365-8.
46. Zhao Y, Pinilla C, Valmori D, et al., *Application of support vector machines for T-cell epitopes prediction*. Bioinformatics, 2003. **19**(15): p. 1978-84.
47. Donnes P and Elofsson A, *Prediction of MHC class I binding peptides, using SVMHC*. BMC Bioinformatics, 2002. **3**: p. 25.
48. Bhasin M and Raghava GP, *SVM based method for predicting HLA-DRB1*0401 binding peptides in an antigen sequence*. Bioinformatics, 2004. **20**(3): p. 421-3.
49. Goodacre R, Vaidyanathan S, Dunn WB, et al., *Metabolomics by numbers: acquiring and understanding global metabolite data*. Trends Biotechnol, 2004. **22**(5): p. 245-52.
50. Chen C, Gonzalez FJ, and Idle JR, *LC-MS-based metabolomics in drug metabolism*. Drug Metab Rev, 2007. **39**(2-3): p. 581-97.
51. Sreekumar A, Poisson LM, Rajendiran TM, et al., *Metabolomic profiles delineate potential role for sarcosine in prostate cancer progression*. Nature, 2009. **457**(7231): p. 910-4.

52. Yin P, Zhao X, Li Q, et al., *Metabonomics study of intestinal fistulas based on ultraperformance liquid chromatography coupled with Q-TOF mass spectrometry (UPLC/Q-TOF MS)*. J Proteome Res, 2006. **5**(9): p. 2135-43.
53. Patterson AD, Li H, Eichler GS, et al., *UPLC-ESI-TOFMS-based metabolomics and gene expression dynamics inspector self-organizing metabolomic maps as tools for understanding the cellular response to ionizing radiation*. Anal Chem, 2008. **80**(3): p. 665-74.
54. Guan W, Zhou M, Hampton CY, et al., *Ovarian cancer detection from metabolomic liquid chromatography/mass spectrometry data by support vector machines*. BMC Bioinformatics, 2009. **10**: p. 259.
55. Li L, Tang H, Wu Z, et al., *Data mining techniques for cancer detection using serum proteomic profiling*. Artif Intell Med, 2004. **32**(2): p. 71-83.
56. Rajapakse JC, Duan KB, and Yeo WK, *Proteomic cancer classification with mass spectrometry data*. Am J Pharmacogenomics, 2005. **5**(5): p. 281-92.
57. Yu JS, Ongarello S, Fiedler R, et al., *Ovarian cancer identification based on dimensionality reduction for high-throughput mass spectrometry data*. Bioinformatics, 2005. **21**(10): p. 2200-9.
58. Shen C, Breen TE, Dobrolecki LE, et al., *Comparison of computational algorithms for the classification of liver cancer using SELDI mass spectrometry: a case study*. Cancer Inform, 2007. **3**: p. 329-39.
59. Wu B, Abbott T, Fishman D, et al., *Comparison of statistical methods for classification of ovarian cancer using mass spectrometry data*. Bioinformatics, 2003. **19**(13): p. 1636-43.
60. Pham TV, van de Wiel MA, and Jimenez CR, *Support vector machine approach to separate control and breast cancer serum samples*. Stat Appl Genet Mol Biol, 2008. **7**(2): p. Article11.
61. Xue R, Lin Z, Deng C, et al., *A serum metabolomic investigation on hepatocellular carcinoma patients by chemical derivatization followed by gas chromatography/mass spectrometry*. Rapid Commun Mass Spectrom, 2008. **22**(19): p. 3061-8.
62. Osl M, Dreiseitl S, Pfeifer B, et al., *A new rule-based algorithm for identifying metabolic markers in prostate cancer using tandem mass spectrometry*. Bioinformatics, 2008. **24**(24): p. 2908-14.
63. Hennege C, Bullinger D, Fux R, et al., *Prediction of breast cancer by profiling of urinary RNA metabolites using Support Vector Machine-based feature selection*. BMC Cancer, 2009. **9**: p. 104.
64. Zhou B, Cheema AK, and Resson HW, *SVM-based spectral matching for metabolite identification*. Conf Proc IEEE Eng Med Biol Soc, 2010. **2010**: p. 756-9.
65. Veropoulos K, Campbell C, and Cristianini N. *Controlling the sensitivity of Support Vector machines*. in *International Joint Conference on Artificial Intelligence*. 1999. Stockholm, Sweden.
66. Brown MP, Grundy WN, Lin D, et al., *Knowledge-based analysis of microarray gene expression data by using support vector machines*. Proc Natl Acad Sci U S A, 2000. **97**(1): p. 262-7.

67. Karchin R, Karplus K, and Haussler D, *Classifying G-protein coupled receptors with support vector machines*. Bioinformatics, 2002. **18**(1): p. 147-59.
68. Wilkins MR, Gasteiger E, Bairoch A, et al., *Protein identification and analysis tools in the ExPASy server*. Methods Mol Biol, 1999. **112**: p. 531-52.
69. Xue Y, Li ZR, Yap CW, et al., *Effect of molecular descriptor feature selection in support vector machine classification of pharmacokinetic and toxicological properties of chemical agents*. J Chem Inf Comput Sci, 2004. **44**(5): p. 1630-8.
70. Al-Shahib A, Breitling R, and Gilbert D, *Feature selection and the class imbalance problem in predicting protein function from sequence*. Appl Bioinformatics, 2005. **4**(3): p. 195-203.
71. Al-Shahib A, Breitling R, and Gilbert D, *FrankSum: new feature selection method for protein function prediction*. Int J Neural Syst, 2005. **15**(4): p. 259-75.
72. Furlanello C, Serafini M, Merler S, et al., *An accelerated procedure for recursive feature ranking on microarray data*. Neural Netw, 2003. **16**(5-6): p. 641-8.
73. Yap CW and Chen YZ, *Prediction of cytochrome P450 3A4, 2D6, and 2C9 inhibitors and substrates by using support vector machines*. J Chem Inf Model, 2005. **45**(4): p. 982-92.
74. Cui J, Han LY, Lin HH, et al., *Prediction of MHC-binding peptides of flexible lengths from sequence-derived structural and physicochemical properties*. Molecular immunology, 2007. **44**(5): p. 866-77.
75. Jorissen RN and Gilson MK, *Virtual screening of molecular databases using a support vector machine*. Journal of chemical information and modeling, 2005. **45**(3): p. 549-61.
76. Glick M, Jenkins JL, Nettles JH, et al., *Enrichment of high-throughput screening data with increasing levels of noise using support vector machines, recursive partitioning, and laplacian-modified naive bayesian classifiers*. Journal of chemical information and modeling, 2006. **46**(1): p. 193-200.
77. Lepp Z, Kinoshita T, and Chuman H, *Screening for new antidepressant leads of multiple activities by support vector machines*. Journal of chemical information and modeling, 2006. **46**(1): p. 158-67.
78. Hert J, Willett P, Wilton DJ, et al., *New methods for ligand-based virtual screening: use of data fusion and machine learning to enhance the effectiveness of similarity searching*. Journal of chemical information and modeling, 2006. **46**(2): p. 462-70.
79. Yap CW and Chen YZ, *Quantitative Structure-Pharmacokinetic Relationships for drug distribution properties by using general regression neural network*. Journal of pharmaceutical sciences, 2005. **94**(1): p. 153-68.
80. Trotter MWB, Buxton BF, and Holden SB, *Support vector machines in combinatorial chemistry*. Meas. Control, 2001. **34**(8): p. 235-239.
81. Burbidge R, Trotter M, Buxton B, et al., *Drug design by machine learning: support vector machines for pharmaceutical data analysis*. Computers & chemistry, 2001. **26**(1): p. 5-14.
82. Czerminski R, Yasri A, and Hartsough D, *Use of support vector machine in pattern classification: Application to QSAR studies*. Quantitative Structure-Activity Relationships, 2001. **20**(3): p. 227-240.

83. Vapnik VN, *The Nature of Statistical Learning Theory*. 1995, New York: Springer-Verlag New York Inc.
84. Vapnik V, *The nature of statistical learning theory*. 1995, New York: Springer.
85. Cristianini N and Shawe-Taylor J, *An introduction to Support Vector Machines : and other kernel-based learning methods*. 2000, New York: Cambridge University Press.
86. Platt JC, *Sequential Minimal Optimization: A fast algorithm for training support vector machines*. Microsoft Research. Technical Report MSR-TR-98-14, 1998.
87. 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: p. 276-285.
88. BURGES CJC, *A Tutorial on Support Vector Machines for Pattern Recognition*. Data Mining and Knowledge Discovery, 1988. **2**: p. 121–167.
89. Aizerman MA, Braverman EM, and er LIR, *Theoretical foundations of the potential function method in pattern recognition and learning*. Automation and Remote Control, 1964. **25**: p. 821--837.
90. Courant R and Hilbert D, *Methods of Mathematical Physics*. 1989: John Wiley & Sons.
91. Baldi P, Brunak S, Chauvin Y, et al., *Assessing the accuracy of prediction algorithms for classification: an overview*. Bioinformatics, 2000. **16**(5): p. 412-24.
92. Cai CZ, Han LY, Ji ZL, et al., *SVM-Prot: Web-based support vector machine software for functional classification of a protein from its primary sequence*. Nucleic acids research, 2003. **31**(13): p. 3692-7.
93. Han LY, Cai CZ, Ji ZL, et al., *Predicting functional family of novel enzymes irrespective of sequence similarity: a statistical learning approach*. Nucleic acids research, 2004. **32**(21): p. 6437-44.
94. Honeyman MC, Brusica V, Stone NL, et al., *Neural network-based prediction of candidate T-cell epitopes*. Nature biotechnology, 1998. **16**(10): p. 966-9.
95. Nielsen M, Lundegaard C, Worning P, et al., *Improved prediction of MHC class I and class II epitopes using a novel Gibbs sampling approach*. Bioinformatics, 2004. **20**(9): p. 1388-97.