摘要

本計畫為「以生物資訊最佳化方法發展電腦輔助疫苗設計系統」之三年期研究計畫，主要目的是發展一套輔助免疫學家設計疫苗之電腦系統。本計畫之核心為發展數個能適用於探勘各種生物免疫現象之重要特徵因子的高性能大量參數最佳化之生物資訊演算法，並結合生物免疫知識庫發展成一整合型重要蛋白質功能分析和註記的應用系統。

在三年的執行期間，每年度都有階段性的成果，並互相相輔相成，而在計畫最後完成一套電腦輔助疫苗設計系統。各年度的成果簡述如下：

第一年：完成高效能重要特徵因子探勘演算法，是疫苗設計系統之演算法核心，並研究細胞毒性T細胞及MHC class I, II產生免疫性的抗原處理及表現路徑之相關致免疫性之重要特徵。

第二年：使用所開發之重要特徵因子探勘演算法探索輔助性T細胞免疫反應之重要特徵，完成輔助性T細胞致免疫性路徑研究及相關預測演算法。

第三年：發展一套富含免疫系統的特徵資料庫，並整合前兩年開發之演算法與重要特徵成果，完成電腦輔助疫苗設計系統。

目前已完成電腦輔助疫苗設計系統，而在三年的研究期間也陸續有相關論文發表，並將核心的重要特徵因素探勘的智慧型基因演算法應用到其他蛋白質功能預測領域，也都有良好的研究成果。此三年計畫中研究進度順利，已達預期目標，並有多篇期刊論文及研討會論文發表，成果豐碩。

關鍵詞：資料探勘、疫苗設計、演化式計算、基因演算法、參數最佳化、蛋白質功能預測、物化特性、免疫反應、分類器、特徵選取、因素分析

Abstract

This is a three-year project titled "Developing computer-aided vaccine design systems using bio-inspired optimization methods." The objectives are to develop computer-aided systems to help immunologists for vaccine designs. The core project is two-fold: 1) to develop various high-performance optimization algorithms for solving large-scale parameter optimization problems of bioinformatics to mine informative physicochemical properties of proteins from known experimental data; and 2) to integrate immune knowledge bases with the developed optimization methodologies to develop application systems of analyzing and annotating informative functions of proteins.

We have good research results to achieve the goal of each year. The individual projects of the three years focus on both study of vaccine designs and algorithm developments, which are fully cooperated. The achievements of each year are described below.

Year 1: Develop high-performance feature mining algorithms which are core algorithms of the vaccine design system, and study cytotoxic T lymphocyte related immune response and informative features related to MHC class I and II binding peptides and their pathways.

Year 2: Study the immune responses of T helper cell, including the infect pathways and features of T cell using the developed optimization algorithms of informative feature mining. A number of prediction algorithms and systems for HTL and CTL immune pathways and epitopes have been proposed.

Year 3: Develop an informative feature database of immune systems and establish the computer-aided vaccine design system by integrating the achievements of the first two years and.

The computer-aided vaccine design systems have been finished and the core high-performance optimization algorithm has been applied to a number of protein function prediction problems. The good achievements were published in several international conference and journal papers.

Keywords: Data Mining, Vaccine Design, Evolutionary Computation, Genetic Algorithm, Parameter Optimization, Protein Function Prediction, Physicochemical Properties, Immune Response, Classifier, Feature Selection, Factor Analysis.

Contents

摘要
Abstract
Contents III
1. Introduction.. 1
2. Objectives 2
3. Literature Review 3
3.1. Introduction of Immune 3
3.1.1. Immunogenicity of MHC class I binding peptides and the pathway.3
3.1.2. Immunogenicity of MHC class II binding peptides and the pathway4
3.2. Introduction of Bioinformatics Tools 6
3.2.1. Biological features 6
3.2.2. Feature selection methods 6
3.2.3. Classifiers 7
4. Methods. 10
4.1. Objectives of three-year project 10
4.2. First year 11
4.2.1. Establishment of core algorithm 11
4.2.2. Studying the immunogenicity of MHC class I 14
4.2.3. Prediction system of cytotoxic T lymphocyte 15
4.3. Second year 17
4.3.1. Developing the useable classifiers of immune reactions 17
4.3.2. Utilizing the varied biological features 18
4.3.3. Study of HTL immune pathway and epitopes 18
4.4. Third year 19
4.4.1. Completing the computer-aided vaccine design system 19
4.4.2. Establishing a database of immune features 19
5. Results and Discussions 20
5.1. Structure of Computer-aided Vaccine Design System 20
5.2. Determining the location of peptides 22
5.2.1. Proposed prediction methods 22
5.2.2. ProLoc-GO – predicting protein subcellular localization 24
5.2.3. S-protein – determining the transport ability of proteins 34
5.3. Detecting the functions of peptide – Virulent-GO 40
5.4. Prediction of T-cell immune response 45
5.4.1. Related works 45

5.4.2. Prediction of ubiquitylation sites 49
5.4.3. Predicting immunogenicity of MHC binding peptides....................62
5.4.4. Identification of T-cell receptor recognition sites 73
6. Conclusions. 84
6.1. Summary...84
6.2. Achievements of the project 85
References. 87
計畫成果自評表 ···98

1. Introduction

Development of vaccine designs is a useful method for curing cancer or other diseases. However, decision of vaccine binding sites to their target cells costs much time and money. The biological factors influence the physiological phenomenon of human body, including the binding site of transcription factor, location of proteins, protein structures, protein functions and evolution of protein. Many of biological factors were used, such as epitope, protein folding and interaction of proteins. Therefore, identification of informative factors is an important issue.

Both modeling of antigen processing pathway and prediction of peptide immunogenicity are necessary to develop a computer-aided system of peptide-based vaccine design that is one goal of immunoinformatics. Numerous studies have dealt with modeling the immunogenic pathway but not the intractable problem of predicting immunogenicity due to complex effects of many intrinsic and extrinsic factors. Moderate affinity of the MHC-peptide complex is essential to induce immune responses, but the relationship between the affinity and peptide immunogenicity is too weak to use for predicting immunogenicity.

The feature selection methods in identifying biological factors for predicting protein functions play an important role in the research field of bioinformatics. The prediction systems can find valuable biological features from previous research results effectively. Utilizing the existing systems, the mechanism of immune systems will be analyzed. How to develop efficient methods for mining effective and interpretable factors from a large number of candidate factors are important in designing computer-aid vaccine system. The core problem forms a large-scale parameter optimization problem.

The projects aims 1) to develop various high-performance optimization algorithms for solving large-scale parameter optimization problems of bioinformatics to mine informative physicochemical properties of proteins from known experimental data; 2) to integrate immune knowledge bases with the developed optimization methodologies to develop application systems of analyzing and annotating informative functions of proteins; 3) to develop a computer-aided vaccine design system to help immunologists for discovering the immune informative factors, and 4) to apply the developed high-performance optimization algorithm to related protein function prediction problems.

2. Objectives

This is a three-year project titled “Developing Computer-aided Vaccine Design Systems Using Bio-inspired Optimization Methods.” The objectives are to develop computer-aided systems to help immunologist for vaccine design and to develop efficient feature mining algorithms to help researchers of bioinformatics for predicting protein functions. The project uses optimization algorithms of bioinformatics to study antigen processing and presentation pathway, epitope, and immunogenicity. In the project, the first objective is to integrate immune knowledge base with the developed optimization methodologies to develop application systems of analyzing and annotating informative functions of proteins. And the final objective is to complete the computer-aided vaccine design system by way of the informative knowledge database.

The project focuses on immune reactions of vaccine design and develops an optimization method. Utilizing the optimization algorithms which are developed by our laboratory, including intelligent genetic algorithm [1], inheritable genetic algorithm [2], orthogonal simulated annealing algorithm, and orthogonal particle swarm algorithm, and combining classifiers, such as Bayesian classifier, artificial neural network, decision tree, and support vector machine, to build up a system of mining bioinformatics informative physicochemical properties. Furthermore, the system is used to study antigen processing and presentation pathway and epitope to discover the important factor of the key reactions.

The main topics of this project including the immune pathway and related biological reactions, described below:

A. Study cytotoxic T lymphocyte related immune response by mining informative features.
B. Study the immune responses of T helper cell, including the infect pathway and feature of T cell.
C. Establish integrated computer-aided vaccine design systems.
D. Developing an informative feature database of immune systems.

After finding out the important feature of key immune reaction, we can design and develop a computer-aided vaccine design system. The system will provide the information of epitope and immunogenicity to help immunologist for vaccine designs.

3. Literature Review

3.1. Introduction of Immune

3.1.1. Immunogenicity of MHC class I binding peptides and the pathway

The most important work of vaccine design is to identify cytotoxic T lymphocyte (CTL) epitopes and investigate their corresponding immunogenicity. The CTL cells play a critical role in protective immunity by recognizing and eliminating self-altered cells, which recognize short peptides derived from intracellular degradation of foreign proteins in combination with major histocompatibility complex (MHC) class I molecules [3]. The immunogenicity of MHC class I binding peptides is their ability to induce CTL responses. Direct approach to predicting the CTL epitopes has been studied initially but its accuracy is fairly low [4]. Instead, indirect approach to predicting the MHC binding peptides is useful because peptides must be processed prior to inducing cellular immune responses. The recent studies of bioinformatics utilized the information about antigen processing pathway to predict the CTL epitopes. The pathway contains three steps described below:

In the first step: the peptides are cleaved by proteasomal cleavage. Several studies elucidating the specificity of proteasome have been presented. To predict proteasomal cleavage sites, NetChop used a neural network method [5] and Pcleavage is based on a support vector machine (SVM) learning model [6].

In the second step: peptide fragments are transported into endoplasmic reticulum by TAP, which is the transporter associated with antigen processing. Some studies of investigating the TAP transport efficiency were presented, such as the affinity prediction of TAP-binding peptides using the cascade SVM [7] and the prediction of TAP transport efficiency of epitope precursors using a simple scoring matrix [8]

In the final step: the peptide fragments that bound to MHC class I molecules are subsequently translocated to the cell surface, where these complexes may active CTL. Some methods have been developed to predict MHC class I binding affinity, such as the SVM-based SVMHC [9] and Gibbs sampling method [10].

Moreover, the hybrid approaches integrated the above mentioned methods like the prediction of proteasomal cleavage, TAP transport efficiency and MHC binding to advance the prediction performance [11, 12].

Figure 3.1.1 explains the MHC class I immune pathway and response of cytotoxic T lymphocyte.

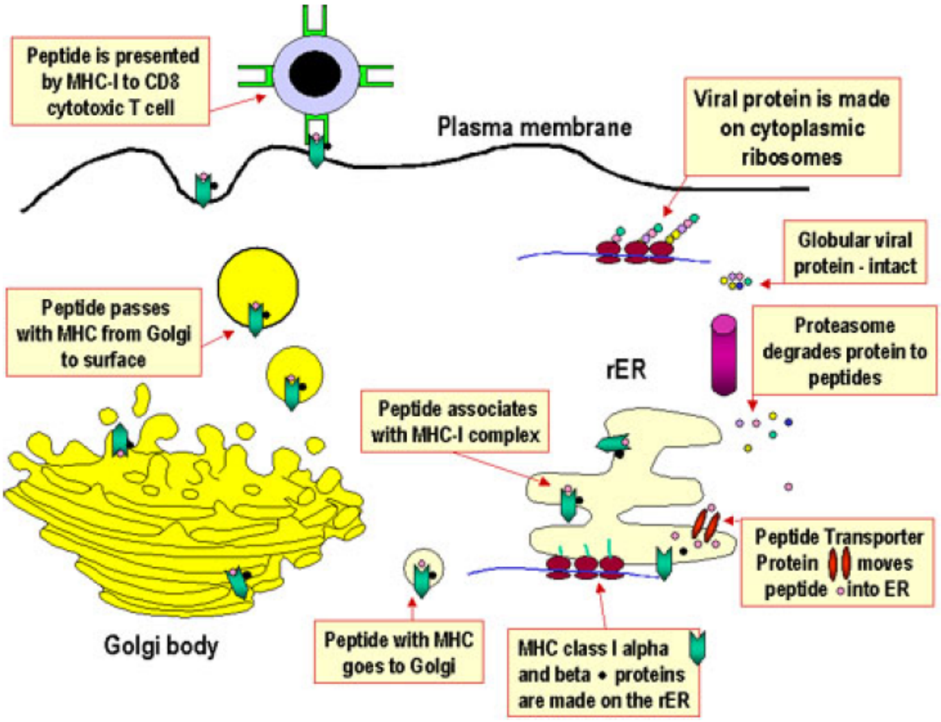

Peptide is presented by MHC-I to CD8 cytotoxic T cell
Plasma membrane
Viral protein is made on cytoplasmic ribosomes
Globular viral protein - intact
Peptide passes with MHC from Golgi to surface
Proteasome degrades protein to peptides
rER
Peptide associates with MHC-I complex
Peptide Transporter Protein moves peptide into ER
Golgi body
Peptide with MHC goes to Golgi
MHC class I alpha and beta • proteins are made on the rER

Figure 3.1.1 The MHC class I immune pathway and response of cytotoxic T lymphocyte [13].

3.1.2. Immunogenicity of MHC class II binding peptides and the pathway

Like the MHC class I molecule, MHC class II binds the peptide fragments and translocated to the cell surface, where these complexes may active helper T cell (HTL).

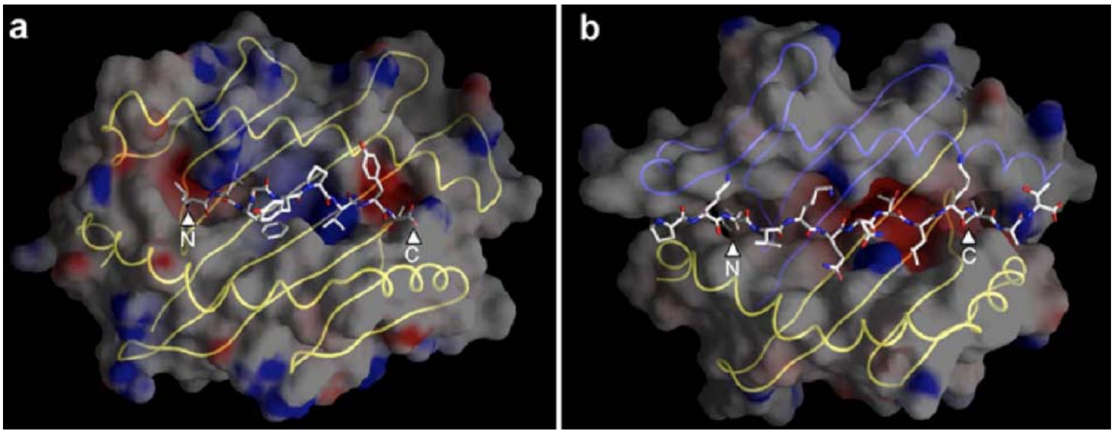


Figure 3.1.2 The difference of MHC class I and MHC class II binding site. (a) is structure of MHC I and (b) is structure of MHC II [14].

Figure 3.1.2 shows the difference of MHC class I and MHC class II, that the binding groove of MHC class II molecules is open at both ends while the corresponding groove on class I molecules is closed at each end. Because the difference of binding groove, peptide fragments of MHC class II have more variable length, which about 9 to 25 residues [15]. As the characteristic of antigen-binding groove, classifiers which have ability to deal with the variable length peptides, such as hidden Markov model (HMM) have better performance [10, 16, 17]. Figure 3.1.3 presents the response of helper T cell and MHC class II immune pathway.

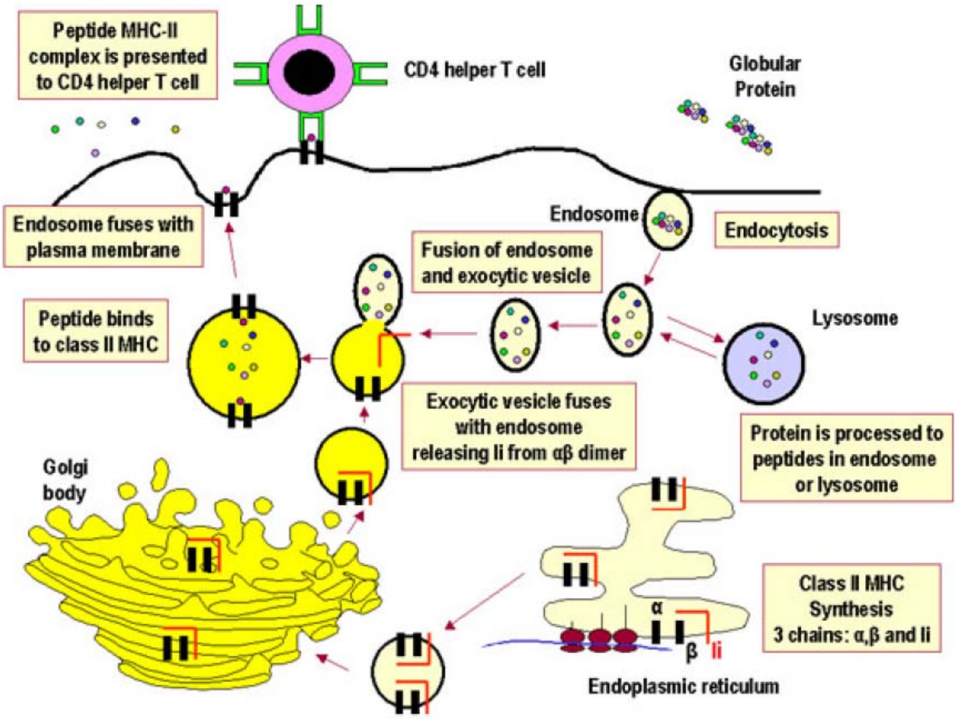

Peptide MHC-II complex is presented to CD4 helper T cell
CD4 helper T cell
Globular Protein
Endosome fuses with plasma membrane
Endosome
Endocytosis
Fusion of endosome and exocytic vesicle
Peptide binds to class II MHC
Lysosome
Exocytic vesicle fuses with endosome releasing Ii from αβ dimer
Protein is processed to peptides in endosome or lysosome
Golgi body
Class II MHC Synthesis 3 chains: α,β and Ii
α
β
Ii
Endoplasmic reticulum

Figure 3.1.3 The MHC class II immune pathway and response of helper T cell [13].

3.2. Introduction of Bioinformatics Tools

3.2.1. Biological features

Feature extraction and selection methods are both important to study the biology by bioinformatics. As the result, describing the features of biological phenomenon completely is necessary for classify and prediction. And the popular usable biological features can be classified into three categories:

1) Features of the amino acid: this kind of features describes the basic function and characteristic of amino acid, for example, hydrophilicity and hydrophobicity, charge, and molecular weight [18-22].

2) Features of the protein sequences: the protein sequences feature explain the composition and functions of protein, such as functional domain, sequence identity, distance of specific residue [18, 23, 24].

3) Features of the protein structures: features of this category have better ability to separate the different protein. The category contain feature, for instance, contact number of residues, solvent accessibility, protein secondary structure [25, 26].

3.2.2. Feature selection methods

1) Knowledge-based

Knowledge-based method means manual selecting the features that related to biological phenomenon. This method selects the feature more accurately but the standard of selection is according to previous researches. As a result, the related works and completeness of the study determine the performance [18].

2) Rank-based Method

Rank-based method is the easiest selection method. The method is ranking the feature and selects the front features. Because of the disregard of interactions between the features, the rank-based method has less ability to select the correct information of biology [22].

3) Simulated Annealing Algorithm

Simulated annealing algorithm involving heating and controlled cooling of a material to increase the size of its solution space and reduce their defects. The heat causes the start atoms to become unstuck from their initial positions and wander randomly through states of higher energy; the slow cooling gives them more chances of finding configurations with lower internal energy than the initial one. When the heat does not have enough energy to move the atoms, the best solution of cooling process is the final solution.

Simulated annealing algorithm has the ability to search the worse solution area and find out the global optimum solution, without the area optimum solutions. However, the start atom is highly related the final solution; otherwise, the heat and cooling also influence the solution. As a result, how to determine the start atoms, the heat temperature, and cooling rate is important.

4) Genetic Algorithm

The concept of genetic algorithm is mentioned by J. Holland that mimics the process of natural evolution. In a genetic algorithm, the chromosomes encode candidate solutions to an optimization problem. The chromosomes can exchange the information and evolves toward better solutions. The genetic algorithm is the evolutionary algorithm that can search the globally optimum solution.

3.2.3. Classifiers

1) Decision Tree

Decision tree is a popular machine method which is used at artificial intelligence and pattern recognition. The method separates the feature space to the subspace by recursion and uses the tree structures present the result. Branches of the decision tree were decided by expected value so the result is fixed that cannot be analyzed in depth.

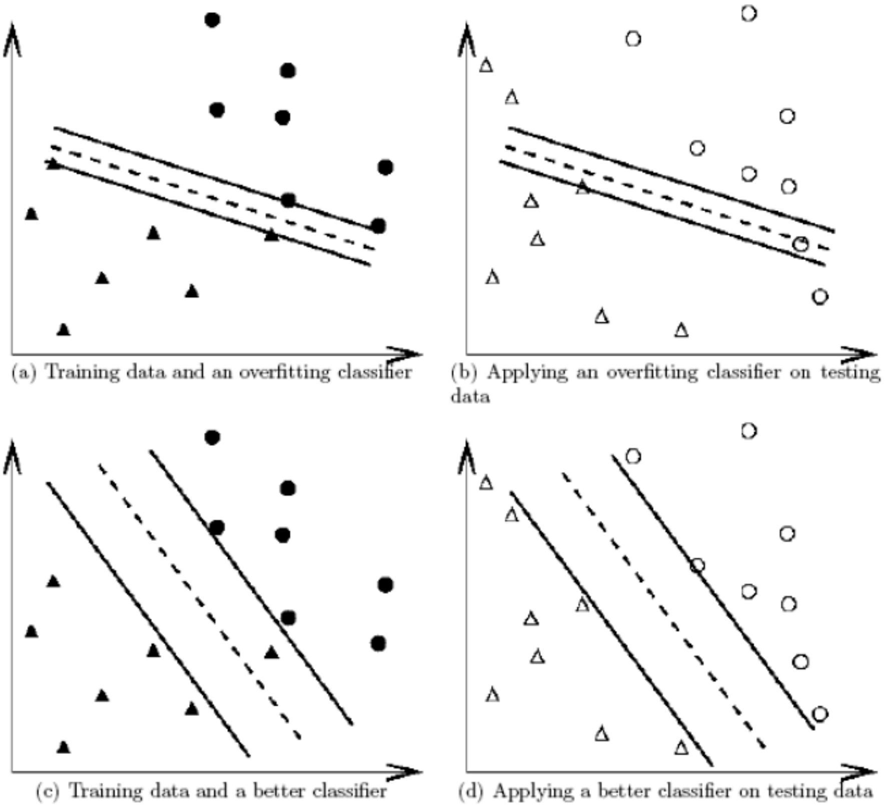
2) Support vector machine

Support vector machine (SVM) is a learning model dealing with binary classification problems. SVM constructs a binary classifier by finding a hyperplane to separate two classes with a maximal distance between margins of two classes consisting of support vectors. In order to make linear separation of samples easier, SVM uses one of various kernel functions to transform the samples into a high-dimensional search space. In this work, the commonly used radial basis function is applied to non-linearly transform the feature space, defined as follows:

$$K(x_i, x_j) = \exp(-\gamma \left\| x_i - x_j \right\|), \gamma > 0$$

The kernel parameter γ determines how the samples are transformed into a high-dimensional search space. The cost parameter C>0 of SVM adjusts the penalty of total error. These two parameters C and γ must be tuned to get the best prediction performance.

The difference of SVM and other classifiers is that SVM can adopt the over fitting. As show in figure 3.2.1a and 3.2.1b, normal classifiers make the best performance of training data, which will lead the false prediction of new data. To avoid the over fitting, SVM finds the hyperplane to separate two classes with a maximal distance that shows in figure 3.2.1c and 3.2.1d.



(a) Training data and an overfitting classifier

(b) Applying an overfitting classifier on testing data

(c) Training data and a better classifier

(d) Applying a better classifier on testing data

Figure 3.2.1 The different training methods of classifiers and problem of over fitting [27].

3) Fuzzy Classifier

Construction of fuzzy classifier including the three steps: first, we need to decide the structure, location, and numbers of rule functions. And then, the rule functions are used to cleavage the feature space with fuzzy. Finally, the fuzzy rules are produced for each feature space. The composition of fuzzy rules contains the antecedent part and consequent part. Antecedent part determine the rule is true or false and the weight of rules. Otherwise, consequent part manages the output form, for example, the consequent part will output the discontinuous label for the classifier.

4) Neural Network

The neural network simulated the structure and function of neuron. Neural network is composed by amount of neuron and each neuron has ability to do the simple computation. A simple structure of neural network is the multi-layered which including input layer, hidden layer, and output layer. Data enter the neural network by input layer, and the hidden layer obtains the data information. Finally, the output layer decides the result of data.

Weighted value is decided by training model and it is used to evaluate the ability of classification. Neural network is hard to analyze the result of each layer, and the output cannot be interpreted by conditional rules. Figure 3.2.2 shows a structure of fuzzy neural network.

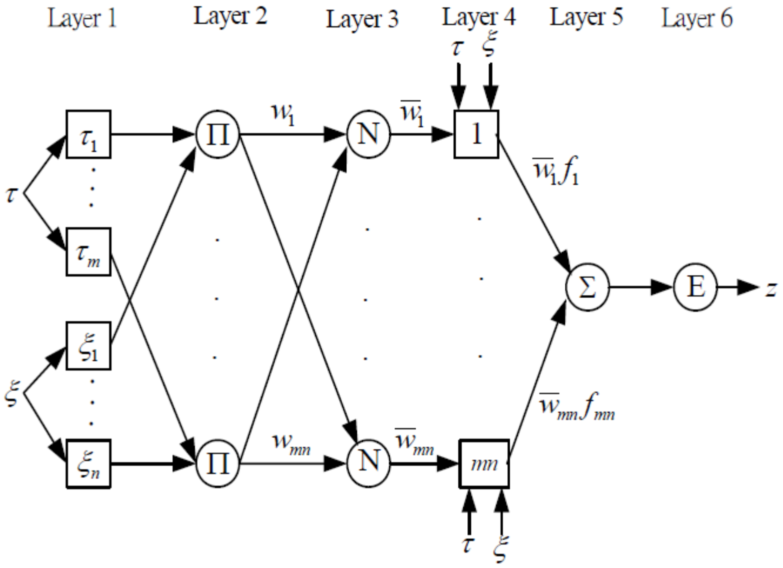


Figure 3.2.2 The illustration of fuzzy neural network.

4. Methods

The main objectives of the project are to develop computer-aided systems to help immunologist for vaccine design. In addition, this project integrates immune knowledge base and application system for analysis and annotation of informative features. We show the proposal of the three years including topic, schedule and methods.

4.1. Objectives of three-year project

To develop high-performance optimization algorithms for solving problems of bioinformatics to mine informative physicochemical properties, we set up the three years plan to finish. The core algorithm of system is the most important part of the project, so we establish the algorithm first, and then we analyse the immune reaction. The topics of each year are described below.

First year: Develop the core high-performance feature mining system, and study cytotoxic T lymphocyte related immune response by mining informative features.

Second year: Study the immune responses of T helper cell, including the infect pathway and feature of T cell, and collect the immune features of key reactions.

Final year: Develop the integrated computer-aided vaccine design systems, and establish an informative feature database of immune systems.

We also apply the developed high-performance optimization algorithm to related protein function prediction problems.

4.2. First year

4.2.1. Establishment of the core algorithm

This study proposes a computational method to mine a feature set of informative physicochemical properties from MHC class I binding peptides to design a support vector machine (SVM) based system that integrate the inheritable bi-objective genetic algorithm for the prediction of peptide immunogenicity.

Selection of important features is completed by high-performance optimization algorithms so this project employs the parameter optimization algorithms, including orthogonal experimental design and inheritable bi-objective genetic algorithm that described below:

1) Orthogonal experimental design, OED

Statistic design of experiments is a process of planning experiments. Orthogonal experimental design with orthogonal array and factor analysis is an efficient method to analyze the effect of several factors simultaneously [28, 29]. The factors are the parameters, which affect response variables, and a discriminative value of a factor is regarded as a level of the factor. A 'complete factorial' experiment would make measurements at each of all possible level combinations. However, the number of level combinations is often so large that this is impractical, and a subset of level combinations must be judiciously selected to be used, resulting in a 'fractional factorial' experiment. Orthogonal experimental design utilizes properties of fractional factorial experiments to efficiently determine the best combination of factor levels to use in design problems.

Orthogonal array is a fractional factorial array, which assures a balanced comparison of levels of any factor. Orthogonal array can reduce the number of level combinations for factor analysis. Each row of an orthogonal array represents the levels of factors in each combination, and each column represents a specific factor that can be changed from each combination. The term 'main effect' of one factor designates the effect on response variables that one can trace to a design parameter, which does not bother the estimation of the main effect of another factor. After proper tabulation of experimental results, the summarized data are analyzed using factor analysis to determine the relative-level effects of factors.

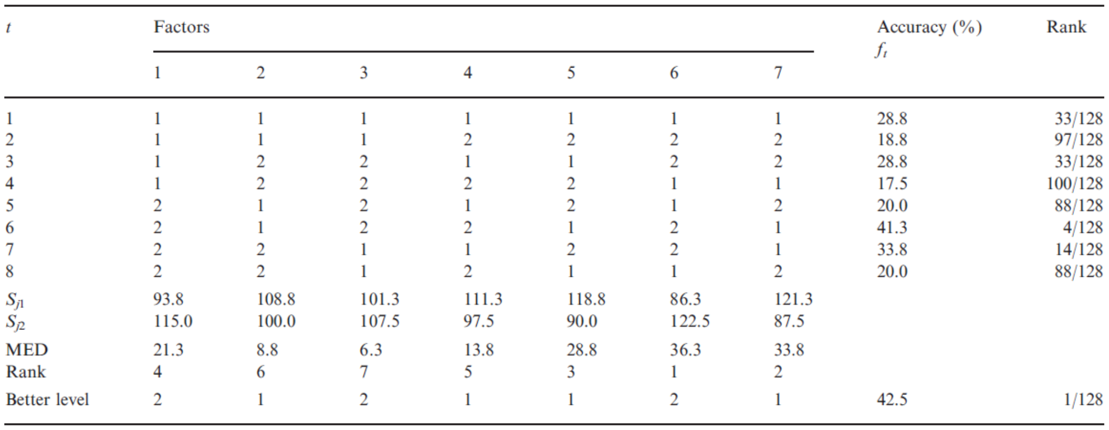
Factor analysis can evaluate the effects of individual factors on the evaluation function, rank the most effective factors, and determine the best level for each factor such that the evaluation function is optimized. Table 2 shows an illustrative example of orthogonal experimental design using a two-level orthogonal array $L_M(2^{M-1})$ with M rows and M-1 columns. In this example of M=8, there are seven factors where each corresponds to a physicochemical property and its two levels correspond to exclusion and inclusion of the feature in the proposed feature selection. Let *ft* denote a function value (prediction accuracy of 10-CV in this study) of the combination t. Define the main effect of factor j with level k as S_{jk} where j=1, . . ., M-1 and k=1, 2:

$$S_{jk} = \sum f_t \cdot F_t, \quad t = 1, \ldots, M,$$

Where F_t=1 if the level of factor j of combination t is k; otherwise, F_t=0. Since the objective function is to be maximized, the level 1 of factor j makes a better contribution to the function than level 2 of factor j does when S_{j1}>S_{j2}. The main effect reveals the individual effect of a factor. After the better one of two levels of each factor is determined, a good combination consisting of all factors with the better levels can be easily reasoned [1].

The rank in Table 4.2.1 shows the rank of the combination t in all 128(=2^7) possible combinations. In this example, the reasoned combination gets the best accuracy with rank 1. Notably, the reasoned combination is not guaranteed to be the best one in general cases. The most effective factor j has the largest main effect difference MED=|S_{j1}-S_{j2}|. The 6th factor having the largest MED 36.3 is the most effective factor.

Table 4.2.1 An illustration example of orthogonal array $L_8(2^7)$ and factor analysis



t	Factors							Accuracy (%) f_t	Rank
	1	2	3	4	5	6	7		
1	1	1	1	1	1	1	1	28.8	33/128
2	1	1	1	2	2	2	2	18.8	97/128
3	1	2	2	1	1	2	2	28.8	33/128
4	1	2	2	2	2	1	1	17.5	100/128
5	2	1	2	1	2	1	2	20.0	88/128
6	2	1	2	2	1	2	1	41.3	4/128
7	2	2	1	1	2	2	1	33.8	14/128
8	2	2	1	2	1	1	2	20.0	88/128
S_{j1}	93.8	108.8	101.3	111.3	118.8	86.3	121.3		
S_{j2}	115.0	100.0	107.5	97.5	90.0	122.5	87.5		
MED	21.3	8.8	6.3	13.8	28.8	36.3	33.8		
Rank	4	6	7	5	3	1	2		
Better level	2	1	2	1	1	2	1	42.5	1/128

2) Inheritable bi-objective genetic algorithm, IBCGA

Selecting a minimal number of informative features while maximizing prediction accuracy is a bi-objective 0/1 combinatorial optimization problem. An efficient inheritable bi-objective genetic algorithm [2] is utilized to solve this optimization problem. IBCGA consists of an intelligent genetic algorithm [1] with an inheritable mechanism. The intelligent genetic algorithm uses a divide-and-conquer strategy and an orthogonal array crossover to efficiently solve large-scale parameter optimization problems. In this study, the intelligent genetic algorithm can efficiently explore and exploit the search space of C(n, r). IBCGA can efficiently search the space of C(n, r±1) by inheriting a good solution in the space of C(n, r) [2]. Therefore, IBCGA can economically obtain a complete set of high-quality solutions in a single run where r is specified in an interesting range such as [5, 45].

The proposed chromosome encoding scheme of IBCGA consists of both binary genes for feature selection and parametric genes for tuning SVM parameters, where the gene and chromosome are commonly used terms of genetic algorithm (GA), named GA-gene and GA-chromosome for discrimination in this article. The GA-chromosome consists of n=531 binary GA-genes bi for selecting informative properties and two 4-bit GA-genes for tuning the parameters C and γ of SVM. If b_i=0, the i_{th} property is excluded from the SVM classifier; otherwise, the i_{th} property is included. This encoding method maps the 16 values of γ and C into [2^{-7}, 2^{-6} ,. . . , 2^{8}]. Figure 1 shows the encoding scheme of GA-chromosome and process of constructing feature vectors for fitness function evaluation using a concise example.

The feature vector for training the SVM classifier is obtained from decoding a GA-chromosome using the following steps. Consider a given peptide sequence, e.g. lysosomal acid lipase (LAL). At first, the index vectors for all selected physicochemical properties (residue volume and molecular weight in this example) are constructed from AAindex for each amino acid. Feature vector of a peptide consists of the selected features whose values are obtained by averaging the values in their corresponding index vectors. Finally, all values of the feature vectors are normalized into [-1, 1] for applying SVM.

Fitness function is the only guide for IBCGA to obtain desirable solutions. To avoid from the prediction bias for some immunogenic levels, the averaged accuracies (AA) of four immunogenic levels is adopted as the fitness function. The performance of selected properties associated with the parameter values of SVM is measured by 10-CV. Therefore, the fitness value of a GA-chromosome is obtained by computing the mean accuracy of 10 runs.

IBCGA with the fitness function f(X) can simultaneously obtain a set of solutions, X_r, where $r=r_{start}$, $r_{start+1}$, ..., r_{end} in a single run. The algorithm of IBCGA with the given values r_{start} and rend is described as follows:

Step 1: (Initiation) Randomly generate an initial population of N_{pop} individuals. All the n binary GA-genes have r 1s and n-r 0s where $r=r_{start}$.

Step 2: (Evaluation) Evaluate the fitness values of all individuals using f(X).

Step 3: (Selection) Use the traditional tournament selection that selects the winner from two randomly selected individuals to form a mating pool.

Step 4: (Crossover) Select $P_c \cdot N_{pop}$ parents from the mating pool to perform orthogonal array crossover on the selected pairs of parents, where P_c is the crossover probability.

Step 5: (Mutation) Apply the swap mutation operator to the randomly selected $P_m \cdot N_{pop}$ individuals in the new population, where P_m is the mutation probability. To prevent the best fitness value from deteriorating, mutation is not applied to the best individual.

Step 6: (Termination test) If the stopping condition for obtaining the solution X_r is satisfied, output the best individual as X_r.

Otherwise, go to Step 2.

Step 7: (Inheritance) If r<rend, randomly change one bit in the binary GA-genes for each individual from 0 to 1; increase the number r by one, and go to Step 2. Otherwise, stop the algorithm.

4.2.2. Studying the immunogenicity of MHC class I

The major work of previous studies for peptide vaccine designs is to identify cytotoxic T lymphocyte (CTL) epitopes and investigate their corresponding immunogenicity. Direct approach to predicting the CTL epitopes has been studied initially but its accuracy is fairly low. Instead, indirect approach to predicting the MHC binding peptides is useful because peptides must be processed prior to inducing cellular immune responses.

It is well recognized that feature selection and classifier design should be optimized simultaneously to maximize prediction accuracy. The SVM-based learning methods are shown effective for various prediction methods from protein sequences. However, internal detection of relevant feature correlation is not offered by conventional SVMs. The project integrates physicochemical properties and evolution information to analyse the immune reactions.

Physicochemical properties of amino acids were extensively and successfully used in sequence-based prediction methods [18, 19, 21, 30]. Because of the weak correlation between peptide immunogenicity and peptide-MHC binding affinity, mining informative physicochemical properties is a potentially good approach to designing a classifier for predicting immunogenicity. This study aims to design an accurate predictor by efficiently selecting a small set of informative physicochemical properties considering the correlated effects

There are 544 physicochemical properties of amino acids extracted from amino acid index database (AAindex), which is a collection of published amino acid indices representing different physicochemical and biological properties of amino acids [31]. Each physicochemical property consists of a set of 20 numerical values for amino acids. The property having the value 'NA' in a value set of amino acid index was discarded. Finally, 531 properties were used for the following mining method.

Support vector machine (SVM) is a learning model dealing with binary classification problems. SVM constructs a binary classifier by finding a hyperplane to separate two classes with a maximal distance between margins of two classes consisting of support vectors. The radial basis function is applied to non-linearly transform the feature space in this study.

For multi-class classification problems, 'one-against-one' strategy is applied to transform the multi-class problem into several binary classification problems. Given h classes, there are $h(h\text{-}1)/2$ classifiers constructed and each one trains the samples from two classes. A voting strategy is applied to give a final prediction for test samples. In this study, h=4 and the used SVM is obtained from LIBSVM package version 2.91[27].

4.2.3. Prediction system of cytotoxic T lymphocyte

Study of indirect approach to predicting the MHC binding peptides is useful because peptides must be processed prior to inducing cellular immune responses. The three reactions of MHC class I are studied:

1) The peptides are cleaved by proteasomal cleavage.

2) TAP transport efficiency of epitope precursors.

3) MHC class I binding affinity.

So the project proposes a predict system of CTL epitope. System integrates the immunogenicity of CTL and MHC class I study to reinforce the model. User can predict the epitope, pathway, immune reaction, and immunogenicity of unknown proteins.

This study integrates the prediction of ubiquitylation sites, TAP transport efficiency, and MHC class I binding affinity, and CTL epitope to establish the CTL prediction system. User can predict the ubiquitylation sites first and peptides which have high affinity can be selected. The system decides the epitope and presents the immunogenicity of peptides.

4.3. Second year

4.3.1. Developing the useable classifiers of immune reactions

We utilized the core algorithm to construct the variable classifiers which response to different part of the computer-aided vaccine design system. The main optimization algorithm in development is intelligent genetic algorithm.

Intelligent Genetic Algorithm (IGA) combines the characteristic of genetic algorithm and orthogonal experimental design, so the algorithm has ability that can converge quickly and obtain the higher accuracy [1]. The differences of intelligent genetic algorithm and traditional genetic algorithm (GA) is that crossover of IGA is combined with orthogonal experimental design. In other words, we consider the crossover as an orthogonal experiment to replace the traditional random crossover. GA-chromosome will crossover according to the orthogonal array and the probability to get better chromosome descendant is promoted. The process shows below:

Step 1: Coding of chromosomes gene make each segment is treated as a factor. Use the first N columns of an $L_{\mathrm{n}}(2^{\mathrm{n-1}})$, where n = $2^{[\log（N+1）].}$

Step 2: Let levels 1 and 2 of factor represent the jth gene segments of chromosomes coming from parents P_1 and P_2, respectively.

Step 3: Compute the fitness value y_t of the combination t, where t = 1, 2, …, n.

Step 4: Compute the main effect S_{jk}, where j = 1, 2, …, N and k = 1, 2.

Step 5: Determine the better one of two levels of each factor based on main effect.

Step 6: The chromosome of first child is formed using the combination of the better gene segments from the derived corresponding parents.

Step 7: The chromosome of second child is formed similarly as first child, except that the factor with the smallest MED adopts the other level.

Except the SVM classifier, different classifiers are considered to be the estimate tools. For example, fuzzy *k*-nearest neighbor classifier is used to deal with the highly interactive data.

Fuzzy *k*-nearest neighbor classifier active according to “Things of one kind come together”, means the same class will gather together. The difference of fuzzy *k*-nearest neighbor and traditional *k*-nearest neighbor is that the traditional *k*-nearest neighbor hires Euclidean distance as the distance function but the fuzzy *k*-nearest neighbor employs the fuzzy C-means clustering method. The fuzzy C-means clustering method is based on fuzzy logic; means there have uncertain area of solution space so the classifier has to inference the possible solutions. Fuzzy *k*-nearest neighbor classifier classes the unknown data by consider the value of *k* nearest neighbor. Furthermore, data in same class has different belong degree, which according to membership function and range from 0 to 1, replace the integer.

4.3.2. Utilizing the varied biological features

The physicochemical properties of AAindex are used at first year. In order to expand the variety of features, we find out the different feature set of sequences, such as functional motif, gene ontology annotation, and amino acid composition.

In previous study, the functional motif is also use to assist the prediction of MHC class II. Functional motif is the conserved sequence fragment that means the sequence fragment is not change during the evolution. The functional motif usually has important information, so the project also employs this property.

Gene Ontology (GO) [32] annotation, which describes the function of genes and gene products across species, has recently been utilized to predict protein subcellular and subnuclear localization. The prediction of protein localization is important for elucidating protein functions involved in various cellular processes. Additionally, the accomplishment of the various genome sequencing projects causes the accumulation of massive amount of gene sequence information.

4.3.3. Study of HTL immune pathway and epitopes

The HTL cells play a critical role in protective immunity by recognizing and eliminating self-altered cells and the immunogenicity of MHC class II binding peptides is their ability to induce HTL responses. The project will study the HTL and MHC class II, and furthermore we select the important physicochemical properties of immunogenicity.

Because the open binding groove, peptide fragments of MHC class II have more variable length, which about 9 to 25 residues. The variable length induces more difficult to analyze MHC class II. Like the MHC class I, the peptides are must cleaved by proteasomal cleavage so we employ the method of first year to develop MHC class II predict systems. The proper usage of epitope information and variable length can promote the prediction accuracy effectively.

4.4. Third year

4.4.1. Completing the computer-aided vaccine design system

We establish an intelligent feature mining system at third year. This system uses the inheritable genetic algorithm and intelligent genetic algorithm as the system core and selects varied features, like gene ontology annotation, amino acid composition, and sequence motifs.

The intelligent feature mining system is used to select the important features from the previous studies. This system can find out the suitable feature for different immune reactions. Utilizing the system, newly features can be discovered and integrated to our immune reaction feature database.

Combination of various high-performance optimization algorithms to the computer-aided vaccine design system and make a proposal. When user input a unknown protein, the computer-aided system can determine the correct prediction tool by the protein sequence information. And the system provides the information of epitope and immunogenicity to help immunologist for vaccine design. We interpret the system at chapter 5.

4.4.2. Establishing a database of immune features

The one important aim is establish a immune feature database. Because the features selected in the project contain a lot of immune information that can be utilized for vaccine design. We collect the features and establish an informative feature database to provide the information to the immunologist.

The project establishes an informative feature database of immune systems. The database contains the selected feature of this project and combines to other immune database. Immunologist can study immune reaction pathway and the mechanism deeply by this database.

5. Results and Discussions

5.1. Structure of Computer-aided Vaccine Design System

We construct the computer-aided vaccine design system during the project. In the section, we introduce the principle and procedure of the system, and then describe how to integrate the achievements of the project to each stage in the system. Figure 5.1.1 shows the procedure of the computer-aided vaccine design system:

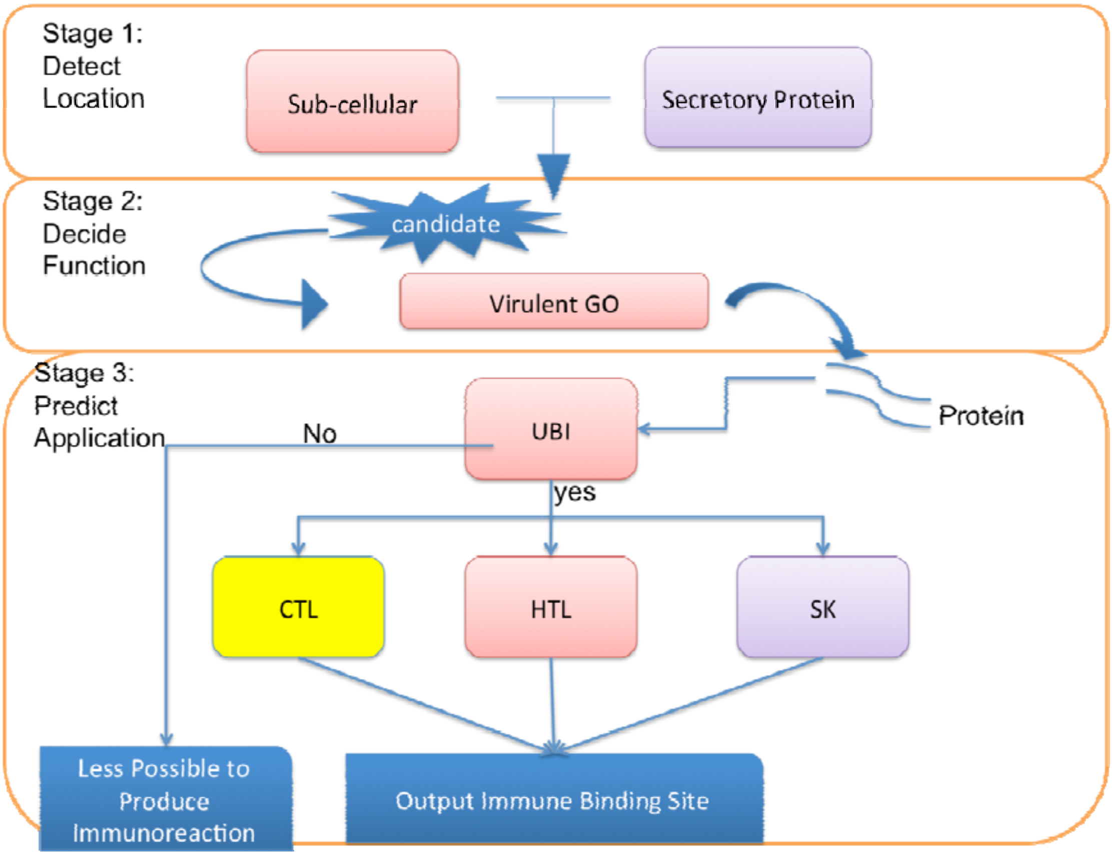


Figure 5.1.1 The procedure of the computer-aided vaccine design system

The computer-aided vaccine design system is constructed by three stages and the core of system is intelligent high-performance feature mining system. We develop different classifiers and prediction systems according the core system in each stage. The tools of each stage will be interpreted and related studies shows in the end. The functions of three stages are described below:

Stage 1: Detecting the location of peptides. The first stage contains the function that detects the peptides is intercellular or intracellular and the transport ability of peptides. According the result, the system can select the peptide that can induce the immune reaction. For example, when the peptides cannot be transported, the immune reaction will not be active.

Stage 2: Predicting the function of peptides. After detect the location of peptide, the system will estimate the function of peptides. In this stage, vaccine design system pick up the functional peptides and accelerate the determination of antigen-binding site of vaccine. The selection of functional peptides also can reduce the number of possible peptides that make vaccine design more accurate.

Stage 3: Predicting the residue binding site and selecting the biological feature of peptides. The final stage will predict the peptides can be ubiquitylated and classify ubiquitylation sites. Because the peptides must be cleaved by proteasomal cleavage to active the immune pathway, the classify ubiquitylation sites is very important. Next, the biological features of the input peptide will be extracted and figure out the peptide is utilized by which T cell base on those features. According to different T cell type, the system chooses the suitable prediction tools to predict the residue binding site accurately and immunogenicity. Finally, the computer-aided vaccine design system provide the possibility of immune reaction, peptide binding site on T cell, and further biological features to assist the vaccine design.

5.2. Determining the location of peptides

The first stage uses the tools, ProLoc-GO and S-protein, to predict the location of peptides. ProLoc-GO has the function that detects the peptides is intercellular or intracellular and S-protein estimates the transport ability of peptides.

5.2.1. Proposed prediction methods

The stage 1 of computer-aided vaccine system is determining where the peptides located. We develop two classifiers to achieve the target, ProLoc-GO and S-protein. Illustration and mechanism of those methods described at next part.

5.2.1.1. ProLoc-GO

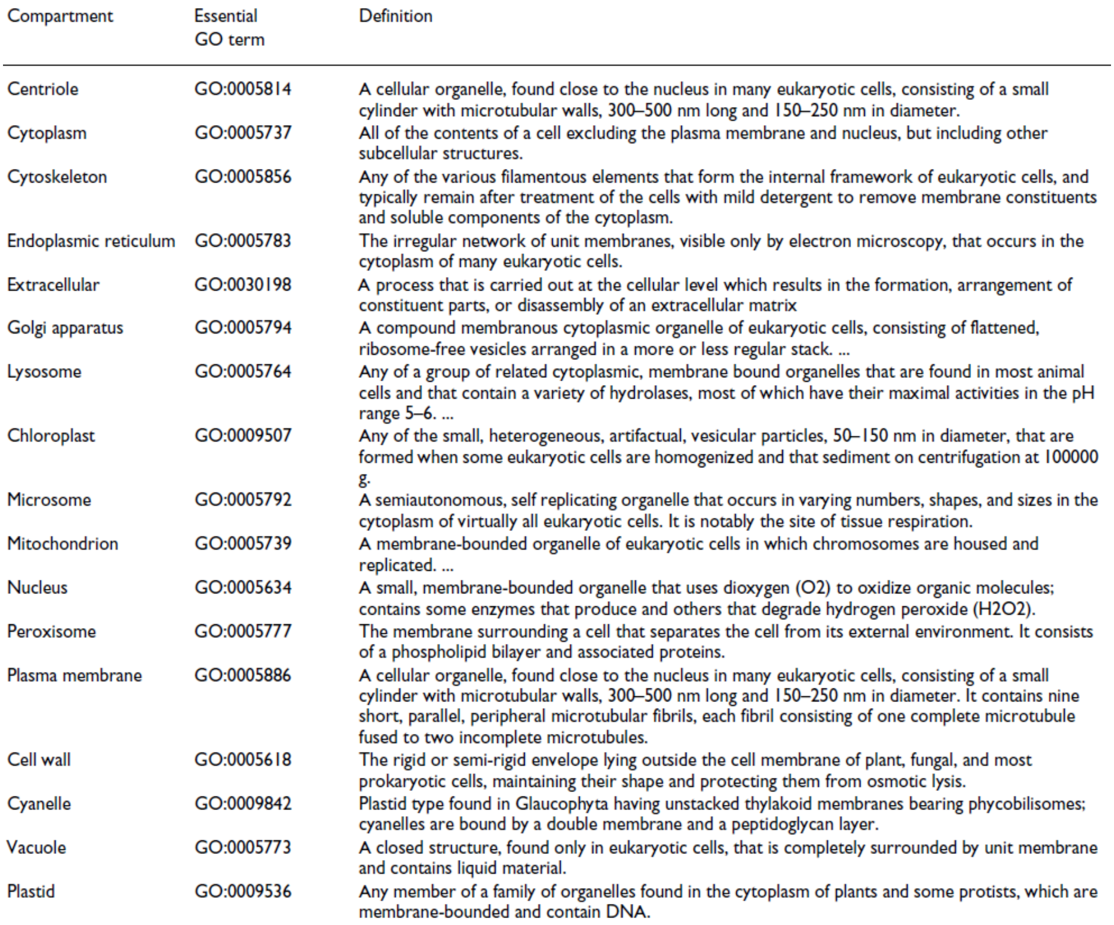
Gene Ontology (GO) [32] annotation, which describes the function of genes and gene products across species, has recently been utilized to predict protein subcellular and subnuclear localization. The prediction of protein localization is important for elucidating protein functions involved in various cellular processes. Additionally, the accomplishment of the various genome sequencing projects causes the accumulation of massive amount of gene sequence information.

Some existing computation methods in literature for predicting protein localization are described which used the classifiers and features. The pSLIP system utilizes five top-rank features of physicochemical properties according to the prediction accuracy of SVM using a single feature [22]. The ProLoc system uses SVM with automatic selection from physicochemical properties to predict protein subnuclear localization [33]. The two efficient GO-based systems Euk-OET-PLoc [34] and Hum-PLoc [35] predict subcellular localization of proteins using their known accession numbers. Those ensemble classifiers Euk-OET-PLoc [34] and Hum-PLoc [35] fuse many basic individual classifiers operated by the engine of k-NN rules, where protein sequences are represented by hybridizing the GO annotation and amphiphilic pseudo amino acid (Pse-AA) composition. However, they cannot work for novel proteins without known accession numbers.

So, we proposes an efficient method, named GOmining, based on an intelligent genetic algorithm (IGA) [1, 2] incorporating an SVM classifier to simultaneously identify a small number m out of a large number n of GO terms as input features, where m <<n. Some GO annotations corresponding to subcellular compartments are

called essential GO terms for subcellular localization prediction, such as GO:0005634 (Nucleus), GO:0005737 (Cytoplasm) and GO:0005856 (Cytoskeleton), shown in Table 5.2.1. These essential GO terms are regarded as domain knowledge to be included in the feature set of m informative GO terms for subcellular localization prediction. A prediction method ProLoc-GO based on GOmining was implemented using the feature set of informative GO terms. This method performed well in predicting protein subcellular localization from input sequences only.

Table 5.2.1 Essential GO terms and their definitions



Compartment	Essential GO term	Definition
Centriole	GO:0005814	A cellular organelle, found close to the nucleus in many eukaryotic cells, consisting of a small cylinder with microtubular walls, 300–500 nm long and 150–250 nm in diameter.
Cytoplasm	GO:0005737	All of the contents of a cell excluding the plasma membrane and nucleus, but including other subcellular structures.
Cytoskeleton	GO:0005856	Any of the various filamentous elements that form the internal framework of eukaryotic cells, and typically remain after treatment of the cells with mild detergent to remove membrane constituents and soluble components of the cytoplasm.
Endoplasmic reticulum	GO:0005783	The irregular network of unit membranes, visible only by electron microscopy, that occurs in the cytoplasm of many eukaryotic cells.
Extracellular	GO:0030198	A process that is carried out at the cellular level which results in the formation, arrangement of constituent parts, or disassembly of an extracellular matrix
Golgi apparatus	GO:0005794	A compound membranous cytoplasmic organelle of eukaryotic cells, consisting of flattened, ribosome-free vesicles arranged in a more or less regular stack. ...
Lysosome	GO:0005764	Any of a group of related cytoplasmic, membrane bound organelles that are found in most animal cells and that contain a variety of hydrolases, most of which have their maximal activities in the pH range 5–6. ...
Chloroplast	GO:0009507	Any of the small, heterogeneous, artifactual, vesicular particles, 50–150 nm in diameter, that are formed when some eukaryotic cells are homogenized and that sediment on centrifugation at 100000 g.
Microsome	GO:0005792	A semiautonomous, self replicating organelle that occurs in varying numbers, shapes, and sizes in the cytoplasm of virtually all eukaryotic cells. It is notably the site of tissue respiration.
Mitochondrion	GO:0005739	A membrane-bounded organelle of eukaryotic cells in which chromosomes are housed and replicated. ...
Nucleus	GO:0005634	A small, membrane-bounded organelle that uses dioxygen (O2) to oxidize organic molecules; contains some enzymes that produce and others that degrade hydrogen peroxide (H2O2).
Peroxisome	GO:0005777	The membrane surrounding a cell that separates the cell from its external environment. It consists of a phospholipid bilayer and associated proteins.
Plasma membrane	GO:0005886	A cellular organelle, found close to the nucleus in many eukaryotic cells, consisting of a small cylinder with microtubular walls, 300–500 nm long and 150–250 nm in diameter. It contains nine short, parallel, peripheral microtubular fibrils, each fibril consisting of one complete microtubule fused to two incomplete microtubules.
Cell wall	GO:0005618	The rigid or semi-rigid envelope lying outside the cell membrane of plant, fungal, and most prokaryotic cells, maintaining their shape and protecting them from osmotic lysis.
Cyanelle	GO:0009842	Plastid type found in Glaucophyta having unstacked thylakoid membranes bearing phycobilisomes; cyanelles are bound by a double membrane and a peptidoglycan layer.
Vacuole	GO:0005773	A closed structure, found only in eukaryotic cells, that is completely surrounded by unit membrane and contains liquid material.
Plastid	GO:0009536	Any member of a family of organelles found in the cytoplasm of plants and some protists, which are membrane-bounded and contain DNA.

5.2.1.2. S-protein

Secreted proteins such as cytokines, chemokines and hormones are potential biomarkers for diagnosis or the evaluation of therapeutic efficiency [36, 37]. They are easy to be detected in body fluids such as serum and urine by noninvasive ways. Discovery of novel human secreted proteins provides possible targets for new drug development and diagnostic technique. With the advances of secretome and proteome researches [38, 39], the identification of novel secreted proteins has made further progress, but the experimentally confirmed secreted proteins are still limited.

Many prediction tools used for subcellular localization such as BaCello, MultiLoc, LOCtree, pTARGET, WoLF PSORT and HSLpred have included the signal peptide or the secreted proteins and regarded as a category of localization [40, 41]. The prediction methods used by these tools include neural network, support vector machine (SVM), Hidden Markov Model (HMM) and k-nearest neighbor [40, 42-44]. Most of these tools rely on the signal peptide to determine whether the protein is secreted or not. Apparently, these tools are not suitable for the prediction of leaderless secreted proteins.

We proposed a novel prediction method combining the informative physicochemical properties of amino acid with SVM to solve the prediction problem of non-classical secreted proteins. The informative physicochemical properties of amino acids indices selected in the method were used as features in designing SVM classifiers. An efficient algorithm inheritable bi-objective genetic algorithm (IBCGA) was used to select significant features which could discriminate the two classes of proteins. The feature sets selected by IBCGA were analyzed carefully to reveal the fundamental differences existed between secreted proteins and non-secreted proteins.

5.2.2. ProLoc-GO – predicting protein subcellular localization

5.2.2.1. Introduction

We propose an efficient sequence-based method (named ProLoc-GO) by mining informative GO terms for predicting protein subcellular localization. For each protein, BLAST is used to obtain a homology with a known accession number to the protein for retrieving the GO annotation. A large number n of all annotated GO terms that have ever appeared are then obtained from a large set of training proteins. A novel genetic algorithm based method (named GOmining) combined with a classifier of support vector machine (SVM) is proposed to simultaneously identify a small number m out of the n GO terms as input features to SVM, where m <<n. The m informative GO terms contain the essential GO terms annotating subcellular compartments such as GO:0005634 (Nucleus), GO:0005737 (Cytoplasm) and GO:0005856 (Cytoskeleton).

Two existing data sets SCL12 (human protein with 12 locations) and SCL16 (Eukaryotic proteins with 16 locations) with <25% sequence identity are used to evaluate ProLoc-GO which has been implemented by using a single SVM classifier with the m = 44 and m = 60 informative GO terms, respectively. ProLoc-GO using input sequences yields test accuracies of 88.1% and 83.3% for SCL12 and SCL16, respectively, which are significantly better than the SVM-based methods, which achieve < 35% test accuracies using amino acid composition (AAC) with acid pairs and AAC with dipedtide composition.

For comparison, ProLoc-GO using known accession numbers of query proteins yields test accuracies of 90.6% and 85.7%, which is also better than Hum-PLoc (85.0%) and Euk-OET-PLoc (83.7%) using ensemble classifiers with hybridization of GO terms and amphiphilic pseudo amino acid composition for SCL12 and SCL16, respectively.

5.2.2.2. Data sets and GO annotation

Two existing data sets SCL12 [35] and SCL16 [34] obtained from UniProtKB/Swiss-Prot database [45] were used to evaluate the proposed method ProLoc-GO. The SCL12 and SCL16 have 2041 human proteins localized in 12 human subcellular compartments and 4150 eukaryotic proteins in 16 subcellular compartments, respectively. The two data sets were operated by a culling program [46] so that those sequences had < 25% sequence identity. The data set SCL12 was divided into two parts, SCL12L and SCL12T, with 919 and 1122 proteins, and the data set SCL16 also consists of two parts, SCL16L for training and SCL16T for independent testing.

We applied the Gene Ontology Annotation (GOA) database [47], which includes GO annotations for nonredundant proteins from many species in the UniProtKB/Swiss-Prot database [45]. The GOA database was downloaded directly from [47] (UniProt 45.0 released in Jan. 2007). The accession numbers of proteins are required for querying the GOA database to obtain GO terms. BLAST [48, 49] was used to obtain a homology with a known accession number to the protein for retrieving the GO terms. The corresponding accession numbers of all protein sequences in SCL12 and SCL16 were obtained by using BLAST with $h = 1$ and $e = 10^{-9}$.

The training data set included the SCL12L and SCL16L were described. For SCL12L, the size of the complete set of all GO terms that appeared was $n = 1714$ from the 919 human proteins. The smallest, largest and mean numbers of GO terms annotated for individual proteins were 0, 35 and 8.3, respectively. The percentage of training proteins whose homologies were not annotated by any GO term (that is, the number of GO terms annotated is zero) was 1.31%. For SCL16L, $n = 2870$ GO terms were obtained from 2423 eukaryotic proteins. The smallest, largest and mean numbers of GO terms annotated were 0, 50 and 7.7, respectively. The percentage of training proteins whose homologies were not annotated was 3.96%. The proteins annotated by GO are often represented as an n-dimensional binary feature vector, where the attribute value is 1 if the corresponding GO term is annotated, and 0 otherwise.

5.2.2.3. Classifiers

Each query protein is first BLASTed with h = 1 and e = 10^{-9} against the Swiss-Prot database to obtain a homology with a known accession number. If no such homology exists, then adjust the threshold value e of BLAST until the desired homology is obtained, where h = 1 and e $\in$ [10^{-9}, 10^{-8},..., 10^{-1}]. The accession number of the homology of each protein sequence in SCL12 and SCL16 was obtained by using BLAST with h = 1 and e = 10^{-9}. This accession number is used as input to the GOA database for retrieving the corresponding k (>1) GO terms: GO:1, GO:2,... GO:k. If none of the k GO terms belongs to the set of m informative GO terms, then the sequence is represented using an n-dimensional binary vector and is predicted by the SVM-GO classifier. Otherwise, the sequence is represented as an m-dimensional binary vector and is predicted by the SVM-IGO classifier. Notably, the SVM-GO classifier predicts only a very small percentage of input sequences. ProLoc-GO is derived from the two major classifiers SVM-GO and SVM-IGO for subcellular localization prediction.

To evaluate the proposed IGA-based feature selection method GOmining, this study implements a classifier SVM-RBS by using SVM with a subset of the n GO terms by the rank-based selection (RBS) method [22, 50]. One previous work on ProLoc [33] showed that this univariate method RBS is inferior to the multivariate feature selection by IGA for selecting physicochemical properties. First, each of all n GO terms (for example, n = 1714 for SCL12L) is ranked according to the accuracy of SVM with the evaluated single feature, where the best values of parameters (C, γ) were determined using a step-wise approach where $\gamma \in$ {2^{-7}, 2^{-6},..., 2^{8}} and C $\in$ {2^{-7}, 2^{-6},..., 2^{8}}. The top-ranking 70 features a_i, i = 1,..., 70 are then picked, and the top-ranking 40 features with r = 40 are used as an initial feature set {b_1,..., b_{40}}. Consequently, the feature set with size r+1 is incrementally established by adding the best feature br+1 (having the highest accuracy of SVM using 10-CV) from the remaining 70-r features into the current feature set.

5.2.2.4. Selected informative GO terms

Selecting a set of m informative GO terms out of n candidate GO terms is a combinatorial optimization problem C(n, m), which can be solved by using the intelligent genetic algorithm with an inheritance mechanism (IGA) [1, 2]. IGA can efficiently search for the solution Sr+1 to C(n, r+1) by inheriting a good solution Sr to C(n, r). This study proposes an efficient algorithm based on IGA, called GOmining, to identify a small set of m informative GO terms including the essential GO terms as features to SVM. The GOmining algorithm incorporates LIBSVM [27] using series of binary classifiers. GOmining aims to maximize the training accuracy of prediction using 10-fold cross-validation (10-CV) when identifying the m informative GO terms.

The SVM classifier based on the selected informative GO terms as features is called SVM-IGO. To evaluate a candidate set of r informative GO terms accompanied with the SVM parameters, the prediction accuracy of 10-CV serves as a fitness function of IGA. Figure 5-2 shows the results of SVM-IGO from r = 40 to 70. Table 5.2.2 lists the m = 44 informative GO terms for SCL12L obtained from the highest accuracy of 89.8% (r = 44), where the SVM parameters (C, γ) = (2^3, 2^{-4}). Table 6 lists the m = 60 informative GO terms for SCL16L, where the highest accuracy was 86.5%, and (C, γ) = (2^5, 2^{-3}).

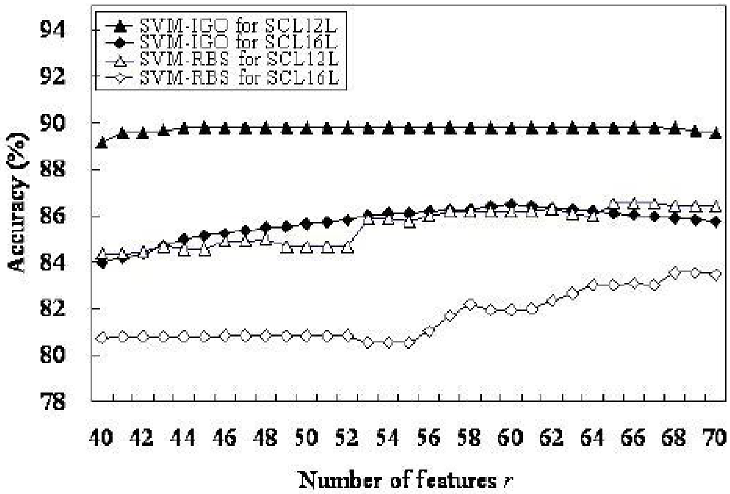
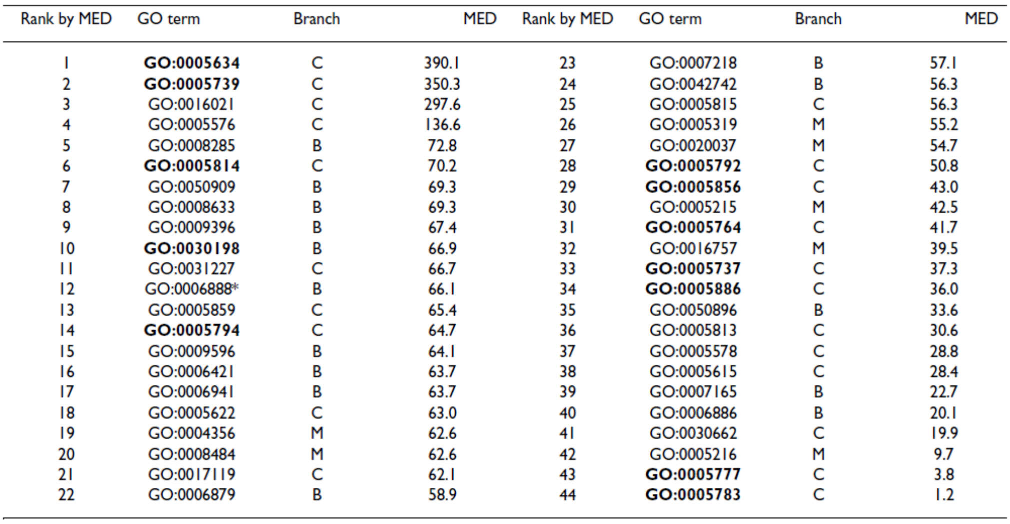


Figure 5.2.1 Training accuracies of SVM-IGO and SVM-RBS performed by using SVM with a number *r* of selected informative GO terms

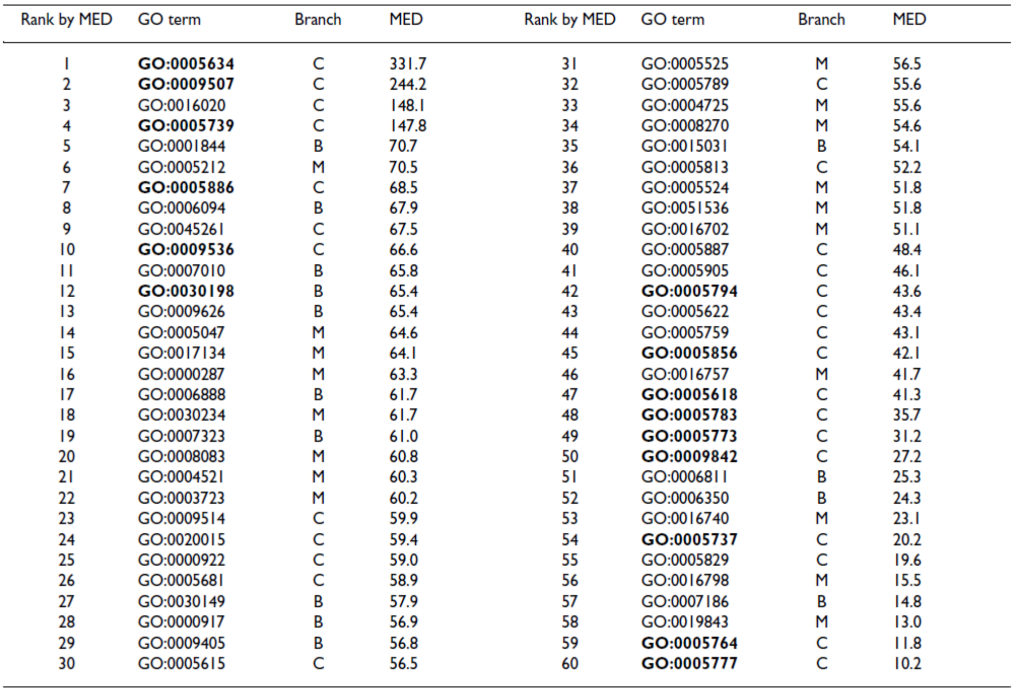
The orthogonal experimental design with orthogonal array and factor analysis used in IGA is an efficient method for simultaneously examining the individual effect of several factors on the evaluative function [51, 52]. The factors are the parameters (GO terms) that manipulate the evaluation function, and a setting of a parameter is regarded as a level of the factor. In this study, the two levels of one factor are the inclusion and exclusion of the ith GO term in the feature selection using IGA. The factor analysis can quantify the effects of individual factors on the evaluation function, rank the most effective factors and determine the best level for each factor to optimize the evaluation function. The most effective factor has the largest main effect difference (MED). Tables 5.2.2 and 5.2.3 show that the essential GO term GO:0005634 (Nucleus) having the largest values of MED is the most effective feature of discrimination. The only essential GO term GO:0030198 (Extracellular matrix organization and biogenesis) belongs to biologic process branch and the other essential GO terms belong to cellular component branch. The abbreviations M, B and C represent the three branches molecular function, biological process, and cellular component, respectively.

Table 5.2.2 The m = 44 informative GO terms by applying GOmining to SCL12L. The GO terms in bold style are essential GO terms



Rank by MED	GO term	Branch	MED	Rank by MED	GO term	Branch	MED
1	**GO:0005634**	C	390.1	23	GO:0007218	B	57.1
2	**GO:0005739**	C	350.3	24	GO:0042742	B	56.3
3	GO:0016021	C	297.6	25	GO:0005815	C	56.3
4	GO:0005576	C	136.6	26	GO:0005319	M	55.2
5	GO:0008285	B	72.8	27	GO:0020037	M	54.7
6	**GO:0005814**	C	70.2	28	**GO:0005792**	C	50.8
7	GO:0050909	B	69.3	29	**GO:0005856**	C	43.0
8	GO:0008633	B	69.3	30	GO:0005215	M	42.5
9	GO:0009396	B	67.4	31	**GO:0005764**	C	41.7
10	**GO:0030198**	B	66.9	32	GO:0016757	M	39.5
11	GO:0031227	C	66.7	33	**GO:0005737**	C	37.3
12	GO:0006888*	B	66.1	34	**GO:0005886**	C	36.0
13	GO:0005859	C	65.4	35	GO:0050896	B	33.6
14	**GO:0005794**	C	64.7	36	GO:0005813	C	30.6
15	GO:0009596	B	64.1	37	GO:0005578	C	28.8
16	GO:0006421	B	63.7	38	GO:0005615	C	28.4
17	GO:0006941	B	63.7	39	GO:0007165	B	22.7
18	GO:0005622	C	63.0	40	GO:0006886	B	20.1
19	GO:0004356	M	62.6	41	GO:0030662	C	19.9
20	GO:0008484	M	62.6	42	GO:0005216	M	9.7
21	GO:0017119	C	62.1	43	**GO:0005777**	C	3.8
22	GO:0006879	B	58.9	44	**GO:0005783**	C	1.2

Table 5.2.3 The *m* = 60 informative GO terms by applying GOmining to SCL16L. The GO terms in bold style are essential GO terms.

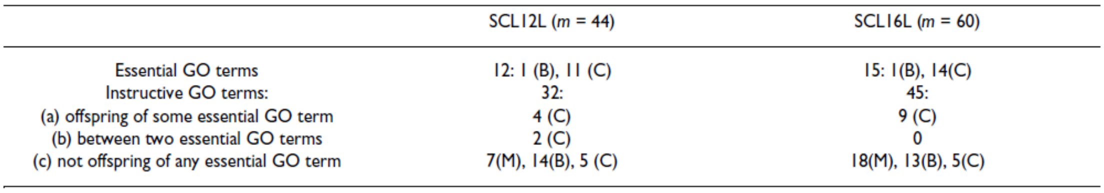


Rank by MED	GO term	Branch	MED	Rank by MED	GO term	Branch	MED
1	**GO:0005634**	C	331.7	31	GO:0005525	M	56.5
2	**GO:0009507**	C	244.2	32	GO:0005789	C	55.6
3	GO:0016020	C	148.1	33	GO:0004725	M	55.6
4	**GO:0005739**	C	147.8	34	GO:0008270	M	54.6
5	GO:0001844	B	70.7	35	GO:0015031	B	54.1
6	GO:0005212	M	70.5	36	GO:0005813	C	52.2
7	**GO:0005886**	C	68.5	37	GO:0005524	M	51.8
8	GO:0006094	B	67.9	38	GO:0051536	M	51.8
9	GO:0045261	C	67.5	39	GO:0016702	M	51.1
10	**GO:0009536**	C	66.6	40	GO:0005887	C	48.4
11	GO:0007010	B	65.8	41	GO:0005905	C	46.1
12	**GO:0030198**	B	65.4	42	**GO:0005794**	C	43.6
13	GO:0009626	B	65.4	43	GO:0005622	C	43.4
14	GO:0005047	M	64.6	44	GO:0005759	C	43.1
15	GO:0017134	M	64.1	45	**GO:0005856**	C	42.1
16	GO:0000287	M	63.3	46	GO:0016757	M	41.7
17	GO:0006888	B	61.7	47	**GO:0005618**	C	41.3
18	GO:0030234	M	61.7	48	**GO:0005783**	C	35.7
19	GO:0007323	B	61.0	49	**GO:0005773**	C	31.2
20	GO:0008083	M	60.8	50	**GO:0009842**	C	27.2
21	GO:0004521	M	60.3	51	GO:0006811	B	25.3
22	GO:0003723	M	60.2	52	GO:0006350	B	24.3
23	GO:0009514	C	59.9	53	GO:0016740	M	23.1
24	GO:0020015	C	59.4	54	**GO:0005737**	C	20.2
25	GO:0000922	C	59.0	55	GO:0005829	C	19.6
26	GO:0005681	C	58.9	56	GO:0016798	M	15.5
27	GO:0030149	B	57.9	57	GO:0007186	B	14.8
28	GO:0000917	B	56.9	58	GO:0019843	M	13.0
29	GO:0009405	B	56.8	59	**GO:0005764**	C	11.8
30	GO:0005615	C	56.5	60	**GO:0005777**	C	10.2

5.2.2.5. Analysis of informative GO terms

The GOmining method identifies a feature set of m effective GO terms, called informative GO terms, to design an accurate SVM-based prediction method. Table 5.2.4 shows the distribution of the m informative GO terms in the GO graph. For SCL12L with m = 44, GOmining selected 12 essential GO terms and 32 instructive GO terms. The 32 instructive GO terms consist of 7 GO terms from the molecular function branch, 14 terms from the biological process branch, and 11 terms from the cellular component branch, denoted as 7(M), 14(B) and 11(C), respectively. Analytical results reveal that all the three branches contain instructive GO terms.

Table 5.2.4 Distribution of the *m* informative GO terms. Most instructive GO terms (80%) are not offspring of the essential GO terms that the ratios are 26/32 and 36/45 for SCL12L and SCL16L, respectively.



	SCL12L (*m* = 44)	SCL16L (*m* = 60)
Essential GO terms	12: 1 (B), 11 (C)	15: 1(B), 14(C)
Instructive GO terms:	32:	45:
(a) offspring of some essential GO term	4 (C)	9 (C)
(b) between two essential GO terms	2 (C)	0
(c) not offspring of any essential GO term	7(M), 14(B), 5 (C)	18(M), 13(B), 5(C)

Due to the high correlation among GO terms in the GO graph, the feature selection of SVM should consider simultaneously a set of informative GO terms, rather than individual GO terms. Since the essential GO terms are always included, GOmining benefits from a confined search space of candidate instructive GO terms.

Considering the position relationships between instructive and essential GO terms in the GO graph, instructive GO terms belonged to one of the three classes: (a) offspring but not ancestor of some essential GO term; (b) between two essential GO terms, and (c) not offspring of any essential GO term. Of the 32 instructive GO terms, 4, 2 and 26 GO terms belonged to the classes (a), (b) and (c), respectively. The 26 GO terms consist of 7(M), 14(B) and 5(C). The GO terms near the root of the GO graphs are considered to be more generic while terms near the leaves are more specific [23]. Of the instructive GO terms, 81.2% (26/32) were not offspring of any essential GO term. These analytical results reveal that the essential GO terms are informative enough in predicting subcellular localization, and are effective in confining the space of searching instructive GO terms. The other six instructive GO terms from the cellular component branch have more specific functions than the essential GO terms in discrimination of the subcellular localization.

Figures 5.2.2, 5.2.3, 5.2.4 illustrate some of the instructive GO terms belonging to the three classes. Three instructive GO terms were found to belong to class (a), namely SCL12L: GO:0031227 (Intrinsic to endoplasmic reticulum membrane, rank 11), GO:30662 (Coated vesicle membrane, rank 41) and GO:0017119 (Golgi transport complex, rank 21), according to Fig. 5.2.2. The two terms belonging to class (b), namely GO:0005815 (Microtubule organizing center, rank 25) and GO:0005813 (Centrosome, rank 36), were found between the essential GO terms GO:0005856 (Cytoskeleton) and GO:0005814 (Centriole), as shown in Fig. 5.2.3. According to Fig. 5.2.4, five instructive GO terms belonging to the class (c) were not offspring of essential GO terms, GO:0016021 (Integral to membrane, rank 3), GO:0005576 (Extracellular region, rank 4), GO:0005622 (intracellular, rank 18), GO:0005578 (Proteinaceous extracellular matrix, rank 37) and GO:0005615 (Extracellular space, rank 38).

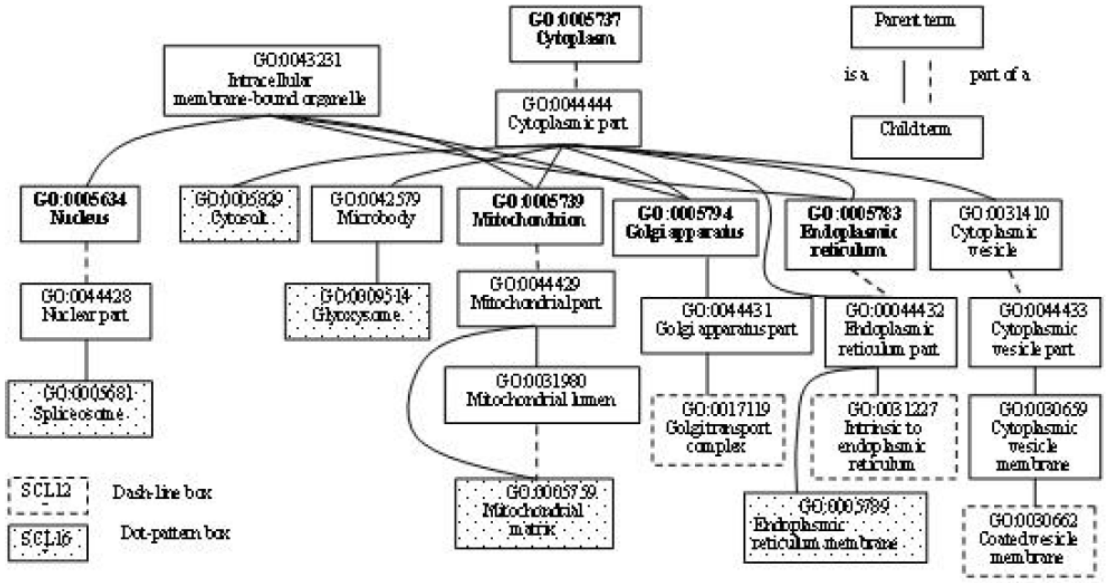


Figure 5.2.2 Some of the selected GO terms which are offspring of essential GO terms. For SCL12L, there are three terms shown: GO:0031227, GO:0030662 and GO:0017119. For SCL16L, five GO terms are shown: GO:0009514, GO:0005681, GO:0005789, GO:0005759 and GO:0005829

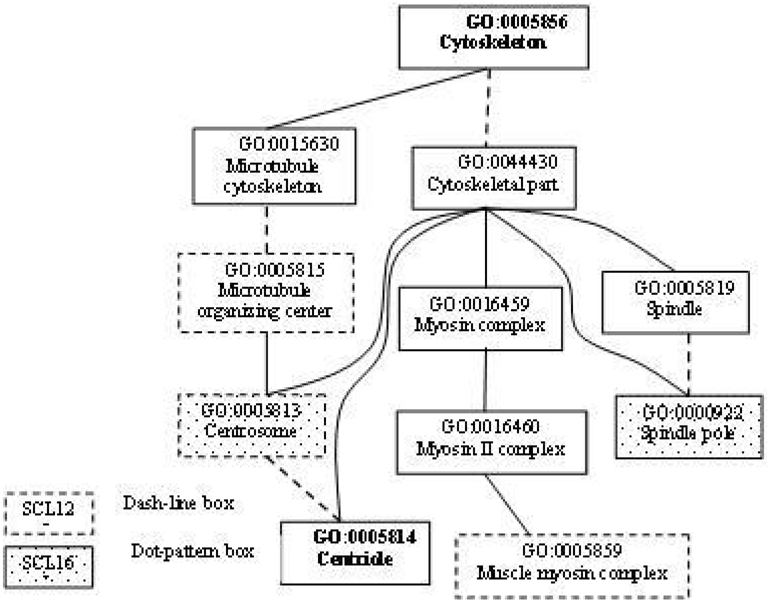


Figure 5.2.3 Some of the selected GO terms which are between two essential GO terms. For SCL12L, the two instructive GO terms GO:0005815 and GO:0005813 are between the essential GO terms GO:0005856 and GO:0005814. For SCL16L, GO:0005813 and GO:0000922 are offspring of the essential GO term GO:0005856, belonging to the class (a). GO:0005814 is not an essential GO term for SCL16L.

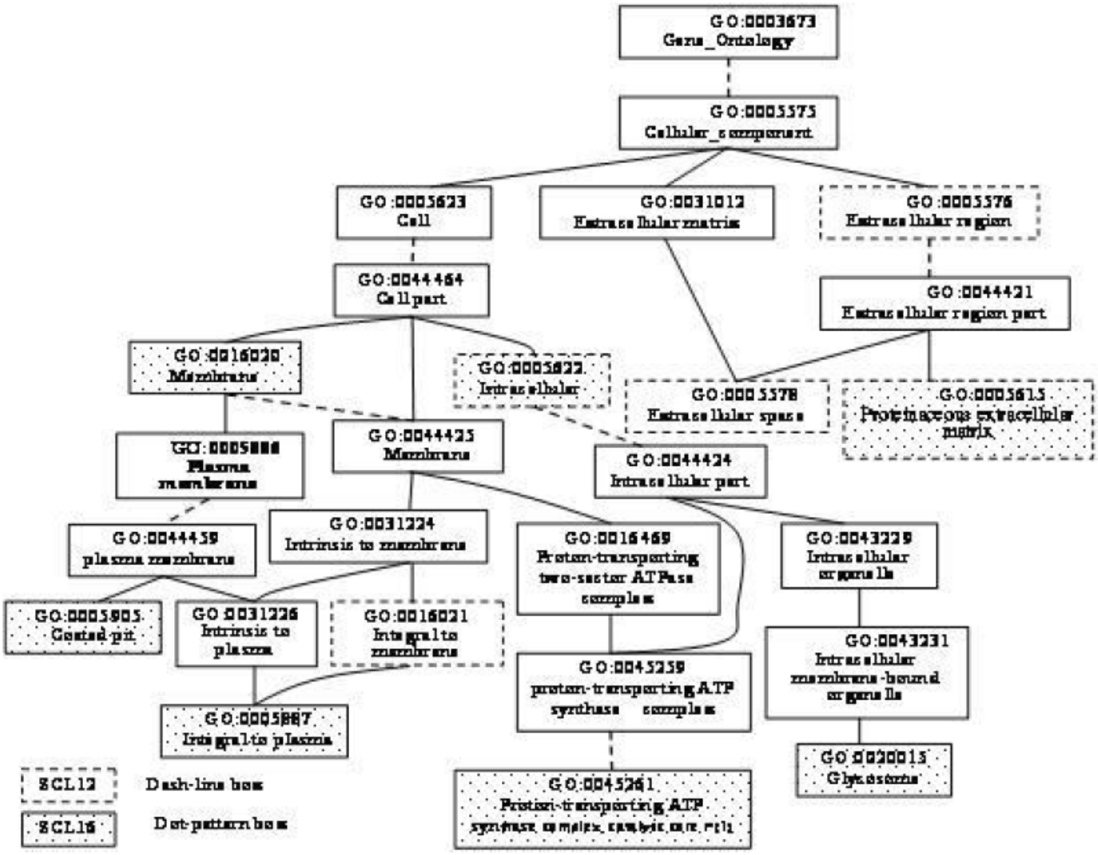


Figure 5.2.4 Some of the selected GO terms are NOT offspring of any essential GO terms. For SCL12L, five instructive GO terms are shown belonging to cellular component branch: GO:0016021, GO:0005576, GO:0005622, GO:0005578 and GO:0005615, which are not offspring of essential GO terms. For SCL16L, five GO terms belonging to the class (c) are shown: GO:0005622, GO:0005615, GO:0020015, GO:0016020 and GO:0045261. GO:0005905 and GO:0005887 belong to the class (a).

The m = 60 informative GO terms for SCL16L comprises 15 essential GO terms and 45 instructive GO terms. The 45 instructive GO terms consisted of 18(M), 13(B) and 14(C). The numbers of instructive GO terms coming from each branch were not significantly different. However, the numbers of instructive GO terms belonging to the three classes (a), (b) and (c) are 9, 0 and 36, respectively, which are very different. 80% (36/45) of the instructive GO terms were not offspring of any essential GO term. The 9 instructive GO terms belonging to the class (a) had 5, 2 and 2 terms, respectively, as shown in Figs. 3, 4 and 5. Class (c) has five GO terms with a dot-pattern box: GO:0005622 (intracellular), GO:0005615 (Extracellular space), GO:0020015 (Glycosome), GO:0016020 (Membrane) and GO:0045261 (Proton-transporting ATP synthase complex, catalytic core F(1)), as revealed by Fig. 5.

5.2.2.6. Predicting system of protein subcellular localization

Computational prediction methods from primary protein sequences are fairly economical in terms of identifying large-scale eukaryotic proteins with unknown functions. The GO annotation, which describes the function of genes and gene products across species, has been used to improve the prediction of protein subcellular localization. The accession numbers of proteins are necessary to query the GOA database to obtain GO terms. Since novel proteins have no known accession numbers, BLAST was used to obtain homologies with known accession numbers to the proteins for the retrieval of GO terms.

GO annotation has grown in size and popularity. However, few studies have explored informative GO terms from the over 20,000 annotations available at present for sequence-based prediction problems. This study proposes a genetic algorithm based method, GOmining, which combines SVM to simultaneously identify a small number m out of the n GO terms as features to SVM, where m <<n. The m GO terms include the essential GO terms annotating subcellular compartments such as GO:0005634 (Nucleus), GO:0005737 (Cytoplasm) and GO:0005856 (Cytoskeleton). ProLoc-GO was evaluated using SVM with the GO-based features from two kinds of input data, sequence and known accession numbers of proteins.

ProLoc-GO was significantly superior to the other methods. And analysis of m informative GO terms in the GO graph reveals that GOmining can consider internal relevant-feature correlation, rather than individual features, by using an efficient global optimization method. GOmining can serve as an efficient tool for mining informative GO terms for various sequence-based predictions of proteins, especially when the GO database grows fast. The prediction system using ProLoc-GO with protein sequence as input data for protein subcellular localization has been implemented.

5.2.3. S-protein – determining the transport ability of proteins

5.2.3.1. Introduction

The prediction of non-classical secreted proteins is a significant problem for drug discovery and development of disease diagnosis. Secreted proteins are secreted from cells into the extracellular space. The secretory process can be classified into two categories: classical and non-classical pathways. In the classical pathway, proteins with signal peptides are processed and transported to the outside of the cells. The signal peptides are usually located in the N-terminal of proteins. In eukaryotic cells, newly translated secreted proteins pass by endoplasmatic reticulum (ER) and Golgi and form secretory vesicles to fuse with the cell membrane [53]. In addition to the classical secretory pathway, there is non-classical secretory pathway which is known as leaderless secretion [54].

Since the characteristic of leaderless secretion makes the prediction of non-classical proteins more difficult and complicated than the classical secreted proteins. We identify a set of informative physicochemical properties of amino acid indices cooperated with support vector machine (SVM) to find discrimination between secreted and non-secreted proteins and to predict non-classical secreted proteins. When the sequence identity of dataset was reduced to 25%, the prediction accuracy on training dataset is 85% which is much better than the traditional sequence similarity-based BLAST or PSI-BLAST tool. The accuracy of independent test is 82%. The most effective features of prediction revealed the fundamental differences of physicochemical properties between secreted and non-secreted proteins. The interpretable and valuable information could be beneficial for drug discovery or the development of new blood biochemical examinations.

5.2.3.2. Analysis of selected feature sets

We analyzed the 30 feature sets of independent IBCGA experiments. IBCGA can select m features from 531 physicochemical properties. The result showed that the number of m is between 13 and 43 (data not shown). It is necessary to design a scoring strategy to select the best set from 30 runs. We hypothesize that if a feature is selected by IBCGA repeatedly, it was considered more significant than other features for the classification of the non-secreted and secreted proteins. Based on this hypothesis, we developed an evaluating strategy to choose the best set of features from the 30 sets of features. The frequency of features selected among 30 runs of training experiments was used as the score of each feature to calculate the score of each set of features. The features with high score (>9) are listed in Table 5.2.5.

Table 5.2.5 Results of IBCGA feature selection of 30 runs

Feature ID	Times	Description
420	19	Normalized positional residue frequency at helix termini C" (Aurora-Rose, 1998)
202	17	AA composition of CYT of single-spanning proteins (Nakashima-Nishikawa, 1992)
192	12	Normalized composition of mt-proteins (Nakashima *et al.*, 1990)
221	12	Optimized average non-bonded energy per atom (Oobatake *et al.*, 1985)
196	11	Normalized composition from fungi and plant (Nakashima *et al.*, 1990)
55	10	Normalized hydrophobicity scales for beta-proteins (Cid *et al.*, 1992)
403	10	Normalized positional residue frequency at helix termini N4'(Aurora-Rose, 1998)
13	9	Retention coefficient in HFBA (Browne *et al.*, 1982)
23	9	Free energy of solution in water, kcal/mole (Charton-Charton, 1982)
89	9	Negative charge (Fauchere *et al.*, 1988)
189	9	AA composition of total proteins (Nakashima *et al.*, 1990)
273	9	Weights for beta-sheet at the window position of -4 (Qian-Sejnowski, 1988)

The total score of each set was divided by the number of features of each set to calculate the average score of each set of features. The top 10 feature sets are listed in Table 5.2.6. The highest average score S_r was 6.7. The feature set with the highest score contains 20 features which are listed in Table 5.2.7. Furthermore, the feature set with the highest average score was used for the independent test. The overall accuracy was 81.94%. The accuracies for the nonsecreted and secreted proteins were 90.07% and 68.60% respectively and the MCC is 0.61.

Table 5.2.6 Results of feature sets scoring

Run No.	No. of features	Score	Score/No.	Ranking of score/No.
17	20	134	6.70	1
9	21	135	6.43	2
4	15	96	6.40	3
14	19	119	6.26	4
23	20	120	6.00	5
28	22	129	5.86	6
1	28	159	5.68	7
18	26	146	5.62	8
2	13	73	5.62	8
30	30	161	5.37	10

Table 5.2.7 Selected features of run 17

Feature ID	Times	Description
5	3	Conformational parameter of inner helix (Beghin-Dirkx, 1975)
13	9	Retention coefficient in HFBA (Browne *et al.*, 1982)
89	9	Negative charge (Fauchere *et al.*, 1988)
99	2	Alpha-helix indices for beta-proteins (Geisow-Roberts, 1980)
101	1	Beta-strand indices (Geisow-Roberts, 1980)
166	1	Frequency of occurrence in beta-bends (Lewis *et al.*, 1971)
190	5	SD of AA composition of total proteins (Nakashima *et al.*, 1990)
196	11	Normalized composition from fungi and plant (Nakashima *et al.*, 1990)
221	12	Optimized average non-bonded energy per atom (Oobatake *et al.*, 1985)
237	3	Normalized frequency of turn in alpha+beta class (Palau *et al.*, 1981)
246	7	Surrounding hydrophobicity in turn (Ponnuswamy *et al.*, 1980)
273	9	Weights for beta-sheet at the window position of -4 (Qian-Sejnowski, 1988)
328	4	Relative preference value at N4 (Richardson-Richardson, 1988)
349	7	Information measure for C-terminal turn (Robson-Suzuki, 1976)
357	8	Loss of Side chain hydropathy by helix formation (Roseman, 1988)
380	8	Bitterness (Venanzi, 1984)
403	10	Normalized positional residue frequency at helix termini N4'(Aurora-Rose, 1998)
407	2	Normalized positional residue frequency at helix termini Nc (Aurora-Rose, 1998)
420	19	Normalized positional residue frequency at helix termini C" (Aurora-Rose, 1998)
502	4	Buriability (Zhou-Zhou, 2004)

The contribution of each feature in the feature set with the highest average score was evaluated with two methods. First, each of 20 features was removed to evaluate the reduction of accuracy (Figure 5.2.5). The decreased accuracies were between 11.56% and 12.26%. Besides, the main effect difference (MED) was also used to analyze the importance of each feature (Figure 5.2.6). The reduced accuracies were between 0.49% and 9.84%. The difference between the results of the two methods is due to the combination of features in different number.

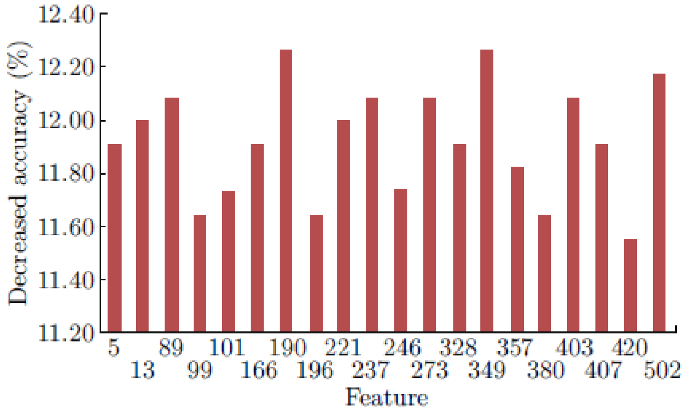


Figure 5.2.5 The significance of each feature selected in run 17 is analyzed by removing each feature orderly to observe the reduction of overall prediction accuracy

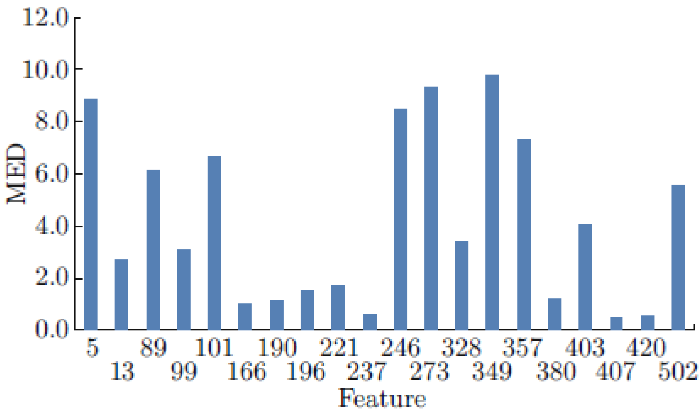


Figure 5.2.6 The significance of each feature selected in run 17 is analyzed by MED analysis

The extracted features from 30 independent IBCGA runs are with strong robustness. Feature 420 and 202 were selected 19 and 17 times respectively. The discrimination ability of only one feature is showed in figure 5.2.7. Obviously, it is difficult to separate the two classes of proteins according to one feature only. But, one small set of features (m = 13) selected by our IBCGA can achieved an accuracy of 82%.

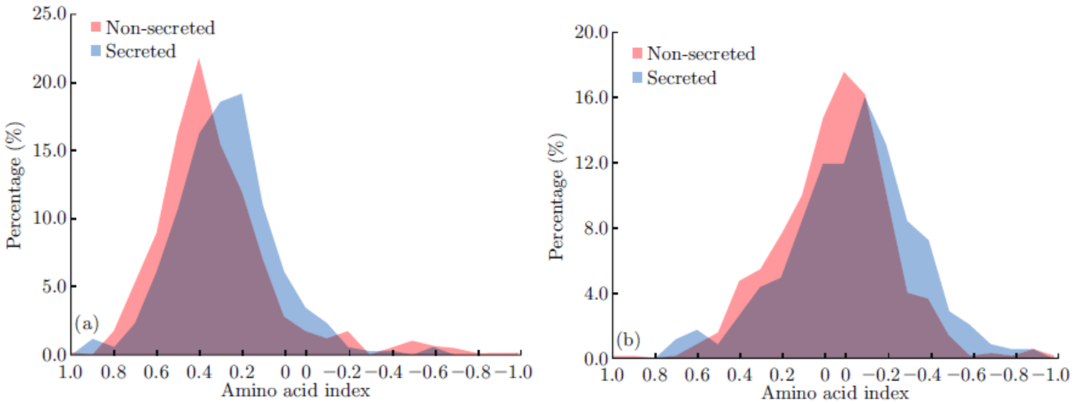
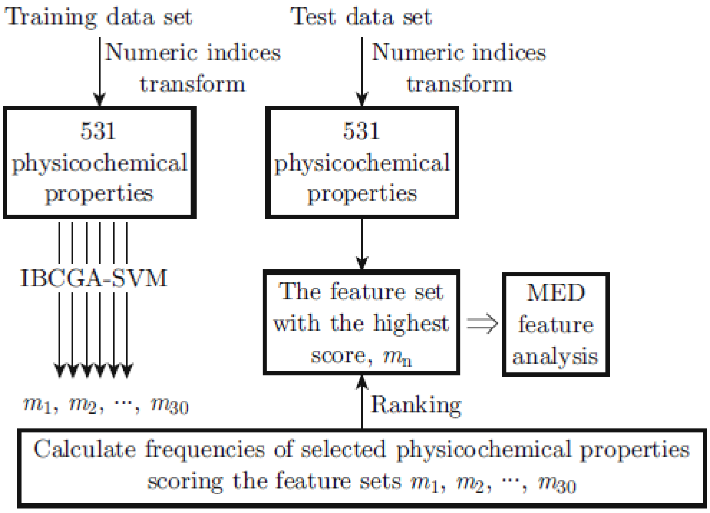


Figure 5.2.7 The distribution of feature (a) 420 and (b) 202. The value is normalized to 1 and −1

Some of the informative physicochemical properties extracted by our algorithm are found to be important in the process of protein secretion [55, 56]. By the analysis of MED, the feature 349 which describes the Information measure for C-terminal turn is the most important. In contrast to the role that N-terminal signal played in the classical secretory pathway, the C-terminal signal may play a more important role in the non-classical secretory pathway. The extracted properties are also relatively more interpretable for biologists. Some of these informative features need to be investigated more detailed

5.2.3.3. Prediction of the secreted proteins

The system flowchart of the prediction method is shown in figure 5.2.8. The selected m physicochemical properties and the associated parameter set of SVM by using IBCGA are used to predict the test data set. The selected physicochemical properties were analyzed to further understand the special properties that are unique for secreted proteins. The prediction system including four parts described as follows:



Training data set
Test data set
Numeric indices transform
Numeric indices transform
531 physicochemical properties
531 physicochemical properties
IBCGA-SVM
The feature set with the highest score, m_n
MED feature analysis
m_1, m_2, ···, m_{30}
Ranking
Calculate frequencies of selected physicochemical properties scoring the feature sets m_1, m_2, ···, m_{30}

Figure 5.2.8 Flowchart of generating feature vectors of instructive GO terms and essential GO terms

1) Numeric Transformation: The sequences in the training data set were transformed into 531-dimensional vectors of numerical values using the AAIndex.

2) Feature selection: IBCGA is performed 30 independent runs where each run the training data set is used as the training data set of 5-CV. There are total 30 sets of m physicochemical properties for each of independent data sets.

3) Scoring the feature sets: The frequencies F(Pi) of the selected physicochemical properties in each feature set were added together and then divided by the number of m to obtain the score Sr for each solution.

$$S_r = (\sum_{i=1}^{m} F(P_i))/m$$

4) Independent test: The set of selected physicochemical properties with a maximal value of Sr was used to calculate the performance of the prediction system.

5) Feature analysis: Each feature in the set of selected physicochemical properties with a maximal value of Sr was analyzed by MED analysis to clarify the importance of each feature. Our method will automatically determine a set of informative physicochemical properties and a SVM model for prediction of secreted

proteins. The robust and informative physicochemical properties were extracted from 30 independent runs of IBCGA to predict the non-classical secreted proteins.

5.3. Detecting the functions of peptide – Virulent-GO

Virulent-GO is employed at this stage. Virulent-GO predicts the function and virulent of peptides and decides the ability of immunogenicity. This stage can select the candidates of vaccine and furthermore the analysis of antigen-binding site can be effective.

5.3.1. Introduction

The identification of novel virulence determinants is a key step of the process to understand how pathogenic bacteria interact with their hosts to produce clinical disease [57]. Multiple virulence factors in bacterial pathogens serve separately or are cooperated each other during a course of stages to infect susceptible hosts. The generic mechanisms shared by these bacterial virulence factors and themselves are adequately discussed in a previous review [58]. These bacterial virulence factors may also serve as targets for vaccine and drug.

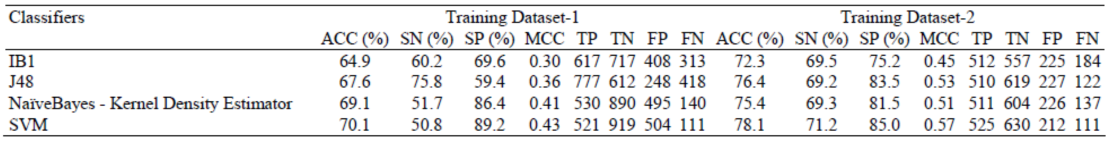
In this study, we propose a sequence-based method Virulent-GO by mining informative GO terms as features for predicting bacterial virulent proteins. The sequences of bacterial pathogens were obtained from SWISS-PROT [59] and VFDB [60]. All the instructive GO terms of these sequences were obtained by using BLAST [49] to obtain its homologies with known accession numbers which are used to query the GOA database [61] consequently. The potential for GO terms to discriminate virulent proteins in bacteria has been demonstrated by distinct differences between virulent and non-virulent proteins. All keywords retrieving from literatures [58] which are associated with categories of virulence factors are also annotated by GO terms. All the GO terms appearing in both sets of instructive GO terms and the GO terms from keywords are denoted as essential GO terms. A point of integrative view from the instructive GO term set and the essential GO term set can reveal a few nature of complexity from virulence factors in bacterial pathogens.

5.3.2. Assessment of Features and Classifiers

To evaluate performance across widely-used classifiers, this study applied four kind of classifiers that are IBk (k-nearest neighbor), J48 (Decision Tree), NaïveBayes and SVM. With five-fold cross-validation, this turned out a strong support for the predictive power orientated form instructive GO terms. The accuracy was archived up to 82.5% (SVM), 80.0% (J48) and 79.5% (NaïveBayes). Even a lazy classifier IBk like could make out an accuracy of 78.6%. On the others hand, using instructive GO terms set to classify virulent proteins turned out a better performance (Accuracy 82.5%) compared to several generic features that are amino acids composition (72.1%),dipeptide composition (71.1%), similarity search (52.1%) and PSSM profile (78.1%).

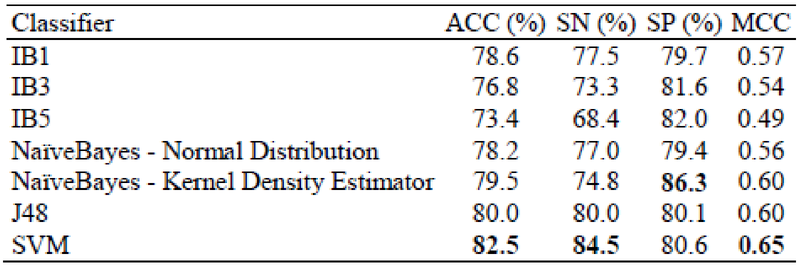
The five-fold cross-validation scheme is also used to evaluating performance for Training Dataset-1 and Training Dataset-2 by combining with widely-used classifiers to demonstrate the efficiency of essential GO terms. Due to 288 non-virulent and 289 virulent proteins have no essential GO terms annotated, they could be recognized as a same class and lead to a lot of false positives or false negatives. These results could be seen from Table 5.3.1. After excluding these proteins, the accurate rate just a little drop against results from training dataset which is annotated by instructive GO terms. These results are shown in Table 5.3.2.

Table 5.3.1 The result of evaluating on only essential GO terms is included with five-fold cross-validation



Classifiers	Training Dataset-1								Training Dataset-2							
	ACC (%)	SN (%)	SP (%)	MCC	TP	TN	FP	FN	ACC (%)	SN (%)	SP (%)	MCC	TP	TN	FP	FN
IB1	64.9	60.2	69.6	0.30	617	717	408	313	72.3	69.5	75.2	0.45	512	557	225	184
J48	67.6	75.8	59.4	0.36	777	612	248	418	76.4	69.2	83.5	0.53	510	619	227	122
NaïveBayes - Kernel Density Estimator	69.1	51.7	86.4	0.41	530	890	495	140	75.4	69.3	81.5	0.51	511	604	226	137
SVM	70.1	50.8	89.2	0.43	521	919	504	111	78.1	71.2	85.0	0.57	525	630	212	111

Table 5.3.2 Training results of instructive GO terms as feature performed on multiple classifiers with five-fold cross-validation

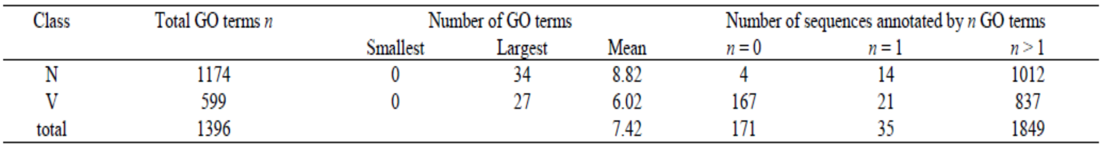


Classifier	ACC (%)	SN (%)	SP (%)	MCC
IB1	78.6	77.5	79.7	0.57
IB3	76.8	73.3	81.6	0.54
IB5	73.4	68.4	82.0	0.49
NaïveBayes - Normal Distribution	78.2	77.0	79.4	0.56
NaïveBayes - Kernel Density Estimator	79.5	74.8	**86.3**	0.60
J48	80.0	80.0	80.1	0.60
SVM	**82.5**	**84.5**	80.6	**0.65**

5.3.3. Analyzing Instructive GO terms and Essential GO Terms

In integrative views with instructive GO terms set and essential GO terms set, the training datasets that are constructed by non-virulent and virulent protein sequences in bacterial pathogens are well-annotated and informative by these two sets. Non-virulent proteins share more diversity of GO terms (1174) to virulent proteins (599) that is shown in Table 5.3.3.

Table 5.3.3 Results of instructive GO annotation for all sequences



Class	Total GO terms *n*	Number of GO terms			Number of sequences annotated by *n* GO terms		
		Smallest	Largest	Mean	*n* = 0	*n* = 1	*n* > 1
N	1174	0	34	8.82	4	14	1012
V	599	0	27	6.02	167	21	837
total	1396			7.42	171	35	1849

Proteins which are recognized as non-virulent in bacterial pathogens annotate with more GO terms (8.82) than virulent proteins (6.02). There are 167 virulent proteins annotated with no GO terms from their homology while only 4 non-virulent proteins have no annotated GO term. In contract to the instructive GO terms, the numbers are similar for non-virulent proteins (288) and virulent proteins (289) annotated without any essential GO terms. Although a wider range of essential GO terms (65 to 60 for virulent proteins) is used to annotated non-virulent proteins, the virulent proteins are annotated by more essential GO terms (2.04) than non-virulent proteins (1.64) in average amount. Moreover, a large numbers of virulent proteins were annotated by several GO terms. This trend could be seen from a frequency-distribution in figure 5.3.1. A clear difference was shown since essential GO terns g get larger than 3.

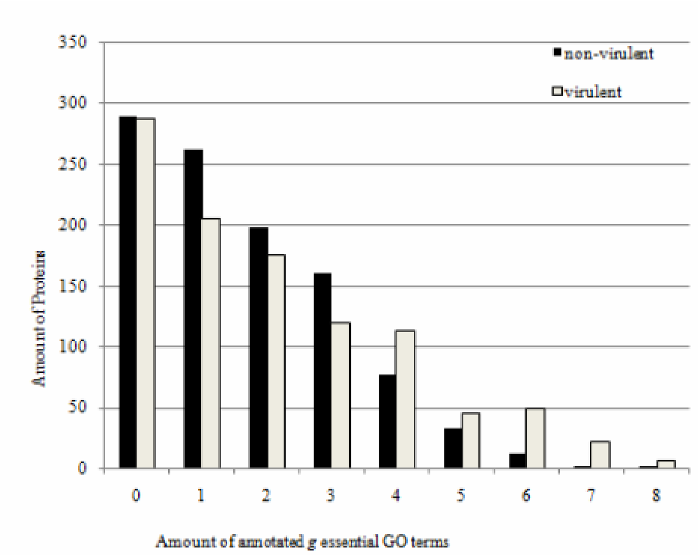


Figure 5.3.1 The number of essential GO terms annotated in each protein is shown in this frequency distribution graph

Most of keywords are successfully accessing to both non-virulent proteins and virulent proteins via retrieving some essential GO terms. Many of they even access hundreds of proteins. The essential GO terms set is constructed across three major branches, and 52 essential GO terms still are shared by both non-virulent proteins and virulent proteins. Thus, a proper classifier should be applied to archive a successful prediction. Although keywords like “Colonization“, “Iron acquisition” and “PhoP/PhoQ two component system” assess to 0 proteins for no GO term own by them could be recognized as a essential GO term, a typical example that is catered to “Iron acquisition”“Siderophore receptor” is querying and access to few proteins. Also, “PhoP/PhoQ two component system” and “ABC transport system” are in the same situation. Besides, there are two keywords retrieved certain GO terms but were failure to intersecting with instructive GO terms. They are “Immune response inhibitor” and “Biofilm”.

5.3.4. Virulent-GO

The design of Virulent-GO is a two-stage approach to classifying virulent proteins in bacterial pathogens utilizing the single kind of GO term features. At the first stage, sequences in the given training dataset are used to obtain their homologies by using BLAST. The accession numbers of homologies were used to query the GOA database to obtain a set of instructive GO terms. All sequences in the training dataset are represented as a vector of instructive GO terms. Additionally, a set of essential GO terms is collected. The flowchart of generating feature vectors of instructive GO terms and essential GO terms is shown in figure 5.3.2. At the second stage, a good classifier for utilizing the instructive GO terms is determined by evaluating some widely-used classifiers. The high-performance classifier determined is further evaluated using an independent test dataset.

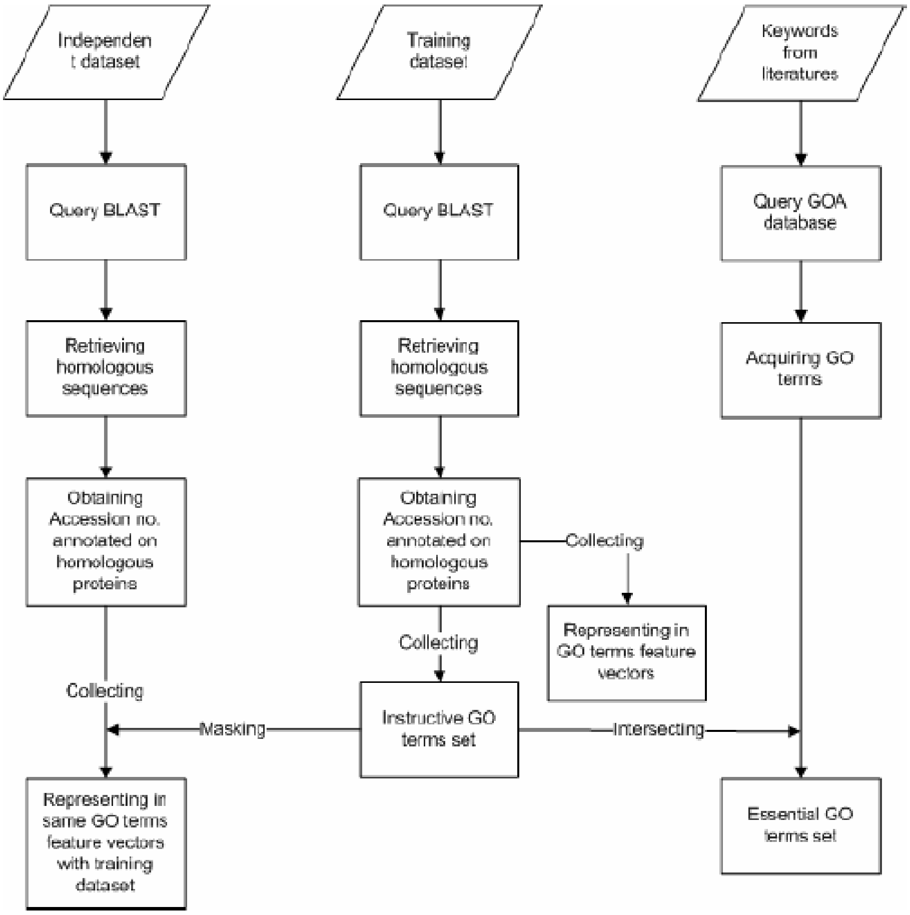


Figure 5.3.2 Flowchart of generating feature vectors of instructive GO terms and essential GO terms

This study proposed an efficient method utilizing instructive GO terms to predict virulent proteins in bacterial pathogens. This method performs well across popular classifiers and also has a significantly better performance than applying features like compositions and evolutionary information. Compared to the existing method, VirulentPred, there is a slight better performance in training that may results from bias originated from applying k-fold cross-validation. While performing on independent test dataset, the Virulent-GO still has a little improvement.

By cooperating instructive GO terms set with some popular features, the performance could be further improved. Furthermore, the ranking of GO terms in the contribution of prediction and a set of interpretable prediction rules provide valuable information for more understanding in a complex virulence mechanism in bacterial pathogens.

5.4. Prediction of T-cell immune response

The final stage is the core of computer-aided vaccine design system. System refers the selected biological features, such as physicochemical properties, GO terms, and sequence motifs, to determine the prediction model. The suitable prediction tools to predict the residue binding site accurately and immunogenicity. Finally, the computer-aided vaccine design system provide the possibility of immune reaction, peptide binding site on T cell, and further biological features to assist the vaccine design.

5.4.1. Related works

5.4.1.1. Highly Ubiquitylated Proteins as Antigen Sources

Ubiquitin-proteasome system is an important mechanism for protein degradation that the ubiquitylated proteins will be degraded by proteasome. The ubiquitin acts as a specific tag for marking proteins for degradation. The proteasome is a major mechanism for cells to regulate the concentration of particular proteins and degrade misfolded proteins. The degradation process produces short peptides of about 7~8 amino acids. The resulting short peptides can be further degraded into amino acids that can be used in protein synthesis [62, 63].

The proteasome plays an important role in the function of the adaptive immune system. The peptide antigens presented on the surface of antigen-presenting cells are produced by proteasomal degradation of pathogen proteins and displayed by MHC class I molecules [64]. A previous study investigated the role of ubiquitin-dependent proteolytic pathway in MHC class I-restricted antigen presentation and concluded that ubiquitin-conjugation (also called ubiquitylation) plays an important role in the presentation of a cytosolic antigen with MHC [65]. Another study found that an amino-terminal modification of a viral protein will promote ubiquitin-dependent degradation and lead to the enhancement of presentation with MHC class I [66].

Some recent studies have similar results that ubiquitin-conjugation will enhance the efficacy of polynucleotide viral vaccines [67] and vaccines against tuberculosis [68]. Another study claimed that the low frequency of memory cytotoxic T lymphocyte and inefficient antiviral protection of DNA immunization with minigenes can be rectified by ubiquitylation [69]. Therefore, accurate prediction of ubiquitylation sites can provide better understandings of ubiquitylation mechanism. The selection of highly ubiquitylated peptides can improve the effectiveness of vaccines. In Chapter 3.2, three kinds of features and three classifiers were assessed for their prediction performances. Subsequently, informative physicochemical property mining algorithm is applied to select informative physicochemical properties and improve the prediction performance. Finally, a prediction system UbiPred was constructed to predict ubiquitylation sites.

5.4.1.2. Immunogenic Pathway of MHC class I

Developing a computer-aided system to design peptide vaccines is one goal of immunoinformatics. The major work of previous studies for peptide vaccine designs is to identify cytotoxic T lymphocyte (CTL) epitopes and investigate their corresponding immunogenicity. The CTL cells play a critical role in protective immunity by recognizing and eliminating self-altered cells, which recognize short peptides derived from intracellular degradation of foreign proteins in combination with major histocompatibility complex (MHC) class I molecules. The immunogenicity of MHC class I binding peptides is their ability to induce CTL responses. Accurate predictions of the CTL epitopes and their corresponding immunogenicity are critical in developing a computer-aided system for vaccine designs.

Direct approach to predicting the CTL epitopes has been studied initially but its accuracy is fairly low [70]. Instead, indirect approach to predicting the MHC-binding peptides is useful because peptides must be processed prior to inducing cellular immunogenicity. The recent studies of bioinformatics utilized the information about antigen processing pathway to predict the CTL epitopes. At first, the peptides are cleaved by proteasomal cleavage. Several studies elucidating the specificity of proteasome have been presented. To predict proteasomal cleavage sites, NetChop used a neural network method [71] and Pcleavage is based on a support vector machine (SVM) learning model [72].

After cleavage, peptide fragments are transported into endoplasmic reticulum by TAP which is the transporter associated with antigen processing. Some studies of investigating the TAP transport efficiency were presented such as the affinity prediction of TAP binding peptides using the cascade SVM [73] and the prediction of TAP transport efficiency of epitope precursors using a simple scoring matrix [74]. Finally, the peptide fragments that bound to MHC class I molecules are subsequently translocated to the cell surface, where these complexes may active CTL. Some methods have been developed to predict MHC class I binding affinity, such as the SVM-based SVMHC [75] and Gibbs sampling method [10]. Moreover, the hybrid approaches integrated the above-mentioned methods like the prediction of proteasomal cleavage, TAP transport efficiency and MHC binding to advance the prediction performance [76, 77].

The problem of predicting immunogenicity of MHC class I binding peptides is crucial to further identify highly immunogenic peptides. The selection of highly immunogenic peptides can save many experimental efforts and accelerate the developing progress. In Chapter 3.3, a prediction system POPI was developed to predict immunogenicity of MHC class I binding peptides. POPI performs better than alignment-based and affinity-based methods.

In Chapter 3.3.8, an improved prediction system POPISK was constructed to predict T-cell responses induced by HLA-A2-restricted peptides. POPISK using string kernels is useful for predict peptide immunogenicity and immunogenicity changes made by single residue modifications that is especially useful for optimizing peptide-based vaccines.

5.4.1.3. Immunogenic Pathway of MHC class II

The immunogenic pathway of MHC class II includes four steps. First, antigens are engulfed by endocytosis forming endosome. Second, endosome fuses with lysosome and is cleaved by peptidase in lysosome. Third, the peptide fragments bound to MHC class II will be translocated to cell surface. Finally, immune responses (also called immunogenicity) will be triggered when helper T lymphocyte (HTL) recognize non-self antigens presented by antigen presenting cell (APC). The activated HTL will induce the resting HTLs to proliferate and differentiate into memory cells or effector cells and provide specific help to CTL, B lymphocytes and phagocytic cells [78, 79].

Previous studies for predicting immunogenic pathway of MHC class II focus on the prediction of MHC class II-restricted peptides (qualitative methods) and the binding affinity of peptide-MHC complex (quantitative methods). Many methods are proposed to predict MHC class II binding peptides. The evolutionary algorithms including ant colony algorithms [80], evolutionary algorithms combined with artificial neural networks [81] and multi-objective evolutionary algorithms [82] are developed for optimizing a matrix for predicting binding affinity. Other methods including the neural network based methods [81, 83, 84], Bayesian neural networks [85], fuzzy neural networks [86], the hidden Markov model [87], Gibbs samplers [10], support vector machines [88-90] and alignment-based method SMM-align that is a stabilization matrix alignment method for predicting MHC class II binding affinity [91].

However, the problem of predicting immunogenicity of MHC class II binding peptides is also important to understand immunogenicity and design effective vaccines. In Chapter 3.3, a prediction system POPI-MHC2 based on informative physicochemical properties was developed to predict immunogenicity of MHC class II binding peptides. The informative physicochemical properties are mined by using the informative physicochemical property mining algorithm. This study shows similar results to POPI that the traditional affinity-based method and alignment-based methods are less effective than the proposed method POPI-MHC2.

5.4.2. Prediction of ubiquitylation sites

5.4.2.1. Introduction

Ubiquitylation (also called ubiquitination) is an important mechanism of post-translational modification that ubiquitin will be linked to specific lysine residues of target proteins by forming isopeptide bonds. Three enzymes including activating enzyme (E1), conjugating enzyme (E2), and ubiquitin ligase (E3) are involved in the ubiquitylation process. Another enzyme E4 can help to stabilize and extend polyubiquitin chain [92, 93]. The first discovered function of ubiquitylation is to target proteins for subsequent degradation by the ATP-dependent ubiquitin-proteasome system. Subsequently, many regulatory functions of ubiquitylation were discovered including the regulation of DNA repair and transcription, control of signal transduction, and implication of endocytosis and sorting [92, 93].

Because of the important regulatory roles of ubiquitylation, numerous methods were developed to purify ubiquitylated proteins [94]. Also, the growing number of studies of large-scale identification of ubiquitylated proteins and analysis of ubiquitin-related proteome reflect the importance of identifying ubiquitylation proteins and sites [95-100]. The three steps affinity purification, proteolytic digestion, and analysis using mass spectrometry were applied in most of these studies [101]. These works cost a lot of experimental efforts. Therefore, developing a prediction system using informative features from protein sequences can not only save experimental efforts but also provide insights into the mechanism of ubiquitylation.

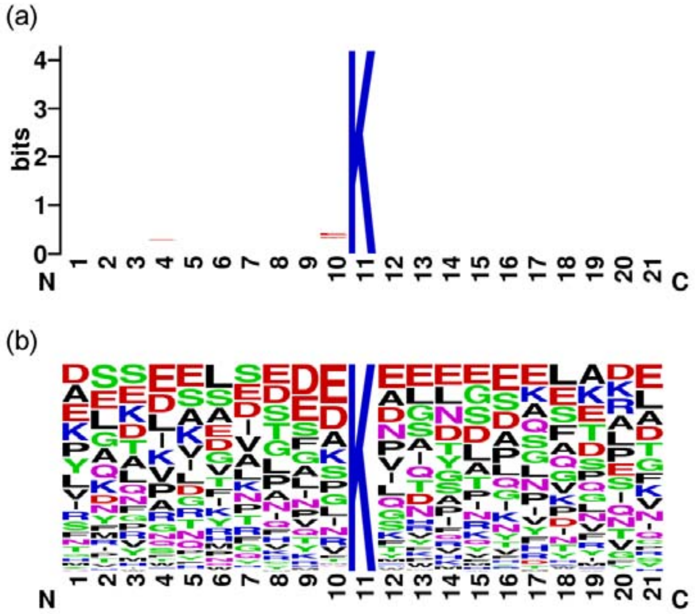


Figure 5.4.1 The sequence logo of the 151 positive samples with *w*=21. (a) information content and (b) frequency plot.

5.4.2.2. Assessment of features and classifiers

This study focuses on the sequence-based prediction of ubiquitylation sites. Therefore, three kinds of useful features which can be extracted from protein sequences and are widely used in bioinformatics studies are evaluated for prediction of ubiquitylation sites: conventional amino acid identity [102], evolutionary information [103, 104], and physicochemical property [33, 52]. For predicting functions of a residue in a protein, it is well recognized that nearby residues will influence the property and structure of a central residue. For machine learning based prediction methods, the environmental information will be useful to enhance prediction accuracy that is extensively used in previous studies [102-104]. The feature representations for applying to the mentioned classifiers are described below.

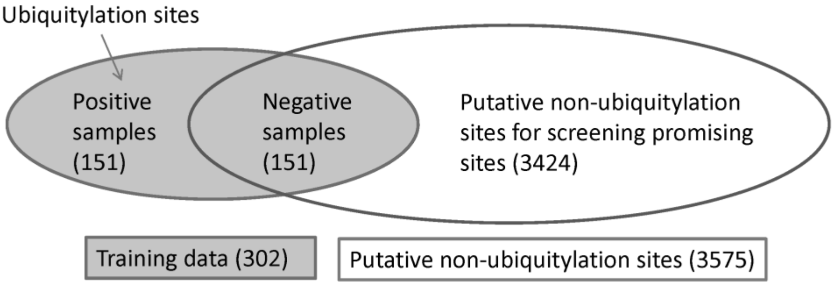

Ubiquitylation sites
Positive samples (151)
Negative samples (151)
Putative non-ubiquitylation sites for screening promising sites (3424)
Training data (302)
Putative non-ubiquitylation sites (3575)

Figure 5.4.2 The schema for the training and an independent of 3424 putative non-ubiquitylation sites in dataset of *w*=21.

The conventional feature representation, amino acid identity, uses 20 binary bits to represent an amino acid [102]. For example, the amino acid A is represented by '00000000000000000001' and R is represented by '00000000000000000010'. To deal with the problem of windows spanning out of N-terminal or C-terminal, one additional bit is appended to indicate this situation. A vector of size (20+1)*w* bits is used for representing a sample where *w* is the window size.

Evolutionary information has been successfully applied in many studies [103, 104]. To prepare evolutionary information for each protein sequence, the corresponding position-specific scoring matrix (PSSM) is obtained by applying PSI-BLAST [105] against non-redundant SWISS-PROT database using 3 iteration and default values of parameters. For each residue, there are 20 values indicating the probabilities of occurrences for 20 amino acids at the position. One additional bit is applied to deal with the terminal spanning windows as used for amino acid identity. A vector of size (20+1)w is used for representing a sample.

Using informative features as well as an appropriate classifier is essential to design an accurate prediction system. Three machine learning classifiers including k-nearest neighbor, NaïveBayes and support vector machine (SVM) are evaluated for predicting ubiquitylation sites. Two extensively used classifiers including IBk for k-nearest neighbor classifier and NaïveBayes classifier that are included in the machine learning tool of WEKA [106] are applied to evaluate prediction performances of features. To optimize the performance of IBk classifier, five numbers of nearest neighbors k=1, 3, …, 9 used to classify samples are evaluated for selecting the best number of k. For NaïveBayes, in addition to normal distribution, a distribution obtained from kernel estimation is used to model numeric attributes.

To find the best kind of feature for SVM-based prediction of ubiquitylation sites, the control parameters C and γ of SVM and associated window size $w \in \{11, 13, …, 29\}$ for each kind of features should be tuned to obtain best performance for comparison. The grid search method is applied to tune parameters C and γ that total 16*16=256 grids are evaluated. The prediction accuracy of 10-CV is used to determine the best parameter values for the three kinds of features for SVM.

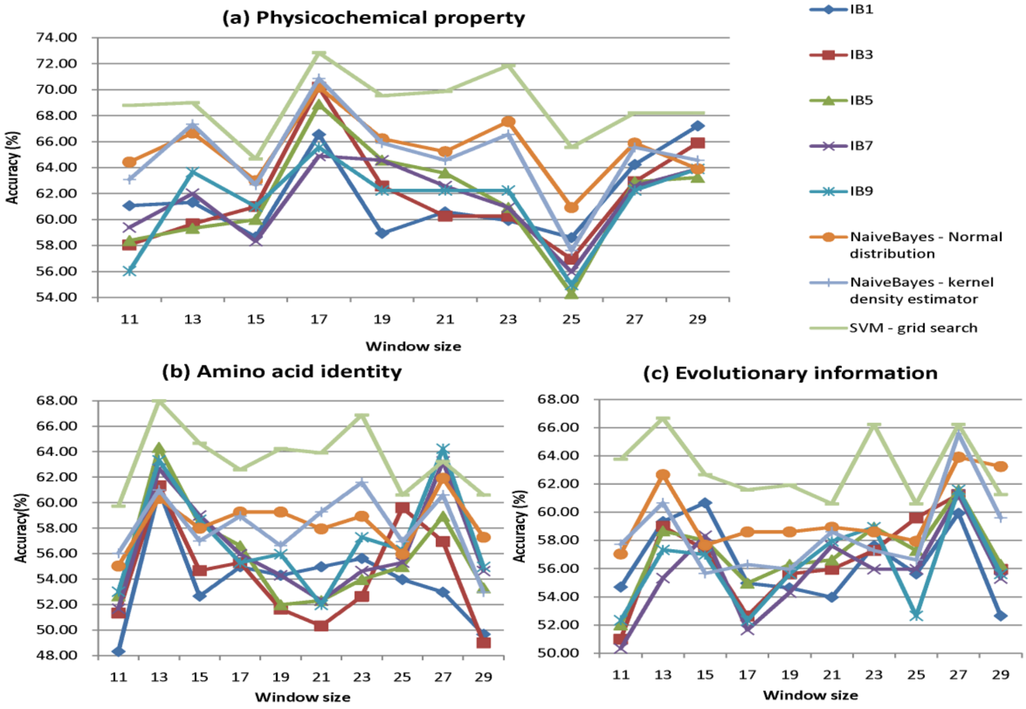


Figure 5.4.3 Performance comparisons among amino acid identity, evolutionary information and physicochemical property with various classifiers.

To evaluate the proposed methods, a positive dataset UBIDATA consisting of 157 ubiquitylation sites from 105 proteins was established by extracting annotated proteins from the UbiProt database [107]. By mapping the ubiquitylation sites to the corresponding 105 protein sequences retrieved from the UniProt Knowledgebase (Swiss-Prot and TrEMBL), the 3676 lysine residues with no annotation of ubiquitylation sites were regarded as putative non-ubiquitylation sites. A sliding window method is applied to the central residue to be predicted for gleaning environment information. A positive sample is denoted as a sequence of size *w* with a central residue lysine which is an ubiquitylation site. If the central residue lysine is not an ubiquitylation site, the sequence is regarded as a negative sample. Only one of the samples with the same sequences and annotation of ubiquitylation sites was used. All the inconsistent samples which have the same sequences but not the same annotation were discarded. The 10 positive datasets were constructed using various values of *w* from UBIDATA, which have 149 samples of *w*=11, 150 samples of *w*=13 and 15, and 151 samples of *w*=17, 19, …, 29. Due to the discard of duplicate and inconsistent samples, different values of *w* would result in different sample numbers of datasets.

For training an SVM classifier, both positive and negative samples are necessary. The dataset of post-translational modification including phosphorylation and ubiquitylation sites is unbalanced that the number of positive samples is much smaller than that of negative samples. The negative samples for training the SVM classifier were selected randomly from the 3676 putative non-ubiquitylation sites. Notably, since the value of *C* for tuning the error penalty (see the next section) is determined subsequently according to the performance measurement of SVM, it is not obligatory to select a matched number of negative peptides for training the SVM classifier. The used datasets of various windows sizes can be publicly downloaded from the web server of UbiPred.

#	Feature	Window size	C	γ	ACC (%)	SEN (%)	SPE (%)	MCC	AUC
1	31 Informative physicochemical properties (UbiPred)	21	4	2^{-1}	84.44	83.44	85.43	0.69	0.85
2	All physicochemical properties	17	1	2^{-4}	72.19	70.86	73.51	0.44	0.74
3	Amino acid identity	13	2	2^{-2}	65.67	57.33	74.00	0.32	0.70

4 Evolutionary information	13	1	2^{-7}	66.33	72.00	60.67	0.33	0.71

Table 5.4.1 Summary of used parameters and LOOCV performances of the methods using informative physicochemical properties (UbiPred), amino acid identity, evolutionary information, and all physicochemical properties.

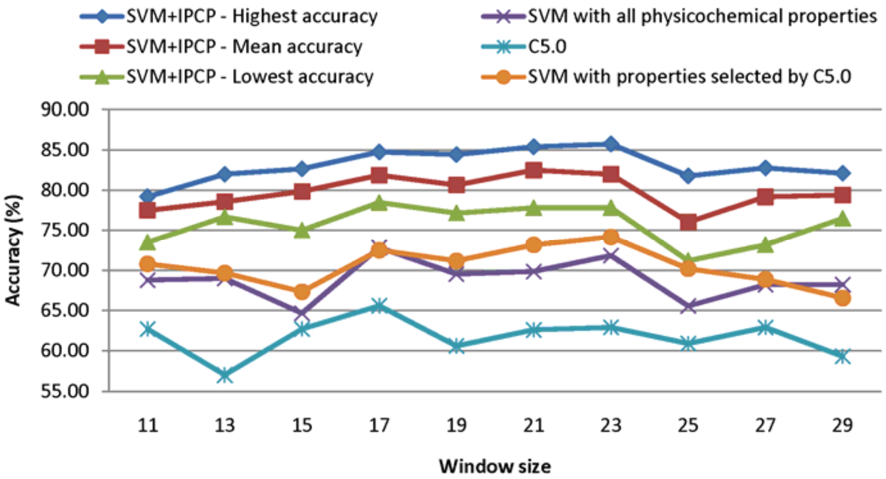


Figure 5.4.4 Performance comparisons between the SVM with informative physicochemical properties (SVM+IPCP) and other compared classifiers.

Figure 5.4.1 shows the sequence logo of the 151 positive samples with *w*=21 generated by the WebLogo tool [108]. The sequence logo with low information content reveals disadvantages of the SVM using the two position-based features, amino acid identity and evolutionary information, compared with the non-position based features, physicochemical properties using averaged measurement of amino acids in a sequence.

We established ten datasets with window sizes 11, 13, …, 29 from UbiProt, a database of ubiquitylated proteins [107], to evaluate the three kinds of features for applying classifiers. The dataset of window size 21 is shown in Figure 5.4.2. According to the prediction accuracies using 10-fold cross-validation (10-CV), the physicochemical property is the best feature to SVM with best performance among all classifiers and all kinds of features shown in Figure 5.4.3.

In order to provide insight into the underlying mechanism of ubiquitylation and improve the prediction accuracy, IPMA is applied to mine physicochemical properties and tune SVM parameters while maximizing the 10-CV accuracy, a set of 31 informative physicochemical properties is obtained. A prediction system UbiPred for identifying ubiquitylation sites is implemented by utilizing the 31 informative physicochemical properties. UbiPred performs well with a prediction accuracy of

84.44% using leave-one-out cross-validation (LOOCV), compared with the SVM-based methods using amino acid identity (65.67%), evolutionary information (66.33%) and all physicochemical properties (72.19%). The performances and area under the ROC curve (AUC) are shown in Table 5.1.1.

5.4.2.3. Informative physicochemical properties

Most of the 531 physicochemical properties may be irrelevant features or even interfere with prediction of the SVM classifier. Therefore, it is important to mine informative physicochemical properties for advancing the prediction accuracy. IPMA determines a feature set of *r* informative physicochemical properties and the values of SVM parameters (*C* and γ) for a given window size *w*. Because of the non-deterministic nature of IPMA, the obtained solutions would be different for each run. To obtain the features with robust performance, 30 independent runs of IPMA were performed for each window size *w*.

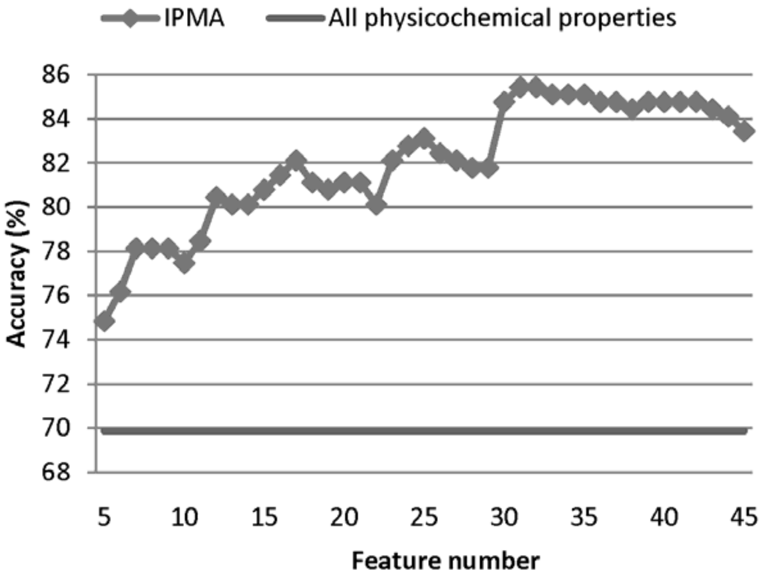


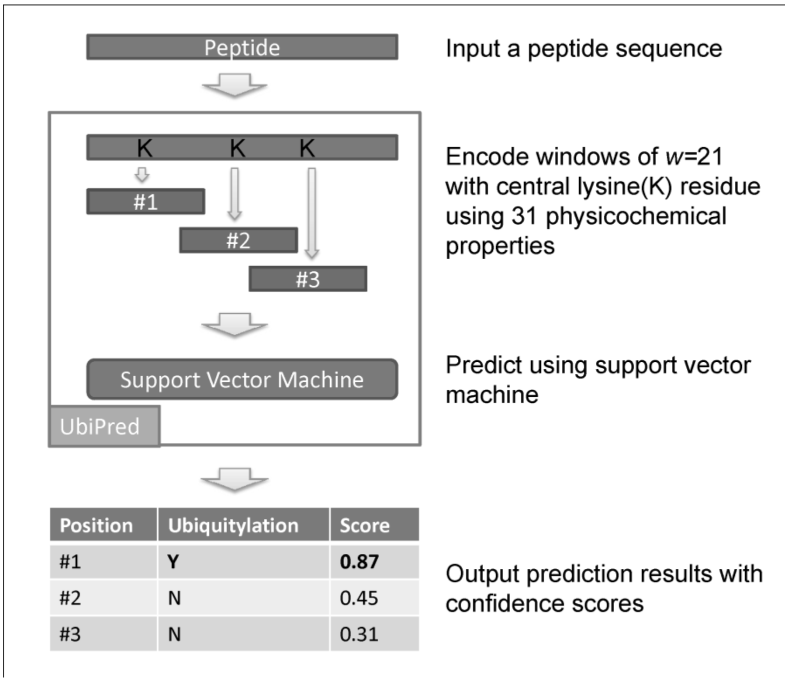
Figure 5.4.5 The best 10-CV accuracies of prediction using SVM with the window size 21 for various numbers of features (properties) selected by IPMA from 30 independent runs.

The highest, mean, and lowest prediction accuracies of IPMA using 10-CV are shown in Figure 5.4.4. For comparison, the decision tree method C5.0 [109] with the ability of feature selection based on information gain was also evaluated. The accuracies of C5.0 and SVM with the properties selected by C5.0 for various window sizes are also given in Figure 5.4.4. For all window sizes, the accuracies of SVM using informative physicochemical properties mined by IPMA are better than those of C5.0, SVM using all 531 physicochemical properties, and SVM using the

C5.0-selected properties. Considering the mean accuracies of SVM with informative physicochemical properties in Figure 5.4.4, the best window size is *w*=21.

Figure 5.4.5 shows the best 10-CV accuracies of using IPMA with *w*=21 for various numbers of features from 30 independent runs. The accuracy of *w*=21 can be improved from 69.87% to 85.43% by using *m*=31 out of *n*=531 physicochemical properties, where the values of SVM parameters are *C*=4 and γ=0.5. The 31 informative physicochemical properties constitute a good feature set obtained by considering the inter-correlation among properties.

The quantified effectiveness of individual physicochemical properties on prediction is useful to characterize the ubiquitylation mechanism by physicochemical properties. Orthogonal experimental design with factor analysis [28] [110] can be used to estimate the individual effects of physicochemical properties according to the value of main effect difference (MED) [52] [33]. The property with the largest value of MED is the most effective in predicting ubiquitylation sites.



Input a peptide sequence

Encode windows of *w*=21 with central lysine(K) residue using 31 physicochemical properties

Predict using support vector machine

Position	Ubiquitylation	Score
#1	**Y**	**0.87**
#2	N	0.45
#3	N	0.31

Output prediction results with confidence scores

Figure 5.4.6 The system flow of prediction system UbiPred.

According to MED, the 31 informative properties are ranked and their descriptions are shown in Table 5.4.1. The most effective property with MED=31.79 is NADH010102 denoting “hydropathy scale based on self-information values in the two-state model of 9% accessibility”. The least effective properties with MED=1.32

are NAKH900101 and QIAN880129 denoting “amino acid composition of total protein” and “weights for coil at the window position of -4”, respectively. The ranked informative physicochemical properties provide valuable information to biologists for further experimental verification.

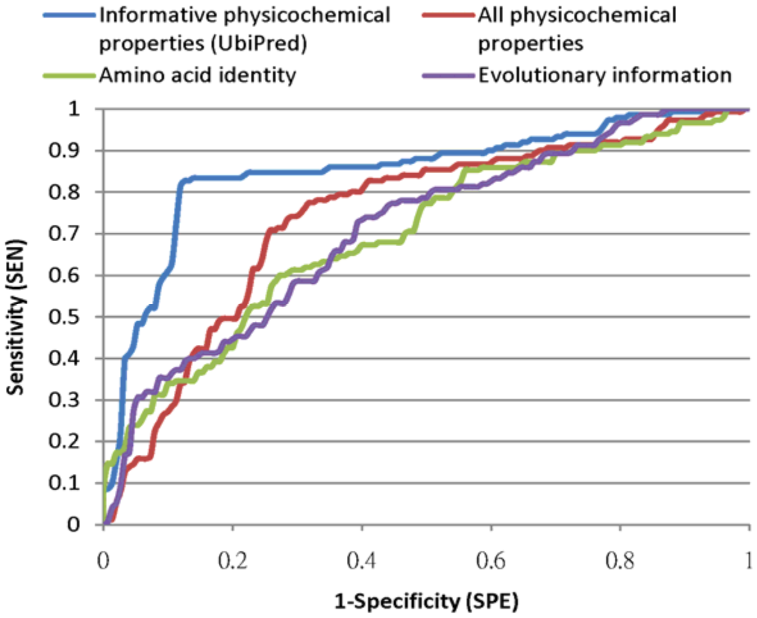


Figure 5.4.7 Comparison of receiver operating characteristic curves among informative physicochemical properties (UbiPred), amino acid identity, evolutionary information and all physicochemical properties.

5.4.2.4. Prediction system UbiPred

To implement a prediction system UbiPred for identifying ubiquitylation sites, the 31 informative physicochemical properties with *w*=21, *C*=4, and γ=0.5 were used. The system flow of UbiPred is shown in Figure 5.4.6. The required input for UbiPred is peptide sequence. UbiPred will automatically encoding the windows with central lysine residue using 31 informative physicochemical properties. Subsequently, the lysine residues will be annotated with SVM predicted result and shown in web page.

The prediction accuracy 84.44% of UbiPred shows good performance, compared with those of SVM with physicochemical property (72.19%), amino acid identity (65.67%) and evolutionary information (66.33%). The SEN, SPE and MCC of UbiPred are 83.44%, 85.43% and 0.69, respectively. To compare the robustness of UbiPred with other methods, the nonparametric method of ROC curve is applied by using the decision value of SVM as a tuning parameter. The area under the ROC curve (AUC) is calculated, as shown in Figure 5.4.7. UbiPred with AUC=0.85 performs well, compared with the SVM-based methods using all physicochemical properties (0.74), amino acid identity (0.70) and evolutionary information (0.71).

Table 5.4.2 The MEDs for 31 mined physicochemical property.

AAindex identity	Description	MED
NADH010102	Hydropathy scale based on self-information values in the two-state model of 9% accessibility	31.79
BROC820102	Retention coefficient in HFBA	29.80
MEIH800102	Average reduced distance for side chain	28.48
LEVM780101	Normalized frequency of alpha-helix, with weights	25.17
GUYH850104	Apparent partition energies calculated from Janin index	23.84
CORJ870101	NNEIG index	23.18
RACS770102	Average reduced distance for side chain	22.52
GEOR030108	Linker propensity from helical (annotated by DSSP) dataset	22.52
HARY940101	Mean volumes of residues buried in protein interiors	21.85
GRAR740102	Polarity	19.87
GUYH850105	Apparent partition energies calculated from Chothia index	19.87
MEIH800103	Average side chain orientation angle	17.88
KRIW790102	Fraction of site occupied by water	17.88
LEVM780106	Normalized frequency of reverse turn, unweighted	14.57
BULH740102	Apparent partial specific volume	13.25
FAUJ880101	Graph shape index	11.92
PUNT030102	Knowledge-based membrane-propensity scale from 3D_Helix in MPtopo databases	10.60
HUTJ700103	Entropy of formation	9.93
EISD840101	Consensus normalized hydrophobicity scale	8.61
CEDJ970105	Composition of amino acids in nuclear proteins (percent)	7.28
ZIMJ680102	Bulkiness	7.28
CEDJ970103	Composition of amino acids in membrane proteins (percent)	5.96
CHOC760103	Proportion of residues 95% buried	5.30
CEDJ970102	Composition of amino acids in anchored proteins (percent)	5.30
ROSM880102	Side chain hydropathy, corrected for solvation	4.64
BROC820101	Retention coefficient in TFA	4.64
FAUJ830101	Hydrophobic parameter pi	1.99
NAKH920101	AA composition of CYT of single-spanning proteins	1.99
ZHOH040102	The relative stability scale extracted from mutation experiments	1.99
NAKH900101	AA composition of total proteins	1.32

QIAN880129	Weights for coil at the window position of -4	1.32

The quantified effectiveness of individual physicochemical properties on prediction is useful to better characterize the ubiquitylation mechanism by physicochemical properties. According to MED, the 31 informative properties are ranked and their descriptions are shown in Table 5.4.2. The ranked informative physicochemical properties provide valuable information to biologists when further performing experimental verification.

5.4.2.5. Knowledge of data mining

Although the prediction accuracy of SVM is rather high compared with the other classifiers evaluated, it is not easy for biologist to interpret the prediction rules. In order to acquire interpretable knowledge from experimental data, C5.0 was applied to construct a compact decision tree by using the 31 informative physicochemical properties selected by IPMA on the whole training dataset. Figure 5.4.7 shows a constructed decision tree by C5.0. By utilizing this decision tree to classify the whole training dataset, the accuracy is 72.5%. This decision tree can be directly converted into a set of eight interpretable rules [109], consisting of three and five if-then rules for ubiquitylation sites and non-ubiquitylation sites, respectively.

Table 5.4.3 Five concise if-then rules with confidence larger than 0.5 obtained by using C5.0 and 31 informative physicochemical properties.

#	Rule	Confidence	Ubiquitylation sites	Covered samples	Misclassified samples
1	MEIH800102 <= 0.95381	0.96	N	23	0
2	HARY940101 > 135.2 AND CORJ870101 > 49.70762	0.90	N	49	4
3	CEDJ970105 > 6.805556	0.85	N	18	2
4	GEOR030108 <= 0.931333	0.75	N	10	2
5	MEIH800102 > 0.95381	0.54	Y	279	128

To obtain rather simple rules for easy interpretation, five concise if-then rules obtained from C5.0 are shown in Table 5.4.3. The first rule with the highest confidence value 0.96 can be interpreted as 'given a peptide with a central residue lysine (*w*=21), if the average reduced distance for side chain [111] (property MEIH800102) is less than or equal to 0.95381, then the residue is a non-ubiquitylation site with a confidence value 0.96'. This rule covers 23 sites in the training dataset and no site is misclassified by this rule.

There is only one of five classification rules for identifying ubiquitylation sites with a moderate confidence value 0.54. This rule means that if the average reduced distance for side chain is larger than 0.95381, then the residue is an ubiquitylation site with a confidence value 0.54. This rule reveals that the ubiquitylation sites are not easily discriminated from non-ubiquitylation sites. Furthermore, the property MEIH800102 plays an important role in predicting ubiquitylation sites. Examining the MED value (28.48) of MEIH800102 in Table 5.4.2, it is rather consistent that MEIH800102 is an informative property with a rank 3.

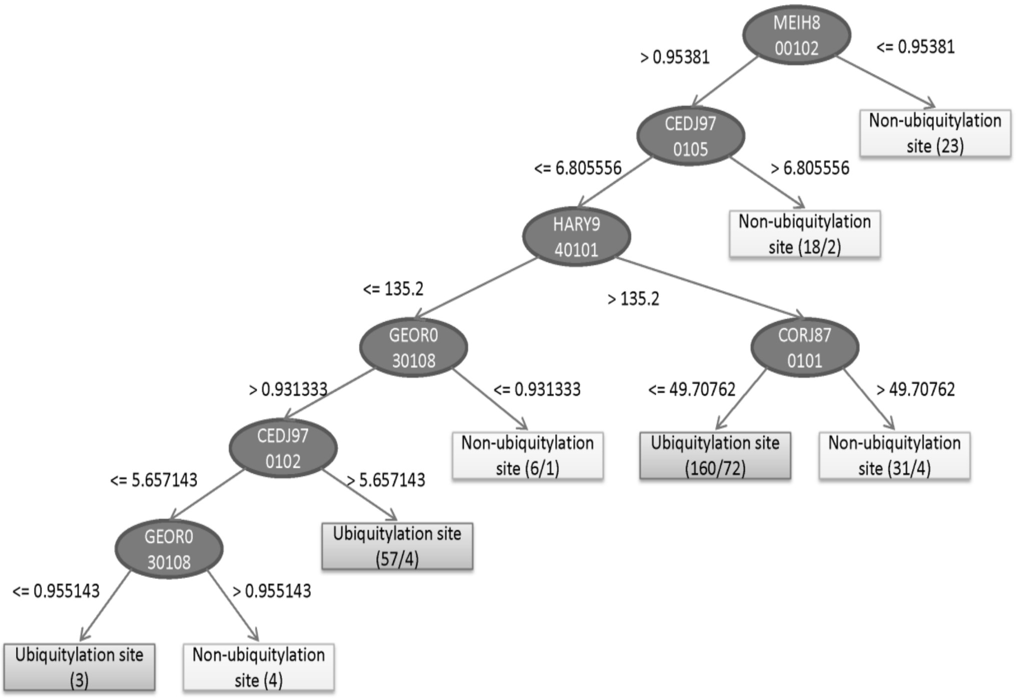


Figure 5.4.8 The derived decision tree by using C5.0 and the features of informative physicochemical properties for classification of ubiquitylation sites.

The second rule means that if the mean volume of residues buried in protein interiors [112] (property HARY940101) is larger than 135.2 and the NNEIG index [113] (property CORJ870101) is larger than 49.70762, then the residue is a non-ubiquitylation site with a confidence value 0.90'. This rule covers 49 samples in the training dataset and 4 of them are misclassified by this rule.

The third rule indicates that if the composition of amino acids in nuclear proteins (percent) [114] is larger than 6.805556, then the residue is a non-ubiquitylation site with a confidence value 0.85'. This rule covers 18 samples in the training dataset and 2 of them are misclassified.

The fourth rule indicates that if the linker propensity from helical (annotated by DSSP) dataset [115] is less than or equal to 0.931333, then the residue is a non-ubiquitylation site with a confidence value 0.75'. This rule covers 10 samples in the training dataset and 2 of them are misclassified.

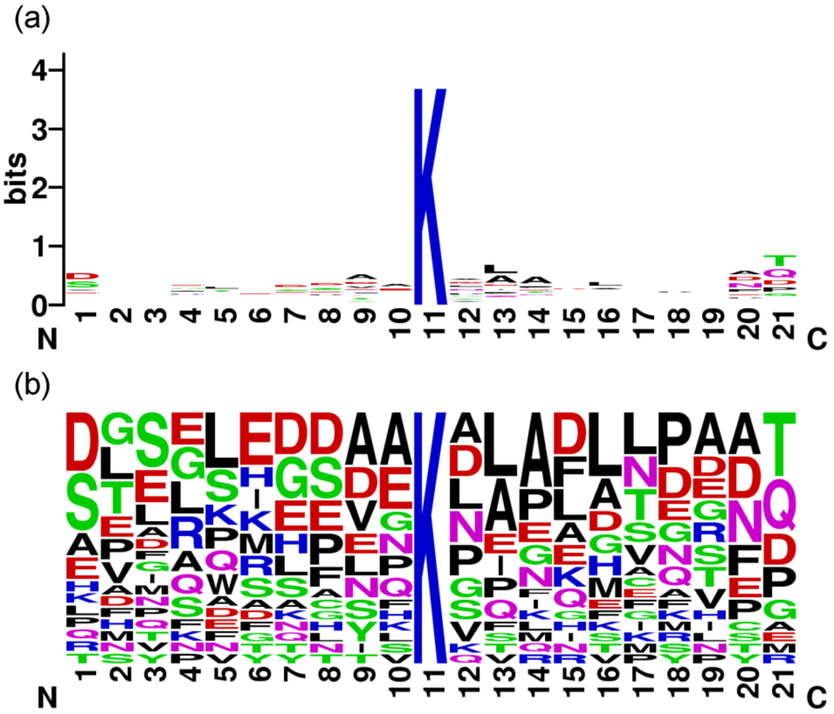


Figure 5.4.9 The sequence logo of the 23 peptides of promising ubiquitylation sites with *w*=21. (a) Information content and (b) Frequency plot.

5.4.2.6. Screening promising ubiquitylation sites

Recently, a new experimental method was proposed with 2.4-fold increase in the number of identified ubiquitylation sites, compared with previous methods [95]. It implies that there may be still many undiscovered ubiquitylation sites. To identify promising ubiquitylation sites from putative non-ubiquitylation sites, a scoring method is designed by normalizing the range of the decision values of SVM obtained from the training dataset of w=21 into the range [0, 1] of prediction scores. Normally, the default threshold value 0 used by the SVM classifier for discriminating ubiquitylation sites from non-ubiquitylation sites is mapped to a prediction score 0.5. The site with a prediction score close to 1 has a high possibility to be an ubiquitylation site. If the high prediction score 0.85 instead of 0.5 was adopted when classifying the peptides in the training dataset for all window sizes, there would be no false positive.

The prediction system UbiPred is applied to score 3424 putative non-ubiquitylation sites in an independent dataset that are not included in the training dataset of w=21, as shown in Figure 5.4.6. There are 1218 putative non-ubiquitylation sites with scores larger than 0.5. There are 23 peptides with scores larger than 0.85, which are the most promising ubiquitylation sites, listed in Table 5.4.4. The detailed information can be found in the website of UbiPred. The sequence logo of the 23 peptides shown in Figure 5.4.9 represents low information content similar to the sequence logo of the 151 positive samples in training dataset.

Table 5.4.4 List of 23 promising ubiquitylation sites identified from an independent dataset of 3424 putative non-ubiquitylation sites.

Accession number	Position	Score	Accession number	Position	Score	Accession number	Position	Score
P19358	114	0.99	P39976	323	0.90	P38080	809	0.87
Q9Y6K9	35	0.96	P38261	147	0.89	P10592	54	0.87
P25694	6	0.96	P25360	846	0.89	P38080	792	0.87
P40087	325	0.95	P09936	195	0.88	P12866	129	0.86
Q08412	232	0.93	P10591	54	0.88	Q05911	460	0.86
P04629	609	0.91	Q06408	156	0.87	P40087	410	0.86
P16603	165	0.91	P37303	283	0.87	P38075	10	0.86
P31539	626	0.91	P32467	38	0.87			

5.4.3. Predicting immunogenicity of MHC binding peptides

5.4.3.1. Introduction

After the prediction of peptides binding to cytotoxic T lymphocyte (CTL) and helper T lymphocyte (HTL), defining peptide immunogenicity is desirable to accurately predict immunogenicity of epitopes (i.e. CTL and HTL responses) for the vaccine design. The peptide immunogenicity is influenced by many factors, including intrinsic physicochemical properties and extrinsic factors such as host immunoglobulin repertoire [116, 117]. Several studies aimed to clarify the relationship between the peptide binding affinity to the MHC molecule and its immunogenicity [118, 119]. These studies revealed that moderate binding affinity of peptide-MHC molecules is essential to induce immunogenicity, but the ability of peptides to induce cytotoxic T lymphocyte and helper T lymphocyte responses does not strongly correlate with their affinity for the MHC molecule. In some extreme cases, a peptide with nearly-undetectable binding affinity of MHC class II molecules can induce strong T-cell responses [120]. Furthermore, peptide-flanking residues other then MHC anchor residues were identified as import factors for MHC class II-restricted T-cell responses [121, 122]. These studies show great importance of modeling T-cell responses.

Physicochemical properties of amino acids were extensively and successfully used in sequence-based prediction methods [18-21, 123]. Because of the weak correlation between peptide immunogenicity and peptide-MHC binding affinity, mining informative physicochemical properties is a potentially good approach to designing a classifier for predicting immunogenicity. Because the number of available physicochemical properties is as large as more than 500, the properties used in previous studies are usually selected according to domain knowledge [20] or the rank-based method [124]. Therefore, these methods cannot be effectively applied to the investigated intractable problems because of limited knowledge or neglect of correlated effects among multiple properties [123]. This study aims to design an accurate predictor by efficiently selecting a small set of informative physicochemical properties considering the correlated effects.

It is well recognized that feature selection and classifier design should be optimized simultaneously to maximize prediction accuracy [51]. The SVM-based learning methods are shown effective for various prediction methods from protein sequences [72, 75]. However, internal detection of relevant-feature correlation is not offered by conventional SVMs; meanwhile, appropriate setting of their control parameters is often treated as another independent problem [125]. Let there be *n* candidates of physicochemical properties of amino acids. To maximize accuracy of the investigated prediction problem by selecting a small number *m* out of *n* properties while cooperating with SVM simultaneously, it is equivalent to solve the binary combinatorial optimization problem having a huge search space of C(*n*, *m*)=*n*!/(*m*!(*n*-*m*)!)). To solve this problem, an informative physicochemical property mining algorithm (IPMA) capable of simultaneous feature selection and classifier design is proposed to mine informative physicochemical properties for predicting CTL and HTL responses.

5.4.3.2. Proposed prediction systems

Two prediction systems named POPI and POPI-MHC2 were proposed to predict immunogenicity of MHC class I and II binding peptides, respectively. High performance of POPI and POPI-MHC2 arises mainly from the inheritable bi-objective genetic algorithm which aims to automatically determine the best number *m* out of 531 physicochemical properties, identify these *m* properties, and tune SVM parameters simultaneously. The datasets of PEPMHCI and PEPMHCII consisting of 428 human MHC class I binding peptides and 226 human MHC class II binding peptides. All the peptides belongs to four classes of immunogenicity and are extracted from MHCPEP, a database of MHC-binding peptides [126]. Table 5.4.5 show the used datasets PEPMHCI and PEPMHCII of peptides associated with human MHC class I and II molecules, respectively. By applying the proposed IPMA to the experimental datasets, two prediction systems of POPI and POPI-MHC2 were constructed by using the selected informative physicochemical properties.

The IPMA is performed to mine informative physicochemical properties using the whole datasets of PEPMHCI and PEPMHCII. In this study, the parameters of IPMA are set as N_{pop}=50, P_c=0.8, P_m=0.05, r_{start}=5 and r_{end}=45. For each feature set with size *r*, IPMA selected a small set of physicochemical properties and parameter values of SVM. Figure 5.4.10 shows a potentially good result for PEPMHCI in terms of averaged accuracy (*AA*) and the number of used features obtained from a single run of IPMA using 10-CV. The result reveals that the best number of selected features is *m*=23 where the SVM classifier with *C*=2 and γ=2 has the best averaged accuracy *AA*=63.67% and overall accuracy *OA*=66.12%.

Table 5.4.5 The dataset PEPMHCI and PEPMHCII of peptides associated with human MHC class I and II molecules

Immunogenicity class	PEPMHCI	PEPMHCII
None	144	45
Little	83	60
Moderate	100	64
High	101	57
Total	428	226

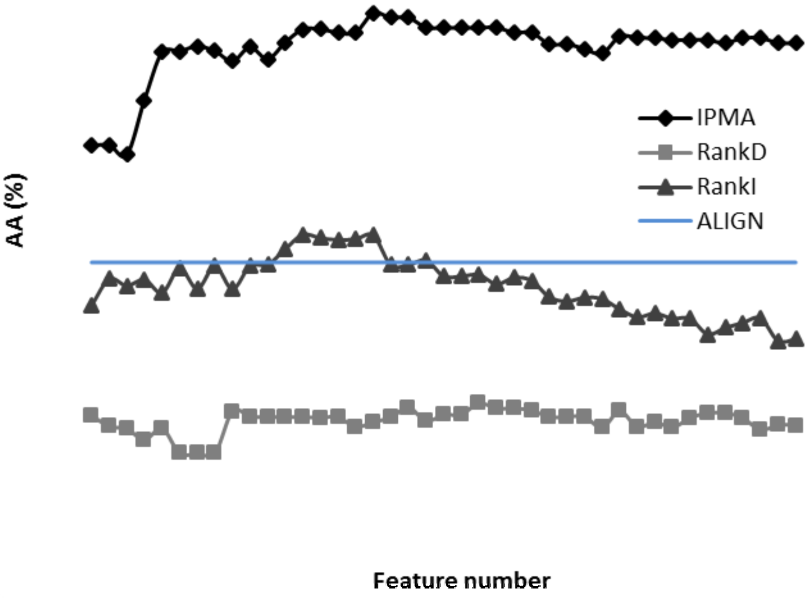


Figure 5.4.10 Averaged accuracies (*AA*s) of 10-CV for IPMA, rank-based methods (RankD and RankI) and the alignment-based method (ALIGN) for MHC class I binding peptides.

To further evaluate the feature selection of IPMA, a traditional rank-based method for evaluating performance of a single feature is also implemented for comparison. The rank-based method suffers from the incapability of finding appropriate values of *C* and *γ* to train SVM classifiers. In order to achieve high performance, two parameter settings of SVM were tested. The first rank-based method named RankD using the default values of SVM parameters that *C*=1 and *γ*=1/*r*. The best performance of RankD is *AA*=36.08% with 21 features. The second rank-based method named RankI using the same values of *C*=2 and *γ*=2 obtained from IPMA. The best performance of RankI is *AA*=48.87% with 18 features. Figure 5.4.10 shows the performance of RankI is better than that of RankD, revealing that the parameter setting of SVM parameters derived from IPMA is effective.

Furthermore, the performance of feature selection of IPMA is much better than that of the rank-based method. This result is well recognized that the feature selection by additionally considering the correlated effects among physicochemical properties can advance prediction performance.

5.4.3.3. POPI for predicting immunogenicity of MHC class I binding peptides

The immunogenicity of a peptide is determined by measuring the concentration of peptides giving 50% of maximum specific lysis by CTLs of target cells displaying the peptide, and is given a descriptive value belonging to the four classes, None, Little, Moderate, High. POPI utilizing the 23 selected properties performs well with the accuracy of 64.72% using leave-one-out cross-validation (LOOCV). For comparison, sequence alignment-based and affinity-based methods were implemented to evaluate the LOOCV performances.

Sequence alignment may be an efficient approach to predicting peptide immunogenicity because similar sequences may have similar peptide immunogenicity. In order to compare the alignment-based prediction methods with POPI, two methods including global sequence alignment tool ALIGN [127] and advanced sequence comparison method PSI-BLAST that is capable of detecting remote homologues [105] were applied to search for similar sequences.

In the past, affinity was considered as an important index to predict peptide immunogenicity. To evaluate the affinity-driven prediction method, an additional dataset was established by extracting MHC class I binding peptides with known activity levels in both fields of ‘BINDING’ and ‘IMMUNOGENICITY’ from the MHCPEP database. However, there are four levels in the field of ‘IMMUNOGENICITY’, but the field of ‘BINDING’ has only three levels without the level ‘none’. To fairly evaluate the prediction performance of the affinity-driven prediction, the immunogenic class None was combined with the class Little. The dataset contains 160 peptides belonging to three classes.

Table 5.4.6 Performance comparisons of ALIGN, PSI-BLAST and POPI-MHC2

Immunogenicity	ALIGN		PSI-BLAST		POPI-MHC2	
	ACC (%)	MCC	ACC (%)	MCC	ACC (%)	MCC
None	68.89	0.74	66.67	0.69	86.67	0.81
Little	46.67	0.34	23.21	0.29	68.33	0.54
Moderate	50.00	0.22	75.86	0.22	57.81	0.53
High	71.93	0.56	38.00	0.31	85.96	0.73
OA	58.41		49.75		73.45	
AA	59.37		50.94		74.69	

To evaluate the affinity-driven prediction method, a prediction system named AFFIPRE to predict peptide immunogenicity was implemented using the following criterion. If the immunogenic level and the affinity level of a peptide are identical, this test is regarded as a successful prediction. Otherwise, this prediction is fail. The four measurements were used to evaluate AFFIPRE, which are the same with those for IPMA.

In contrast to the existing affinity-based methods of predicting immunogenicity by way of predicting MHC-binding peptides, POPI is the first computational system based on physicochemical properties to predict peptide immunogenicity using epitopes associated with human MHC class I molecules, which has been implemented as a web server (http://iclab.life.nctu.edu.tw/POPI). Up to date, there are >18,690 visits from >20 countries, and >20,000 sequences were analyzed.

5.4.3.4. POPI-MHC2 for predicting immunogenicity of MHC class II binding peptides

The 21 informative physicochemical properties and SVM parameters selected by IBCGA are applied to construct POPI-MHC2, an SVM-based prediction system for immunogenicity of MHC class II binding peptides. The web server has also been implemented and is available at http://iclab.life.nctu.edu.tw/POPI. POPI-MHC2 performs well with accuracy of 73.45% using leave-one-out cross-validation, compared with two alignment-based methods ALIGN (58%) and PSI-BLAST (<49.75%) shown in Table 5.4.7.

Table 5.4.7 Performance comparison between AFFIPRE and POPI-MHC2

Immunogenicity class	Peptides	AFFIPRE		POPI-MHC2	
		ACC (%)	MCC	ACC (%)	MCC
None and Little	21	23.81	0.30	42.86	0.49
Moderate	6	33.33	-0.08	0.00	-0.07
High	42	50.00	0.16	92.86	0.41
OA		40.58		69.57	
AA		35.71		45.24	

For comparing with affinity-based prediction, another dataset consisting of 69 peptides with annotated binding and immunogenicity level was constructed. POPI-MHC2 (69.57%) performs better than the affinity-based method (40.5%) shown in Table 5.4.7. The poor performance of AFFIRE (OA=40.58 and AA=35.71%) implies that affinity is not the deterministic factor for peptide immunogenicity of MHC class II binding peptide. Instead, physicochemical properties might play more important roles for determining the immunogenicity.

Users can use POPI-MHC2 by entering either a sequence or a file of sequences of MHC binding peptides. The predicted immunogenicity levels will be shown in the web page. POPI-MHC2 is publicly available at http://iclab.life.nctu.edu.tw/POPI

5.4.3.5. Analysis of informative physicochemical properties

After identification of informative physicochemical properties, it is desired to analyze and interpret the obtained knowledge. Revealing individual effects of identified physicochemical properties on immunogenicity of MHC class II-restricted peptides is important for immunologist to further investigate immunogenic problems. Factor analysis of the orthogonal experimental design used in IPMA can efficiently estimate effects of an individual feature by evaluating its main effect difference (*MED*). The property with the largest *MED* value is the most effective property.

Because IPMA is a non-deterministic algorithm and SVM parameter values will slightly affect prediction accuracy, the identified feature sets with the highest accuracy obtained from multiple independent runs would be not the same. In order to obtain a robust feature set, 60 independent runs of IPMA were performed for identifying informative physicochemical properties. The largest, mean and smallest numbers *m* of selected features are 45, 28.63 and 12, respectively. The highest, mean and lowest *AA* accuracies in the training phase are 76.84%, 73.64% and 69.68%, respectively. The statistic result reveals that a small set of effective properties is more stable in each run of IPMA.

Table 5.4.8 and Table 5.4.9 show the typical feature sets with MED values considering both training accuracy and selection frequency for MHC class I and II binding peptide, respectively. For CTL immune response, the property of AAindex identity GEIM800103 is the most effective property with *MED*=33.29, which corresponds to ‘Alpha-helix indices for beta-proteins’ [128]. The least effective property is MIYS850101 with MED=0.80 which corresponds to ‘Effective partition energy’ [129]. For HTL immune response, the AAindex identity KUHL950101 is the most effective property (denoting ‘Hydrophilicity scale’) with MED=46.06 [130]. The AAindex identity DESM900102 with the smallest MED value of 4.11 denoting ‘Average membrane preference: AMP07’ [131].

5.4.3.6. Comparison of physicochemical properties responsible for CTL and HTL responses

It is interesting to know similarity and difference between the two property sets responsible for HTL and CTL responses. To analyze compositions of informative physicochemical properties, physicochemical properties of each set are categorized into four classes, hydrophobicity, structure, volume and others. Properties with obvious annotation of hydrophobicity-, secondary structure- and volume-related words can be easily categorized first. For each of uncategorized properties, its correlation coefficients (CCs) to the categorized properties are measured. The same class of the categorized property is assigned to the uncategorized property with the CC value larger than or equal to 0.85.

Figure 11 shows pie-chart representations of the property compositions in terms of the four classes for CTL and HTL responses. As expected, hydrophobicity-related properties play an important role in both HTL (43%) and CTL (17%) immune responses in immunogenicity that is consistent with our knowledge that hydrophobicity is important for biomolecular recognition [132, 133]. Recent studies [134, 135] have reported importance of antigen structures in influencing T-cell dominance. It is also consistent that structure propensity-related properties has a large proportion for both HTL (33%) and CTL (57%) immune responses (Figure 24).

Table 5.4.8 Individual effects of identified properties for CTL responses in terms of main effect difference (*MED*).

ID of AAindex	Description	*MED*	Class
GEIM800103	Alpha-helix indices for beta-proteins	33.29	S
OOBM770104	Average non-bonded energy per residue	31.97	O
PALJ810115	Normalized frequency of turn in alpha+beta class	24.91	S
QIAN880132	Weights for coil at the window position of -1	23.90	S
OOBM850102	Optimized propensity to form reverse turn	17.09	S
NADH010106	Hydropathy scale based on self-information values in the two-state model (36% accessibility)	14.79	H
RADA880106	Accessible surface area	11.64	V
QIAN880112	Weights for alpha-helix at the window position of 5	10.71	S
WEBA780101	RF value in high salt chromatography	10.65	O
QIAN880125	Weights for beta-sheet at the window position of 5	10.63	S
JOND750101	Hydrophobicity	9.27	H
QIAN880124	Weights for beta-sheet at the window position of 4	9.06	S
MUNV940101	Free energy in alpha-helical conformation	7.44	S
HUTJ700102	Absolute entropy	6.62	V
MITS020101	Amphiphilicity index	5.10	H
KARP850103	Flexibility parameter for two rigid neighbors	4.63	O
FAUJ880113	pK-a(RCOOH)	4.37	S
ISOY800106	Normalized relative frequency of helix end	4.31	S
RACS820113	Value of theta(i)	3.25	S
GEOR030105	Linker propensity from small dataset (linker length is less than six residues)	3.05	S
QIAN880114	Weights for beta-sheet at the window position of -6	2.99	S
DIGM050101	Hydrostatic pressure asymmetry index, PAI	1.60	O
MIYS850101	Effective partition energy	0.80	H

H: hydrophobicity; S: structure; V: volume; O: others

Table 5.4.9 Individual effects of identified properties for HTL responses in terms of main effect difference (*MED*).

ID of AAindex	Description	*MED*	Class
KUHL950101	Hydrophilicity scale	46.06	H
WERD780103	Free energy change of alpha(Ri) to alpha(Rh)	37.10	O
KHAG800101	The Kerr-constant increments	32.78	O
VHEG790101	Transfer free energy to lipophilic phase	31.92	H
BIOV880102	Information value for accessibility; average fraction 23%	31.20	H
ENGD860101	Hydrophobicity index	27.79	H
WOLR810101	Hydration potential	26.18	H
JOND750102	pK (-COOH)	25.03	H
GEIM800109	Aperiodic indices for alpha-proteins	23.66	O
AURR980103	Normalized positional residue frequency at helix termini N"	22.46	S
ROBB760111	Information measure for C-terminal turn	16.96	S
YUTK870104	Activation Gibbs energy of unfolding, pH9.0	15.93	O
PALJ810113	Normalized frequency of turn in all-alpha class	15.36	S
RACS820114	Value of theta(i-1)	14.21	S
MAXF760104	Normalized frequency of left-handed alpha-helix	12.83	S
KUMS000103	Distribution of amino acid residues in the alpha-helices in thermophilic proteins	11.13	S
CHOC750101	Average volume of buried residue	9.12	V
RICJ880106	Relative preference value at N3	8.75	H
FASG760105	pK-C	7.95	H
ISOY800108	Normalized relative frequency of coil	5.27	S
DESM900102	Average membrane preference: AMP07	4.11	H

H: hydrophobicity; S: structure; V: volume; O: others

The situation is similar that all the hydrophobicity- and structure-related properties take a large proportion (close to 75%) among all properties. The major difference is that the categorized properties with the largest proportion for HTL (43%) and CTL (57%) responses are the hydrophobicity and structure classes, respectively. In other words, hydrophobicity-related properties are more important for HTL responses, compared with CTL responses. In contrast, structure-related properties are more important for CTL than HTL responses.

The great importance of structure- and hydrophobicity-related properties for CTL and HTL responses, respectively, can also be observed by the *MED*-based analysis for ranking individual effects of informative physicochemical properties. For CTL responses, the most effective property of AAindex identity GEIM800103 with *MED*=33.29 is 'Alpha-helix indices for beta-proteins' [128] (Table 5.4.8). In contrast, the property of AAindex identity KUHL950101 denoting 'Hydrophilicity scale' [130] is the most effective property with *MED*=46.06 for HTL responses (Table 5.4.9).

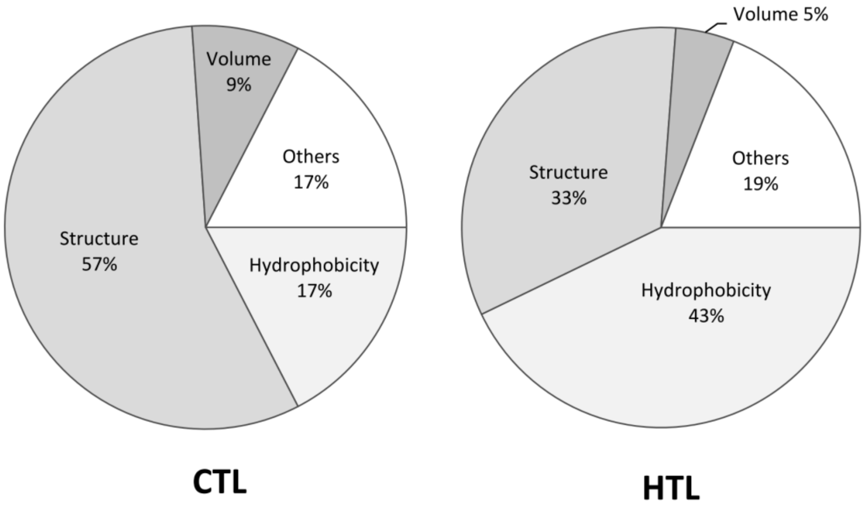


Figure 5.4.11 Pie-chart representations of compositions of categorized physicochemical properties of peptides responsible for CTL and HTL responses

From the perspective of similarity, the CTL response-related property of AAindex identity MIYS850101 denoting 'Effective partition energy' highly correlate with two HTL response-related properties of AAindex identities BIOV880102 and DESM900102 denoting 'Information value for accessibility; average fraction 23%' and 'Average membrane preference: AMP07' with CC values of 0.93 and 0.83, respectively. All three properties are hydrophobicity-related properties. Both volume-related properties of AAindex identities RADA880106 and HUTJ700102 denoting ‘Accessible surface area’ and ‘Absolute entropy’, respectively, for CTL responses highly correlate with volume-related property of AAindex identity CHOC750101 denoting ‘Average volume of buried residue’ for HTL responses (CC=0.87 and 0.80, respectively). Structure-related properties of RACS820114 and MUNV940101 for HTL and CTL responses denoting ‘Value of theta(i-1)’ and ‘Free energy in alpha-helical conformation’ also show high correlation with CC=0.83. Altogether, informative physicochemical properties for CTL and HTL responses share a few similar properties of all three major classes except for the class, others.

5.4.3.7. Peptides capable of inducing both CTL and HTL responses

An epitope capable of inducing both CTL and HTL responses is considered as a good candidate for peptide-based vaccine designs [136, 137]. An interesting question is whether peptides capable of inducing one kind of HTL and CTL responses necessarily induce the other kind of responses. The POPI 2.0 prediction system is used to reveal an answer to the question. For all peptides annotated with known categorized immunogenicity ‘High’ for one kind of HTL and CTL responses, its ability to induce the other kind of CTL and HTL responses is predicted by using the POPI 2.0 server.

All the test peptides are obtained from the PEPMHCII and PEPMHCI datasets. Table 5.4.10 shows results that 69% and 37% of peptides inducing CTL and HTL responses were predicted as no inducing capability for HTL and CTL responses, respectively. Only 21% of peptides with high immunogenicity for HTL responses can induce high immunogenicity of CTL responses. There is no peptide with high CTL responses can induce high immunogenicity of HTL responses. Results reveal that there exists no obvious necessary conduction between peptides inducing the two kinds of responses. It is consistent to the general observation that only a small proportion of peptides inducing both HTL and CTL responses [120]. This result provides a good reason to build a prediction system to quickly select peptide candidates inducing both CTL and HTL responses.

Table 5.4.10 Predicted levels of peptides to induce both CTL and HTL responses.

Predicted level	High	Moderate	Little	None	Total
Peptides with high-level CTL response	0	17	14	70	101
Peptides with high-level HTL response	12	16	8	21	57

5.4.4. Identification of T-cell receptor recognition sites

5.4.4.1 Introduction

Compared to the knowledge of anchor positions of peptides for MHC binding, previous studies for identifying T-cell receptor (TCR) recognition positions were based on small-scale analyses using only a few peptides and concluded different recognition positions. Large-scale analyses are necessary to better characterize and predict a peptide’s T-cell reactivity (and thus immunogenicity). The identification and characterization of important positions influencing T-cell reactivity will provide insights into the underlying mechanism of immunogenicity. In Chapter 5.4.3, the POPI prediction systems are proposed to predict peptide immunogenicity with reasonably high accuracy. However, the effect of MHC alleles on immunogenicity was not considered. Also, it is hard to identify T-cell receptor recognition sites because of the used averaged features. In this chapter, a weighted degree string kernel is proposed to identify T-cell receptor recognition sites and improve prediction performances by considering the effects of positions and MHC alleles.

The first predictor for T-cell reactivity published is POPI [52]. POPI is a support vector machine (SVM)-based method trained on 23 informative physicochemical properties of MHC class I binding peptides. While POPI performs reasonably well, it uses averaged physicochemical properties to represent peptides independent of their length. It thus does not allow for identifying relevant positions of the peptide for T-cell reactivity. The method thus cannot yield structural insights into T-cell reactivity.

In previous studies on the formation of the TCR-peptide-MHC complex, crystal structures have been analyzed [138-140] to correlate structural features of the TCR with immunogenicity and to identify TCR recognition positions. However, due to the low number of available crystal structures of the ternary complex, these are just case studies, with limited potential for generalization. For example, two studies found different important positions of HLA-A2 binding peptides for TCR recognition (position 8 [140]; positions 4 and 6 [138]). As an alternative approach to T-cell reactivity, experiments with substitutions and cytotoxicity assays have been performed for HLA-B27 [141]. However, so far results are based on only a few peptides. Large-scale analyses are thus desirable to better characterize the important positions of MHC binding peptides for immunogenicity.

In this work, a systematic statistical approach is proposed for the prediction of T-cell reactivity. This study presents a more advanced machine learning study considering the effects of MHC restriction on immunogenicity. In order to better characterize the immunogenicity induced by MHC class I binding peptides, we employ support vector machines (SVMs) using string kernels (SK) that have been successfully applied in many classification tasks [142-146]. This method was applied (1) to predict peptide immunogenicity and (2) to identify important positions of MHC binding peptides for immunogenicity. The present study is based on a large dataset IMMA2, which contains data from databases of MHCPEP [126], SYFPEITHI [147, 148] and IEDB [149].

The prediction system POPISK for predicting peptide immunogenicity of HLA-A2 binding peptides was built on this machine learning approach. POPISK performs well achieving an overall performance of 0.68 for accuracy (ACC) and 0.74 for area under the receiver operating characteristic curve (AUC). This is significantly better than POPI on the same dataset (0.60 for ACC and 0.64 for AUC) IMMA2. In an analysis of seven HLA-A2-binding peptides with known crystal structures, POPISK accurately predicts the immunogenicity for the majority of peptides and successfully predicted the immunogenicity change of single residue modifications reported in previous studies [150, 151]. We also analyzed the importance of amino acid positions of the peptides by selecting positions whose deletion significantly decrease prediction performance. This technique shows that six positions (1, 4, 5, 6, 8 and 9) of HLA-A2 binding peptides are the most important for T-cell reactivity and thus immunogenicity. Three of these positions were reported in previous studies (position 8 [140]; positions 4 and 6 [138]). As a confirmation, graphical analyses using two sample logos [152] identified nearly identical important positions 4, 6, 8 and 9.

5.4.4.2. Datasets

We first extracted peptide binders of length 9 with associated human MHC class I alleles and the corresponding immunogenicity data from MHCPEP [126], SYFPEITHI [147, 148] and IEDB [149]. For the MHCPEP database, the peptide sequences and their associated MHC alleles, binding and immunogenicity data are extracted from the fields of 'SEQUENCE', 'MHC MOLECULE', 'BINDING' and 'ACTIVITY', respectively. The 'BINDING' field annotates a peptide as either a binder or a non-binder. The peptide immunogenicity in MHCPEP is defined by its PD_{50} value, which is the peptide concentration giving 50% maximal specific lysis by cytotoxic T-cells of target cells displaying the MHC-peptide complex. According to MHCPEP, a peptide with PD_{50} value (obtained from the field 'ACTIVITY') larger than 10 μM is considered a non-immunogenic peptide, all others are considered immunogenic. For the SYFPEITHI database, the data of binders and immunogenic peptides associated with various MHC alleles is extracted from the field 'Natural ligands' and 'T-Cell epitopes', respectively. For the IEDB database, the peptide sequences and their associated MHC alleles, qualitative binding and qualitative immunogenicity data are extracted from the fields of ‘Epitope’, ‘MHC Restriction’, ‘MHC binding’, ‘T cell response’, respectively.

Only peptides with positive binding annotation were selected for analyses. These peptide sequences were grouped into allele-specific datasets according to their associated HLA supertypes [153]. In order to utilize all available data for analyses, peptides with contradictory annotations (immunogenic and non-immunogenic) were regarded as immunogenic peptides. After removing duplicate entries, the dataset of allele HLA-A2 (named IMMA2) consists of 558 immunogenic and 527 non-immunogenic peptides. The IMMA2 dataset is available at http://iclab.life.nctu.edu.tw/POPISK/download.php. This study focuses on HLA-A2 because it is one of the best known allele. It is easy to compare results obtained from this study and previous knowledge. Also, due to the small number of peptides associated with the other alleles, it is hard to create robust models for the other alleles.

5.4.4.3. Weighted degree string kernel

An effective weighted degree string kernel [145, 154, 155] counting the numbers of matched subsequences of length *p* at corresponding positions of two sequences is applied to transform samples to high-dimensional space to make linear separation easier. Given two sequences s_i and s_j of equal length *L* and degree *d*, the weighted degree string kernel computes the total numbers of matched subsequences of length *p* ∈ {1, …, *d*} at corresponding positions *l* of two sequences, defined as follows:

$$k(\mathrm{s}_i,\mathrm{s}_j)=\sum_{p=1}^{d}\beta_p\sum_{l=1}^{L-p+1}I(u_{p,l}(\mathrm{s}_i)=u_{p,l}(\mathrm{s}_j)), \tag{1}$$

where $I(h)=1$ if *h* is true; otherwise, $I(h)=0$, $u_{p,l}(\mathrm{s})$ is the subsequence of length *p* starting from position *l* of peptide sequence s, and β_p are weighted coefficients. In this study, sequence length *L* is 9. The fixed values of $\beta_p=2(d\text{-}p+1)/(d(d+1))$ are adopted as used in previous study [145]. Shogun [156] release 0.6.7 was used and LIBSVM [125] was chosen for the implementation of the predictor.

5.4.4.4. Prediction of peptide immunogenicity

To accurately predict immunogenicity of HLA-A2 binding peptides, it is necessary to tune two parameters (cost parameter *C* of the SVM and degree *d* of the weighted degree kernel) to build an accurate SVM classifier. In this study, a nested 10-fold cross-validation (10-CV) procedure was adopted to evaluate the prediction performance of our string kernel-based SVM classifier as it provides an almost unbiased estimate of the prediction error [157].

The nested 10-CV consists of two cross-validation loops: an inner loop for tuning SVM parameters and an outer loop for evaluating the prediction performance of tuned SVM classifiers. First, the IMMA2 dataset was randomly divided into ten subsets of approximately equal size. For each iteration *m* (outer loop), the *m*-th subset is left out for testing the tuned SVM classifier trained by using the selected optimal parameters giving highest AUC performance using 10-CV on the remaining dataset (inner loop). The grid search method is applied to tune the parameters $C \in \{2^{-4}, 2^{-3}, \ldots, 2^{4}\}$ and $d \in \{1, 2, \ldots, 9\}$.

To obtain a robust statistical estimation of prediction performances, a total of 20 runs of nested 10-CV procedure were applied to calculate the mean values of performance measurements as final prediction performances. The best values of *C* and *d* having the highest AUC value on the inner 10-CV loop are always 1 and 9, respectively. The mean prediction performances and corresponding standard deviation (SD) values of nested 10-CV on the IMMA2 dataset are 0.68 and 0.007 for ACC, 0.74 and 0.004 for AUC and 0.37 and 0.013 for MCC, respectively (Figure 5.4.12). All nine string kernels and five complex string kernels provided by Shogun were evaluated. Most of them perform similarly to or slightly worse than the weighted degree string kernel. Except for cost parameters *C* and degree parameter *d*, the above-mentioned results were obtained by using default values of parameters. All kernels might thus perform better by carefully tuning the respective parameters.

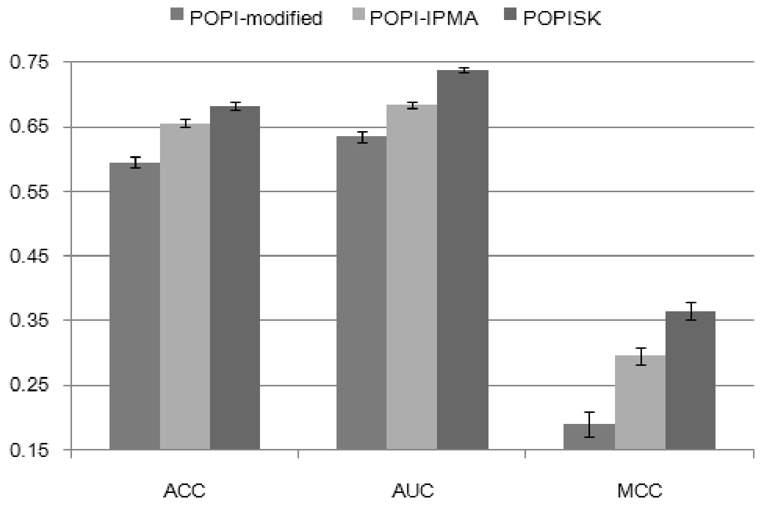


Figure 5.4.12 Comparison of nested 10-CV performances of POPISK and POPI-modified and POPI-IPMA

5.4.4.5. Comparison to POPI

POPI is an SVM-based method using radial basis function kernel and 23 informative physicochemical properties mined by using an inheritable bi-objective genetic algorithm. It is not fair to directly compare the results of POPISK with POPI because POPI is a four-class prediction method that predicts a peptide as highly, medium, little and not immunogenic. Furthermore, POPI is based on a smaller dataset. In order to perform a comparison, a modified POPI method (POPI-modified) was constructed using the same dataset IMMA2 and the 23 informative physicochemical properties for binary prediction problem of immunogenic and non-immunogenic peptides.

The evaluation procedures of POPI-modified are described as follows. First, the 23 informative physicochemical properties were used to encode peptides of IMMA2 dataset. Subsequently, 20 runs of nested 10-CV were applied as follows. The grid search method was applied to tune the cost parameter $C \in \{2^{-4}, 2^{-3}, \ldots, 2^{4}\}$ and the kernel parameter $\gamma \in \{2^{-4}, 2^{-3}, \ldots, 2^{4}\}$ in the inner 10-CV loop. The SVM classifiers trained by using the selected parameters giving highest AUC performance in inner 10-CV loop are used to evaluate the prediction performances in the outer 10-CV loop.

Due to the difference of datasets and assays for measuring immunogenicity between the original POPI method and POPISK, another comparison using IPMA method to reselect informative physicochemical properties can provide better insights into the advantage of used string kernel method POPISK. However, due to the time-consuming nature of genetic algorithm, it is difficult to do 200 runs of IPMA. Considering the balance of preliminary results for comparisons and experiment efforts, 20 runs of IPMA is applied to give a rough performance for comparison with POPISK. The evaluation procedures of POPI-IPMA are similar with POPI-modified. The only difference is that POPI-IPMA reselect informative physicochemical according to the validation performance instead of using 23 informative physicochemical properties selected by previous POPI method.

The comparison of nested 10-CV performances of POPISK, POPI-modified and POPI-IPMA is shown in Figure 3.12. Obviously, POPISK dominates POPI-modified with 10% improvements of ACC and AUC. Although the performance of POPISK is 2-5% better than POPI-IPMA, note that the POPI-IPMA utilize average feature could be further improved by changing the position-independent feature to consider the position effects of physicochemical properties. The nested 10-CV performances and corresponding SD values of POPI-modified are 0.60 and 0.009 for ACC, 0.64 and 0.009 for AUC and 0.19 and 0.018 for MCC, respectively. The nested 10-CV performances and corresponding SD values of POPI-IPMA are 0.65 and 0.017 for ACC, 0.68 and 0.147 for AUC and 0.30 and 0.033 for MCC, respectively. By collecting more data, POPISK is expected to perform better and can be applied to analyze immunogenicity of peptides associated with other MHC alleles.

5.4.4.6. Identification of important positions for immunogenicity

Compared to well-known MHC binding motifs, T-cell recognition positions of MHC binding peptides are still not fully understood. Some studies have aimed to identify the T-cell recognition positions. However, these studies were based on only a few crystal structures and identified different recognition positions [138-140]. The computational identification of important positions for immunogenicity will shed light on the mechanism of T-cell recognition and accelerate the development of peptide-based vaccines. To assess the individual contributions of each position of MHC-binding peptides to the prediction performance, we proposed an efficient method to estimate the importance of positions that is described as follows.

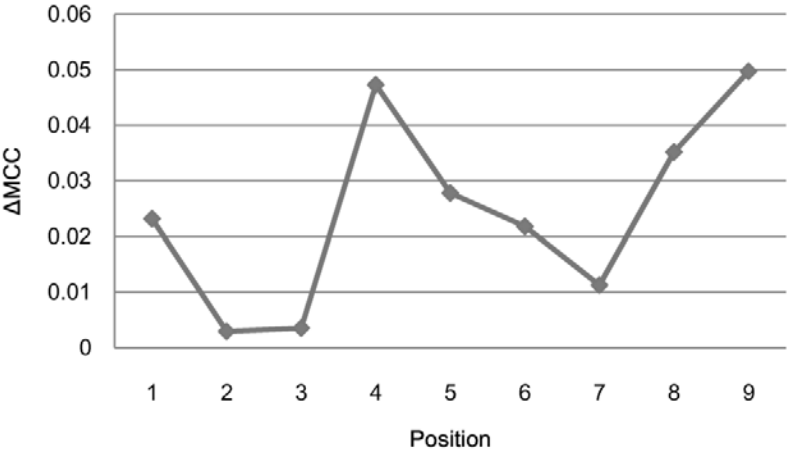


Figure 5.4.13 The decrease in MCC performances evaluated on datasets without using residues in specific positions.

The proposed method uses the decrease in prediction performance resulted from removing the sequence information on a specific position within the peptide to designate the importance for each position. The larger the decrease in performance, the greater the importance of the position is. The change in prediction performance is evaluated as follows. First, nine additional datasets for nine positions were created by removing residues in the corresponding positions from the IMMA2 dataset. Subsequently, for each of the nine datasets, 20 runs of nested 10-CV were performed as described above to evaluate prediction performances. For the parameter tuning process, the maximum value of degree parameter *d* is set to 8 (the same as the remaining peptide length). The decreases in performance as measured by MCC (ΔMCC) for these datasets are shown in Figure 5.2.13. Other performance measures (AUC, ACC) yield similar results (data not shown). Six positions (1, 4, 5, 6, 8 and 9) are identified as important positions since those of the prediction performance on datasets where the corresponding positions have been removed decreased significantly.

To further investigate over- and underrepresented amino acids in corresponding positions, two-sample logos [152] are computed to graphically represent the differences between immunogenic and non-immunogenic peptides of all peptides in IMMA2. Statistically significant residues selected by using a two-sample *t*-test with $p < 0.05$ are represented in the logo. In addition, a widely used multiple-comparison correction (Bonferroni correction) is applied to eliminate false positives by adjusting the significance level. Figure 5.4.14 shows the resulting two-sample logo representations. The residues overrepresented in immunogenic peptides (shown in the upper half of Figure 3.14) are glycine, valine and threonine at positions 4, 6 and 8, respectively. On the other hand, the residues underrepresented in immunogenic peptides (shown in the lower half of Figure 5.4.14) are threonine and isoleucine at positions 6 and 9, respectively.

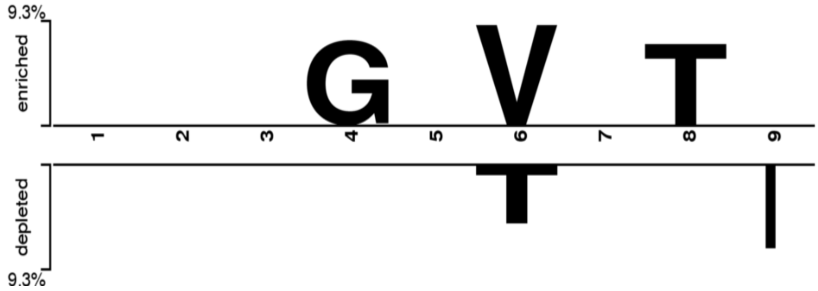


Figure 5.4.14 Two Sample Logo representation of over- (upper half) and underrepresented (lower half) residues in immunogenic peptides

Our method successfully identified previously reported TCR recognition positions (4, 6 and 8) for HLA-A2 binding peptides from an analysis of crystal structures [138, 140]. Notably, the underrepresented residue isoleucine in position 9 is the anchor residue for peptides binding to HLA-A2 molecules[158]. However, position 2, the primary anchor position of HLA-A2 binding peptides [158, 159], is not important to immunogenicity. These findings of unimportance of MHC anchor residues for immunogenicity might explain the observation that peptides with high binding affinity to MHC class I molecules do not always induce immune responses [118, 119]. It is noteworthy to note that the average predicted affinity of non-immunogenic peptides is significantly stronger than that of immunogenic peptides ($p < 0.05$, *t*-test) in IMMA2. This result confirms the idea that binding affinity is not strongly correlated with peptide immunogenicity [118, 119].

5.4.4.7. Analysis of physicochemical properties

Physicochemical properties play an important role in biomolecular recognition. The identification of important physicochemical properties will provide insights into the underlying mechanism of immunogenicity. To further investigate the position-dependent effect of important physicochemical properties, two properties were selected to encode amino acids of IMMA2 peptides to two three-alphabet sequences (small (S), medium (M) and large (L)): hydrophobicity (thresholds 0.5 and 2.5) [160] and normalized van der Waals volume (thresholds 2.0 and 6.0) [161]. The encoded sequences yielded the two-sample logos shown in Figure 5.4.15. Both primary and secondary anchor positions for MHC binding (positions 2 and 9, respectively) and position 6 prefer residues of medium hydrophobicity (Figure 5.4.15A). Positions 4, 5, 7 and 8 prefer residues of small hydrophobicity. Positions 1 and 4 prefer residues with small van der Waals volume (Figure 5.4.15B) whereas position 9 prefers medium volume residues. The logos obtained by using the other volume-related properties are similar to Figure 5.4.15B.

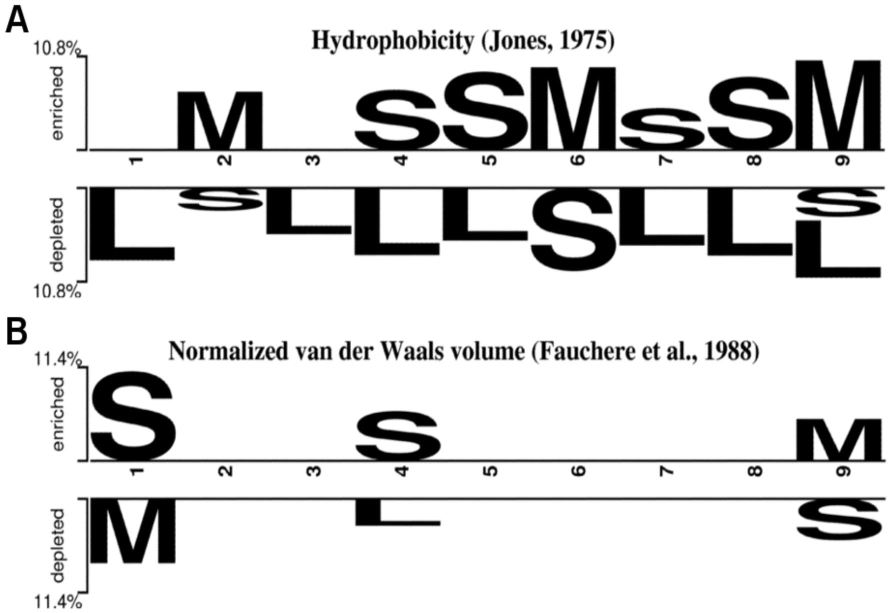


Figure 5.4.15 The over- (upper half) and underrepresented (lower half) position-specific properties in immunogenic peptides. (A) Hydrophobicity. (B) Normalized van der Waals volume. The symbols S, M and L indicate residues with small, medium and large hydrophobicity/volume, respectively.

5.4.4.8. POPISK

The prediction system named POPISK (Prediction Of Peptide Immunogenicity using String Kernels) was implemented by training an SVM classifier using weighted degree string kernel (parameters C=1 and d=9) on the whole dataset IMMA2. Users can either input a peptide sequence of length 9 that binds to HLA-A2 molecules or upload a file of multiple 9-mer sequences. POPISK will output the predicted immunogenicity (immunogenic or non-immunogenic) accompanied with a score (decision value of SVM) for the strength of immunogenicity. Peptides with a decision value larger than zero are considered immunogenic. The web server of POPISK is publicly available at http://iclab.life.nctu.edu.tw/POPISK.

To evaluate the prediction and analysis abilities of POPISK, a total of 17 crystal structures consisting of TCR, peptide of length 9 and HLA-A2 molecule were extracted from the Protein Data Bank (PDB) [162]. By removing entries with duplicate peptide sequences or modified amino acids, seven crystal structures (PDB ID: 1qrn, 1qse, 1qsf, 1ao7, 1oga, 2bnr and 2bnq) are used for the following analyses. These peptides are classified as immunogenic (1qse, 1ao7, 1oga, 2bnr and 2bnq) or non-immunogenic (1qrn and 1qsf) according to the original publications [140, 150, 151].

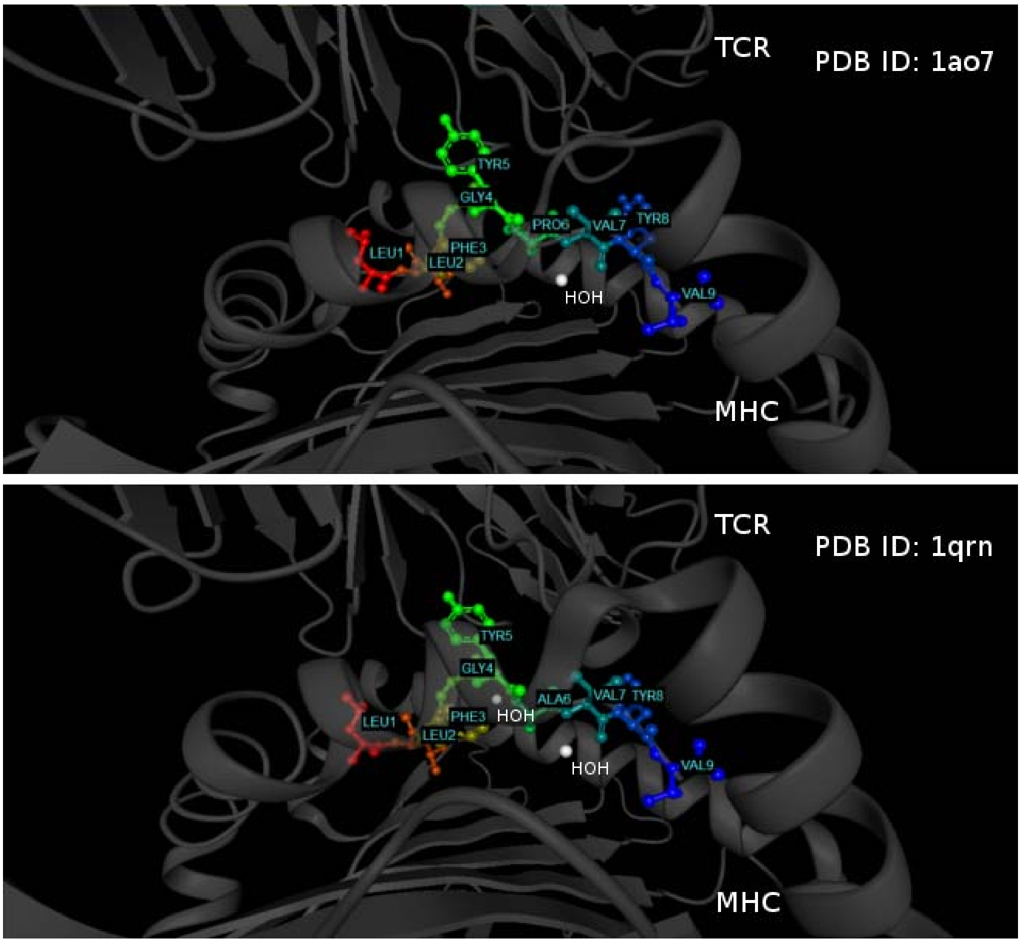


Figure 5.4.16 Structures of PDB IDs 1ao7 and 1qrn. Structures of PDB IDs 1ao7 and 1qrn share high structural similarity presenting complexes of TCR-peptide-MHC.

First, POPISK was trained by using a modified dataset that excludes peptides of the seven test peptides from IMMA2. Subsequently, POPISK was applied to predict the seven peptides. POPISK classified 5 out of 7 peptides correctly. Although the peptide of 1ao7 is misclassified, its score (-0.04) is very close to the decision threshold (0). The scores predicted by POPISK are useful for predicting the immunogenicity change made by single residue modifications. For example, the predicted results show that modified cancer/testis antigen with valine in position 9 (POPISK score: 1.36) is more immunogenic than the original antigen (POPISK score: 1.11) and are consistent with a previous study [150]. Also, compared to original Tax protein of human T-lymphotropic virus (POPISK score: -0.04), the reduced immunogenicity of three modified Tax proteins (POPISK scores: -0.07, -0.14 and -0.26) as shown in a previous study [151] is successfully predicted.

Among the seven TCR-peptide-MHC structures taken for our analyses, three different TCRs, the A6 TCR (1qrn, 1qse, 1qsf, 1ao7), the $V_{\beta}17V_{\alpha}10.2$ TCR from the T-cell clone JM22 (1oga), and the 1G4 TCR (2bnr, 2bnq) are present. Hence, a comparison from the structural perspective can only be performed for each type of TCR individually. Most interesting here is the A6 TCR, where structures with immunogenic as well as non-immunogenic peptides are available. The very high structural similarity among the structures with the A6 TCR has been stressed by Ding *et al.* [151]. These authors did not see any correlation between the overall shape of the complexes or rearrangements at the interface and immunogenicity. The overall structural similarity of complexes with the immunogenic peptide LLFGYPVYV (wild-type, 1ao7) with a POPISK score of -0.04 and the non-immunogenic peptide LLFGYAVYV (P6A, 1qrn) with a POPISK score of -0.26 was found to be highest. Also, between these two peptides no difference in their solvent-accessible surface areas could be determined. Figure 5.4.16 generated with BALLView 1.3 [163, 164] shows the two crystal structures of 1ao7 and 1qrn.

There is only one significant difference of the enlarged cavity at position 6 of the non-immunogenic peptide LLFGYAVYV in the 1qrn complex, compared with the immunogenic peptide LLFGYPVYV in the 1ao7 complex. An ordered water molecule entered this cavity, leading to some rearrangements of amino acids to accommodate the water. However, the formation of a cavity, the small rearrangements and the entropic loss due to the conserved water account for only a fraction of the difference in complex dissociation constants [151]. A second difference was evident from shape complementarity analyses, showing a hole in the interface of P6A and a decrease in complementarity [165] affecting binding to residue at position 5. These findings show that even an in-depth structural analysis of the ternary complexes can only give hints on the immunogenicity of peptides, stressing the importance of large-scale statistical studies.

6. Conclusions

6.1. Summary

We have developed the computer-aided vaccine design system in the project. To achieve the goal, the core algorithms are developed for identifying and analyzing informative features of peptides and protein sequences. Utilizing these core algorithms, various prediction tools are established to analyze the immune reactions and protein functions. Integrating prediction tools into the computer-aided vaccine design system has been finished, and furthermore, the selected physicochemical properties of prediction specific functions of proteins are collected to establish the relative feature databases of the immune system and specific protein functions. In conclusion, the immunologist can use the computer-aided vaccine design system to determine the epitope site and immunogenicity of unknown peptides and get important biological features from the feature database of immune system. The developed efficient feature mining algorithms can help researchers of bioinformatics predict protein functions.

We have good research results to achieve the goal of each year. The individual projects of the three years focus on both study of vaccine designs and algorithm developments, which are fully cooperated. The achievements of each year are described below.

Year 1: Develop high-performance feature mining algorithms which are core algorithms of the vaccine design system, and study cytotoxic T lymphocyte related immune response and informative features related to MHC class I and II binding peptides and their pathways.

Year 2: Study the immune responses of T helper cell, including the infect pathways and features of T cell using the developed optimization algorithms of informative feature mining. A number of prediction algorithms and systems for HTL and CTL immune pathways and epitopes have been proposed.

Year 3: Develop an informative feature database of immune systems and establish the computer-aided vaccine design system by integrating the achievements of the first two years and.

6.2. Achievements of the project

The computer-aided vaccine design systems have been finished and the core high-performance optimization algorithm has been applied to a number of protein function prediction problems. In the three years, we have published many international journal papers and conference papers. The publication lists are given in Tables 6.2.1 and 6.2.2. The main part of the computer-aid design system – prediction of adaptive T-cell immune response is referred to the PH.D. dissertation of C.-W. Tung [166].

Table 6.2.1 The list of international journal papers

Paper title	Journal	Year
POPI: Predicting immunogenicity of MHC class I binding peptides by mining informative physicochemical properties	Bioinformatics, issue 8	2007
ProLoc: Prediction of protein subnuclear localization using SVM with automatic selection from physicochemical composition features	BioSystems, issue 2	2007
ProLoc-GO: Utilizing informative Gene Ontology terms for sequence-based prediction of protein subcellular localization	BMC Bioinformatics, vol. 9	2008
Computational identification of ubiquitylation sites from protein sequences	BMC Bioinformatics, vol. 9	2008
Predicting protein subnuclear localization using GO-amino-acid composition features	BioSystems, issue 2	2009
Prediction of non-classical secreted proteins using informative physicochemical properties	Interdiscip Sci Comput Life Sci	2010
Improving protein secondary structure prediction based on short subsequences with local structure similarity	Accepted by BMC Genomics	2010
Predicting and analyzing DNA-binding domains using a systematic approach to identifying a set of informative physicochemical and biochemical properties	Accepted by BMC Bioinformatics	2010
Benchmark method for evaluating prediction methods on subcellular localization of eukaryotic and prokaryotic proteins	PLoS ONE (under revision)	2010
POPISK: T-cell reactivity prediction using support vector	PLoS ONE	2010

machines and string kernels	(under revision)	

Table 6.2.2 The list of international conference papers

Paper title	Conference	Year
Mining physicochemical properties for predicting immunogenicity of MHC class II binding peptides	18th International Conference on Genome Informatics. Biopolis, Singapore	2007
ProLoc-rGO: Using rule-based knowledge with Gene Ontology terms for prediction of protein subnuclear localization	IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology. (CIBCB)	2008
Virulent-GO: Prediction of virulent proteins inbacterial pathogens utilizing Gene Ontology terms	International Conference on Bioinformatics and Bioengineering (ICBB)	2009
Analysis of physicochemical properties on prediction of R5, X4 and R5X4 HIV-1 coreceptor usage	International Conference on Bioinformatics and Bioengineering (ICBB)	2009
Prediction of non-classical secreted proteins using informative physicochemical properties	The International Conference on Computational and Systems Biology (ICBBA)	2009
Human Pol II promoter prediction by using nucleotide property composition features	The International Symposium on Biocomputing (ISB)	2010
Sequence-based Prediction Of Gamma-turn Types Using A Physicochemical Property-based Decision Tree Method	International Conference on Computational Biology (ICCB2010)	2010
Prediction of protein subchloroplast locations using Random Forests	International Conference on Computational Biology (ICCB2010)	2010

References

1. Ho, S.Y., L.S. Shu, and J.H. Chen, *Intelligent evolutionary algorithms for large parameter optimization problems.* Ieee Transactions on Evolutionary Computation, 2004. **8**(6): p. 522-541.
2. Ho, S.Y., J.H. Chen, and M.H. Huang, *Inheritable genetic algorithm for biobjective 0/1 combinatorial optimization problems and its applications.* IEEE Trans Syst Man Cybern B Cybern, 2004. **34**(1): p. 609-20.
3. Ha¨ mmerling, G.J., *Antigen processing and presentation – towards the millennium.* Immunol. Rev, 1999. **172**.
4. Deavin, A.J., T.R. Auton, and P.J. Greaney, *Statistical comparison of established T-cell epitope predictors against a large database of human and murine antigens.* Mol Immunol, 1996. **33**(2): p. 145-55.
5. Kesmir, C., et al., *Prediction of proteasome cleavage motifs by neural networks.* Protein Eng, 2002. **15**(4): p. 287-96.
6. Bhasin, M. and G.P. Raghava, *Pcleavage: an SVM based method for prediction of constitutive proteasome and immunoproteasome cleavage sites in antigenic sequences.* Nucleic Acids Res, 2005. **33**(Web Server issue): p. W202-7.
7. Bhasin, M. and G.P. Raghava, *Analysis and prediction of affinity of TAP binding peptides using cascade SVM.* Protein Sci, 2004. **13**(3): p. 596-607.
8. Peters, B., et al., *Identifying MHC class I epitopes by predicting the TAP transport efficiency of epitope precursors.* J Immunol, 2003. **171**(4): p. 1741-9.
9. Donnes, P. and A. Elofsson, *Prediction of MHC class I binding peptides, using SVMHC.* BMC Bioinformatics, 2002. **3**: p. 25.
10. Nielsen, M., 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.
11. Donnes, P. and O. Kohlbacher, *Integrated modeling of the major events in the MHC class I antigen processing pathway.* Protein Sci, 2005. **14**(8): p. 2132-40.
12. Larsen, M.V., et al., *An integrative approach to CTL epitope prediction: a combined algorithm integrating MHC class I binding, TAP transport efficiency, and proteasomal cleavage predictions.* Eur J Immunol, 2005. **35**(8): p. 2295-303.
13. William, B.M., et al., *Acetylator phenotyping in patients with malignant lymphomas, using caffeine as the metabolic probe.* Pol J Pharmacol, 2004.

56(4): p. 445-9.

14. Reche, P.A., et al., *Enhancement to the RANKPEP resource for the prediction of peptide binding to MHC molecules using profiles.* Immunogenetics, 2004. **56**(6): p. 405-19.

15. Chang, S.T., et al., *Peptide length-based prediction of peptide-MHC class II binding.* Bioinformatics, 2006. **22**(22): p. 2761-7.

16. Karpenko, O., J. Shi, and Y. Dai, *Prediction of MHC class II binders using the ant colony search strategy.* Artif Intell Med, 2005. **35**(1-2): p. 147-56.

17. Murugan, N. and Y. Dai, *Prediction of MHC class II binding peptides based on an iterative learning model.* Immunome Res, 2005. **1**: p. 6.

18. Cao, Y., et al., *Prediction of protein structural class with Rough Sets.* BMC Bioinformatics, 2006. **7**: p. 20.

19. Idicula-Thomas, S., et al., *A support vector machine-based method for predicting the propensity of a protein to be soluble or to form inclusion body on overexpression in Escherichia coli.* Bioinformatics, 2006. **22**(3): p. 278-84.

20. Liu, W., et al., *Quantitative prediction of mouse class I MHC peptide binding affinity using support vector machine regression (SVR) models.* BMC Bioinformatics, 2006. **7**: p. 182.

21. Nanni, L. and A. Lumini, *An ensemble of K-local hyperplanes for predicting protein-protein interactions.* Bioinformatics, 2006. **22**(10): p. 1207-10.

22. Sarda, D., et al., *pSLIP: SVM based protein subcellular localization prediction using multiple physicochemical properties.* BMC Bioinformatics, 2005. **6**: p. 152.

23. Delorenzi, M. and T. Speed, *An HMM model for coiled-coil domains and a comparison with PSSM-based predictions.* Bioinformatics, 2002. **18**(4): p. 617-25.

24. Reche, P.A., J.P. Glutting, and E.L. Reinherz, *Prediction of MHC class I binding peptides using profile motifs.* Hum Immunol, 2002. **63**(9): p. 701-9.

25. Haste Andersen, P., M. Nielsen, and O. Lund, *Prediction of residues in discontinuous B-cell epitopes using protein 3D structures.* Protein Sci, 2006. **15**(11): p. 2558-67.

26. Petrova, N.V. and C.H. Wu, *Prediction of catalytic residues using Support Vector Machine with selected protein sequence and structural properties.* BMC Bioinformatics, 2006. **7**: p. 312.

27. Chang, C.-C. and C.-J. Lin, *LIBSVM : a library for support vector machines.* 2001.

28. Dey, A., *Orthogonal fractional factorial designs*. 1985, New York: Wiley. viii, 133 p.

29. Wu, Q., *On the optimality of orthogonal experimental design.* Acta Math.Appl. Sin., 1978. **1**: p. 16.

30. Blythe, M.J. and D.R. Flower, *Benchmarking B cell epitope prediction: underperformance of existing methods.* Protein Sci, 2005. **14**(1): p. 246-8.

31. Kawashima, S., et al., *AAindex: amino acid index database, progress report 2008.* Nucleic Acids Res, 2008. **36**(Database issue): p. D202-5.

32. Ashburner, M., et al., *Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.* Nat Genet, 2000. **25**(1): p. 25-9.

33. Huang, W.L., et al., *ProLoc: prediction of protein subnuclear localization using SVM with automatic selection from physicochemical composition features.* Biosystems, 2007. **90**(2): p. 573-81.

34. Chou, K.C. and H.B. Shen, *Predicting eukaryotic protein subcellular location by fusing optimized evidence-theoretic K-Nearest Neighbor classifiers.* J Proteome Res, 2006. **5**(8): p. 1888-97.

35. Chou, K.C. and H.B. Shen, *Hum-PLoc: a novel ensemble classifier for predicting human protein subcellular localization.* Biochem Biophys Res Commun, 2006. **347**(1): p. 150-7.

36. Bonin-Debs, A.L., et al., *Development of secreted proteins as biotherapeutic agents.* Expert Opin Biol Ther, 2004. **4**(4): p. 551-8.

37. Damas, J.K., L. Gullestad, and P. Aukrust, *Cytokines as new treatment targets in chronic heart failure.* Curr Control Trials Cardiovasc Med, 2001. **2**(6): p. 271-277.

38. Chevallet, M., et al., *Toward a better analysis of secreted proteins: the example of the myeloid cells secretome.* Proteomics, 2007. **7**(11): p. 1757-70.

39. Grimmond, S.M., et al., *The mouse secretome: functional classification of the proteins secreted into the extracellular environment.* Genome Res, 2003. **13**(6B): p. 1350-9.

40. Emanuelsson, O., et al., *Locating proteins in the cell using TargetP, SignalP and related tools.* Nat Protoc, 2007. **2**(4): p. 953-71.

41. Klee, E.W. and C.P. Sosa, *Computational classification of classically secreted proteins.* Drug Discov Today, 2007. **12**(5-6): p. 234-40.

42. Bendtsen, J.D., et al., *Feature-based prediction of non-classical and leaderless protein secretion.* Protein Eng Des Sel, 2004. **17**(4): p. 349-56.

43. Bendtsen, J.D., et al., *Improved prediction of signal peptides: SignalP 3.0.* J Mol Biol, 2004. **340**(4): p. 783-95.

44. Pierleoni, A., et al., *BaCelLo: a balanced subcellular localization predictor.* Bioinformatics, 2006. **22**(14): p. e408-16.

45. Bairoch, A., et al., *The Universal Protein Resource (UniProt).* Nucleic Acids Res,

2005. **33**(Database issue): p. D154-9.

46. Wang, G. and R.L. Dunbrack, Jr., *PISCES: a protein sequence culling server.* Bioinformatics, 2003. **19**(12): p. 1589-91.

47. Camon, E., et al., *The Gene Ontology Annotation (GOA) Database: sharing knowledge in Uniprot with Gene Ontology.* Nucleic Acids Res, 2004. **32**(Database issue): p. D262-6.

48. Altschul, S.F., et al., *Basic local alignment search tool.* J Mol Biol, 1990. **215**(3): p. 403-10.

49. Altschul, S.F., et al., *Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.* Nucleic Acids Res, 1997. **25**(17): p. 3389-402.

50. Li, T., C. Zhang, and M. Ogihara, *A comparative study of feature selection and multiclass classification methods for tissue classification based on gene expression.* Bioinformatics, 2004. **20**(15): p. 2429-37.

51. Ho, S.Y., et al., *Interpretable gene expression classifier with an accurate and compact fuzzy rule base for microarray data analysis.* Biosystems, 2006. **85**(3): p. 165-76.

52. Tung, C.W. and S.Y. Ho, *POPI: predicting immunogenicity of MHC class I binding peptides by mining informative physicochemical properties.* Bioinformatics, 2007. **23**(8): p. 942-9.

53. Klumperman, J., *Transport between ER and Golgi.* Curr Opin Cell Biol, 2000. **12**(4): p. 445-9.

54. Nickel, W., *Unconventional secretory routes: direct protein export across the plasma membrane of mammalian cells.* Traffic, 2005. **6**(8): p. 607-14.

55. Duong, F., A. Lazdunski, and M. Murgier, *Protein secretion by heterologous bacterial ABC-transporters: the C-terminus secretion signal of the secreted protein confers high recognition specificity.* Mol Microbiol, 1996. **21**(3): p. 459-70.

56. Tang, J. and J.S. Bond, *Maturation of secreted meprin alpha during biosynthesis: role of the furin site and identification of the COOH-terminal amino acids of the mouse kidney metalloprotease subunit.* Arch Biochem Biophys, 1998. **349**(1): p. 192-200.

57. Wu, H.J., A.H. Wang, and M.P. Jennings, *Discovery of virulence factors of pathogenic bacteria.* Curr Opin Chem Biol, 2008. **12**(1): p. 93-101.

58. Finlay, B.B. and S. Falkow, *Common themes in microbial pathogenicity revisited.* Microbiol Mol Biol Rev, 1997. **61**(2): p. 136-69.

59. Bairoch, A. and R. Apweiler, *The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000.* Nucleic Acids Res, 2000. **28**(1): p. 45-8.

60. Chen, L., et al., *VFDB: a reference database for bacterial virulence factors.*

Nucleic Acids Res, 2005. **33**(Database issue): p. D325-8.

61. Barrell, D., et al., *The GOA database in 2009--an integrated Gene Ontology Annotation resource.* Nucleic Acids Res, 2009. **37**(Database issue): p. D396-403.

62. Ciechanover, A., *Early work on the ubiquitin proteasome system, an interview with Aaron Ciechanover. Interview by CDD.* Cell Death Differ., 2005. **12**(9): p. 1167-77.

63. Hershko, A., *Early work on the ubiquitin proteasome system, an interview with Avram Hershko. Interview by CDD.* Cell Death Differ., 2005. **12**(9): p. 1158-61.

64. Wang, J. and M.A. Maldonado, *The ubiquitin-proteasome system and its role in inflammatory and autoimmune diseases.* Cell. Mol. Immunol., 2006. **3**(4): p. 255-61.

65. Michalek, M.T., et al., *A role for the ubiquitin-dependent proteolytic pathway in MHC class I-restricted antigen presentation.* Nature, 1993. **363**(6429): p. 552-4.

66. Townsend, A., et al., *Defective presentation to class I-restricted cytotoxic T lymphocytes in vaccinia-infected cells is overcome by enhanced degradation of antigen.* J. Exp. Med., 1988. **168**(4): p. 1211-24.

67. Liu, W.J., et al., *Polynucleotide viral vaccines: codon optimisation and ubiquitin conjugation enhances prophylactic and therapeutic efficacy.* Vaccine, 2001. **20**(5-6): p. 862-9.

68. Wang, Q.M., et al., *Epitope DNA vaccines against tuberculosis: spacers and ubiquitin modulates cellular immune responses elicited by epitope DNA vaccine.* Scand. J. Immunol., 2004. **60**(3): p. 219-25.

69. Rodriguez, F., et al., *DNA immunization with minigenes: low frequency of memory cytotoxic T lymphocytes and inefficient antiviral protection are rectified by ubiquitination.* J. Virol., 1998. **72**(6): p. 5174-81.

70. Deavin, A.J., T.R. Auton, and P.J. Greaney, *Statistical comparison of established T-cell epitope predictors against a large database of human and murine antigens.* Mol. Immunol., 1996. **33**(2): p. 145-55.

71. Keşmir, C., et al., *Prediction of proteasome cleavage motifs by neural networks.* Protein Eng., 2002. **15**(4): p. 287-296.

72. Bhasin, M. and G.P.S. Raghava, *Pcleavage: an SVM based method for prediction of constitutive proteasome and immunoproteasome cleavage sites in antigenic sequences.* Nucleic Acids Res., 2005. **33**(Web Server issue): p. W202-W207.

73. Bhasin, M. and G.P. Raghava, *Analysis and prediction of affinity of TAP binding peptides using cascade SVM.* Protein Sci., 2004. **13**(3): p. 596-607.

74. Peters, B., et al., *Identifying MHC class I epitopes by predicting the TAP transport efficiency of epitope precursors.* J. Immunol., 2003. **171**(4): p. 1741-9.

75. Dönnes, P. and A. Elofsson, *Prediction of MHC class I binding peptides, using SVMHC.* BMC Bioinformatics, 2002. **3**: p. 25.

76. Dönnes, P. and O. Kohlbacher, *Integrated modeling of the major events in the MHC class I antigen processing pathway.* Protein Sci., 2005. **14**(8): p. 2132-2140.

77. Larsen, M.V., et al., *An integrative approach to CTL epitope prediction: a combined algorithm integrating MHC class I binding, TAP transport efficiency, and proteasomal cleavage predictions.* Eur. J. Immunol., 2005. **35**(8): p. 2295-303.

78. Goldsby, R.A., et al., *Immunology*. 5th ed. 2003, New York: W.H. Freeman. 551 p.

79. van Bergen, J., et al., *Get into the groove! Targeting antigens to MHC class II.* Immunol. Rev., 1999. **172**: p. 87-96.

80. Karpenko, O., J. Shi, and Y. Dai, *Prediction of MHC class II binders using the ant colony search strategy.* Artif. Intell. Med., 2005. **35**(1-2): p. 147-56.

81. Brusic, V., et al., *Prediction of MHC class II-binding peptides using an evolutionary algorithm and artificial neural network.* Bioinformatics, 1998. **14**(2): p. 121-30.

82. Rajapakse, M., et al., *Predicting peptides binding to MHC class II molecules using multi-objective evolutionary algorithms.* BMC Bioinformatics, 2007. **8**(1): p. 459.

83. Bisset, L.R. and W. Fierz, *Using a neural network to identify potential HLA-DR1 binding sites within proteins.* J. Mol. Recognit., 1993. **6**(1): p. 41-8.

84. Honeyman, M.C., et al., *Neural network-based prediction of candidate T-cell epitopes.* Nat. Biotechnol., 1998. **16**(10): p. 966-9.

85. Burden, F.R. and D.A. Winkler, *Predictive Bayesian neural network models of MHC class II peptide binding.* J. Mol. Graph. Model., 2005. **23**(6): p. 481-9.

86. Noguchi, H., et al., *Fuzzy neural network-based prediction of the motif for MHC class II binding peptides.* J. Biosci. Bioeng., 2001. **92**(3): p. 227-31.

87. Noguchi, H., et al., *Hidden Markov model-based prediction of antigenic peptides that interact with MHC class II molecules.* J. Biosci. Bioeng., 2002. **94**(3): p. 264-70.

88. Bhasin, M. and G.P. Raghava, *SVM based method for predicting HLA-DRB1*0401 binding peptides in an antigen sequence.* Bioinformatics, 2004. **20**(3): p. 421-3.

89. Cui, J., et al., *Prediction of MHC-binding peptides of flexible lengths from sequence-derived structural and physicochemical properties.* Mol. Immunol., 2007. **44**(5): p. 866-77.

90. Wan, J., et al., *SVRMHC prediction server for MHC-binding peptides.* BMC Bioinformatics, 2006. **7**: p. 463.

91. Nielsen, M., C. Lundegaard, and O. Lund, *Prediction of MHC class II binding affinity using SMM-align, a novel stabilization matrix alignment method.* BMC Bioinformatics, 2007. **8**: p. 238.

92. Herrmann, J., L.O. Lerman, and A. Lerman, *Ubiquitin and ubiquitin-like proteins in protein regulation.* Circulation Res., 2007. **100**(9): p. 1276-91.

93. Welchman, R.L., C. Gordon, and R.J. Mayer, *Ubiquitin and ubiquitin-like proteins as multifunctional signals.* Nat. Rev. Mol. Cell. Biol., 2005. **6**(8): p. 599-609.

94. Tomlinson, E., et al., *Methods for the purification of ubiquitinated proteins.* Proteomics, 2007. **7**(7): p. 1016-22.

95. Denis, N.J., et al., *Tryptic digestion of ubiquitin standards reveals an improved strategy for identifying ubiquitinated proteins by mass spectrometry.* Proteomics, 2007. **7**(6): p. 868-74.

96. Hitchcock, A.L., et al., *A subset of membrane-associated proteins is ubiquitinated in response to mutations in the endoplasmic reticulum degradation machinery.* Proc. Natl. Acad. Sci. U.S.A., 2003. **100**(22): p. 12735-40.

97. Jeon, H.B., et al., *A proteomics approach to identify the ubiquitinated proteins in mouse heart.* Biochem. Biophys. Res. Commun., 2007. **357**(3): p. 731-6.

98. Kirkpatrick, D.S., et al., *Proteomic identification of ubiquitinated proteins from human cells expressing His-tagged ubiquitin.* Proteomics, 2005. **5**(8): p. 2104-11.

99. Matsumoto, M., et al., *Large-scale analysis of the human ubiquitin-related proteome.* Proteomics, 2005. **5**(16): p. 4145-51.

100. Peng, J., et al., *A proteomics approach to understanding protein ubiquitination.* Nat. Biotechnol., 2003. **21**(8): p. 921-6.

101. Denison, C., D.S. Kirkpatrick, and S.P. Gygi, *Proteomic insights into ubiquitin and ubiquitin-like proteins.* Curr. Opin. Chem. Biol., 2005. **9**(1): p. 69-75.

102. Xue, Y., et al., *NBA-Palm: prediction of palmitoylation site implemented in Naive Bayes algorithm.* BMC Bioinformatics, 2006. **7**: p. 458.

103. Jones, D.T., *Improving the accuracy of transmembrane protein topology prediction using evolutionary information.* Bioinformatics, 2007. **23**(5): p. 538-44.

104. Kaur, H. and G.P. Raghava, *A neural network method for prediction of beta-turn types in proteins using evolutionary information.* Bioinformatics, 2004. **20**(16): p. 2751-8.

105. Altschul, S.F., et al., *Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.* Nucleic Acids Res., 1997. **25**(17): p. 3389-402.

106. Witten, I.H. and E. Frank, *Data Mining: Practical machine learning tools and techniques*. 2nd ed. 2005, San Francisco: Morgan Kaufmann.

107. Chernorudskiy, A.L., et al., *UbiProt: a database of ubiquitylated proteins.* BMC Bioinformatics, 2007. **8**: p. 126.

108. Crooks, G.E., et al., *WebLogo: a sequence logo generator.* Genome Res, 2004. **14**(6): p. 1188-90.

109. Quinlan, J.R., *C4.5: programs for machine learning*. 1993, Morgan Kaufmann: San Mateo, CA.

110. Wu, Q., *On the optimality of orthogonal experimental design.* Acta Math. Appl. Sinica, 1978. **1**: p. 283-299.

111. Meirovitch, H., Rackovsky, S. and Scheraga, H.A., *Empirical studies of hydrophobicity. 1. Effect of protein size on the hydrophobic behavior of amino acids.* Macromolecules, 1980. **13**: p. 1398-1405.

112. Harpaz, Y., M. Gerstein, and C. Chothia, *Volume changes on protein folding.* Structure, 1994. **2**(7): p. 641-9.

113. Cornette, J.L., et al., *Hydrophobicity scales and computational techniques for detecting amphipathic structures in proteins.* J Mol Biol, 1987. **195**(3): p. 659-85.

114. Cedano, J., et al., *Relation between amino acid composition and cellular location of proteins.* J Mol Biol, 1997. **266**(3): p. 594-600.

115. George, R.A. and J. Heringa, *An analysis of protein domain linkers: their classification and role in protein folding.* Protein Eng, 2002. **15**(11): p. 871-9.

116. Kanduc, D., *Peptimmunology: immunogenic peptides and sequence redundancy.* Curr. Drug. Discov. Technol., 2005. **2**(4): p. 239-44.

117. Van Regenmortel, M.H., *Antigenicity and immunogenicity of synthetic peptides.* Biologicals, 2001. **29**(3-4): p. 209-13.

118. Feltkamp, M.C., et al., *Efficient MHC class I-peptide binding is required but does not ensure MHC class I-restricted immunogenicity.* Mol. Immunol., 1994. **31**(18): p. 1391-401.

119. Ochoa-Garay, J., et al., *The ability of peptides to induce cytotoxic T cells in vitro does not strongly correlate with their affinity for the H-2Ld molecule: implications for vaccine design and immunotherapy.* Mol. Immunol., 1997. **34**(3): p. 273-81.

120. Dow, C., et al., *Lymphocytic choriomeningitis virus infection yields overlapping CD4+ and CD8+ T-cell responses.* J. Virol. , 2008. **82**(23): p. 11734-41.

121. Arnold, P.Y., et al., *The majority of immunogenic epitopes generate CD4+ T cells that are dependent on MHC class II-bound peptide-flanking residues.* J. Immunol., 2002. **169**(2): p. 739-49.

122. Conant, S.B. and R.H. Swanborg, *MHC class II peptide flanking residues of exogenous antigens influence recognition by autoreactive T cells.* Autoimmun. Rev., 2003. **2**(1): p. 8-12.

123. Blythe, M.J. and D.R. Flower, *Benchmarking B cell epitope prediction: underperformance of existing methods.* Protein Sci., 2005. **14**(1): p. 246-8.

124. Sarda, D., et al., *pSLIP: SVM based protein subcellular localization prediction using multiple physicochemical properties.* BMC Bioinformatics, 2005. **6**: p. 152.

125. Chang, C.C. and C.J. Lin, *LIBSVM : a library for support vector machines. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.* 2001.

126. Brusic, V., G. Rudy, and L.C. Harrison, *MHCPEP, a database of MHC-binding peptides: update 1997.* Nucleic Acids Res., 1998. **26**(1): p. 368-71.

127. Myers, E.W. and W. Miller, *Optimal alignments in linear space.* Comput. Appl. Biosci., 1988. **4**(1): p. 11-7.

128. Geisow, M.J. and R.D.B. Roberts, *Amino acid preferences for secondary structure vary with protein class.* Int. J. Biol. Macromol., 1980. **2**: p. 387-389.

129. Miyazawa, S. and R.L. Jernigan, *Estimation of effective interresidue contact energies from protein crystal structures: Quasi-chemical approximation.* Macromolecules, 1985. **18**: p. 534-552.

130. Kuhn, L.A., et al., *Atomic and residue hydrophilicity in the context of folded protein structures.* Proteins, 1995. **23**(4): p. 536-47.

131. Degli Esposti, M., M. Crimi, and G. Venturoli, *A critical evaluation of the hydropathy profile of membrane proteins.* Eur. J. Biochemistry, 1990. **190**(1): p. 207-19.

132. Rajesh, S., et al., *Ubiquitin binding interface mapping on yeast ubiquitin hydrolase by NMR chemical shift perturbation.* Biochemistry, 1999. **38**(29): p. 9242-53.

133. Sundberg, E.J., et al., *Estimation of the hydrophobic effect in an antigen-antibody protein-protein interface.* Biochemistry, 2000. **39**(50): p. 15375-87.

134. Melton, S.J. and S.J. Landry, *Three dimensional structure directs T-cell epitope dominance associated with allergy.* Clin. Mol. Allergy, 2008. **6**: p. 9.

135. Mirano-Bascos, D., M. Tary-Lehmann, and S.J. Landry, *Antigen structure*

influences helper T-cell epitope dominance in the human immune response to HIV envelope glycoprotein gp120. Eur. J. Immunol., 2008. **38**(5): p. 1231-7.

136. Jager, E., et al., *Recombinant vaccinia/fowlpox NY-ESO-1 vaccines induce both humoral and cellular NY-ESO-1-specific immune responses in cancer patients.* Proc. Natl. Acad. Sci. U.S.A., 2006. **103**(39): p. 14453-8.

137. Odunsi, K., et al., *Vaccination with an NY-ESO-1 peptide of HLA class I/II specificities induces integrated humoral and T cell responses in ovarian cancer.* Proc. Natl. Acad. Sci. U.S.A., 2007. **104**(31): p. 12837-42.

138. Rudolph, M.G., J.G. Luz, and I.A. Wilson, *Structural and thermodynamic correlates of T cell signaling.* Annu. Rev. Biophys. Biomol. Struct., 2002. **31**: p. 121-49.

139. Silver, M.L., et al., *Atomic structure of a human MHC molecule presenting an influenza virus peptide.* Nature, 1992. **360**(6402): p. 367-9.

140. Stewart-Jones, G.B., et al., *A structural basis for immunodominant human T cell receptor recognition.* Nat. Immunol., 2003. **4**(7): p. 657-63.

141. Bowness, P., R.L. Allen, and A.J. McMichael, *Identification of T cell receptor recognition residues for a viral peptide presented by HLA B27.* Eur. J. Immunol. , 1994. **24**(10): p. 2357-63.

142. Boisvert, S., et al., *HIV-1 coreceptor usage prediction without multiple alignments: an application of string kernels.* Retrovirology, 2008. **5**: p. 110.

143. El-Manzalawy, Y., D. Dobbs, and V. Honavar, *Predicting linear B-cell epitopes using string kernels.* J. Mol. Recognit., 2008. **21**(4): p. 243-55.

144. Jacob, L. and J.P. Vert, *Efficient peptide-MHC-I binding prediction for alleles with few known binders.* Bioinformatics, 2008. **24**(3): p. 358-66.

145. Rätsch, G., S. Sonnenburg, and B. Scholkopf, *RASE: recognition of alternatively spliced exons in C.elegans.* Bioinformatics, 2005. **21 Suppl 1**: p. i369-77.

146. Sonnenburg, S., et al., *POIMs: positional oligomer importance matrices--understanding support vector machine-based signal detectors.* Bioinformatics, 2008. **24**(13): p. i6-14.

147. Rammensee, H., et al., *SYFPEITHI: database for MHC ligands and peptide motifs.* Immunogenetics, 1999. **50**(3-4): p. 213-9.

148. Schuler, M.M., M.D. Nastke, and S. Stevanovikc, *SYFPEITHI: database for searching and T-cell epitope prediction.* Meth. Mol. Biol., 2007. **409**: p. 75-93.

149. Peters, B., et al., *The immune epitope database and analysis resource: from vision to blueprint.* PLoS Biology, 2005. **3**(3): p. e91.

150. Chen, J.L., et al., *Structural and kinetic basis for heightened immunogenicity of T cell vaccines.* J. Exp. Med., 2005. **201**(8): p. 1243-55.

151. Ding, Y.H., et al., *Four A6-TCR/peptide/HLA-A2 structures that generate very*

different T cell signals are nearly identical. Immunity, 1999. **11**(1): p. 45-56.

152. Vacic, V., L.M. Iakoucheva, and P. Radivojac, *Two Sample Logo: a graphical representation of the differences between two sets of sequence alignments.* Bioinformatics, 2006. **22**(12): p. 1536-7.

153. Lund, O., et al., *Definition of supertypes for HLA molecules using clustering of specificity matrices.* Immunogenetics, 2004. **55**(12): p. 797-810.

154. Rätsch, G. and S. Sonnenburg, *Accurate Splice Site Prediction for Caenorhabditis Elegans*, in *MIT Press MIT Press series on Computational Molecular Biology*. 2003. p. 277-298.

155. Sonnenburg, S., et al., *Accurate splice site prediction using support vector machines.* BMC Bioinformatics, 2007. **8 Suppl 10**: p. S7.

156. Sonnenburg, S., et al., *Large scale multiple kernel learning.* J. Mach. Learn. Res., 2006. **7**: p. 1531-1565.

157. Varma, S. and R. Simon, *Bias in error estimation when using cross-validation for model selection.* BMC Bioinformatics, 2006. **7**: p. 91.

158. Hunt, D.F., et al., *Characterization of peptides bound to the class I MHC molecule HLA-A2.1 by mass spectrometry.* Science, 1992. **255**(5049): p. 1261-3.

159. Falk, K., et al., *Allele-specific motifs revealed by sequencing of self-peptides eluted from MHC molecules.* Nature, 1991. **351**(6324): p. 290-6.

160. Jones, D.D., *Amino acid properties and side-chain orientation in proteins: a cross correlation appraoch.* J Theor Biol, 1975. **50**(1): p. 167-83.

161. Fauchere, J.L., et al., *Amino acid side chain parameters for correlation studies in biology and pharmacology.* Int. J. Pept. Protein Res., 1988. **32**(4): p. 269-78.

162. Berman, H.M., et al., *The Protein Data Bank.* Nucleic Acids Res, 2000. **28**(1): p. 235-42.

163. Kohlbacher, O. and H.P. Lenhof, *BALL--rapid software prototyping in computational molecular biology. Biochemicals Algorithms Library.* Bioinformatics, 2000. **16**(9): p. 815-24.

164. Moll, A., et al., *BALLView: an object-oriented molecular visualization and modeling framework.* J. Comput. Aided Mol. Des., 2005. **19**(11): p. 791-800.

165. Baker, B.M., et al., *Structural, biochemical, and biophysical studies of HLA-A2/altered peptide ligands binding to viral-peptide-specific human T-cell receptors.* Cold Spring Harbor Symposia on Quantitative Biology, 1999. **64**: p. 235-41.

166. Tung, C.-W., *Prediction of adaptive T-cell immune response.* PH.D. Dissertation of Institute of Bioinformatics and Systems Biology, National Chiao Tung University, 2010.

國科會補助專題研究計畫成果報告自評表

請就研究內容與原計畫相符程度、達成預期目標情況、研究成果之學術或應用價值（簡要敘述成果所代表之意義、價值、影響或進一步發展之可能性）、是否適合在學術期刊發表或申請專利、主要發現或其他有關價值等，作一綜合評估。

1. 請就研究內容與原計畫相符程度、達成預期目標情況作一綜合評估

■ 達成目標

□ 未達成目標（請說明，以 100 字為限）

　□ 實驗失敗

　□ 因故實驗中斷

　□ 其他原因

說明：

2. 研究成果在學術期刊發表或申請專利等情形：

論文：■已發表 □未發表之文稿 □撰寫中 □無

專利：□已獲得 □申請中 ■無

技轉：□已技轉 □洽談中 ■無

其他：（以 100 字為限）

3. 請依學術成就、技術創新、社會影響等方面，評估研究成果之學術或應用價值（簡要敘述成果所代表之意義、價值、影響或進一步發展之可能性）（以 500 字為限）

本計畫以三年時間完成一套電腦輔助疫苗設計系統，包括多功能的各種知識挖掘演算法，在學術及技術方面，都有相當良好的成果。在本計畫執行期間，已有多篇期刊論文及研討會論文發表，其中不乏發表在生物資訊領域中具領導地位的期刊，如 Bioinformatics 和 BMC Bioinformatics，另有 2 篇 PLoS ONE 複審中，本計畫優越的成果表現在所發表論文之質與量上。

在技術層面，本計畫所開發之高效能生物特徵篩選演算法，除了應用在電腦輔助疫苗設計系統的核心，亦有許多延伸之應用，如 DNA 結合蛋白質之預測、人體免疫缺失病毒協同受體種類之預測等，都有相當好的成果。

而本計畫開發之電腦輔助疫苗設計系統，不僅可以加速候選胜肽的篩選，還具有提供相關生物資訊的功能，協助疫苗設計。免疫學家可藉由此系統所提供之生物特徵資訊，快速且準確地去設計出疫苗的結構及相關特性。對於研發所需之時間和金錢，藉由此系統皆能大幅的減少，已達到用最少的資源獲得最大產能的目標。