國立交通大學

生物資訊及系統生物研究所

博士論文

預測T細胞後天免疫反應

Prediction of adaptive T-cell immune response

研究生：童俊維

指導教授：何信瑩 教授

中華民國九十九年七月

預測T細胞後天免疫反應

Prediction of adaptive T-cell immune response

研究生：童俊維	Student：Chun-Wei Tung
指導教授：何信瑩	Advisor：Shinn-Ying Ho

國立交通大學
生物資訊及系統生物研究所
博士論文

A Thesis
Submitted to Institute of Bioinformatics and Systems Biology
College of Biological Science and Technology
National Chiao Tung University
in partial Fulfillment of the Requirements
for the Degree of Doctor of Philosophy

in

Bioinformatics and Systems Biology

July 2010

Hsinchu, Taiwan, Republic of China

中華民國九十九年七月

預測T細胞後天免疫反應

學生：童俊維　　　　　　　　　　指導教授：何信瑩 教授

國立交通大學生物資訊及系統生物研究所博士班

摘　要

發展電腦輔助疫苗設計系統能幫助免疫學家能快速有效的辨識候選疫苗，並且是免疫資訊學的終極目標之一。而精準的預測T細胞後天免疫反應是發展電腦輔助疫苗設計系統的關鍵。本研究之核心為發展能適用於探勘致免疫路徑（immunogenic pathways）中各種反應之重要物化特性的高性能大量參數最佳化演算法。此重要物理化學特性探勘系統之研發過程包含了三個重要的步驟：(1) 蒐集各種能夠有效解釋生物現象之物理化學特性；(2) 結合生物知識與演算法技巧來建立最佳化問題；(3) 發展特定的高性能演算法來解決最佳化設計問題。重要特徵探勘系統可從大量資訊中探勘重要特徵來解釋各種免疫反應，並幫助建立T細胞後天免疫反應預測系統。

T細胞後天免疫反應包括有細胞毒性與輔助T細胞免疫反應。對於預測T細胞後天免疫反應，過去的研究多專注於建立主要組織相容性複合物(MHC)第一及第二型分子的抗原處理與表現路徑之預測模型。然而被主要組織相容性複合物結合的胜肽抗原並不一定能引起免疫反應。對於更複雜的T細胞免疫反應需要有更深入的研究並建立其預測模型。另外，對於抗原表現有重要影響的蛋白質泛素化（Ubiquitylation），至今仍未有預測模型。高度泛素化的蛋白因較容易被裂解，因而容易產生可供T細胞辨認用的抗原。因此準確的蛋白質泛素化預測將有助於辨識容易引起免疫反應的蛋白質抗原。

本研究專注在研究抗原的內生性物化特性，研發出第一套使用物化特性來預測與主要組織相容性複合物結合之蛋白質引起的T細胞免疫反應預測系統POPI與泛素化預測系統UbiPred。並發現過去普遍認同的抗原與主要組織相容性複合物的結合親和力並不足以準確預測T細胞免疫反應。針對影響細胞毒性與輔助T細胞免疫反應的重要物化特性之分析比較對於了解免疫反應有極大助益。本研究接著提出基於字串核函數的細胞毒性T細胞免疫反應預測模型 POPISK。藉由融入主要組織相容性複合物與胺基酸位置的資訊，

POPISK 不僅能加強細胞毒性 T 細胞免疫反應之預測，同時也能準確預測由單一胺基酸突變引起的免疫反應變化。本研究並利用 POPISK 之特性來研究蛋白抗原上被 T 細胞辨認的重要位置。本篇研究結果將能幫助了解免疫系統並加速新疫苗的發展。

關鍵字：致免疫性路徑，物理化學特性，智慧型基因演算法，疫苗設計

Prediction of adaptive T-cell immune response

Student: Chun-Wei Tung　　Advisor: Shinn-Ying Ho

Institute of Bioinformatics and Systems Biology
National Chiao Tung University

Abstract

The development of computer-aided vaccine design systems is a goal of immunoinformatics that can largely accelerate the design of vaccines. Accurate prediction of adaptive T-cell immune response is the critical step to develop computer-aided vaccine design systems. The core of this study is to develop high-performance optimization algorithms for solving large-scale parameter optimization problems of bioinformatics to mine informative physicochemical properties from known experimental data for predicting immunogenic pathway. The development of these algorithms involves three major phases: (a) collection of physicochemical properties for encoding peptide sequences; (b) formulation of optimization problems using domain knowledge and computing techniques and, and (c) development of efficient optimization algorithms for solving optimization problems. The developed informative feature mining algorithms can be used to mine informative physicochemical properties for predicting peptide immunogenicity.

There are two major T cells including cytotoxic and helper T cells. For the prediction of adaptive T-cell immune response, previous studies mainly focused on modeling antigen processing and presentation pathways of MHC class I and II. However, the prediction of T-cell response is much harder and less addressed because of the complex nature of T-cell response. Moreover, because over-ubiquitylated protein correlated with its half life, ubiquitylation plays an important role in providing antigen sources. Accurate prediction of ubiquitylation sites is helpful to identify immunogenic peptides.

This study proposed the first prediction systems POPI and UbiPred for predicting T-cell response and ubiquitylation sites, respectively. The poor performance of a well recognized affinity-based method shows that binding affinity only is not sufficient for predicting T-cell response. The informative physicochemical properties for cytotoxic and helper T cells are identified and analyzed. Subsequently, an improved prediction system POPISK is proposed to predict cytotoxic T-cell response. The POPISK prediction system incorporating MHC allele information is used to identify important positions for T-cell recognition, and can predict immunogenicity changes made by single residue modifications. This study yields insights into the mechanism of immune response and can accelerate the development of vaccines.

Keywords: immunogenic pathway; physicochemical properties; intelligent genetic algorithm; vaccine design.

To my family

Acknowledgement

First of all, I would like to thank my supervisor, Prof. Shinn-Ying Ho, whose guidance and support enabled me to pursue this interesting research. He is the principal investigator of Intelligent Computing Lab (ICLAB). Without his profound knowledge in research experience, it is impossible for me to finish this dissertation. His patience and kindness are greatly appreciated. I learned so much from him in every aspect, especially in the presentation of research ideas and professional ethics.

I am very grateful to my supervisor, Prof. Oliver Kohlbacher, during the time of my research visits to Germany. He showed me different ways to rethink a problem. He always encouraged me and supported me to pursue my research interests. I greatly enjoyed our discussions and learned a lot from his valuable guidance and open mind.

Besides my supervisors, I would like to show my gratitude to my dissertation committees, Prof. Jenn-Kang Hwang, Prof. Jinn-Moon Yang, Prof. Hsien-Da Huang, Prof. Hsueh-Fen Juan and Prof. Hsuan-Cheng Huang for giving me valuable suggestions and criticisms.

Furthermore, I would like to thank the National Science Council (NSC) of Taiwan and the German Academic Exchange Service (DAAD) for supporting the Scholarship of Sandwich Program for research visits to Germany (97-2911-I-009-016-2).

I would like to thank all my Taiwan labmates Chyn Liaw, Prof. Wen-Ling Huang, Prof. Hui-Ling Huag, Kuan-Wei Chen, Shiou-Bang Chang, Jhong-Chi Guo, Chih-Guo Yeh, Yi-Hsiung Chen, Kai-Ti Hsu, Chia-Ta Tsai, Kuo-Ching Kao, Yu-Hsiang Lin, Chiung-Hui Hung, Yi-Jer Lin, Te-Fen Kao, Yin-Yin Yu, Pin-Chen Lion, Chih-Hung Hsieh, Chia-Yun Chang, and Fu-Chieh Yu for nice collaborations and discussions.

For giving me enthusiastic support in Germany, I am very grateful to Andreas Bertsch, Sebastian Briesemeister, Magdalena Feldhahn, Nina Fischer, Sandra Gesing, Andreas Kaemper, Erhan Kenar, Sven Nahnsen, Lars Nilse, Marc Roettig, Marc Sturm, Nico, Pfeifer, Nora Toussaint, Matthias Ziehm, Jan Schulze, and Claudia Walter. There is so much to thank them for their valuable research and daily life supports.

I owe my deepest gratitude to my Sandwich Program partners Shu-Hsien Li, Chun-Chi Lee, Tung-Hui Hsu, Chih-Chun Hsieh, Ning-Ying Wang, Win-Bin Huang, Tzai-Tang Tsai, Yu-Jen Huang, Hsien-Yi Hsiao, Hsiao-Ching Li, Hui-Yi Shiao, Guang-Yin Chen, Hao-Peng Lin for their daily life support in Germany.

Last, but no least, I am deeply grateful to my parents for their unconditional support and encouragement to pursue my interests.

Chun-Wei Tung

Table of Contents

List of Figures

List of Tables

List of Abbreviations

AUC	Area under Receiver Operating Characteristic Curve
APC	Antigen Presenting Cell
ER	Endoplasmic Reticulum
GA	Genetic Algorithm
DNA	Deoxyribonucleotide Acid
CTL	Cytotoxic T Lymphocytes
HTL	Helper T Lymphocytes
HLA	Human Leukocyte Antigen
IBCGA	Inheritable Bi-objective Genetic Algorithm
IPMA	Informative Physicochemical Property Mining Algorithm
MHC	Major Histocompatibility Complex
MED	Main Effect Difference
SVM	Support Vector Machine
TAP	Transporter associated with Antigen Processing

Chapter 1

Introduction

This study aims to develop computer-aided vaccine design systems to help the design of vaccines. For peptide-based vaccine design, the most critical step is to select immunogenic peptides capable of inducing immune responses. Therefore, accurate prediction of adaptive T-cell immune response can greatly accelerate vaccine designs. For this propose, efficient optimization algorithms were developed and applied to mine informative features for the predictions of adaptive T-cell immune response. This dissertation presents a comprehensive study of the developed prediction systems for identifying peptide vaccine candidates.

1.1 Motivation

The most significant advance of medicine is the utilization of vaccines against diseases. Vaccines can help to prevent infections and prolong peoples' life. There are mainly five types of vaccines: live attenuated vaccines, killed vaccines, purified subunit vaccines, recombinant subunit vaccines and gene-based vaccines [1-4].The live attenuated vaccines consist of the pathogens with reduced toxicity to prevent the risk of infections. However, it is hard to develop live attenuated vaccines with high safety. The killed vaccines consist of killed or deactivated pathogens have higher safety, compared to live attenuated vaccines.

The killed vaccines suffered from incapability of replication result in low immunogenicity. The deficiency of these two types of vaccines is caused by using whole pathogens as vaccines that will dilute the immunogenicity of vaccines. Thus, the subunit vaccines using identified protective antigens are safer and more efficient to focus the immune response on specific target. For preventing or curing cancers,

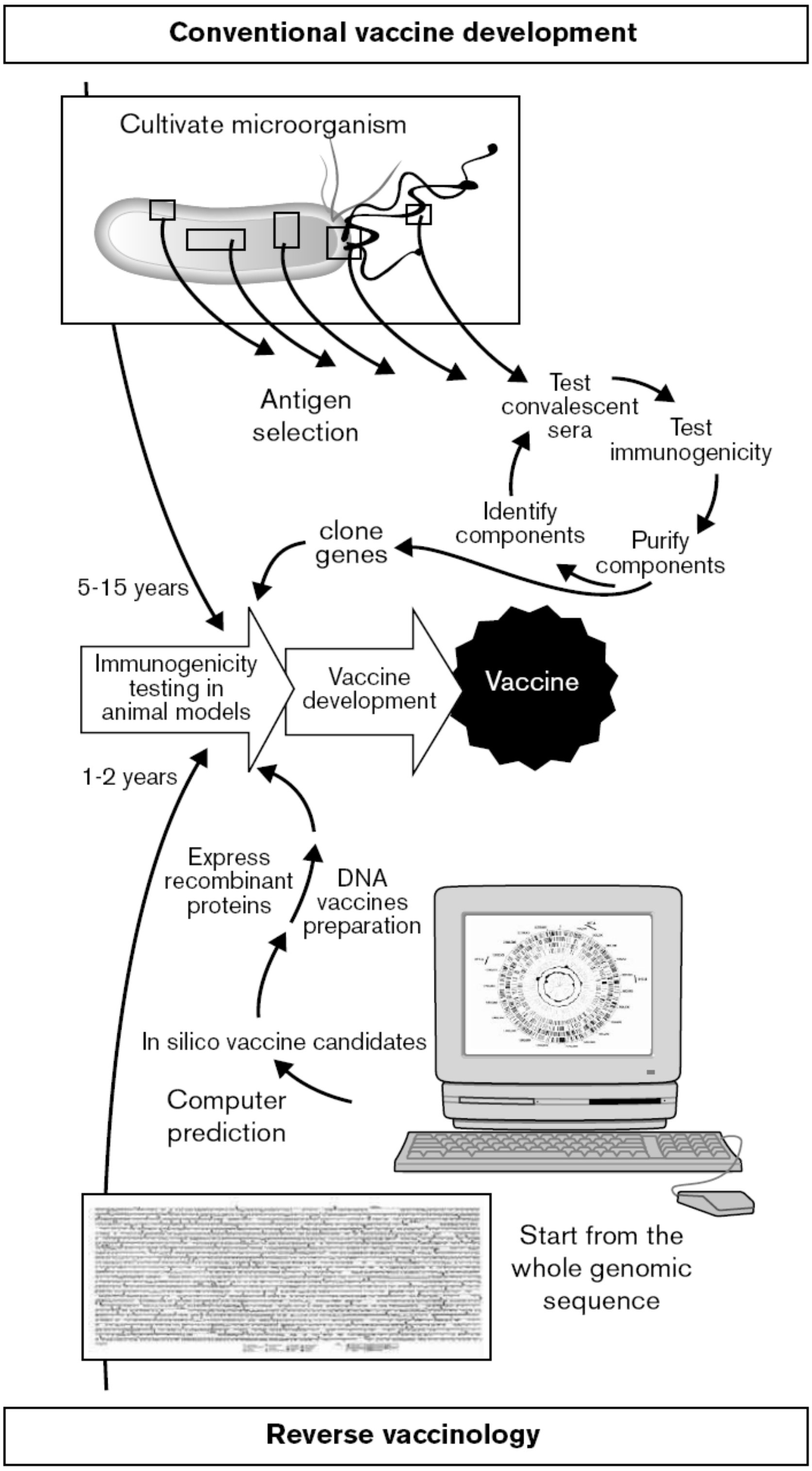


Figure 1.1 Comparison between conventional vaccine development and reverse vaccinology [7].

the traditional vaccines described above are less effective. By applying deoxyribonucleotide acid (DNA) as vaccines, it can provide effective protections against cancers with high immunogenicity of cytotoxic T lymphocyte (CTL) [1].

For developing subunit vaccines and DNA vaccines, traditional experiment methods to identify protective antigens cost a lot and is time consuming with often five to fifteen years duration. The growing needs of identifying protective antigen to develop vaccines result in the emergence of reverse vaccinology (shown in Figure1.1). Vaccine design using bioinformatics methods can largely reduce the cost of time and money [2-4].

Peptide immunogenicity, its ability to induce immune responses, determines the effectiveness of vaccines. T-cell activation, one kind of immune responses, plays important roles in developing adaptive immunity. An immunogenic peptide should be processed and presented to a cell surface by antigen processing and presentation pathway, and then induce T-cell responses. Major histocompatibility complex (MHC) molecules are responsible for both recognition of antigens and presentation of antigens to T cells. MHC class I molecules can present processed endogenous peptides of antigen to cytotoxic T lymphocytes (CTL), while processed exogenous peptides of antigen are presented to helper T lymphocytes (HTL) by MHC class II molecules.

Immune responses will be triggered when CTL or HTL recognize immunogenic antigens. CTL response is mainly characterized by killing target cells (e.g. tumor cell and infected cell), presenting immunogenic antigens by the activated CTL. In contrast, the activated HTL will induce resting HTLs to proliferate and differentiate into memory cells or effector cells, and provide specific help for CTL, B lymphocytes and phagocytic cells, as known as HTL response. A simple illustration of helper and cytotoxic T cell responses is shown in Figure 1.2.

For computer-based vaccine design, previous studies pocus on modeling antigen processing and presentation pathways of MHC class I and II. The works for modeling the pathway of MHC class I (shown in reaction 2-4 of Figure1.3) include predictions of antigen proteasomal cleavage sites, binding affinities of peptides and the transporter associated with antigen processing (TAP) and binding affinities of peptides and MHC class I molecules. The major work for modeling the pathway of MHC class II (shown in reaction 6 of Figure1.3) is the prediction of binding affinities between peptides and MHC class II molecules.

The above studies assume that peptide binding affinity to MHC molecules correlates with its immunogenicity. The prediction problems of CTL and HTL immune responses are rarely studied (reaction 5 and 7 of Figure1.3). However, recent studies

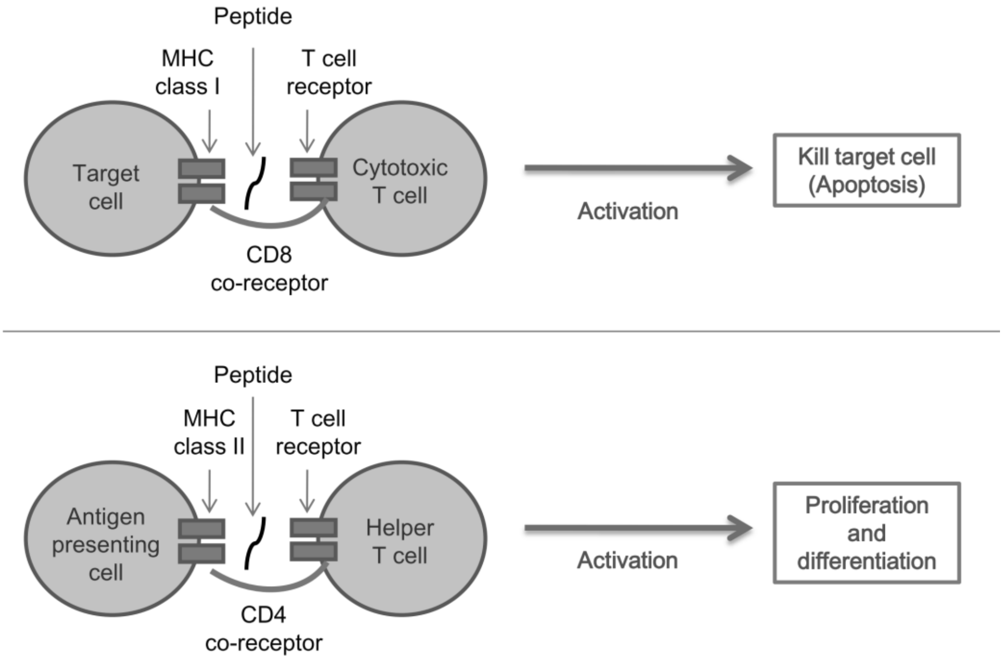


Figure 1.2 A simple illustration of helper and cytotoxic T cell responses.

show that peptide binding affinity to MHC molecules is required but not strongly correlated to the strength of immunogenicity. Also, the prediction of protein ubiquitylation sites is crucial for the prediction of peptide immunogenicity because ubiquitylation plays key roles in antigen supply (reaction 1 of Figure1.3). The investigation of these problems is necessary for accurate prediction of adaptive T-cell immune response and development of computer-aided vaccine design systems.

1.2 Overview of the research

This dissertation presents a comprehensive prediction system for computer-aided vaccine design whose architecture is shown in Figure 1.4. The proposed system is based on several state-of-the-art methods for predicting antigen processing and presentation pathways and newly developed prediction systems for T-cell responses and protein ubiquitylation. To develop prediction systems for T-cell responses and protein ubiquitylation, An informative physicochemical property mining algorithm is proposed to mine informative physicochemical properties for predicting reactions 1, 5 and 7 in this system that are described as follows.

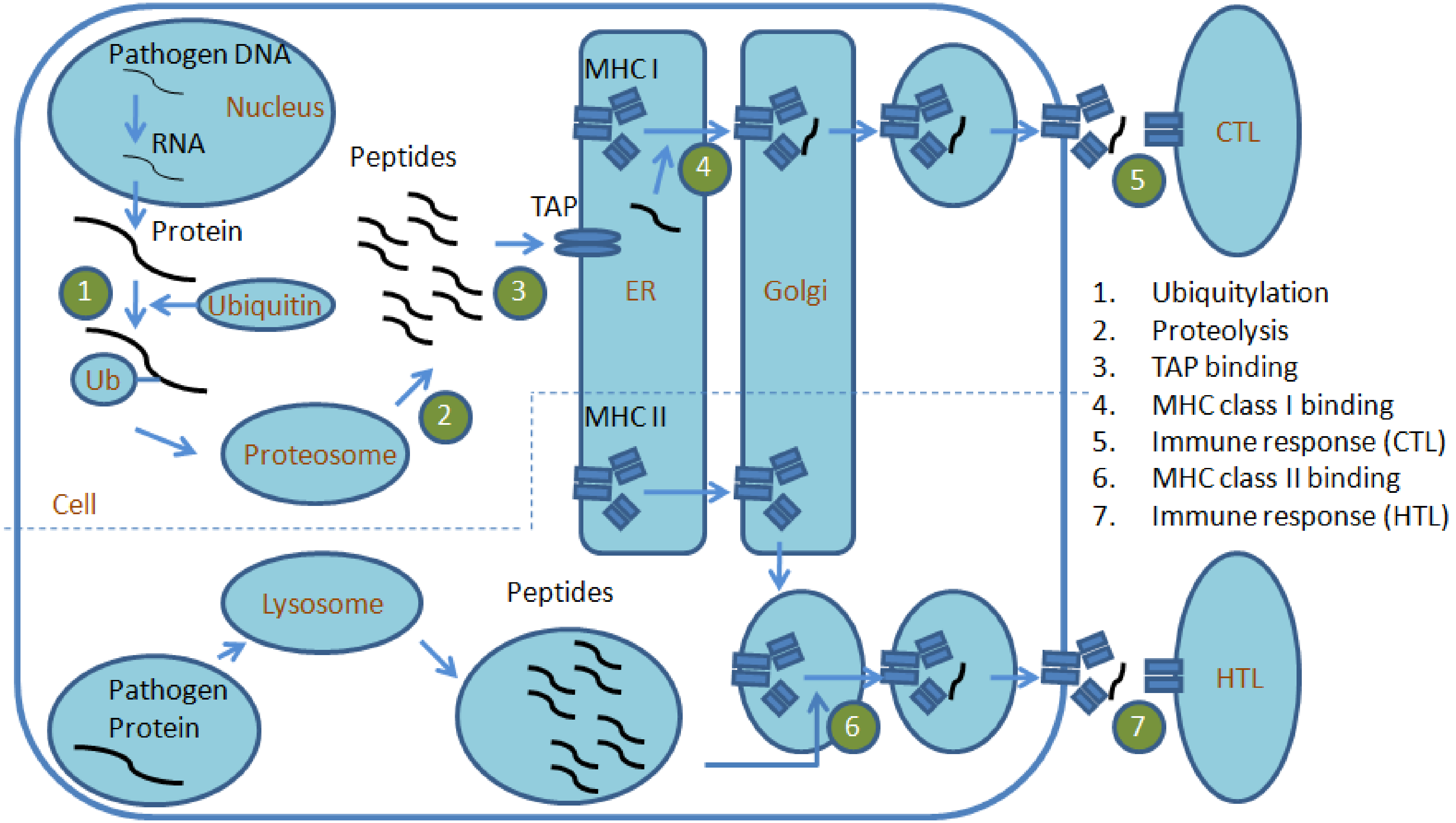


Figure 1.3 The immunogenic pathways of MHC class I and II.

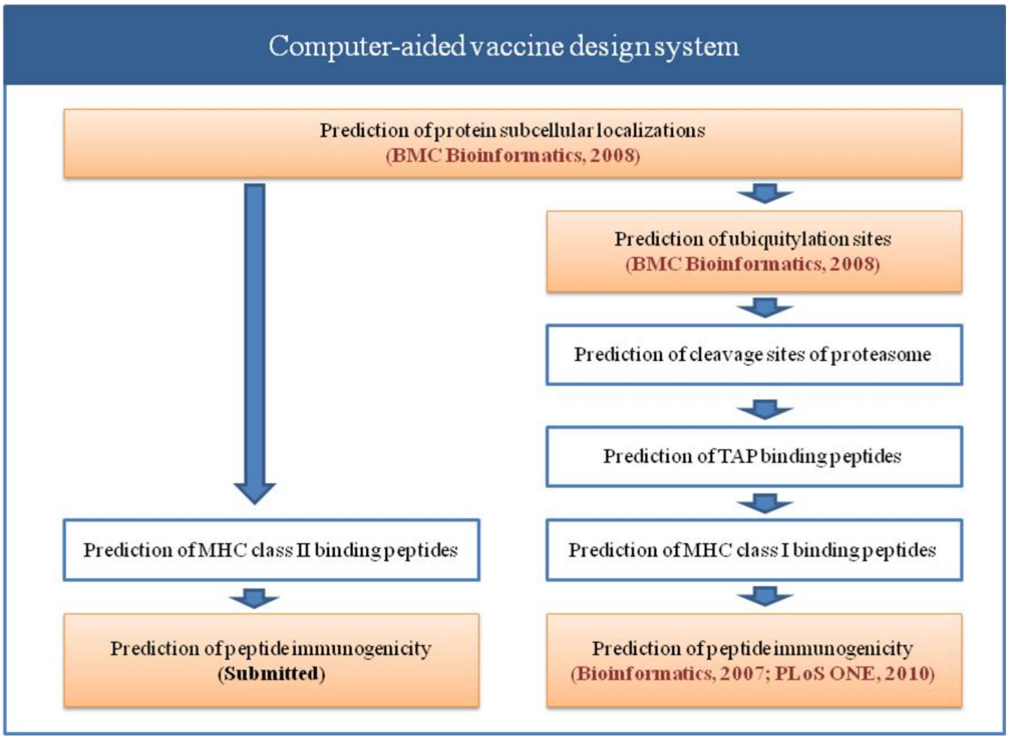

Computer-aided vaccine design system
Prediction of protein subcellular localizations
(BMC Bioinformatics, 2008)
Prediction of ubiquitylation sites
(BMC Bioinformatics, 2008)
Prediction of cleavage sites of proteasome
Prediction of TAP binding peptides
Prediction of MHC class II binding peptides
Prediction of MHC class I binding peptides
Prediction of peptide immunogenicity
(Submitted)
Prediction of peptide immunogenicity
(Bioinformatics, 2007; PLoS ONE, 2010)

Figure 1.4 The architecture of proposed peptide immunogenicity prediction system for computer-aided vaccine design.

1) **(reaction 5 in Figure 1.3)** The prediction of immunogenicity of MHC class I binding peptides (CTL response) is important to understand immune systems and accelerate vaccine design. Previous studies show that moderate binding affinity of peptides to MHC molecules is required but is not the deterministic factor. Because of the complex effects of intrinsic factors like physicochemical properties and extrinsic factors of MHC repertoire, it is hard to predict immunogenicity. For solving this problem, an informative physicochemical property mining algorithm was proposed to simultaneously mine informative physicochemical properties from existing experimental data and design a support vector machine classifier. By mining a subset of 23 informative physicochemical properties from 531 physicochemical properties, a prediction system of POPI was constructed. POPI performs better than alignment-based methods and traditional affinity-based methods. For

HLA-A2-restricted peptides, an improved prediction system POPISK is proposed for predicting immunogenicity using string kernels. The prediction and analysis ability of POPISK gives insights into the underlying mechanism of T-cell recognition. Important positions for T-cell responses are identified and analyzed. POPISK can accurately predict immunogenicity changes made by single residue modifications.

2) **(reaction 7 in Figure 1.3)** For the prediction of immunogenicity of MHC class II binding peptides (HTL response), the developed informative physicochemical property mining algorithm was applied to mine informative physicochemical properties from experimental data. A prediction system POPI-MHC2 for predicting immunogenicity of MHC class II binding peptides was implemented by using 21 informative physicochemical properties. The same as POPI, POPI-MHC2 performs much better than alignment-based methods and traditional affinity-based methods.

3) **(reaction 1 in Figure 1.3)** Three kinds of features were assessed for their performances of ubiquitylation site prediction. For classifier selection, three classifiers including k-nearest neighbor classifier, NaïveBayes and support vector machines (SVM) were assessed. Results show that SVM using physicochemical properties performs best. Moreover, the informative physicochemical property mining algorithm was applied to mine 31 informative physicochemical properties from all 531 physicochemical properties. A prediction system UbiPred constructed by using 31 informative physicochemical properties shows large improvement in prediction performance.

This dissertation presents the first prediction systems for CTL and HTL responses and protein ubiquitylation. The obtained informative physicochemical properties yield insights into the immune systems and are helpful to develop prediction systems for vaccine designs.

1.3 Organization

In summary, this dissertation focuses on predicting adaptive T-cell immune responses for computer-aided vaccine design. For solving optimization problems of mining informative physicochemical properties for the peptide immunogenicity, efficient evolutionary algorithms were proposed to develop efficient vaccine design system. The rest of this dissertation is organized as follows. Chapter 2 addresses the related works of this dissertation. The proposed algorithm for mining informative physicochemical properties is presented in Chapter 3. Chapter 4 presents the prediction system for predicting ubiquitylation sites. Chapter 5 describes the proposed prediction systems for predicting immunogenicity of MHC class I and II binding peptides. The improved prediction of peptide immunogenicity using string kernels is presented in Chapter 6. Finally, conclusions are given in Chapter 7.

Chapter 2

Related Work

This chapter presents related works for predicting adaptive T-cell immune response including prediction of ubiquitylation sites, immunogenicity prediction of MHC class I binding peptides, and immunogenicity prediction of MHC class II binding peptides.

2.1 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 [2, 3].

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 [4]. 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 [5]. 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 [6].

Some recent studies have similar results that ubiquitin-conjugation will enhance the efficacy of polynucleotide viral vaccines [7] and vaccines against tuberculosis [8]. 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 [9]. 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 4, 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.

2.2 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 [10]. 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 [11] and Pcleavage is based on a support vector machine (SVM) learning model [12].

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 [13] and the prediction of TAP transport efficiency of epitope precursors using a simple scoring matrix [14]. 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 [15] and Gibbs sampling method [16]. 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 [17, 18].

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 5, 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 6, 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.

2.3 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 [19, 20].

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 [21], evolutionary algorithms combined with artificial neural networks [22] and multi-objective evolutionary algorithms [23] are developed for optimizing a matrix for predicting binding affinity. Other methods including the neural network based methods [22, 24, 25], Bayesian neural networks [26], fuzzy neural networks [27], the hidden Markov model [28], Gibbs samplers [16], support vector machines [29-31] and alignment-based method SMM-align that is a stabilization matrix alignment method for predicting MHC class II binding affinity [32].

However, the problem of predicting immunogenicity of MHC class II binding peptides is also important to understand immunogenicity and design effective vaccines. In Chapter 5, 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 (described in Chapter 3). 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.

Chapter 3

Informative Physicochemical Property Mining Algorithm

For mining informative physicochemical properties from experimental data, a genetic algorithm based method was proposed to simultaneously determine optimal subset of physicochemical properties and design a support vector machine classifier.

3.1 Physicochemical properties

Physicochemical properties of amino acids were extensively and successfully used in sequence-based prediction methods [33-37]. There are 544 physicochemical properties of amino acids extracted from amino acid index database version 9.0 (AAindex), which is a collection of published amino acid indices representing different physicochemical and biological properties of amino acids [38, 39]. 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.

To encode an input vector from peptide sequences for machine learning classifiers, a two-step method is utilized. The first step determines a vector D_t of 531 index values for each amino acid of peptides. A peptide of length *l* has 531 *l*-dimensional vectors that can be defined as:

$$\mathrm{D}_t = \left(d_{t1}, d_{t2}, \ldots, d_{tl}\right), t = 1, \ldots, 531,$$

where t denotes the t-th physicochemical properties. In the second step, a vector V of 531 mean values is obtained by averaging these l attributes in each vector, defined as follows:

$$\mathrm{V} = \left(\bar{v}_1, \bar{v}_2, \ldots, \bar{v}_t\right),$$

$$\bar{v}_t = \frac{1}{l}\sum_{i=1}^{l} d_{ti}\ , \tag{3-1}$$

where $\bar{v}_t$ is the averaged value of elements in $\mathrm{D}_{t,}$.

3.2 Support vector machines

Support vector machines (SVMs) are powerful tools in the field of machine learning. SVMs cope well with the over-fitting problem arising from a small training dataset by finding a linear separation hyperplane that maximizes the distance between two classes to create a classifier. SVMs can efficiently deal with classification, prediction, and regression problems. Given training vectors $\mathbf{x}_i \in \mathrm{R}^n$ and their class values $y_i \in \{-1, 1\}$, $i = 1, \ldots, N$, an SVM solves the following problem:

$$\min \quad \frac{1}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w} + C\sum_{i=1}^{N}\xi_i\ ,$$

$$\text{subject to} \quad y_i(\mathbf{w}^{\mathrm{T}}\mathbf{x}_i + b) \geq 1 - \xi_i\ , \tag{3-2}$$

$$\xi_i \geq 0\ ,$$

where $\mathbf{w}$ is a normal vector perpendicular to the hyperplane and ξ_i are slack variables allowing for some misclassifications. The cost parameter $C > 0$ controls the trade-off between the margin and the training error. Larger values of C will lead to a higher error penalty.

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 nonlinearly transform the feature space, defined as follows:

$$K(x_i, x_j) = \exp(-\gamma\left\|x_i - x_j\right\|^2), \gamma > 0\ . \tag{3-3}$$

The kernel parameter γ determines how the samples are transformed into a high-dimensional search space. These two parameters C and γ must be tuned to get the best prediction performance.

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 $h(h-1)/2$ 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, the used SVM is obtained from LIBSVM package version 2.81 [40].

3.3 Orthogonal experimental design

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 [41, 42]. 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 3.1 shows an illustrative example of orthogonal experimental design using a two-level orthogonal array $\mathrm{L}_M(2^{M\text{-}1})$ with M rows and M-1 columns. In this example of M=8, there are 7 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 f_t 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:

$$S_{jk}=\sum f_t \cdot F_t \ , \ t = 1, \ \ldots, \ M, \qquad (3\text{-}4)$$

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 [43].

The Rank in Table 3.1 shows the rank of the combination t in all 128 (=27)

Table 3.1 An illustration example of orthogonal array $L_8(2^7)$ and factor analysis.

t	Factors							Accuracy(%)	Rank
	1	2	3	4	5	6	7	f_t	
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

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 main effect difference 36.3 is the most effective factor.

3.4 Inheritable bi-objective genetic algorithm

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 (IBCGA, [43]) is utilized to solve this optimization problem. IBCGA consists of an intelligent genetic algorithm [44] 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*) [43]. 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 paper. The GA-chromosome consists of n=531 binary GA-genes b_i 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 $\{2^{-7}, 2^{-6}\ldots, 2^{8}\}$. Figure 3.1 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., 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

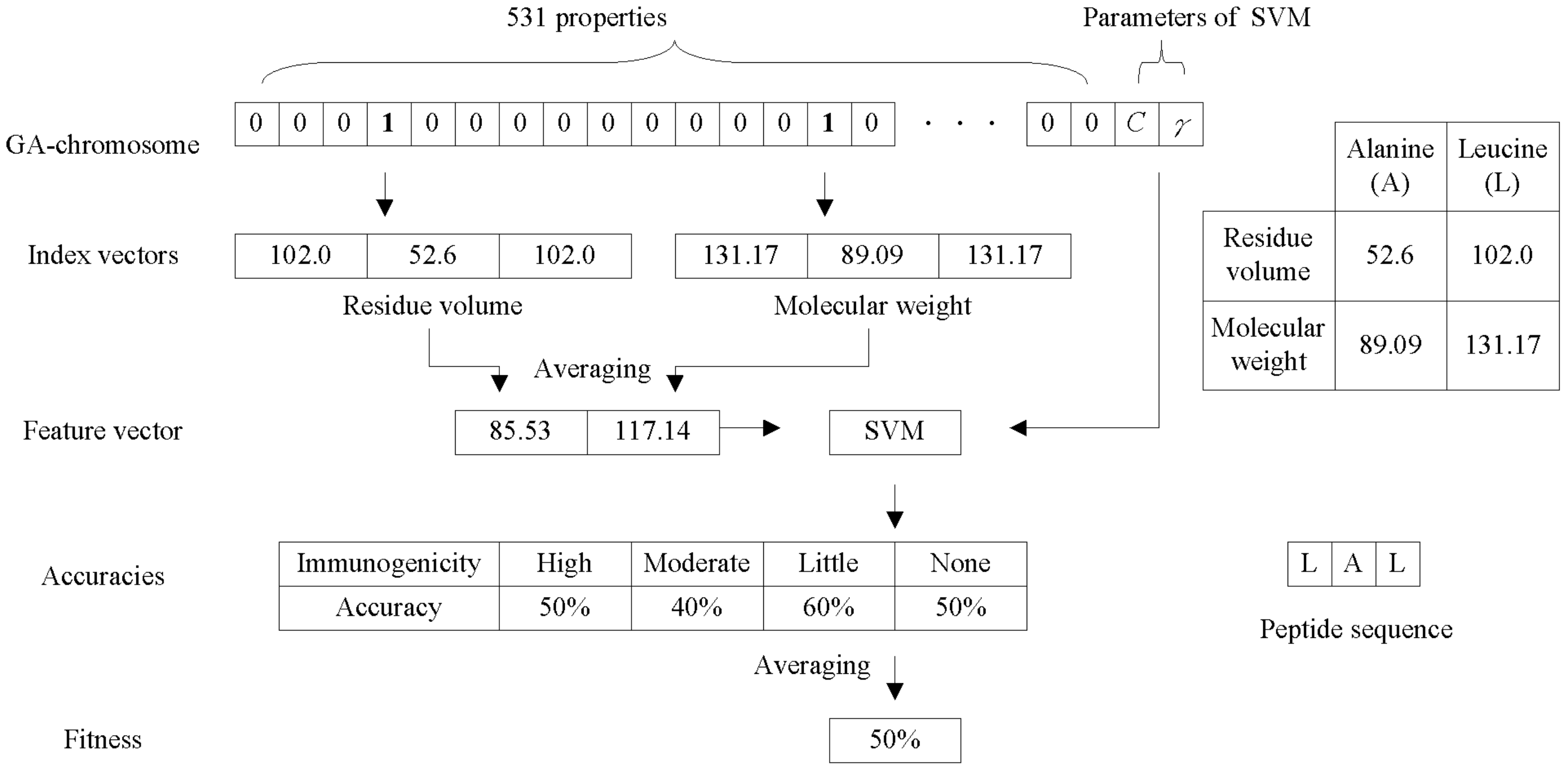


Figure 3.1 An illustration example of fitness function evaluation from decoding a GA-chromosome.

avoid from the prediction bias for some classes, the averaged accuracies (*AA*) of all classes, defined in (3-10), 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, Xr, where r=r_{start}, r_{start}+1, …, rend in a single run. The algorithm of IBCGA with the given values r_{start} and r_{end} is described as follows:

Step 1) (Initiation) Randomly generate an initial population of N_{pop} individuals. All the n binary GA-genes have r 1's and n-r 0's 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 < r_{end}, 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.

3.5 Performance evaluations

Four measurements are applied to evaluate developed prediction systems including accuracy (*ACC*) and Matthew's correlation coefficient (*MCC*) for each class, and overall accuracy (*OA*) and averaged accuracy (*AA*) for all classes, defined as follows:

$$ACC_i = \frac{TP_i}{TP_i + FN_i} \times 100\% , \tag{3-5}$$

$$MCC_i = \frac{TP_i \times TN_i - FP_i \times FN_i}{\sqrt{\left(TP_i + FN_i\right) \times \left(TP_i + FP_i\right) \times \left(TN_i + FP_i\right) \times \left(TN_i + FN_i\right)}} , \tag{3-6}$$

$$OA = \sum \frac{TP_i}{N} , \tag{3-7}$$

$$AA = \sum \frac{ACC_i}{h} , \tag{3-8}$$

where *i* is the number of classes and TP_i, TN_i, FP_i and FN_i are the numbers of true positives, true negatives, false positives and false negatives, respectively. *N* is the total number of sequences and *h* is the number of classes.

Chapter 4

Predicting of ubiquitylation sites

Ubiquitylation plays an important role in regulating protein functions. Recently, experimental methods were developed toward effective identification of ubiquitylation sites. To efficiently explore more undiscovered ubiquitylation sites, this stud aims to develop an accurate sequence-based prediction method to identify promising ubiquitylation sites.

4.1 Motivation

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 [45, 46]. 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 [45, 46].

Because of the important regulatory roles of ubiquitylation, numerous methods were developed to purify ubiquitylated proteins [47]. Also, the growing number of studies of large-scale identification of ubiquitylated proteins and analysis of ubiqui-

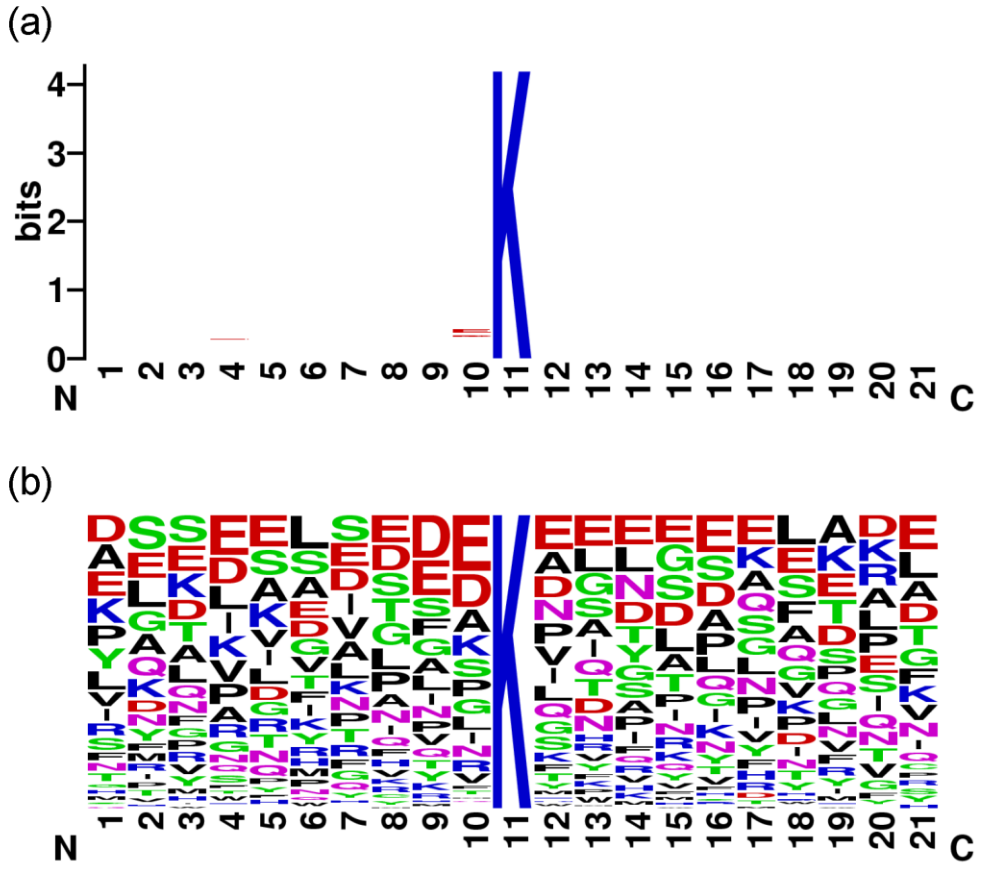


Figure 4.1 The sequence logo of the 151 positive samples with *w*=21. (a) information content and (b) frequency plot.

tin-related proteome reflect the importance of identifying ubiquitylation proteins and sites [48-53]. The three steps affinity purification, proteolytic digestion, and analysis using mass spectrometry were applied in most of these studies [54]. 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.

4.2 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

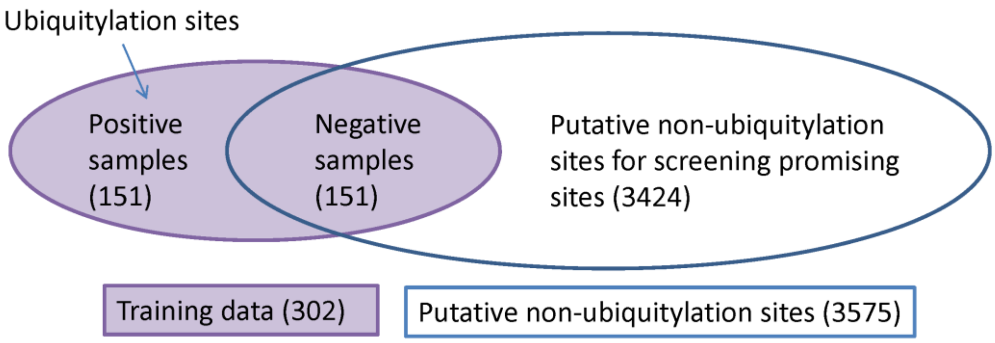

Ubiquitylation sites
Positive samples (151)
Negative samples (151)
Putative non-ubiquitylation sites for screening promising sites (3424)
Training data (302)
Putative non-ubiquitylation sites (3575)

Figure 4.2 The schema for the training and an independent of 3424 putative non-ubiquitylation sites in dataset of *w*=21.

and are widely used in bioinformatics studies are evaluated for prediction of ubiquitylation sites: conventional amino acid identity [55], evolutionary information [56, 57], and physicochemical property [58, 59]. 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 [55-57]. The feature representations for applying to the mentioned classifiers are described below.

The conventional feature representation, amino acid identity, uses 20 binary bits to represent an amino acid [55]. 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 [56, 57]. To prepare evolutionary information for each protein sequence, the corresponding position-specific scoring matrix (PSSM) is obtained by applying PSI-BLAST [60] 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

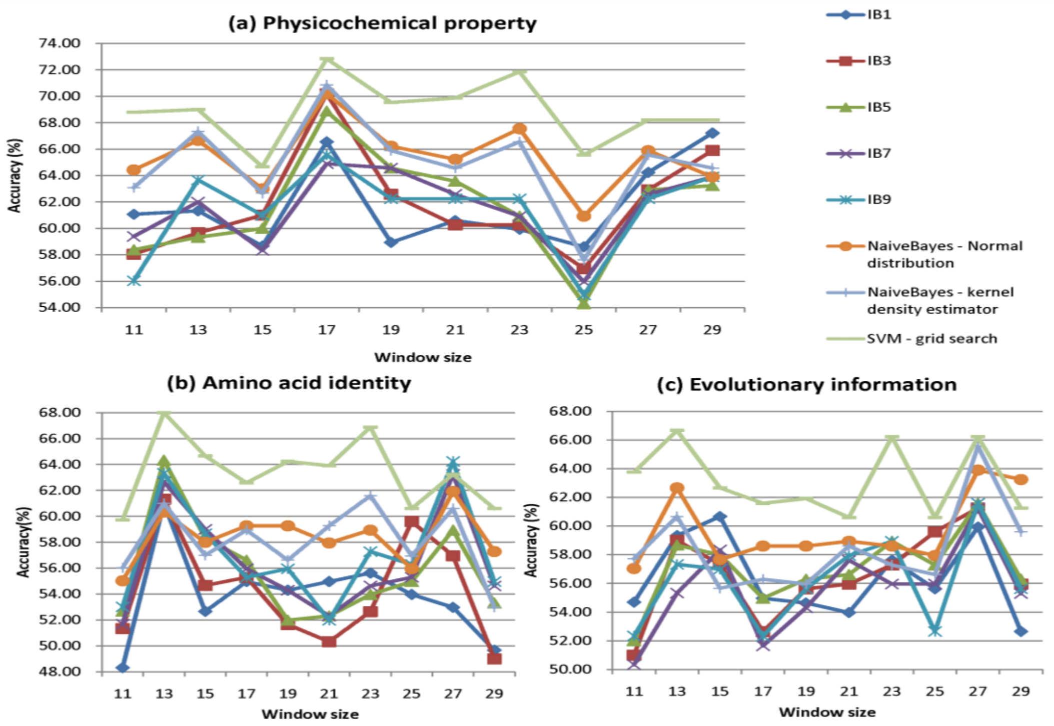


Figure 4.3 Performance comparisons among amino acid identity, evolutionary information and physicochemical property with various classifiers.

(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 [61] 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 $w \in \{11, 13, \ldots, 29\}$ 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.

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 [62]. 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.

Table 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.

#	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

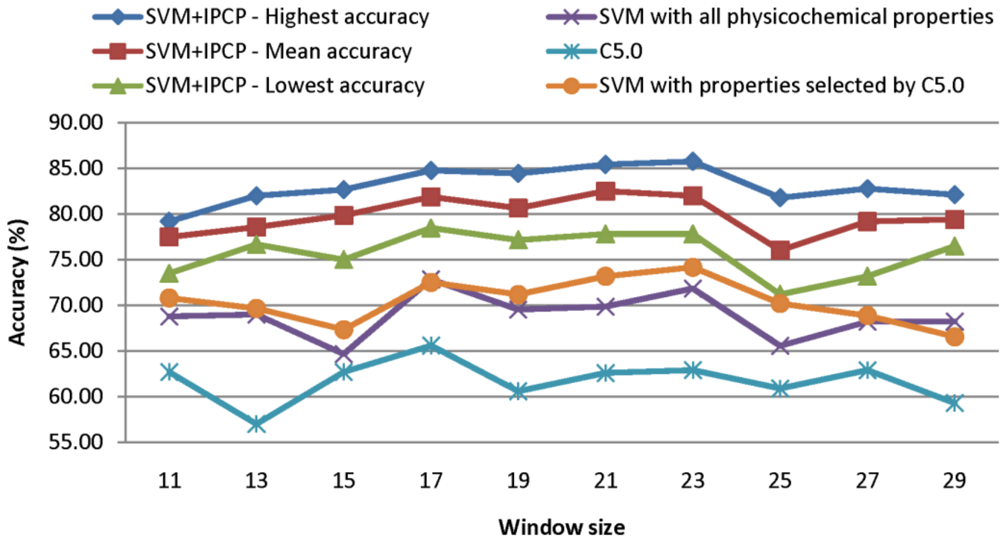


Figure 4.4 Performance comparisons between the SVM with informative physicochemical properties (SVM+IPCP) and other compared classifiers.

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. In this study, the number of negative samples is the same with that of positive samples in the dataset. For example, there are 151 negative samples in the dataset of *w*=21. The rest (e.g., 3424 samples with no annotation of ubiquitylation sites for *w*=21) are formed as an independent dataset to be scored for identifying promising ubiquitylation sites (see Fig. 8). 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.

Figure 4.1 shows the sequence logo of the 151 positive samples with *w*=21 generated by the WebLogo tool [63]. 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 [62], to evaluate the three kinds of features for applying classifiers. The dataset of window size 21 is shown in Figure 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 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 4.1

4.3 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*.

The highest, mean, and lowest prediction accuracies of IPMA using 10-CV are shown in Figure 4.4. For comparison, the decision tree method C5.0 [64] 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 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

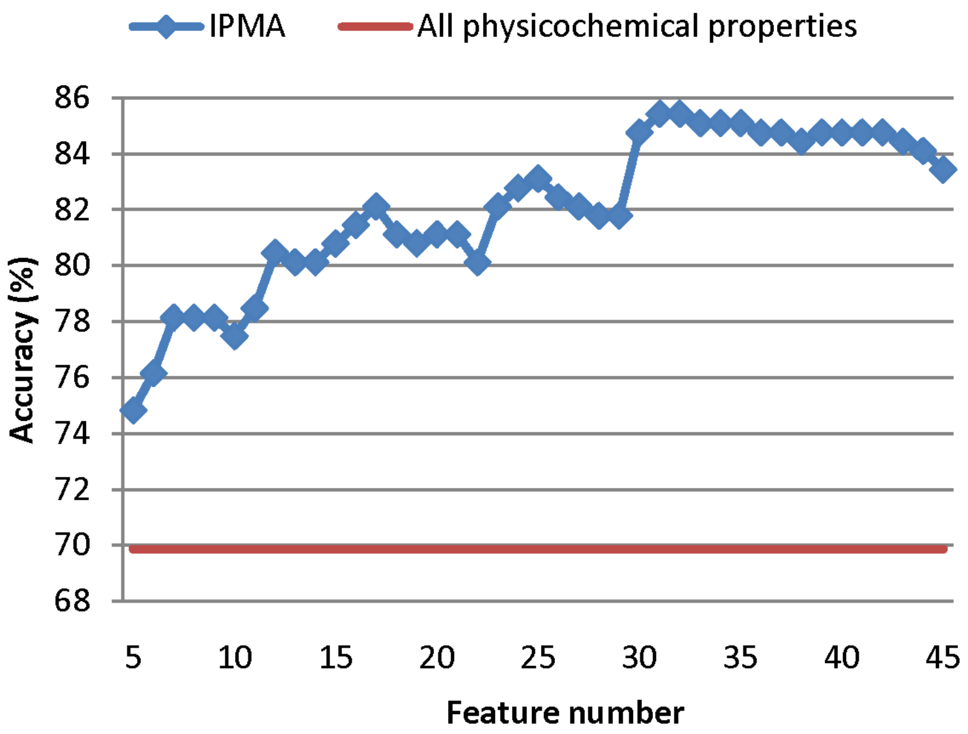
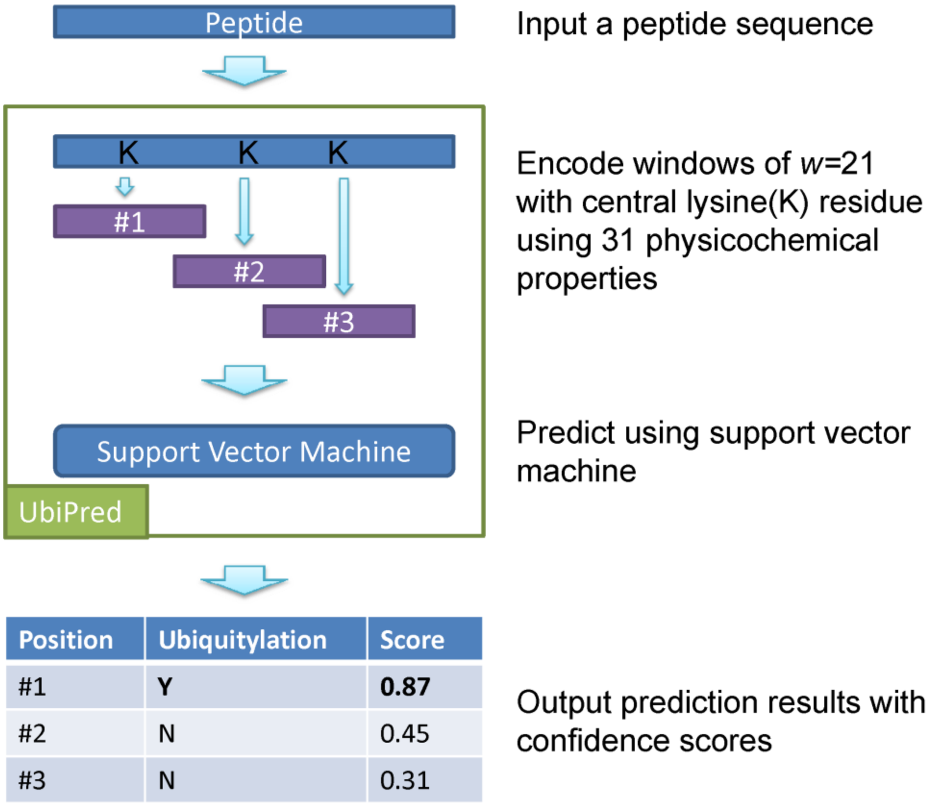


Figure 4.5 The best 10-CV accuracies of prediction using SVM with the window size 21 for various numbers of features (properties) selected by IPMA from 30 independent runs.

properties. Considering the mean accuracies of SVM with informative physicochemical properties in Figure 4.4, the best window size is w=21.

Figure 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 [41] [42] can be used to estimate the individual effects of physicochemical properties according to the val-



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

Input a peptide sequence

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

Predict using support vector machine

Output prediction results with confidence scores

Figure 4.6 The system flow of prediction system UbiPred.

ue of main effect difference (MED) [59] [58]. The property with the largest value of MED is the most effective in predicting ubiquitylation sites.

According to MED, the 31 informative properties are ranked and their descriptions are shown in Table 4.2. 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.

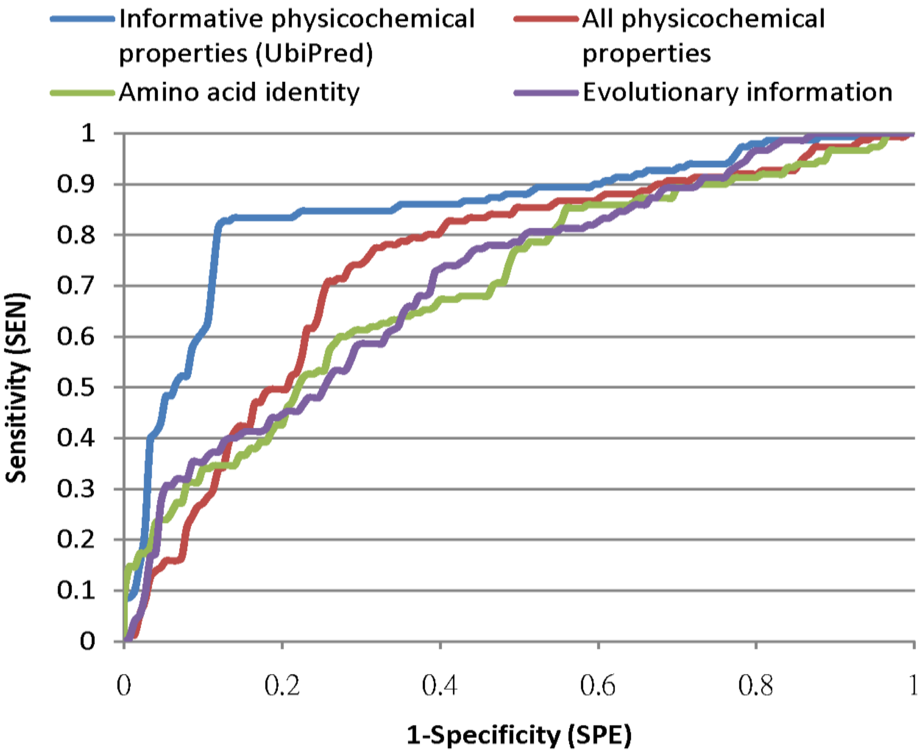


Figure 4.7 Comparison of receiver operating characteristic curves among informative physicochemical properties (UbiPred), amino acid identity, evolutionary information and all physicochemical properties.

4.4 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 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

Table 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

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 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).

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 4.2. The ranked informative physicochemical properties provide valuable information to biologists when further performing experimental verification.

The problem of sequence redundancy may result in overestimation of prediction performance. To address this issue, six thresholds of sequence identity (90%, 80%, …, 40%) were applied to construct six additional datasets from the dataset of *w*=21 by using CD-HIT [65]. The numbers of positive and negative samples of datasets with various sequence identity thresholds are shown in Table 4.3. By using the strictest threshold 40%, there are only 36 redundant samples and the resulting dataset consists of 145 negative samples and 121 positive samples. By applying LOOCV to evaluate prediction accuracies on these datasets, good performance (>79%) was obtained by using SVM with the mined 31 informative physicochemical properties and

Table 4.3 The LOOCV performances of the SVM with 31 informative physicochemical properties on datasets of various sequence identity thresholds.

Sequence identity threshold	Accuracy(%)	Number of positive samples	Number of negative samples
100%	84.44	151	151
90%	82.71	145	150
80%	81.72	141	149
70%	80.63	136	148
60%	81.23	131	146
50%	80.80	130	146
40%	79.70	121	145

SVM parameters (shown in Table 4.3). The results show the effectiveness of the proposed UbiPred.

4.5 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 4.8 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 [64], consisting of three and five if-then rules for ubiquitylation sites and non-ubiquitylation sites, respectively.

To obtain rather simple rules for easy interpretation, five concise if-then rules obtained from C5.0 are shown in Table 4.4. 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 [66] (property MEIH800102)

Table 4.4 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

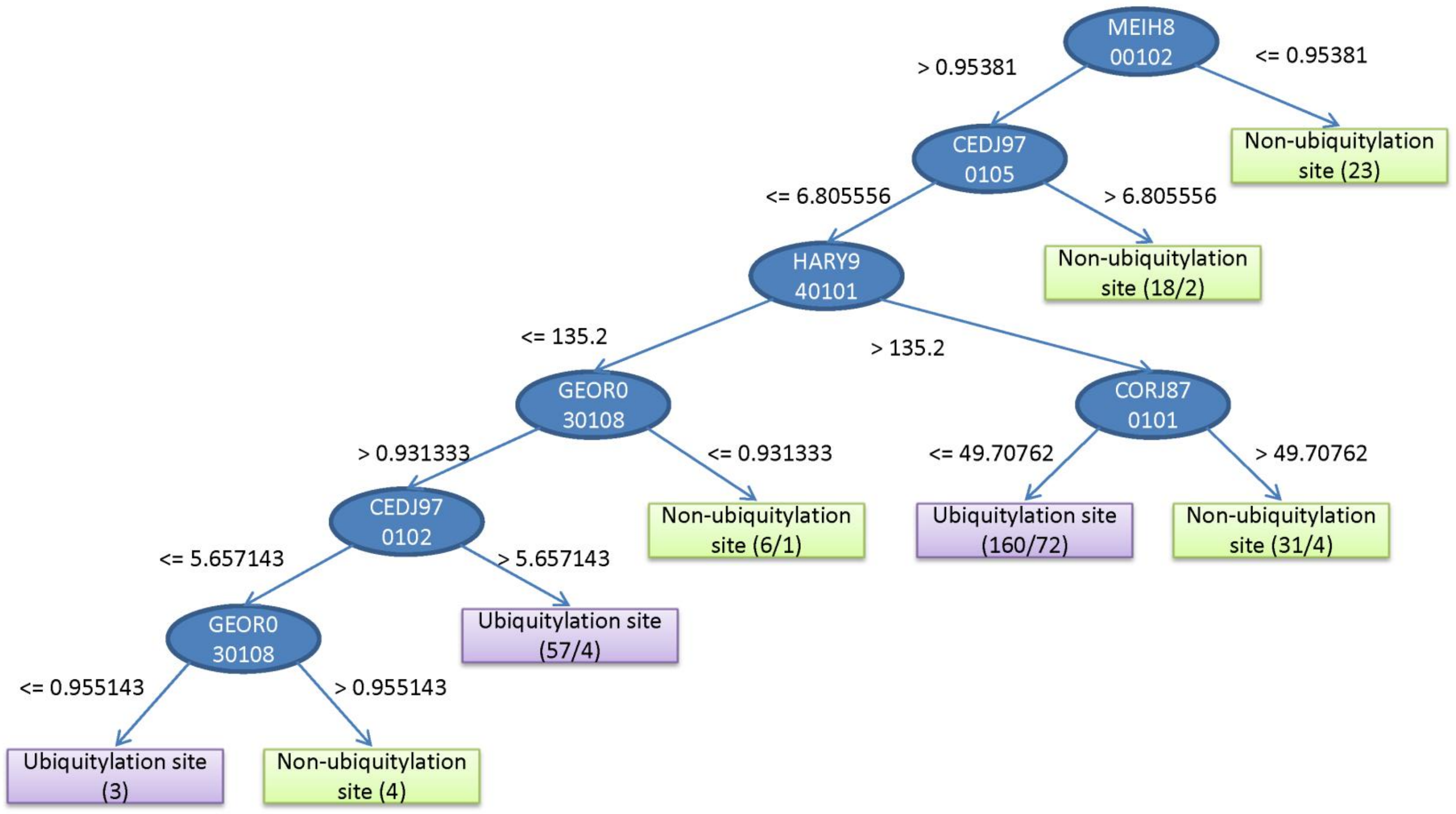


Figure 4.8 The derived decision tree by using C5.0 and the features of informative physicochemical properties for classification of ubiquitylation sites.

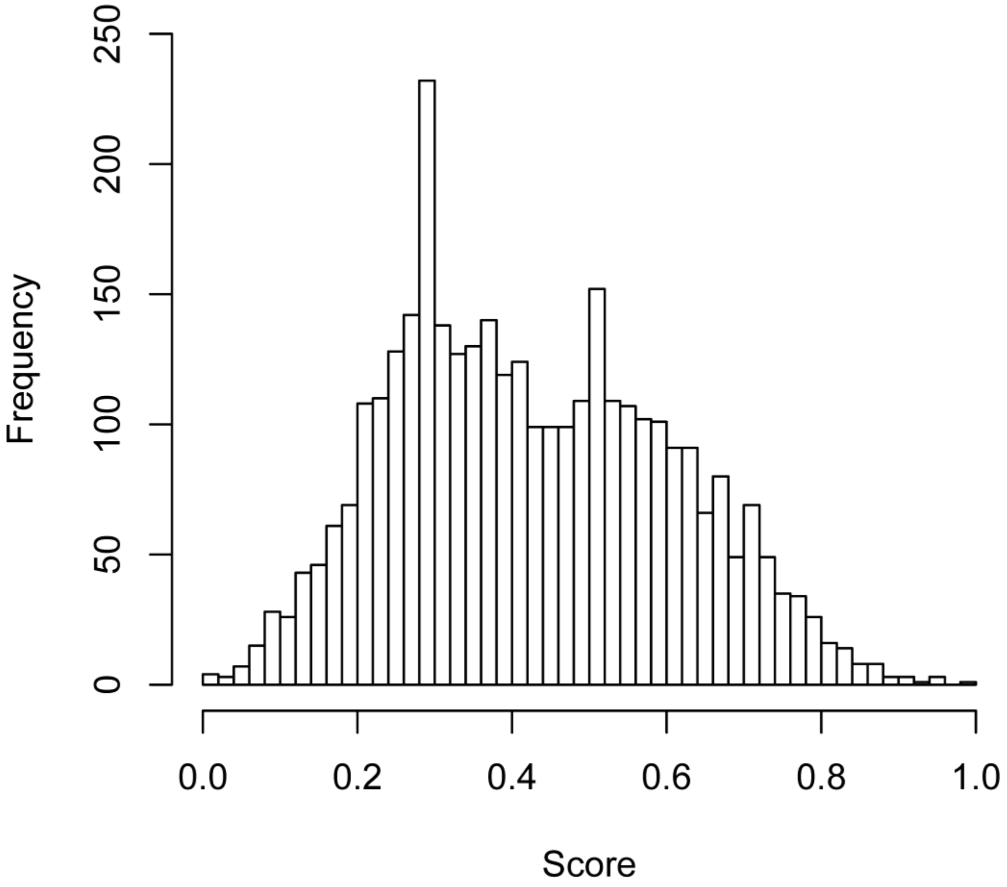


Figure 4.9 Histogram result of UbiPred using prediction scores from evaluating 3424 putative non-ubiquitylation sites in an independent dataset. The site with a score close to 1 has a high possibility to be an ubiquitylation site.

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

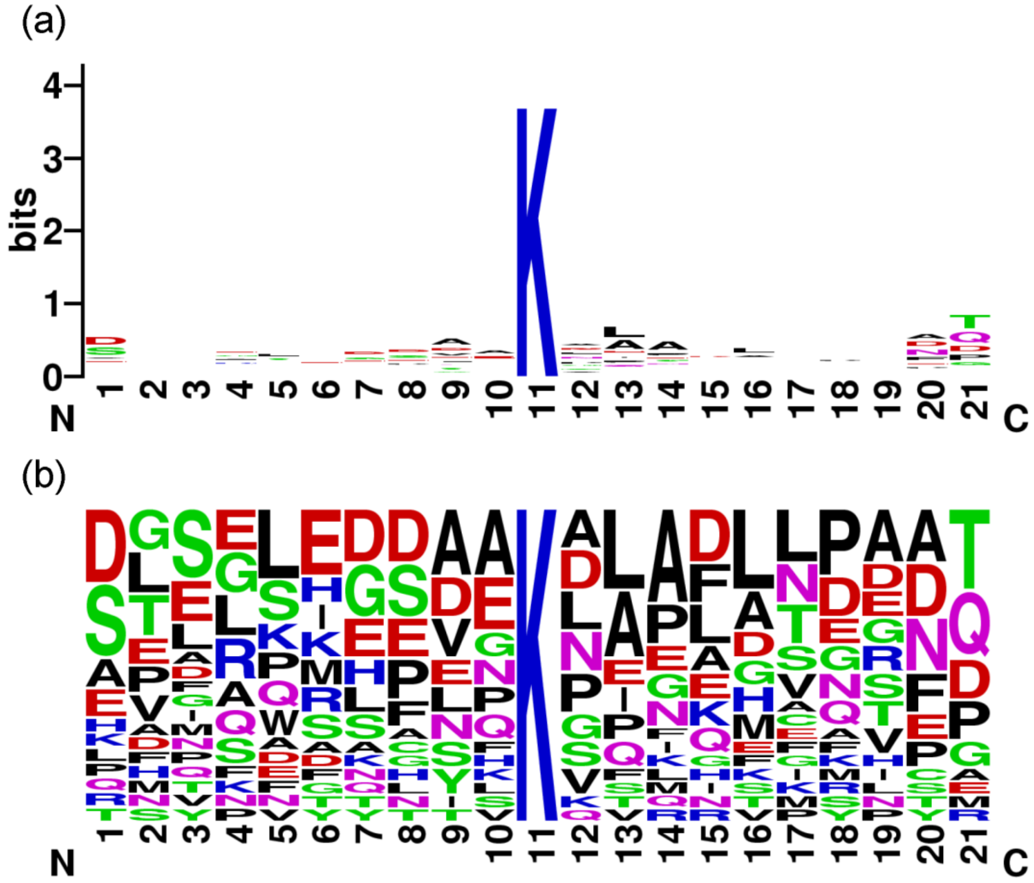


Figure 4.10 The sequence logo of the 23 peptides of promising ubiquitylation sites with *w*=21. (a) Information content and (b) Frequency plot.

the MED value (28.48) of MEIH800102 in Table 4.2, it is rather consistent that MEIH800102 is an informative property with a rank 3.

The second rule means that if the mean volume of residues buried in protein interiors [67] (property HARY940101) is larger than 135.2 and the NNEIG index [68] (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) [69] 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 [70] 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.

4.6 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 [48]. 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

Table 4.5 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			

non-ubiquitylation sites in an independent dataset that are not included in the training dataset of w=21, as shown in Figure 4.6. The screening result is shown in Figure 4.9 using a histogram of prediction scores. 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 4.5. The detailed information can be found in the website of UbiPred. The sequence logo of the 23 peptides shown in Figure 4.10 represents low information content similar to the sequence logo of the 151 positive samples in training dataset.

For further validating the 23 peptides as ubiquitylation sites, the five prediction rules obtained from C5.0 (shown in Table 4.3) were applied to the 23 peptides. Results show that all the 23 promising peptides are classified as ubiquitylation sites. For example, the average value of property MEIH800102 for the 23 peptides is 1.001 which is larger than the threshold of 0.95. This value is close to that (1.007) of the 151 positive samples in training dataset. Note that the smallest and largest index values of MEIH800102 for 20 amino acids are 0.73 and 1.23, respectively. The prediction system UbiPred can predict ubiquitylation sites with prediction scores to identify the most promising ubiquitylation sites for experimental verification or future research.

4.7 Follow-up works

Two prediction methods were published after our work. The first method is UbPred [71]. UbPred is trained on two datasets. One of the datasets is the same as our study, and the other dataset contains 141 new ubiquitination sites identified by using a combination of liquid chromatography, mass spectrometry, and mutant yeast strains. In their assessment, the 141 new sites identified from short-lived proteins are used to independently test our UbiPred server. Note that although it is unfair to test our server using the 141 new sites because our training dataset does not focus on short-lived proteins, our prediction server can still identify ubiquitylation sites in short-lived protein with accuracy of 53%. The second method is based on neural network [72]. They use the same dataset and window size as ours and conclude only slightly better performance of AUC=0.88, compared to our method of AUC=0.85. The follow-up works show the importance of this work.

Furthermore, a published study used our prediction system UbiPred to predict ubiquitylation sites and found that 43% of high-confidence lysine methylation sites were also predicted to be ubiquitination sites that is consistent with previous study [73]. Their analysis provides additional confidence in the usability of UbiPred.

4.8 Summary

Ubiquitylation plays many important regulatory roles in the physiology of eukaryotic cell. Nowadays, many experimental studies are working on identifying ubiquitylated proteins and their ubiquitylation sites. To accurately predict ubiquitylation sites by computational methods is helpful to save experimental efforts. In this study, an SVM-based method is presented to assess three kinds of features, including amino acid identity, evolutionary information and physicochemical property, in predicting ubiquitylation sites. The ubiquitylation datasets extracted from the UbiProt database are established to evaluate the proposed methods. Results show that physicochemical property is the best kind of features for the SVM-based prediction method.

It is well recognized that irrelevant information will interfere with classifiers. This study proposes an algorithm IPMA for mining a small set of informative physicochemical properties to advance the prediction performance. The 31 informative physicochemical properties improve the prediction accuracy from 72.19% to 84.44%, and their individual effectiveness is ranked for further understanding the ubiquitylation mechanism. Finally, the system UbiPred for predicting ubiquitylation sites is designed by using 31 informative physicochemical properties. The web server of UbiPred has been implemented and is available at http://iclab.life.nctu.edu.tw/ubipred.

Chapter 5

Predicting immunogenicity of MHC binding peptides

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

This study focuses on mining informative physicochemical properties from known experimental immunogenicity data to understand immunogenicity and predict immunogenicity of MHC-binding peptides accurately.

5.1 Motivations

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 [74, 75]. Several studies aimed to clarify the relationship between the peptide binding affinity to the MHC molecule and its immunogenicity [76, 77]. 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 [78]. Furthermore, peptide-flanking residues other then MHC anchor residues were identified as import factors for MHC class II-restricted T-cell responses [79, 80]. 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 [33-37]. 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 [36] or the rank-based method [81]. Therefore, these methods cannot be effectively applied to the investigated intractable problems because of limited knowledge or neglect of correlated effects among multiple properties [33]. 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 [82]. The SVM-based learning methods are shown effective for various prediction methods from protein sequences [12, 15]. 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 [40]. 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 (described in Chapter 3) is proposed to mine informative physicochemical properties for predicting CTL and HTL responses.

5.2 The 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 [83]. Table 5.1 and Table 5.2 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 5.1 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 5.1 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

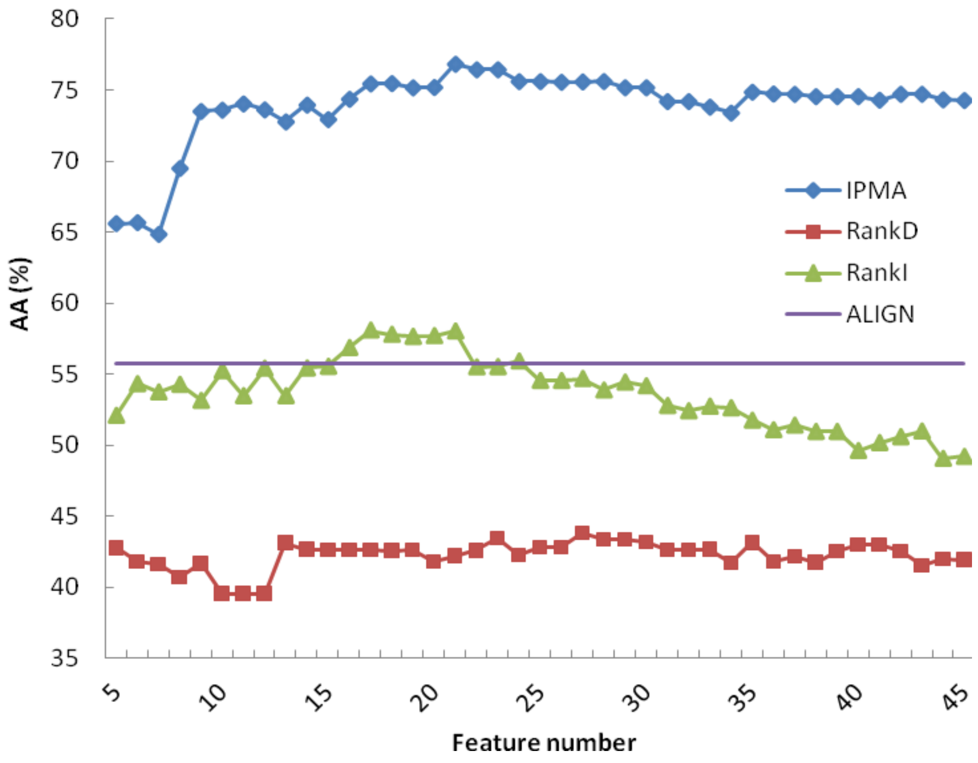


Figure 5.1 Averaged accuracies (*AA*s) of 10-CV for IPMA, rank-based methods (RankD and RankI) and the alignment-based method (ALIGN) for MHC class I binding peptides.

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 5.1 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. The results of mining informative physicochemical properties for PEPMHCI2 is similar to PEPMHCI that shown in Figure5.2.

5.3 POPI for predicting immunogenicity of

Table 5.2 Performance comparisons of ALIGN, PSI-BLAST and POPI using LOOCV on the whole dataset PEPMHCI.

Immunogenicity class	ALIGN		PSI-BLAST		POPI	
	ACC (%)	MCC	ACC (%)	MCC	ACC (%)	MCC
None	69.44	0.61	82.14	0.59	83.33	0.63
Little	39.76	0.32	45.59	0.40	50.60	0.44
Moderate	39.00	0.22	34.67	0.12	55.00	0.47
High	62.38	0.37	46.99	0.37	59.41	0.49
OA	54.91		53.23		64.72	
AA	52.64		52.35		62.09	

Table 5.3 Performance comparisons between AFFIPRE and POPI.

Immunogenicity class	Number of peptides	AFFIPRE		POPI	
		ACC (%)	MCC	ACC (%)	MCC
None and Little	87	35.63	0.17	80.46	0.39
Moderate	31	32.26	0.01	25.81	0.23
High	42	52.38	0.15	45.24	0.27
OA		39.38		60.63	
AA		40.09		50.50	

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 [84] and advanced sequence comparison method PSI-BLAST that is capable of detecting remote homologues [60] were applied to search for similar sequences. For each tested peptide, ALIGN and PSI-BLAST using three iterations were applied separately to search for its homologues. Results are shown in Table 5.2.

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.

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.

The results shown in Table 5.3 reveal that POPI performs well, compared with two sequence alignment-based prediction methods ALIGN (54.91%) and PSI-BLAST (53.23%). The poor performance of AFFIPRE reveals that the affinity only can not be directly used to predict peptide immunogenicity and this result is consistent with previous studies that the affinity of peptide-MHC molecules is not the main factor for predicting peptide immunogenicity [76, 77].

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 vis-

Table 5.4 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	

Table 5.5 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	

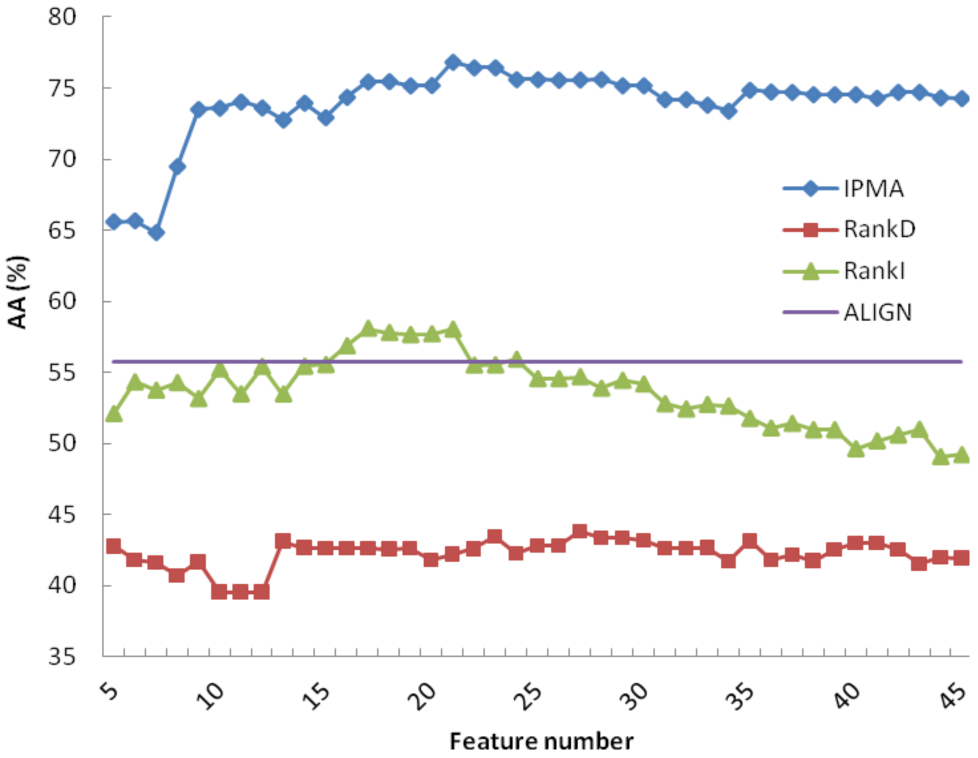


Figure 5.2 Averaged accuracies (*AA*s) of 10-CV for IPMA, rank-based methods (RankD and RankI) and the alignment-based method (ALIGN) for MHC class II binding peptides.

its from >20 countries, and >20,000 sequences were analyzed.

5.4 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.

For comparing with affinity-based prediction, another dataset consisting of 69 peptides with annotated binding and immunogenicity level was constructed. PO-

PI-MHC2 (69.57%) performs better than the affinity-based method (40.5%) shown in Table 5.5.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

5.5 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 5.6 and Table 5.7 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' [85]. The least effective property is MIYS850101 with *MED*=0.80 which corresponds to 'Effective partition energy' [86]. For HTL immune response, the AAindex identity KUHL950101 is the most effective property (denoting 'Hydrophilicity scale') with *MED*=46.06 [87]. The AAindex

Table 5.6 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 5.7 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 identity DESM900102 with the smallest MED value of 4.11 denoting ‘Average membrane preference: AMP07’ [88].

5.6 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 5.3 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 [88, 89]. Recent studies [90, 91] 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 5.3).

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’ [85] (Table 5.6). In contrast, the property of AAindex identity KUHL950101 denoting ‘Hydrophilicity scale’ [87]

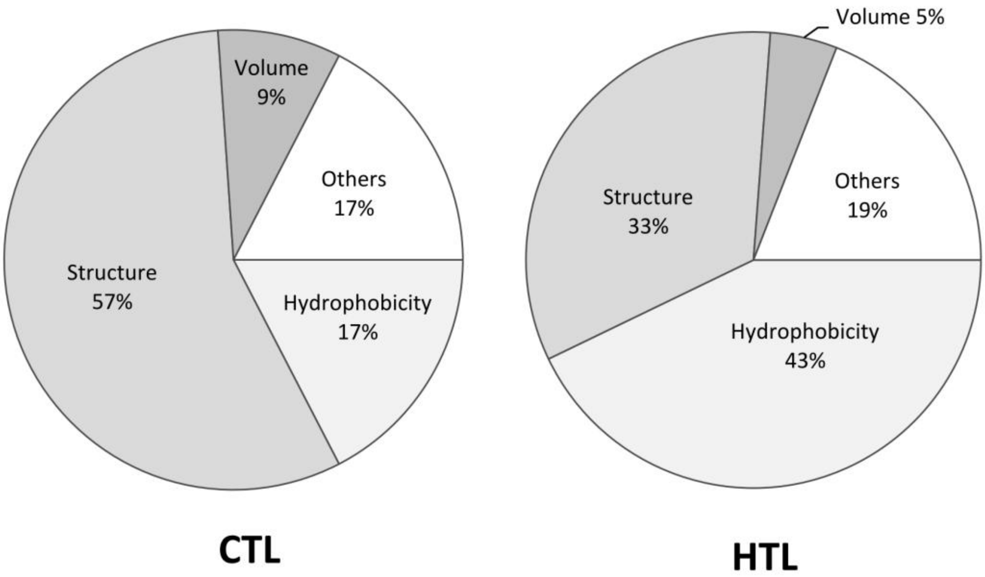


Figure 5.3 Pie-chart representations of compositions of categorized physicochemical properties of peptides responsible for CTL and HTL responses.

is the most effective property with *MED*=46.06 for HTL responses (Table 5.7).

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.

5.7 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 [92, 93]. 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 5.8 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 [78]. This result provides a good reason to build a prediction system to quickly select peptide candidates inducing both CTL and HTL responses.

5.8 Independent test performance of POPI-MHC2

For testing the informative physicochemical properties mined from PEPMHCII dataset, we extracted an additional independent test dataset IEDB1500 from IEDB database [94] which is a largest collection of immune epitopes. The IEDB1500 consists of all T-cell response data using proliferation assays, human host, and naturally processed peptides restricted by HLA class II molecules. Peptides of human protein source are removed because this study attempts to model normal immune systems instead of host with autoimmune disease. Note that the T cell response data is qualitative. A peptide is annotated as either immunogenic or non-immunogenic. After removing duplicate and inconsistence records, the numbers of immunogenic and non-immunogenic peptides of IEDB1500 are 1301 and 199, respectively.

All peptides in IEDB1500 were encoded using the 21 informative physicochemical properties. Due to the huge difference of immunogenic levels and dataset sizes between datasets PEPMHCII and IEDB1500, it is hard and not fair to directly test PEPMHCII-derived model on IEDB1500. To evaluate the prediction performance of the 21 informative physicochemical properties, jackknife test is applied to predict peptides in IEDB1500 with default SVM parameters of C=1 and γ=1/21. The area under ROC (receiver operating characteristic) curve (AUC) is a robust and nonparametric performance measurement for binary-class problems and is widely used for comparison of prediction methods. Finally, a reasonable high performance of POPI-MHC2 with AUC=0.67 using jackknife test is obtained with a highly unbalanced dataset which is different from PEPMHCII.

The previous section 5.4 already showed the poor performance of affinity-based method AFFIPRE on PEPMHCII. However, an additional performance comparison between POPI-MHC2 and affinity-based methods on IEDB1500 is desirable to show the robustness of POPI-MHC2. Due to the lack of annotated MHC binding affinities for peptides in IEDB1500, two state-of-the-art methods of ARB method [95] in IEDB analysis resource [96] and NetMHCIIpan [97] are applied to predict binding affinity of a peptide-MHC complex. The ARB method is based on an average relative binding matrix and can directly predict IC50 values. The matrix is trained on a large number of quantitative peptide binding data of IEDB and regularly updated with new data. Also, it is benchmarked as one of the best methods for binding affinity prediction [96, 98]. The NetMHCIIpan based on neural networks allows pan-specific predictions of peptide binding affinity to many HLA-DR molecules and is ranked as best individual predictor [99]. Therefore, comparing POPI-MHC2 with ARB can provide meaningful results.

Because most peptides are not annotated with complete supertype and subtype information of restricted MHC alleles, the ARB and NetMHCIIpan methods are not

Table 5.8 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

able to predict their binding affinity. Therefor, two datasets of TEST163 and TEST 320 consisting of only 163 and 320 peptides restricted by ARB and NetMHCIIpan support alleles were isolated from the IEDB1500 dataset, respectively. For each peptide in TEST163, its corresponding binding affinity was predicted by ARB. The scores for calculating ROC curve are minus predicted IC50 values because a large IC50 value means a weak binder. The binding score predicted by NetMHCIIpan represents the binding strength of each peptide in TEST320 and is used to calculate ROC curve. For POPI-MHC2, jackknife test using the 21 informative physicochemical properties and default SVM parameters is again used to evaluate the prediction performance on TEST163.

Due to the small number of peptides in TEST163 and TEST320, POPI-MHCII performs slightly worse. However, POPI-MHCII with AUC=0.60 is still much better than the affinity prediction method ARB with AUC=0.34 for TEST163. The NetMHCIIpan method with AUC=0.43 is worse than POPI-MHCII with AUC=0.59 for TEST320. The poor performances of affinity prediction methods are reasonable because they do not intend to directly predict T-cell responses. The results confirm the idea that the binding affinity alone is not sufficient for predicting T-cell responses.

5.9 Follow-up works

A recently published study utilize our POPI prediction server to analyze their the secretome of *Candida albicans* [100]. The *Candida albicans* is a pathogenic fungus and secrets a large number of proteins. To select candidates for vaccine developments, they applied mass spectrometry to identify secretory proteins and applied our POPI server to predict peptide immunogenicity. Finally, 29 highly immunogenic peptides originating from 18 proteins were identified as candidates for vaccine development.

A work done in University of Tübingen, Germany tried to improve our work by constructing a larger datasets and transforming the usage of averaged values of informative physicochemical properties to consider the position effects [101].

5.10 Summary

The effectiveness of vaccination depends on peptide immunogenicity in designing peptide-based vaccines. Accurate prediction of peptide immunogenicity will decrease many experimental efforts. This study investigates the prediction problem of peptide immunogenicity and proposes two efficient prediction systems POPI and PO-

PI-MHC2 to predict immunogenicity of peptides with variable lengths. POPI and POPI-MHC2 are SVM-based classifiers with a set of informative features selected by the proposed informative physicochemical property mining algorithm (IPMA).

In this study, two datasets PEPMHCI and PEPMHCI2 of peptides associated with human MHC class I and II molecules extracted from MHCPEP was established, respectively. Considering the correlated effects among physicochemical properties and the cooperation with the SVM classifier, both feature selection and parameter tuning are simultaneously optimized using IPMA. A feature set consisting of 23 and 21 physicochemical properties was selected to implement the prediction system POPI and POPI-MHC2.

To our knowledge POPI and POPI-MHC2 is the first computational system for prediction of peptide immunogenicity based on physicochemical properties. The feature selection method was compared with a rank-based selection method and the selected properties were analyzed using the factor analysis of orthogonal experimental design. Simulation results show that IPMA can select a small set of informative properties considering the correlated effects, compared with the rank-based method.

Three prediction methods were tested for comparison, namely the alignment-based methods ALIGN and PSI-BLAST, and the affinity-driven prediction method AFFIPRE. Because the reference dataset is not sufficiently large, ALIGN and PSI-BLST cannot work well. This poor performance of AFFIPRE shows that affinity is not suitable to predict peptide immunogenicity directly. This result is consistent with previous studies that the peptide immunogenicity does not strongly correlate with its affinity for the MHC molecule [76, 77].

To cope with the small size of the training dataset in mining informative physicochemical properties, the proposed method can provide each selected property with the effectiveness according to its main effect difference in discriminating immunogenic levels and the robustness in terms of selection frequency. The valuable information is helpful in determining a best set of features to implement an accurate prediction system, as well as to further understand immunogenicity from the informative physicochemical properties.

Chapter 6

Identification of T-cell receptor recognition sites

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, 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.

6.1 Motivation

The first predictor for T-cell reactivity published is POPI [59] (Chapter 5). 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 [102-104] 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 [104]; positions 4 and 6 [102]). As an alternative approach to T-cell reactivity, experiments with substitutions and cytotoxicity assays have been performed for HLA-B27 [105]. 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 [106-110]. 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 [83], SYFPEITHI [111, 112] and IEDB [94].

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 [113, 114]. 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

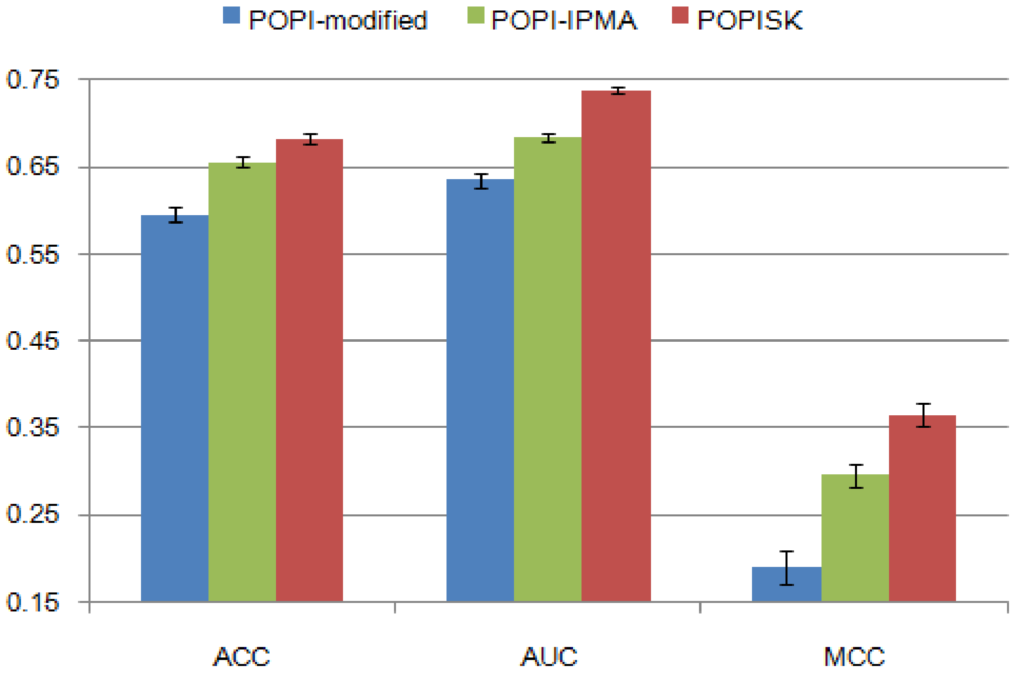


Figure 6.1 Comparison of nested 10-CV performances of POPISK and POPI-modified and POPI-IPMA.

immunogenicity. Three of these positions were reported in previous studies (position 8 [104]; positions 4 and 6 [102]). As a confirmation, graphical analyses using two sample logos [115] identified nearly identical important positions 4, 6, 8 and 9.

6.2 Datasets

We first extracted peptide binders of length 9 with associated human MHC class I alleles and the corresponding immunogenicity data from MHCPEP [83], SYFPEITHI [111, 112] and IEDB [94]. 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 [116]. 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.

6.3 Weighted degree string kernel

An effective weighted degree string kernel [109, 117] 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 $p \in \{1, \ldots, d\}$ at corresponding positions *l* of two sequences, defined as follows:

$$k(s_i, s_j) = \sum_{p=1}^{d} \beta_p \sum_{l=1}^{L-p+1} I(u_{p,l}(s_i) = u_{p,l}(s_j)), \qquad (1)$$

where *I(h)*=1 if *h* is true; otherwise, *I(h)*=0, $u_{p,l}$(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 $\beta_p=2(d-p+1)/(d(d+1))$ are adopted as used in previous study [109]. Shogun [118] release 0.6.7 was used and

LIBSVM [40] was chosen for the implementation of the predictor.

6.4 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 [119].

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 6.1). 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.

6.5 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,

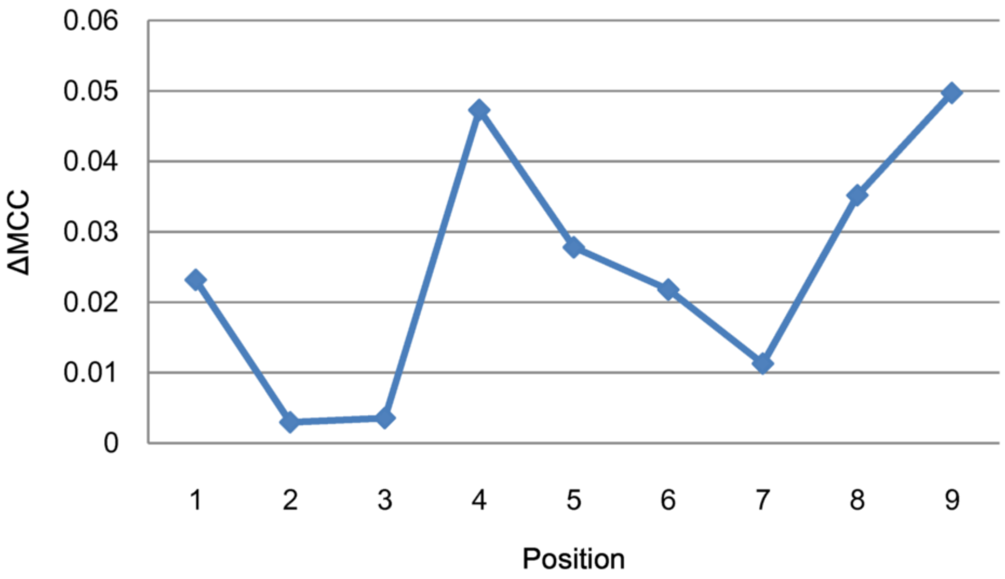


Figure 6.2 The decrease in MCC performances evaluated on datasets without using residues in specific positions.

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 6.1. 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.

6.6 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 [102-104]. 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.

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

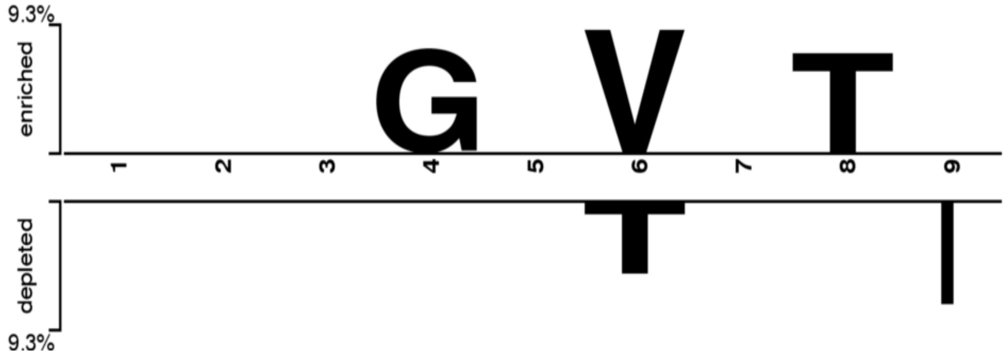


Figure 6.3 Two Sample Logo representation of over- (upper half) and underrepresented (lower half) residues in immunogenic peptides

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 6.2. 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 [115] 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 6.3 shows the resulting two-sample logo representations. The residues overrepresented in immunogenic peptides (shown in the upper half of Figure 6.3) 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 6.3) are threonine and isoleucine at positions 6 and 9, re-

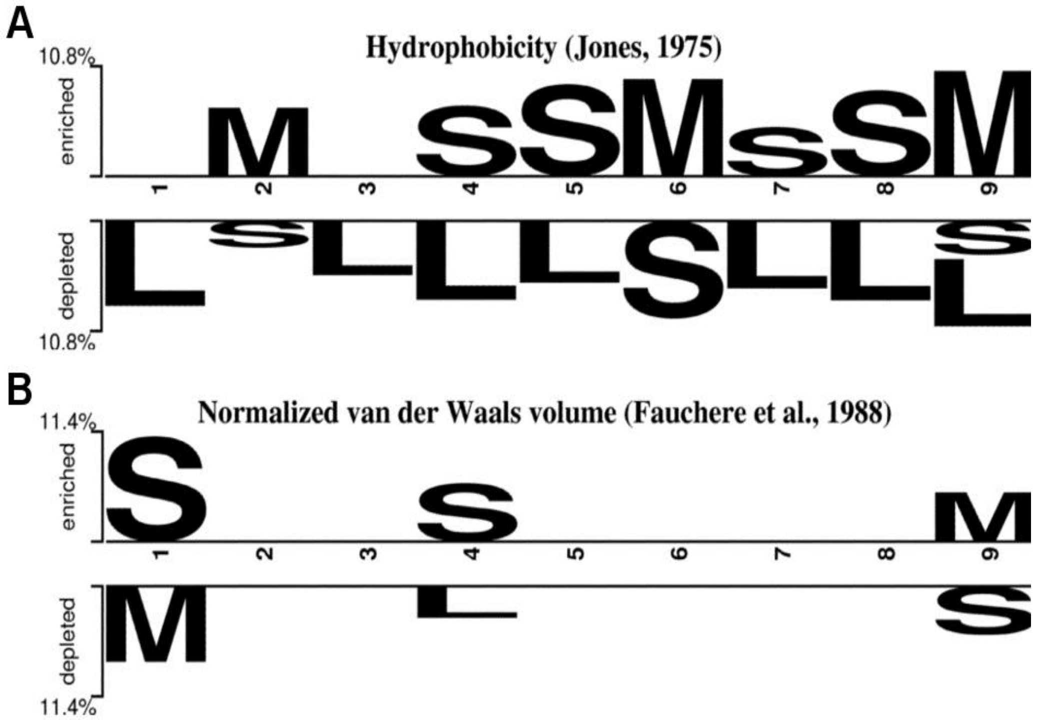


Figure 6.4 The over- (upper half) and underrepresented (lower half) position-specific properties in immunogenic peptides. (A) Hydrophobicity. (B) Normalized van der Waals volume. The symbols S, M and L indicate residues with small, medium and large hydrophobicity/volume, respectively.

spectively.

Our method successfully identified previously reported TCR recognition positions (4, 6 and 8) for HLA-A2 binding peptides from an analysis of crystal structures [102, 104]. Notably, the underrepresented residue isoleucine in position 9 is the anchor residue for peptides binding to HLA-A2 molecules [120]. However, position 2, the primary anchor position of HLA-A2 binding peptides [120, 121], 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 [76, 77]. 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 [76, 77].

6.7 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 identify the important position-independent physicochemical properties, all HLA-A2 binding peptides were encoded as feature vectors with 531 mean values of physicochemical properties. Subsequently, C5.0 was applied to build a decision tree using the whole IMMA2 dataset. The feature usage obtained from C5.0 can be used to rank the physicochemical properties. Table 6.1 shows physicochemical properties with usage larger than 50%.

Hydrophobicity (AAindex IDs MEEJ800102, CASG920101, NAKH900110, and FASG760105) is obviously a major contributor to peptide immunogenicity. Another property with AAindex ID WOLS870102 is correlated with molecular weight and residue volume and probably relates to the limited space between MHC and TCR. Three properties (QIAN880127, RACS820108 and TANS770109) are re-

Table 6.1 Physicochemical properties with feature usage larger than 50%

Usage	AAindex ID	Physicochemical properties
100%	MEEJ800102	Retention coefficient in HPLC, pH2.1
91%	WOLS870102	Principal property value z2
87%	CASG920101	Hydrophobicity scale from native proteins
84%	NAKH900110	Normalized composition of membrane proteins
81%	FASG760105	pK-C
79%	FAUJ880105	STERIMOL minimum width of the side chain
76%	CHAM830107	A parameter of charge transfer capability
61%	QIAN880127	Weights for coil at the window position of -6
59%	RACS820108	Average relative fractional occurrence in AR (i-1)
58%	DIGM050101	Hydrostatic pressure asymmetry index, PAI
56%	TANS770109	Normalized frequency of coil

lated to secondary structure propensities and most likely indicate structural preferences of the peptide backbone.

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) [122] and normalized van der Waals volume (thresholds 2.0 and 6.0) [123]. The encoded sequences yielded the two-sample logos shown in Figure 6.4. Both primary and secondary anchor positions for MHC binding (positions 2 and 9, respectively) and position 6 prefer residues of medium hydrophobicity (Figure 6.4A). Positions 4, 5, 7 and 8 prefer residues of small hydrophobicity. Positions 1 and 4 prefer residues with small van der Waals volume (Figure 6.4B) whereas position 9 prefers medium volume residues. The logos obtained by using the other volume-related properties are similar to Figure 6.4B.

6.8 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.

6.9 Prediction and analysis using 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) [124]. 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 [104, 113, 114].

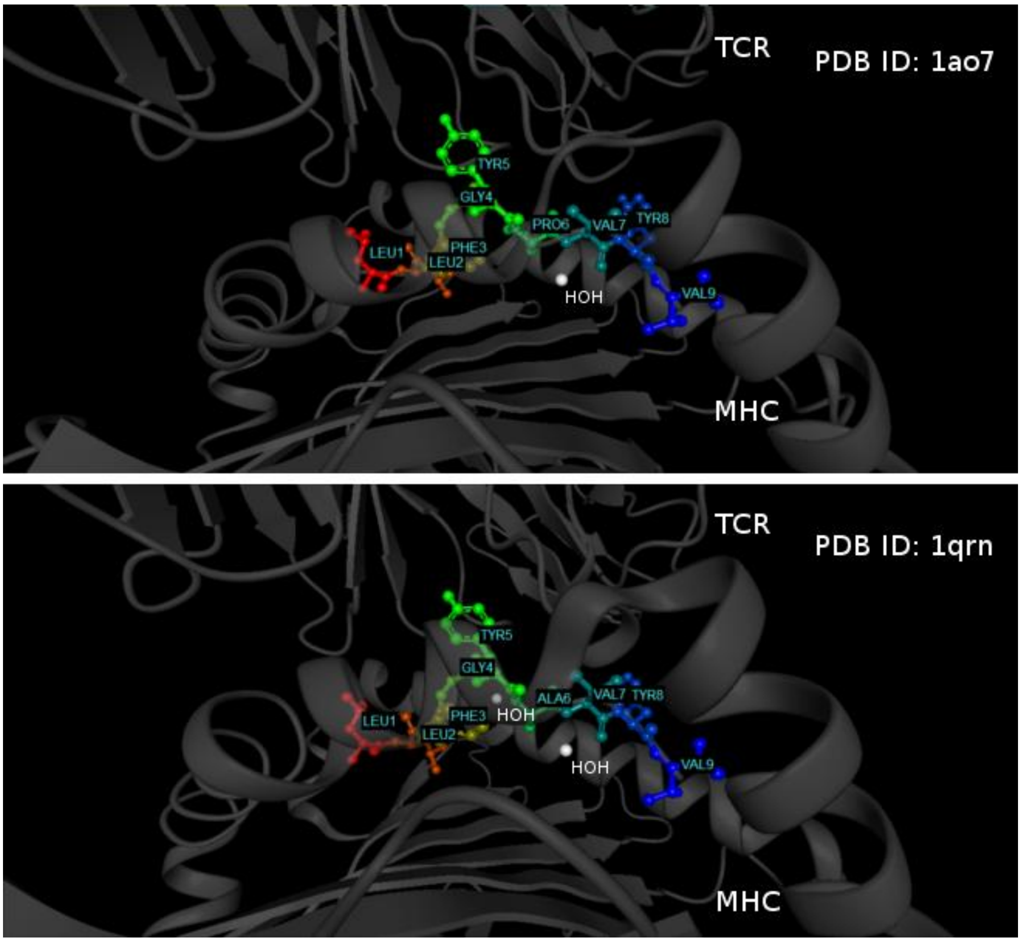


Figure 6.5 Structures of PDB IDs 1ao7 and 1qrn. Structures of PDB IDs 1ao7 and 1qrn share high structural similarity presenting complexes of TCR-peptide-MHC.

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 [113]. 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 [114] 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.* [114]. 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 6.5 generated with BALLView 1.3 [125, 126] 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 [114]. A second difference was evident from shape complementarity analyses, showing a hole in the interface of P6A and a decrease in complementarity [127] 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.

6.10 Summary

The immunogenicity of peptides affected by intrinsic physicochemical properties and the extrinsic immunoglobulin repertoire determines the effectiveness of peptide vaccines and therapeutic peptides. Prediction of peptide immunogenicity will be valuable to the development of peptide vaccines. This study proposes a computational method based on string kernels and support vector machines to predict peptide immunogenicity. Compared to the only published predictor of T-cell reactivity, POPI [59], the new method yields insights into the relevance of specific sequence positions of the peptide for immunogenicity.

A total of three central positions (4, 5 and 6) and three terminal positions (1, 8 and 9) of HLA-A2 binding peptides are identified as important positions for immunogenicity. Positions 4, 6 and 8 are consistent with previously reported T-cell recognition positions [102, 104]. Physicochemical properties of peptides play important roles in determining immunogenic strength. Finally, a prediction system POPISK is constructed and successfully predicts the immunogenicity changes made by single residue modifications. By collecting more data, POPISK is expected to perform better and can be applied to analyze immunogenicity of peptides associated with the other MHC alleles.

Chapter 7

Conclusions

7.1 Summary

Accurate prediction of adaptive T-cell immune response can accelerate the design of vaccines. Previous studies focused on the prediction of antigen processing and presentation pathways and assumed that peptide-MHC binding affinity determines peptide immunogenicity. However, recent studies suggested that binding affinity is required but do not strongly correlate with the strength of immunogenicity. To accurately predict immunogenicity, it is necessary to clarify the relation between binding affinity and immunogenicity and construct more accurate prediction systems. The analysis of prediction model provides insights into the mechanism of T-cell immune responses.

In this dissertation, an informative physicochemical property mining algorithm (IPMA) was developed for extracting information from experimental data. In order to develop a comprehensive computer-aided vaccine design system, three important problems that are rarely addressed because of the huge complexity were investigated including the predictions of immunogenicity induced by MHC class I and II binding peptides and ubiquitylation sites. For predicting peptide immunogenicity, informative physicochemical properties are mined from experimental immunogenicity data using IPMA. Two prediction systems of POPI and POPI-MHC2 were constructed by using the informative physicochemical properties for predicting immunogenicity of MHC class I and II binding peptides, respectively. Both prediction systems perform better than alignment-based and traditional affinity-based methods. The similarity and difference are also analyzed to yield insights into the mechanism of T-cell responses.

Subsequently, a string kernel and MHC allele information are utilized to improve the prediction accuracy of immunogenicity. The developed POPISK prediction system capable of accurately predicting immunogenicity changes made by single residue modifications is utilized to identify T-cell receptor recognition sites. For predicting ubiquitylation sites, three kinds of features and three classifiers were assessed. Results show that the SVM classifier based on physicochemical properties performs best. A large improvement of prediction performances is obtained by further selecting informative physicochemical properties using IPMA. Finally, a prediction system UbiPred was constructed.

The proposed systems for predicting immunogenicity and ubiquitylation and existing methods for predicting antigen processing and presentation pathways provide an efficient way to identify promising epitopes for vaccine design and are expected to accelerate the development of new vaccines.

7.2 Future works

This dissertation presents a novel informative physicochemical property mining algorithm (IPMA) and applied IPMA to mine informative physicochemical properties for predicting immunogenicity. Three prediction systems are proposed as first methods for predicting CTL and HTL responses and protein ubiquitylation sites. While the proposed systems perform so well, further improvement of the proposed systems can provide better assistance for vaccine design.

Future works to improve the proposed systems are shown as follows.

1) The prediction performances and robustness of constructed models can be improved by collecting more data. It also enables the application of string kernels to analyze important positions for T-cell receptor recognition in an allele-specific manner.

2) The proposed prediction systems based on support vector machines are so-called black-box methods. It is hard to interpret how the classifier makes decisions. However, the traditional decision tree methods suffering from their low prediction accuracies are not suitable to be used to predict the complex immune responses. The incorporation of interpretable method of fuzzy rule-based classifiers proposed by professor Ho [82] is expected to provide intuitive fuzzy rules with high prediction accuracies for better understanding immune systems.

3) The position effect of physicochemical was not considered by IPMA because of the various length property of used datasets PEPMHCI and PEPMHCII. However, the position effect is an important feature to determine immunogenicity. The IPMA method can be improved by using position-dependent physicochemical properties instead of using position-independent physicochemical properties. By mining and analyzing position-dependent informative physicochemical properties, the prediction performance is expected to perform better than the proposed method, and it will yield better insights into the T-cell response.

4) While the use of position-dependent physicochemical properties is expected to have better prediction performances, it can be further improved by incorporating weighted physiochemical properties. The weight for each position can be obtained by applying the same elimination-and-test method as shown in Chapter 6. The analysis result can provide better understanding of T-cell immune response, and can be utilized to further improve the method.

Reference

[1] Ulmer, J. B., Valley, U., Rappuoli, R., Vaccine manufacturing: challenges and solutions. *Nat. Biotechnol.* 2006, *24*, 1377-1383.

[2] Ciechanover, A., Early work on the ubiquitin proteasome system, an interview with Aaron Ciechanover. Interview by CDD. *Cell Death Differ.* 2005, *12*, 1167-1177.

[3] Hershko, A., Early work on the ubiquitin proteasome system, an interview with Avram Hershko. Interview by CDD. *Cell Death Differ.* 2005, *12*, 1158-1161.

[4] Wang, J., Maldonado, M. A., The ubiquitin-proteasome system and its role in inflammatory and autoimmune diseases. *Cell. Mol. Immunol.* 2006, *3*, 255-261.

[5] Michalek, M. T., Grant, E. P., Gramm, C., Goldberg, A. L., Rock, K. L., A role for the ubiquitin-dependent proteolytic pathway in MHC class I-restricted antigen presentation. *Nature* 1993, *363*, 552-554.

[6] Townsend, A., Bastin, J., Gould, K., Brownlee, G., *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*, 1211-1224.

[7] Liu, W. J., Zhao, K. N., Gao, F. G., Leggatt, G. R., *et al.*, Polynucleotide viral vaccines: codon optimisation and ubiquitin conjugation enhances prophylactic and therapeutic efficacy. *Vaccine* 2001, *20*, 862-869.

[8] Wang, Q. M., Sun, S. H., Hu, Z. L., Zhou, F. J., *et al.*, Epitope DNA vaccines against tuberculosis: spacers and ubiquitin modulates cellular immune responses elicited by epitope DNA vaccine. *Scand. J. Immunol.* 2004, *60*, 219-225.

[9] Rodriguez, F., An, L. L., Harkins, S., Zhang, J., *et al.*, DNA immunization with minigenes: low frequency of memory cytotoxic T lymphocytes and inefficient anti-

viral protection are rectified by ubiquitination. *J. Virol.* 1998, *72*, 5174-5181.

[10] Deavin, A. J., Auton, T. R., Greaney, P. J., Statistical comparison of established T-cell epitope predictors against a large database of human and murine antigens. *Mol. Immunol.* 1996, *33*, 145-155.

[11] Keşmir, C., Nussbaum, A. K., Schild, H., Detours, V., Brunak, S., Prediction of proteasome cleavage motifs by neural networks. *Protein Eng.* 2002, *15*, 287-296.

[12] Bhasin, M., Raghava, G. P. S., Pcleavage: an SVM based method for prediction of constitutive proteasome and immunoproteasome cleavage sites in antigenic sequences. *Nucleic Acids Res.* 2005, *33*, W202-W207.

[13] Bhasin, M., Raghava, G. P., Analysis and prediction of affinity of TAP binding peptides using cascade SVM. *Protein Sci.* 2004, *13*, 596-607.

[14] Peters, B., Bulik, S., Tampe, R., Van Endert, P. M., Holzhutter, H. G., Identifying MHC class I epitopes by predicting the TAP transport efficiency of epitope precursors. *J. Immunol.* 2003, *171*, 1741-1749.

[15] Dönnes, P., Elofsson, A., Prediction of MHC class I binding peptides, using SVMHC. *BMC Bioinformatics* 2002, *3*, 25.

[16] Nielsen, M., Lundegaard, C., Worning, P., Hvid, C. S.*, et al.*, Improved prediction of MHC class I and class II epitopes using a novel Gibbs sampling approach. *Bioinformatics* 2004, *20*, 1388-1397.

[17] Dönnes, P., Kohlbacher, O., Integrated modeling of the major events in the MHC class I antigen processing pathway. *Protein Sci.* 2005, *14*, 2132-2140.

[18] Larsen, M. V., Lundegaard, C., Lamberth, K., Buus, S.*, 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*, 2295-2303.

[19] Goldsby, R. A., Kindt, T. J., Osborne, B. A., Kuby, J., *Immunology*, W.H. Freeman, New York 2003.

[20] van Bergen, J., Ossendorp, F., Jordens, R., Mommaas, A. M.*, et al.*, Get into the groove! Targeting antigens to MHC class II. *Immunol. Rev.* 1999, *172*, 87-96.

[21] Karpenko, O., Shi, J., Dai, Y., Prediction of MHC class II binders using the ant

colony search strategy. *Artif. Intell. Med.* 2005, *35*, 147-156.

[22] Brusic, V., Rudy, G., Honeyman, G., Hammer, J., Harrison, L., Prediction of MHC class II-binding peptides using an evolutionary algorithm and artificial neural network. *Bioinformatics* 1998, *14*, 121-130.

[23] Rajapakse, M., Schmidt, B., Feng, L., Brusic, V., Predicting peptides binding to MHC class II molecules using multi-objective evolutionary algorithms. *BMC Bioinformatics* 2007, *8*, 459.

[24] Bisset, L. R., Fierz, W., Using a neural network to identify potential HLA-DR1 binding sites within proteins. *J. Mol. Recognit.* 1993, *6*, 41-48.

[25] Honeyman, M. C., Brusic, V., Stone, N. L., Harrison, L. C., Neural network-based prediction of candidate T-cell epitopes. *Nat. Biotechnol.* 1998, *16*, 966-969.

[26] Burden, F. R., Winkler, D. A., Predictive Bayesian neural network models of MHC class II peptide binding. *J. Mol. Graph. Model.* 2005, *23*, 481-489.

[27] Noguchi, H., Hanai, T., Honda, H., Harrison, L. C., Kobayashi, T., Fuzzy neural network-based prediction of the motif for MHC class II binding peptides. *J. Biosci. Bioeng.* 2001, *92*, 227-231.

[28] Noguchi, H., Kato, R., Hanai, T., Matsubara, Y.*, et al.*, Hidden Markov model-based prediction of antigenic peptides that interact with MHC class II molecules. *J. Biosci. Bioeng.* 2002, *94*, 264-270.

[29] Bhasin, M., Raghava, G. P., SVM based method for predicting HLA-DRB1*0401 binding peptides in an antigen sequence. *Bioinformatics* 2004, *20*, 421-423.

[30] Cui, J., Han, L. Y., Lin, H. H., Zhang, H. L.*, et al.*, Prediction of MHC-binding peptides of flexible lengths from sequence-derived structural and physicochemical properties. *Mol. Immunol.* 2007, *44*, 866-877.

[31] Wan, J., Liu, W., Xu, Q., Ren, Y.*, et al.*, SVRMHC prediction server for MHC-binding peptides. *BMC Bioinformatics* 2006, *7*, 463.

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

[33] Blythe, M. J., Flower, D. R., Benchmarking B cell epitope prediction: underperformance of existing methods. *Protein Sci.* 2005, *14*, 246-248.

[34] Cao, Y., Liu, S., Zhang, L., Qin, J*., et al.*, Prediction of protein structural class with Rough Sets. *BMC Bioinformatics* 2006, *7*, 20.

[35] Idicula-Thomas, S., Kulkarni, A. J., Kulkarni, B. D., Jayaraman, V. K., Balaji, P. V., 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*, 278-284.

[36] Liu, W., Meng, X., Xu, Q., Flower, D. R., Li, T., Quantitative prediction of mouse class I MHC peptide binding affinity using support vector machine regression (SVR) models. *BMC Bioinformatics* 2006, *7*, 182.

[37] Nanni, L., Lumini, A., An ensemble of K-local hyperplanes for predicting protein-protein interactions. *Bioinformatics* 2006, *22*, 1207-1210.

[38] Kawashima, S., Kanehisa, M., AAindex: amino acid index database. *Nucleic Acids Res.* 2000, *28*, 374.

[39] Kawashima, S., Pokarowski, P., Pokarowska, M., Kolinski, A*., et al.*, AAindex: amino acid index database, progress report 2008. *Nucleic Acids Res.* 2008, *36*, D202-205.

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

[41] Dey, A., *Orthogonal fractional factorial designs*, Wiley, New York 1985.

[42] Wu, Q., On the optimality of orthogonal experimental design. *Acta Math. Appl. Sinica* 1978, *1*, 283-299.

[43] Ho, S. Y., Chen, J. H., Huang, M. H., Inheritable genetic algorithm for biobjective 0/1 combinatorial optimization problems and its applications. *IEEE Trans. Syst. Man. Cybern. B Cybern.* 2004, *34*, 609-620.

[44] Ho, S. Y., Shu, L. S., Chen, J. H., Intelligent evolutionary algorithms for large parameter optimization problems. *IEEE Trans. Evol. Comput.* 2004, *8*, 522-541.

[45] Herrmann, J., Lerman, L. O., Lerman, A., Ubiquitin and ubiquitin-like proteins in protein regulation. *Circulation Res.* 2007, *100*, 1276-1291.

[46] Welchman, R. L., Gordon, C., Mayer, R. J., Ubiquitin and ubiquitin-like proteins as multifunctional signals. *Nat. Rev. Mol. Cell. Biol.* 2005, *6*, 599-609.

[47] Tomlinson, E., Palaniyappan, N., Tooth, D., Layfield, R., Methods for the purification of ubiquitinated proteins. *Proteomics* 2007, *7*, 1016-1022.

[48] Denis, N. J., Vasilescu, J., Lambert, J. P., Smith, J. C., Figeys, D., Tryptic digestion of ubiquitin standards reveals an improved strategy for identifying ubiquitinated proteins by mass spectrometry. *Proteomics* 2007, *7*, 868-874.

[49] Hitchcock, A. L., Auld, K., Gygi, S. P., Silver, P. A., 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*, 12735-12740.

[50] Jeon, H. B., Choi, E. S., Yoon, J. H., Hwang, J. H.*, et al.*, A proteomics approach to identify the ubiquitinated proteins in mouse heart. *Biochem. Biophys. Res. Commun.* 2007, *357*, 731-736.

[51] Kirkpatrick, D. S., Weldon, S. F., Tsaprailis, G., Liebler, D. C., Gandolfi, A. J., Proteomic identification of ubiquitinated proteins from human cells expressing His-tagged ubiquitin. *Proteomics* 2005, *5*, 2104-2111.

[52] Matsumoto, M., Hatakeyama, S., Oyamada, K., Oda, Y.*, et al.*, Large-scale analysis of the human ubiquitin-related proteome. *Proteomics* 2005, *5*, 4145-4151.

[53] Peng, J., Schwartz, D., Elias, J. E., Thoreen, C. C.*, et al.*, A proteomics approach to understanding protein ubiquitination. *Nat. Biotechnol.* 2003, *21*, 921-926.

[54] Denison, C., Kirkpatrick, D. S., Gygi, S. P., Proteomic insights into ubiquitin and ubiquitin-like proteins. *Curr. Opin. Chem. Biol.* 2005, *9*, 69-75.

[55] Xue, Y., Chen, H., Jin, C., Sun, Z., Yao, X., NBA-Palm: prediction of palmitoylation site implemented in Naive Bayes algorithm. *BMC Bioinformatics* 2006, *7*, 458.

[56] Jones, D. T., Improving the accuracy of transmembrane protein topology prediction using evolutionary information. *Bioinformatics* 2007, *23*, 538-544.

[57] Kaur, H., Raghava, G. P., A neural network method for prediction of beta-turn types in proteins using evolutionary information. *Bioinformatics* 2004, *20*, 2751-2758.

[58] Huang, W. L., Tung, C. W., Huang, H. L., Hwang, S. F., Ho, S. Y., ProLoc: Prediction of protein subnuclear localization using SVM with automatic selection from physicochemical composition features. *Biosystems* 2007, *90*, 573-581.

[59] Tung, C. W., Ho, S. Y., POPI: predicting immunogenicity of MHC class I binding peptides by mining informative physicochemical properties. *Bioinformatics* 2007, *23*, 942-949.

[60] Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., *et al.*, Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 1997, *25*, 3389-3402.

[61] Witten, I. H., Frank, E., *Data Mining: Practical machine learning tools and techniques*, Morgan Kaufmann, San Francisco 2005.

[62] Chernorudskiy, A. L., Garcia, A., Eremin, E. V., Shorina, A. S., *et al.*, UbiProt: a database of ubiquitylated proteins. *BMC Bioinformatics* 2007, *8*, 126.

[63] Crooks, G. E., Hon, G., Chandonia, J. M., Brenner, S. E., WebLogo: a sequence logo generator. *Genome Res.* 2004, *14*, 1188-1190.

[64] Quinlan, J. R., Morgan Kaufmann, San Mateo, CA 1993.

[65] Li, W., Godzik, A., Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* 2006, *22*, 1658-1659.

[66] 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*, 1398-1405.

[67] Harpaz, Y., Gerstein, M., Chothia, C., Volume changes on protein folding. *Structure* 1994, *2*, 641-649.

[68] Cornette, J. L., Cease, K. B., Margalit, H., Spouge, J. L., *et al.*, Hydrophobicity scales and computational techniques for detecting amphipathic structures in proteins. *J. Mol. Biol.* 1987, *195*, 659-685.

[69] Cedano, J., Aloy, P., Perez-Pons, J. A., Querol, E., Relation between amino acid composition and cellular location of proteins. *J. Mol. Biol.* 1997, *266*, 594-600.

[70] George, R. A., Heringa, J., An analysis of protein domain linkers: their classification and role in protein folding. *Protein Eng.* 2002, *15*, 871-879.

[71] Radivojac, P., Vacic, V., Haynes, C., Cocklin, R. R., *et al.*, Identification, analysis, and prediction of protein ubiquitination sites. *Proteins* 2009, *78*, 365-380.

[72] Edwards, Y. J., Lobley, A. E., Pentony, M. M., Jones, D. T., Insights into the regulation of intrinsically disordered proteins in the human proteome by analyzing sequence and gene expression data. *Genome Biol.* 2009, *10*, R50.

[73] Pang, C. N., Gasteiger, E., Wilkins, M. R., Identification of arginine- and lysine-methylation in the proteome of Saccharomyces cerevisiae and its functional implications. *BMC Genomics* 2010, *11*, 92.

[74] Kanduc, D., Peptimmunology: immunogenic peptides and sequence redundancy. *Curr. Drug. Discov. Technol.* 2005, *2*, 239-244.

[75] Van Regenmortel, M. H., Antigenicity and immunogenicity of synthetic peptides. *Biologicals* 2001, *29*, 209-213.

[76] Feltkamp, M. C., Vierboom, M. P., Kast, W. M., Melief, C. J., Efficient MHC class I-peptide binding is required but does not ensure MHC class I-restricted immunogenicity. *Mol. Immunol.* 1994, *31*, 1391-1401.

[77] Ochoa-Garay, J., McKinney, D. M., Kochounian, H. H., McMillan, M., 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*, 273-281.

[78] Dow, C., Oseroff, C., Peters, B., Nance-Sotelo, C., *et al.*, Lymphocytic choriomeningitis virus infection yields overlapping CD4+ and CD8+ T-cell responses. *J. Virol.*

2008, *82*, 11734-11741.

[79] Arnold, P. Y., La Gruta, N. L., Miller, T., Vignali, K. M., *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*, 739-749.

[80] Conant, S. B., Swanborg, R. H., MHC class II peptide flanking residues of exogenous antigens influence recognition by autoreactive T cells. *Autoimmun. Rev.* 2003, *2*, 8-12.

[81] Sarda, D., Chua, G. H., Li, K.-B., Krishnan, A., pSLIP: SVM based protein subcellular localization prediction using multiple physicochemical properties. *BMC*

Bioinformatics 2005, *6*, 152.

[82] Ho, S. Y., Hsieh, C. H., Chen, H. M., Huang, H. L., Interpretable gene expression classifier with an accurate and compact fuzzy rule base for microarray data analysis. *Biosystems* 2006, *85*, 165-176.

[83] Brusic, V., Rudy, G., Harrison, L. C., MHCPEP, a database of MHC-binding peptides: update 1997. *Nucleic Acids Res.* 1998, *26*, 368-371.

[84] Myers, E. W., Miller, W., Optimal alignments in linear space. *Comput. Appl. Biosci.* 1988, *4*, 11-17.

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

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

[87] Kuhn, L. A., Swanson, C. A., Pique, M. E., Tainer, J. A., Getzoff, E. D., Atomic and residue hydrophilicity in the context of folded protein structures. *Proteins* 1995, *23*, 536-547.

[88] Rajesh, S., Sakamoto, T., Iwamoto-Sugai, M., Shibata, T.*, et al.*, Ubiquitin binding interface mapping on yeast ubiquitin hydrolase by NMR chemical shift perturbation. *Biochemistry* 1999, *38*, 9242-9253.

[89] Sundberg, E. J., Urrutia, M., Braden, B. C., Isern, J.*, et al.*, Estimation of the hydrophobic effect in an antigen-antibody protein-protein interface. *Biochemistry* 2000, *39*, 15375-15387.

[90] Melton, S. J., Landry, S. J., Three dimensional structure directs T-cell epitope dominance associated with allergy. *Clin. Mol. Allergy* 2008, *6*, 9.

[91] Mirano-Bascos, D., Tary-Lehmann, M., Landry, S. J., Antigen structure influences helper T-cell epitope dominance in the human immune response to HIV envelope glycoprotein gp120. *Eur. J. Immunol.* 2008, *38*, 1231-1237.

[92] Jager, E., Karbach, J., Gnjatic, S., Neumann, A.*, 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*, 14453-14458.

[93] Odunsi, K., Qian, F., Matsuzaki, J., Mhawech-Fauceglia, P., *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*, 12837-12842.

[94] Peters, B., Sidney, J., Bourne, P., Bui, H. H., *et al.*, The immune epitope database and analysis resource: from vision to blueprint. *PLoS Biology* 2005, *3*, e91.

[95] Bui, H. H., Sidney, J., Peters, B., Sathiamurthy, M., *et al.*, Automated generation and evaluation of specific MHC binding predictive tools: ARB matrix applications. *Immunogenetics* 2005, *57*, 304-314.

[96] Zhang, Q., Wang, P., Kim, Y., Haste-Andersen, P., *et al.*, Immune epitope database analysis resource (IEDB-AR). *Nucleic Acids Res* 2008, *36*, W513-518.

[97] Nielsen, M., Lundegaard, C., Blicher, T., Peters, B., *et al.*, Quantitative predictions of peptide binding to any HLA-DR molecule of known sequence: NetMHCIIpan. *PLoS Comput Biol* 2008, *4*, e1000107.

[98] Wang, P., Sidney, J., Dow, C., Mothe, B., *et al.*, A systematic assessment of MHC class II peptide binding predictions and evaluation of a consensus approach. *PLoS Comput Biol* 2008, *4*, e1000048.

[99] Lin, H. H., Zhang, G. L., Tongchusak, S., Reinherz, E. L., Brusic, V., Evaluation of MHC-II peptide binding prediction servers: applications for vaccine research. *BMC Bioinformatics* 2008, *9 Suppl 12*, S22.

[100] Sorgo, A. G., Heilmann, C. J., Dekker, H. L., Brul, S., *et al.*, Mass spectrometric analysis of the secretome of Candida albicans. *Yeast* 2010.

[101] Ziehm, M., *Prediction of peptide immunogenicity using T cell selection modelling*, University of Tübingen, Diplomarbeit, Tübingen 2009.

[102] Rudolph, M. G., Luz, J. G., Wilson, I. A., Structural and thermodynamic correlates of T cell signaling. *Annu. Rev. Biophys. Biomol. Struct.* 2002, *31*, 121-149.

[103] Silver, M. L., Guo, H. C., Strominger, J. L., Wiley, D. C., Atomic structure of a human MHC molecule presenting an influenza virus peptide. *Nature* 1992, *360*, 367-369.

[104] Stewart-Jones, G. B., McMichael, A. J., Bell, J. I., Stuart, D. I., Jones, E. Y., A structural basis for immunodominant human T cell receptor recognition. *Nat.*

Immunol. 2003, *4*, 657-663.

[105] Bowness, P., Allen, R. L., McMichael, A. J., Identification of T cell receptor recognition residues for a viral peptide presented by HLA B27. *Eur. J. Immunol.*

1994, *24*, 2357-2363.

[106] Boisvert, S., Marchand, M., Laviolette, F., Corbeil, J., HIV-1 coreceptor usage prediction without multiple alignments: an application of string kernels. *Retrovirology* 2008, *5*, 110.

[107] El-Manzalawy, Y., Dobbs, D., Honavar, V., Predicting linear B-cell epitopes using string kernels. *J. Mol. Recognit.* 2008, *21*, 243-255.

[108] Jacob, L., Vert, J. P., Efficient peptide-MHC-I binding prediction for alleles with few known binders. *Bioinformatics* 2008, *24*, 358-366.

[109] Rätsch, G., Sonnenburg, S., Scholkopf, B., RASE: recognition of alternatively spliced exons in C.elegans. *Bioinformatics* 2005, *21 Suppl 1*, i369-377.

[110] Sonnenburg, S., Zien, A., Philips, P., Ratsch, G., POIMs: positional oligomer importance matrices--understanding support vector machine-based signal detectors. *Bioinformatics* 2008, *24*, i6-14.

[111] Rammensee, H., Bachmann, J., Emmerich, N. P., Bachor, O. A., Stevanovic, S., SYFPEITHI: database for MHC ligands and peptide motifs. *Immunogenetics* 1999, *50*, 213-219.

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

[113] Chen, J. L., Stewart-Jones, G., Bossi, G., Lissin, N. M.*, et al.*, Structural and kinetic basis for heightened immunogenicity of T cell vaccines. *J. Exp. Med.* 2005, *201*, 1243-1255.

[114] Ding, Y. H., Baker, B. M., Garboczi, D. N., Biddison, W. E., Wiley, D. C., Four A6-TCR/peptide/HLA-A2 structures that generate very different T cell signals are nearly identical. *Immunity* 1999, *11*, 45-56.

[115] Vacic, V., Iakoucheva, L. M., Radivojac, P., Two Sample Logo: a graphical representation of the differences between two sets of sequence alignments. *Bioinformatics* 2006, *22*, 1536-1537.

[116] Lund, O., Nielsen, M., Kesmir, C., Petersen, A. G., *et al.*, Definition of supertypes for HLA molecules using clustering of specificity matrices. *Immunogenetics* 2004, *55*, 797-810.

[117] Rätsch, G., Sonnenburg, S., *MIT Press MIT Press series on Computational Molecular Biology* 2003, pp. 277-298.

[118] Sonnenburg, S., Ratsch, G., Schafer, C., Scholkopf, B., Large scale multiple kernel learning. *J. Mach. Learn. Res.* 2006, *7*, 1531-1565.

[119] Varma, S., Simon, R., Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics* 2006, *7*, 91.

[120] Hunt, D. F., Henderson, R. A., Shabanowitz, J., Sakaguchi, K., *et al.*, Characterization of peptides bound to the class I MHC molecule HLA-A2.1 by mass spectrometry. *Science* 1992, *255*, 1261-1263.

[121] Falk, K., Rotzschke, O., Stevanovic, S., Jung, G., Rammensee, H. G., Allele-specific motifs revealed by sequencing of self-peptides eluted from MHC molecules. *Nature* 1991, *351*, 290-296.

[122] Jones, D. D., Amino acid properties and side-chain orientation in proteins: a cross correlation appraoch. *J. Theor. Biol.* 1975, *50*, 167-183.

[123] Fauchere, J. L., Charton, M., Kier, L. B., Verloop, A., Pliska, V., Amino acid side chain parameters for correlation studies in biology and pharmacology. *Int. J. Pept. Protein Res.* 1988, *32*, 269-278.

[124] Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., *et al.*, The Protein Data Bank. *Nucleic Acids Res.* 2000, *28*, 235-242.

[125] Kohlbacher, O., Lenhof, H. P., BALL--rapid software prototyping in computational molecular biology. Biochemicals Algorithms Library. *Bioinformatics* 2000, *16*, 815-824.

[126] Moll, A., Hildebrandt, A., Lenhof, H. P., Kohlbacher, O., BALLView: an object-oriented molecular visualization and modeling framework. *J. Comput. Aided Mol. Des.* 2005, *19*, 791-800.

[127] Baker, B. M., Ding, Y. H., Garboczi, D. N., Biddison, W. E., Wiley, D. C., Structural, biochemical, and biophysical studies of HLA-A2/altered peptide ligands binding to viral-peptide-specific human T-cell receptors. *Cold Spring Har-*

bor Symposia on Quantitative Biology 1999, *64*, 235-241.

Curriculum Vitae

➢ Education

Year	Degree	Institute
2001-2005	B.S.	Department of Biology, National Cheng Kung University
2005-2006	Master student	Institute of Bioinformatics, National Chiao Tung University
2006-2010	PhD	Institute of Bioinformatics, National Chiao Tung University

➢ Experience

Year	Position	Institute
2008-2009	Visiting scholar	Wilhelm Schickard Institute for Computer Science, Eberhard Karls University Tübingen, Tübingen, Germany (Prof. Oliver Kohlbacher)

➢ Project

Year	Title	Funder
2004-2005	Identification of microRNAs from Phalaenopsis equestris by bioinformatics	NSC College Student Research Project (NSC 93-2815-C-006 -084 -B)

➢ Professional certification and awards

Year	Certification or Awards
2004	SUN Certified JAVA Programmer (SCJP 1.4)
2007	Research Excellence Award (by College of Biological Science and Technology, NCTU, Taiwan)
2008	Research Excellence Award (by College of Biological Science and Technology, NCTU, Taiwan)
2008	Scholarship of Sandwich Program for research visits to Germany (supported by DAAD of Germany and NSC of Taiwan)
2010	Research Excellence Award (by College of Biological Science and Technology, NCTU, Taiwan)

➢ Academic service

Year	Description
2009	Program committee, The special session "Evolutionary Computation in Bioinformatics and Computational Biology" of 2009 IEEE Congress on Evolutionary Computation (IEEE CEC 2009). Trondheim, Norway, May 18-21, 2009
2008-2010	Reviewer for Journal of Proteomics & Bioinformatics
2008	Reviewer for Bioinformatics and Biology Insights
2010	Reviewer for 2010 IEEE Congress on Evolutionary Computation (IEEE CEC 2010). Barcelona, Spain, July 18-23, 2010

Publications

➢ Journal papers

1. **Tung, C.W.**, Ziehm, M., Kämper, A., Ho, S.Y. and Kohlbacher, O. (2010) POPISK: T-cell reactivity prediction using support vector machines and string kernels. *PLoS ONE*. (under revision)
2. **Tung, C.W.** and Ho, S.Y. (2010) Predicting immunogenicity of MHC class II-restricted peptides. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*. (under review)
3. Huang, W.L., **Tung, C.W.** and Ho, S.Y. (2010) Predicting promoters by identifying and analyzing an informative feature set of DNA sequence descriptors. *BMC Bioinformatics* (under review)
4. Huang, W.L., **Tung, C.W.**, Huang, H.L. and Ho, S.Y. (2009) Predicting protein subnuclear localization using GO-amino-acid composition features. *Bio Systems*, 98, 73-79.
5. Hsu, K.T., Huang, H.L., **Tung, C.W.**, Chen, Y.H. and Ho, S.Y.(2009) Analysis of physicochemical properties on prediction of R5, X4, and R5X4 HIV-1 coreceptor usage. *International Journal of Biological and Life Sciences*, 5, 208-215.
6. **Tung, C.W.** and Ho, S.Y. (2008) Computational identification of ubiquitylation sites from protein sequences. *BMC* Bioinformatics, 9, 310. **(Highly accessed)**
7. Huang, W.L., **Tung, C.W.**, Ho, S.W., Hwang, S.F. and Ho, S.Y. (2008) ProLoc-GO: Utilizing informative Gene Ontology terms for sequence-based prediction of protein subcellular localization. *BMC Bioinformatics*, 9, 80.
8. **Tung, C.W.** and Ho, S.Y. (2007) POPI: predicting immunogenicity of MHC class I binding peptides by mining informative physicochemical properties. *Bioinformatics*, 23, 942-949.

9. Huang, W.L., **Tung, C.W.**, Huang, H.L., Hwang, S.F. and Ho, S.Y. (2007) ProLoc: Prediction of protein subnuclear localization using SVM with automatic selection from physicochemical composition features. *Bio Systems*, 90, 573-581.
10. Tsai, W.C., Hsiao, Y.Y., Lee, S.H., **Tung, C.W.**, Wang, D.P., Wang, H.C., Chen, W.H. and Chen, H.H. (2006) Expression analysis of the ESTs derived from the flower buds of_*Phalaenopsis equestris*. *Plant Science*, 170, 426-432.

Paper under preparation:

11. **Tung, C.W.** and Ho, S.Y. (2010) Towards a consensus feature set for survival prediction of hepatic cancer patients.
12. Huang, W.L., **Tung, C.W.** and Ho, S.Y. (2010) Informative GO-amino-acid composition features for predicting subcellular localization of both eukaryotic and prokaryotic proteins.

➢ International conferences

1. Liaw, C., **Tung, C.W.**, Ho, S.J. and Ho, S.Y. (2010) Sequence-based Prediction Of Gamma-turn Types Using A Physicochemical Property-based Decision Tree Method, *International Conference on Computational Biology (ICCB2010)*, Tokyo, Japan. (EI)
2. **Tung, C.W.**, Liaw, C., Ho, S.J. and Ho, S.Y. (2010) Prediction of protein subchloroplast locations using Random Forests, *International Conference on Computational Biology (ICCB2010)*, Tokyo, Japan. (EI)
3. Huang, W.L., **Tung, C.W.** and Ho, S.Y. (2010) Human Pol II promoter prediction by using nucleotide property composition features, *International Symposium on Biocomputing (ISB2010)*, Calicut, Kerala, India.
4. Hsu, K.T., Huang, H.L., **Tung, C.W.**, Chen, Y.H. and Ho, S.Y.(2009) Analysis of physicochemical properties on prediction of R5, X4, and R5X4 HIV-1 coreceptor usage. *International Conference on Bioinformatics and Bioengineering (ICBB2009)*, Tokyo, Japan, 53, 1120-1127. (EI)
5. Huang, W.L., **Tung, C.W.**, Ho, S.W. and Ho, S.Y. (2008) 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 (CIBCB2008)*, Sun Valley, Idaho, USA, 201-206.
6. **Tung, C.W.** and Ho, S.Y. (2007) Mining physicochemical properties for predicting immunogenicity of MHC class II binding peptides. *18th Interna-*

tional Conference on Genome Informatics (GIW2007), Biopolis, Singapore.

7. Hsiao, Y.Y., Tsai, W.C., **Tung, C.W.**, Chiu, Y.F., Pan, Z.J., Chen, W.H. and Chen, H.H. (2004) Gene expression during *Phalaenopsis* embryo development. *International Symposium on Agricultural Genomics and Biotechnology*, Tainan, Taiwan.

Conferences hold by NCTU

1. **Tung, C.W.** and Ho, S.Y. (2010) Prediction of adaptive T-cell dependent immune response. *Inter-discipline biotechnology symposium*, National Chiao Tung University, Hsinchu, Taiwan.
2. Liaw C., **Tung, C.W.** and Ho, S.Y. (2010) Sequence-based Prediction Of Gamma-turn Types Using A Physicochemical Property-based Decision Tree Method. *Inter-discipline biotechnology symposium*, National Chiao Tung University, Hsinchu, Taiwan.
3. Yu, Y.Y., Tsai, C.T., **Tung, C.W.** and Ho, S.Y. (2009) POCP: Prediction of Cyclin Proteins by Mining Informative Physicochemical Properties. *Competition of Academic Posters*, National Chiao Tung University, Hsinchu, Taiwan.
4. **Tung, C.W.** and Ho, S.Y. (2008) Computational identification of ubiquitylation sites from protein sequences. *Competition of Academic Posters*, National Chiao Tung University, Hsinchu, Taiwan.
5. **Tung, C.W.** and Ho, S.Y. (2007) The impact of physicochemical properties on predicting immune responses induced by MHC binding peptides. *Competition of academic posters*, National Tsing Hua University, Hsinchu, Taiwan.
6. Tsai, C.T., **Tung, C.W.** and Ho, S.Y. (2006) Predicting continuous B-cell epitopes by mining informative physicochemical properties. *Academic Symposium of Biotechnology and Competition of Academic Posters*, National Chiao Tung University, Hsinchu, Taiwan.
7. **Tung, C.W.** and Ho, S.Y. (2006) POPI: Predicting immunogenicity of MHC class I binding peptides by mining informative physicochemical properties. *Academic Symposium of Biotechnology and Competition of Academic Posters*, National Chiao Tung University, Hsinchu, Taiwan.

Invited talk

1. **Tung, C.W.** and Ho, S.Y. (2008) POPI: predicting immunogenicity of MHC class I binding peptides by mining informative physicochemical properties. *Academic Symposium of Biotechnology*, National Chiao Tung University, Hsinchu, Taiwan.