Epitope and T-cell Reactivity Prediction Using Machine Learning Approaches (機械学習アプローチを用いたエピトープ予測およびT細胞反応性予測)

SAETHANG THAMMAKORN

Graduate School of Natural Science and Technology
Kanazawa University

Abstract

The development of new vaccines is very necessary to protect human populations from deadly infectious pathogens. Epitope identification is one of the most important steps in the vaccine development since epitopes play an essential role in the activation of the immune response. Epitopes are conventionally identified by immunological experiments. However, such approaches are time-consuming and laborious. Therefore, computational methods have been applied to speed up the process of the vaccine development by searching for novel epitopes. This application is commonly called epitope prediction.

Currently, most of successful methods for epitope prediction used machine learning techniques. In this dissertation, a novel epitope prediction method named EpicCapo^{+REF} was developed. Nonapeptides, peptides with nine amino acids, were encoded numerically using our peptide-encoding scheme and then input to the support vector machine (SVM). This scheme utilized the information of amino acid pairwise contact potentials (referred to as AAPPs throughout this dissertation) and peptide-MHC (pMHC) contact sites. The predictive performance of EpicCapo^{+REF} outperformed other state-of-the-art methods in many datasets. Furthermore, EpicCapo^{+REF} was applied to identify candidates of promiscuous epitopes from influenza viruses. Many predicted candidates were consistent with previous immunological experiments.

Additionally, we develop a new T-cell reactivity prediction method named PAAQD since recent studies shown unreliable results of epitope prediction methods. In PAAQD, nonapeptides were encoded numerically, using the combined information of AAPPs and quantum topological molecular similarity (QTMS) descriptors and then input to the random forest. We found that PAAQD provided high predictive performance and stability.

We speculate that EpicCapo^{+REF} and PAAQD may be useful in the development of new vaccines.

Chapter 1 Introduction

1.1 Human immune system

The immune system is mechanisms of biological components that work together to defend an organism from “foreign” invaders. All living organisms possess such mechanisms and the human immune system is the most sophisticate one. The human immune system is able to detect various pathogens such as bacteria, fungi, viruses, and other infectious agents. This system consists of numerous types of cells and proteins, each of which has a specific function in the defense system.

There are two major subdivisions of the immune system: the innate immune system and the adaptive immune system. In humans, the immune system is layered lines of defense. The first line of defense is the innate immune system which includes physical barriers such as skin, various types of white blood cells, and proteins. If pathogens successfully breach the innate immune system, they will engage with the second line of defense, the adaptive immune system. Responses of the innate immune system are immediate whereas responses of the adaptive immune system are slower. However, responses of the adaptive immune system are more specific and superior. This system also provides the immunological memory. This memory allows the adaptive immune system to act faster and more effective when the memorized pathogen is encountered [1]. Although these two lines of defense function differently, there are interactions between these systems. For examples, some components of the innate immune system can activate or support the adaptive immune system and vice versa.

The adaptive immune system comprises of lymphocytes which are a specific type of white blood cells. Similar to leukocytes, lymphocytes can freely move around our body via the blood and lymph system. The major lymphocytes in the adaptive immune system are T and B cells which are produced by stem cells in the bone marrow [2]. There are two subtypes of T cells: cytotoxic T-lymphocyte (CTL) and helper T-lymphocyte (Th).

1.2 Vaccines and immune system

The adaptive or acquired immune system is the main target for the vaccine development since long-term protection can be established. Vaccines are agents that stimulate the protective immunity against pathogens and the diseases they cause. This protective immunity is an established immunogenic memory ready for the future encounter with the infectious pathogen. The term vaccine derives from Edward Jenner in 1796 when cowpox was inoculated into humans resulting in protection against smallpox. The word “vacca” means cow in Latin [3].

According to T and B cells, only a part of the pathogen is used in the activation of the immune response. This part is called antigen and it is a large molecule, usually. An antigen introduces several surface and molecular features that are the sites of interactions with CTLs, Th cells, B cells, and antibodies. Each feature defines as an epitope. Epitopes can be used to create new vaccines instead of using the entire cell of a pathogen. These vaccines can be designed to specifically activate responses of CTL, Th, or B cells. In this dissertation, we are focused on CTL epitopes.

The vaccine development is very essential for mankind. From many past decades until now, vaccination saves countless life around the world and prevents suffering from diseases and permanent disabilities. Therefore, the vaccine development is necessary and should be concerned by the governments as the top priority in the public health plans.

1.3 Applications of machine learning in CTL epitope prediction

When cells are infected by a pathogen, epitopes are transported to the cell membrane and presented to T cells via major histocompatibility complex molecules (MHCs). MHCs are classified into three main subclasses: class I, II, and III. MHC genes are highly polymorphic and have many variants. MHC class I (MHC-I) found on all nucleated cells. MHC-I presents epitopes to CTLs. MHC class II (MHC-II) presents epitopes to Th cells and normally found on macrophages, B cells, and dendritic cells. In humans, MHC is referred to as human leukocyte antigen (HLA) [4].

Generally, epitope is a small peptide consists of 8-12 amino acids for MHC-I and 15-24 amino acids for MHC-II. The complexes of peptide-MHC (pMHC) are shown in Figure 1.1. Binding clefts of MHC-I and II consist of two α-helices and one β-sheet, but both terminals of the MHC-I cleft are closed whereas those of the MHC-II are open. Since the groove is closed, the length of epitopes is rather fixed for MHC-I. In contrast, the length of epitopes bond with MHC-II is varying because of the opened groove [5].

To develop CTL vaccines, known epitopes are required. The identification of epitope is a non-trivial task since it is possible that a large number of surface and molecular features are presented on an antigen. The intensive physicochemical experiments are required to identify epitopes. However, such approach is time-consuming and laborious. Therefore, machine learning techniques have been applied to search for epitopes [6].

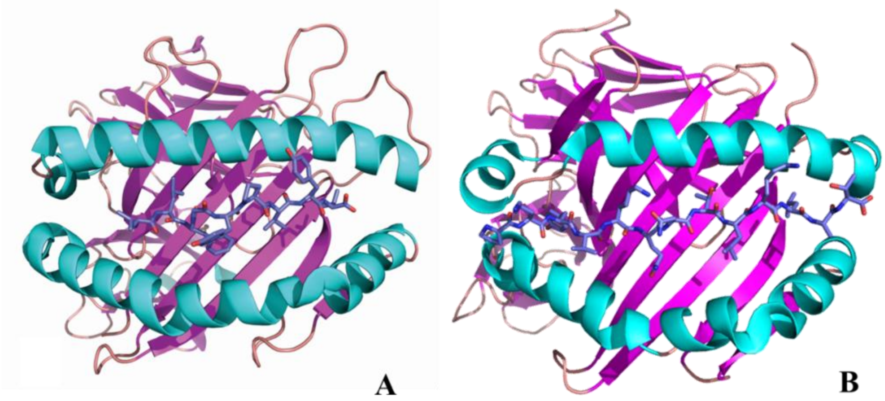

Figure 1.1 Visualization of pMHC complexes.

(A) MHC-I (PDB entry 1DUZ [7]). (B) MHC-II (PDB entry 1DLH [8]).

In this study, we focus on MHC-I on humans that is HLA-I. The presentation of epitopes on HLA-I mainly targets to stimulate CTLs responses. There are three subdivisions of HLA-I: HLA-A, HLA-B, and HLA-C. Most of early epitope binding prediction methods concentrated on the HLA-A*02:01 allele because it is the most frequent allele of the A2 supertype in the Northeast Asian and Caucasian populations [9]. In addition, peptides composed of 9 amino acids known as nonapeptides have been popularly studied.

Currently, most of successful methods for epitope prediction used machine learning algorithms. Examples of these methods are NetMHC [5], NetMHCpan [10], NetCTL [11], NetCTLpan [12], and SVRMHC [13]. The use of machine learning techniques usually requires a large number of training data. In case of epitope prediction, a large number of training peptides is recommended. Therefore, specific databases are needed. The most important database is the Immune Epitope Database (IEDB) [14] which is the largest one. In addition, there are also other available databases such as SYFPEITHI [15], FIMM [16], MHCPEP [17], MHCBN [18], and AntiJen [19].

Machine learning-based epitope prediction techniques significantly accelerate the process of the vaccine development. However, the effectiveness of these techniques depends on the amount of experimental data used for training. In some rare HLA alleles, there are only small numbers of experimented epitopes available. Therefore, the increase in experimental data will improve the accuracy of epitope prediction [20].

1.4 Epitope prediction versus T-cell reactivity prediction

Recent experiments show that predicted epitopes by epitope prediction methods are not always activate T-cell responses [21]. In addition, other biological factors were more strongly correlated to T-cell responses than MHC binding affinities [22]. Therefore, immunogenicity of peptides cannot be accurately inferred from the result of epitope prediction.

T-cell reactivity prediction is more sophisticate than epitope prediction since many biological factors are needed to be concerned. This complication is difficult to be learned by machine learning approaches [23–25]. The first published method for T-cell reactivity prediction is POPI [26]. POPI used physicochemical properties from the AAindex database [27] to encode peptides into numerical vectors. These vectors are then input to the support vector machine (SVM). Afterwards, POPISK [25] was developed. POPISK simply used the SVM with the string kernels. In this dissertation, besides developed new epitope prediction method, we also proposed new T-cell reactivity prediction method.

1.5 Objectives

The main objectives of this dissertation are as follows:

(1) To develop a novel epitope prediction method

(2) To develop a new T-cell reactivity prediction method

1.6 Contribution

According to the above objectives, the main contributions of this thesis are summarized as follows:

(1) A new epitope prediction method which we called EpicCapo and its variants, EpicCapo^{+} and EpicCapo^{+REF} were developed. EpicCapo^{+REF} achieved high performance and outperformed other methods in many datasets of HLA alleles. In some datasets, although there are small numbers of training peptides, EpicCapo^{+REF} still provided the high performance. Therefore, this method is a promising tool for the development of new vaccines.

(2) A new T-cell reactivity prediction method which we called PAAQD was developed. The performance of PAAQD is at least comparable with the previous high performance T-cell reactivity prediction method. In addition, our method shows high predictive stability when tested with the blinded dataset.

Chapter 2 Review of machine learning in immunoinformatics

2.1 The major usages of machine learning algorithms in immunoinformatics

The immune system is composed of many networks of interacting molecules. To understand complicated mechanisms in the immune system, immunologists have been using high throughput experimental techniques. By the use of these techniques, large amount of data was generated. The development of new computational techniques is required for collecting and analyzing these data. This has given rise to a new field called immunoinformatics. Immunoinformatics is one branch of bioinformatics that focused on *in silico* analysis and modeling of immunological data and problems [28, 29].

Most immunoinformatics researches are related to prediction of potential B- and T-cell epitopes. The most successful B- and T-cell epitope prediction methods applied machine learning algorithms. Hereby, the main streams of these researches are categorized as follows.

2.1.1 Artificial neural network

The artificial neural networks (ANNs) are mathematical models inspired by biological neural networks. ANNs are capable of finding relationships and describing nonlinear data [30]. The examples of T-cell epitope prediction method that used ANN are NetMHC [5], NetMHCpan [10], NetCTL [11], and NetCTLpan [12].

2.1.2 Support vector machine

The support vector machine (SVM) is a supervised learning method that has been used for data analysis and pattern recognition. The SVM was first developed by Vapnik [31]. The SVM is described as a non-probabilistic binary classifier and belongs to the group of the kernel-based approaches [32]. The examples of epitope prediction method that used ANN are SVRMHC [13], TAPPred [33], Pcleavage [34], and COBEpro [35].

2.1.3 Hidden Markov models

The hidden Markov models (HMMs) were described by Baum et al. [36]. The examples of epitope prediction method that used HMM are PredTAP [37].

2.2 Immunoinformatics databases

Nowadays, there are many immunoinformatics databases. Most of them are related to T- or B-cell epitopes. Each database has specific features and purposes. Some databases include 3D structures of MHC molecules or peptides and also provide epitope prediction tools. Table 2-1 describes available immunoinformatics databases.

Table 2-1 The description of the datasets

Type	**Name**	**URL**	**Ref.**
T-cell epitopes	JenPep	http://www.darrenflower.info/jenpep/	[38]
	SYFPEITHI	http://www.syfpeithi.de	[15]
	FRED	http://www-bs.informatik.uni-tuebingen.de/Software/FRED	[39]
	MHCBN	http://www.imtech.res.in/raghava/mhcbn/	[18]
B-cell epitopes	CED	http://immunet.cn/ced/	[40]
	Bcipep	http://www.imtech.res.in/raghava/bcipep	[41]
	Epitome	http://cubic.bioc.columbia.edu/services/epitome/	[42]
Both T- and B- cell epitopes	IEDB	http://www.iedb.org/	[14]
	IMGT	http://www.imgt.org/	[43]
	MHCPEP	http://wehih.wehi.edu.au/mhcpep/	[17]
	AntiJen	http://www.ddg-pharmfac.net/antijen/AntiJen/antijenhomepage.htm	[19]
Allergen	Database of IUIS	http://www.allergen.org	[44]
	Allergen Pro	http://www.niab.go.kr/nabic/	[45]
	SDAP	http://fermi.utmb.edu/SDAP/	[46]
Information related to molecular evolution of immune system components	ImmTree	http://bioinf.uta.fi/ImmTree	[47]
	Immunome database	http://bioinf.uta.fi/Immunome/	[48]
	ImmunomeBase	http://bioinf.uta.fi/ImmunomeBase	[49]
	Immunome Knowledge Base	http://bioinf.uta.fi/IKB/	[50]

Chapter 3 EpicCapo

3.1 Introduction

In the last decade, there are outbreaks appeared in the human populations such as SARS in 2003, avian flu (H5N1) in 2006, and swine flu (H1N1) in 2009. These outbreaks cause high mortality around the world. To prevent human populations from future outbreaks, the novel vaccine development is necessary. From the reviews in the chapter 1, epitope prediction methods are used to accelerate the process of the vaccine development. In this chapter, we would like to introduce our novel epitope prediction method named EpicCapo and its variants, EpicCapo^{+} and EpicCapo$^{+\mathrm{REF}}$.

3.2 Methods

3.2.1 Peptide data encoding

We propose a novel peptide-encoding scheme for machine learning algorithms. This scheme utilized the information of pMHC contact sites retrieved from the international ImMunoGeneTics information system, IMGT [43], the allele-specific positional scoring matrices developed by SMM$^{\mathrm{PMBEC}}$ [51], and the amino acid pairwise contact potentials (AAPPs) from AAindex [27]. We define a score $S_{k,i}{}^{(n)}$ for the i^{th} amino acid of the nonapeptide n under a k^{th} type of AAPP as follows:

$$S_{k,i}^{(n)} = T_i\left(u_i^{(n)}\right) \cdot \left(\sum_{j=1}^{L} \delta_{ij} E_k\left(u_i^{(n)}, v_j\right) \middle/ \sum_{j=1}^{L} \delta_{ij} \right),$$

Here, we denote the i^{th} amino acid of the nonapeptide n and the j^{th} amino acid of HLA by $u_i{}^{(n)}$ and v_j, respectively. L is the length of the HLA protein, $T_i(a)$ is the i^{th} position score of the amino acid a for the nonapeptides described by SMM$^{\mathrm{PMBEC}}$, and δ_{ij} is an indicator variable that takes the value of 1 if the i^{th} amino acid of a nonapeptide and the j^{th} amino acid of HLA contact each other, and 0 otherwise. The positional scoring matrix $T_i(a)$ is trained based on training data and multiplied by -1 to reverse the order of values (a high positive value denotes high preference between an amino acid and the position) and scaled into the range of 1 to 10 since we need to avoid loss of information when $T_i(a)$ equals zero. In fact, any range that does not include zero can be used; in this study, it is the range of 1 to 10. Intuitively, this score represents average pair-potential of contact sites, weighted by the position-specific amino acid score for nonapeptides. Let K be the number of AAPPs available, and M be the length of the peptide, set to 9 throughout this study. Using this scoring scheme, we transform a nonapeptide n into a $M \times K$-dimensional numerical vector, whose $(M(k\text{–}1) + i)^{\mathrm{th}}$ element is $S_{k,i}{}^{(n)}$. For example, the encoded nonapeptides consist of 9 features if one AAPP is used and 360 features if 40 AAPPs are used. Figure 3.1 illustrates an example of the data-encoding scheme for the first position of the nonapeptide.

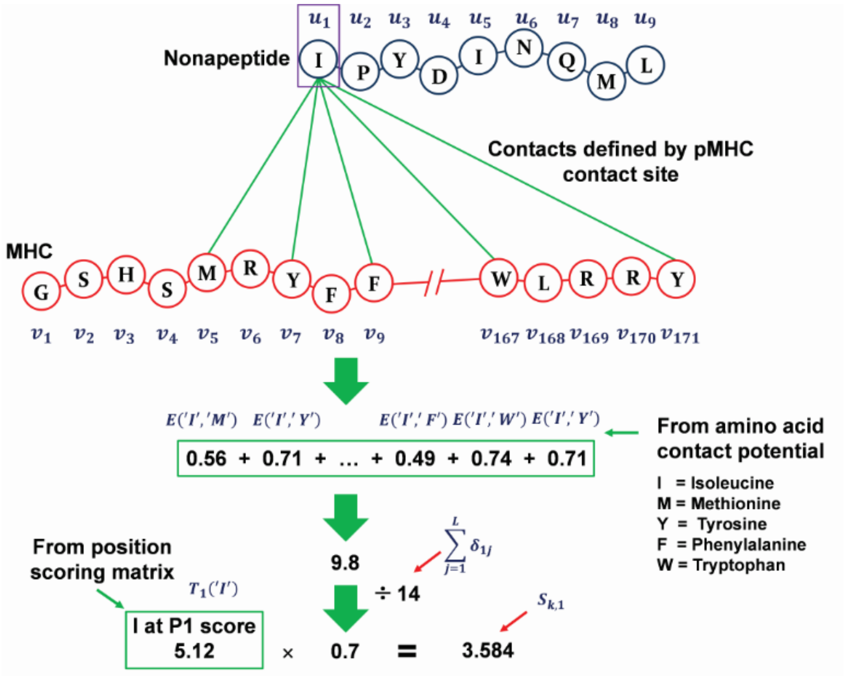
u_1 u_2 u_3 u_4 u_5 u_6 u_7 u_8 u_9
Nonapeptide
I P Y D I N Q M L
Contacts defined by pMHC contact site
MHC
G S H S M R Y F F W L R R Y
v_1 v_2 v_3 v_4 v_5 v_6 v_7 v_8 v_9 v_{167} v_{168} v_{169} v_{170} v_{171}
$E('I','M')$ $E('I','Y')$ $E('I','F')$ $E('I','W')$ $E('I','Y')$
0.56 + 0.71 + … + 0.49 + 0.74 + 0.71
From amino acid contact potential
I = Isoleucine
M = Methionine
Y = Tyrosine
F = Phenylalanine
W = Tryptophan
9.8
$\sum_{j=1}^{L} \delta_{1j}$
÷ 14
From position scoring matrix
$T_1('I')$
I at P1 score 5.12
× 0.7 = 3.584
$S_{k,1}$

Figure 3.1 Our peptide data-encoding scheme, using the first position of a nonapeptide as an example.

3.2.2 Benchmark datasets

We used the benchmark datasets of 34 MHC-I alleles provided by Peters et al. [52]. After encoding these datasets using our scheme, encoded data were then input to the SVM implemented in the R package kernlab [53]. The performances were evaluated in classification tasks, using a 5-fold cross validation. The predictive performance is evaluated using area under receiver operating characteristic curve (AUC). We compared the results of our method with those of ARB, NetMHC, SMM, and SMM$^{\text{PMBEC}}$. We named our method EpicCapo which is the combination of the encoding scheme with the SVM.

3.2.3 EpicCapo^{+} and EpicCapo^{+REF}

EpicCapo was further developed as EpicCapo^{+} by selecting important AAPPs. After that, EpicCapo^{+} was improved to EpicCapo$^{+\text{REF}}$ by employing Relief-F algorithm to remove irrelevant features.

3.2.4 Identification of candidates of promiscuous epitopes

EpicCapo$^{+\text{REF}}$ was further tested to identify candidates of promiscuous epitopes—i.e., nonapeptides that were predicted to be MHC binders for various HLA alleles—from the protein sequences of four influenza A viral subtypes: H1N1 (A/PR/8/34), H3N2 (A/Aichi/2/68), H1N1 (A/New York/4290/2009), and H5N1 (A/Hong Kong/483/97). The identified epitopes were validated by cross-checking with the results of immunological experiments.

3.3 Results and discussion

3.3.1 Classification results of benchmark datasets

As seen in Table 3-1, the performance of EpicCapo^{+} was higher than EpicCapo and comparable with NetMHC. The overall performance of EpicCapo^{+} is significantly higher than that of other methods according to a paired *t*-test (two-tailed) comparison of average AUCs from all alleles.

Table 3-1 Classification results of 34 allele datasets.

MHC	**# of peptides**	**AUC**					
		ARB	**SMM**	**SMM^{PMBEC}**	**NetMHC**	**EpicCapo**	**EpicCapo^{+}**
HLA-A*01:01	1157	0.964	0.980	0.977	0.982	0.972 ± 0.004	0.977 ± 0.003
HLA-A*02:01	3089	0.934	0.952	0.946	0.957	0.950 ± 0.004	0.951 ± 0.004
HLA-A*02:02	1447	0.875	0.899	0.899	0.900	0.901 ± 0.004	**0.909** ± 0.004
HLA-A*02:03	1443	0.884	0.916	0.916	0.921	0.920 ± 0.003	0.923 ± 0.003
HLA-A*02:06	1437	0.872	0.914	0.916	0.927	0.925 ± 0.004	0.927 ± 0.004
HLA-A*03:01	2094	0.908	0.940	0.928	0.937	0.934 ± 0.004	0.938 ± 0.003
HLA-A*11:01	1985	0.918	0.948	0.939	0.951	0.945 ± 0.004	0.951 ± 0.002
HLA-A*24:02	197	0.718	0.780	0.801	0.825	**0.853** ± 0.012	**0.865** ± 0.011
HLA-A*26:01	672	0.907	0.931	0.924	0.956	0.941 ± 0.005	0.957 ± 0.007
HLA-A*29:02	160	0.755	0.911	0.916	0.935	**0.944** ± 0.008	**0.945** ± 0.010
HLA-A*31:01	1869	0.909	0.930	0.925	0.928	0.930 ± 0.002	**0.935** ± 0.003
HLA-A*33:01	1140	0.892	0.925	0.925	0.915	0.926 ± 0.004	**0.934** ± 0.004
HLA-A*68:01	1141	0.840	0.885	0.885	0.883	**0.891** ± 0.003	**0.899** ± 0.003
HLA-A*68:02	1434	0.865	0.898	0.889	0.899	0.901 ± 0.005	0.907 ± 0.003
HLA-B*07:02	1262	0.952	0.964	0.960	0.965	0.960 ± 0.004	0.964 ± 0.002
HLA-B*08:01	708	0.936	0.943	0.956	0.955	0.942 ± 0.005	0.951 ± 0.004
HLA-B*15:01	978	0.900	0.952	0.940	0.941	0.940 ± 0.006	0.950 ± 0.005
HLA-B*18:01	118	0.573	0.853	0.880	0.838	0.886 ± 0.013	**0.911** ± 0.009
HLA-B*27:05	969	0.915	0.940	0.941	0.938	**0.949** ± 0.005	**0.958** ± 0.003
HLA-B*35:01	736	0.851	0.889	0.889	0.875	0.900 ± 0.004	**0.907** ± 0.007
HLA-B*40:02	118	0.541	0.842	0.843	0.754	0.811 ± 0.007	**0.912** ± 0.011
HLA-B*44:02	119	0.533	0.740	0.739	0.778	**0.798** ± 0.009	**0.861** ± 0.013
HLA-B*44:03	119	0.461	0.770	0.753	0.763	**0.813** ± 0.010	**0.871** ± 0.008
HLA-B*51:01	244	0.822	0.868	0.895	0.886	**0.930** ± 0.012	**0.948** ± 0.015
HLA-B*53:01	254	0.871	0.882	0.885	0.899	**0.916** ± 0.008	**0.940** ± 0.008
HLA-B*54:01	255	0.847	0.921	0.935	0.903	0.927 ± 0.008	0.938 ± 0.006
HLA-B*57:01	59	0.428	0.871	0.843	0.826	0.792 ± 0.009	0.854 ± 0.010
HLA-B*58:01	988	0.889	0.964	0.945	0.961	0.959 ± 0.005	0.964 ± 0.004
H-2 Db	303	0.865	0.912	0.901	0.933	**0.940** ± 0.014	**0.968** ± 0.006
H-2 Dd	85	0.696	0.853	0.837	0.925	**0.956** ± 0.016	**0.985** ± 0.017
H-2 Kb	223	0.792	0.810	0.833	0.850	0.844 ± 0.021	**0.880** ± 0.017
H-2 Kd	176	0.798	0.936	0.931	0.939	**0.950** ± 0.015	**0.966** ± 0.009
H-2 Kk	164	0.758	0.770	0.793	0.790	**0.883** ± 0.009	**0.926** ± 0.008
H-2 Ld	102	0.551	0.924	0.942	0.977	**0.984** ± 0.012	**0.992** ± 0.013
Average		0.801	0.895	0.895	0.900	0.912	0.931
t-test|ARB		NA	4.37E-5	3.69E-5	1.25E-5	5.21E-6	2.64E-6
t-test|SMM			NA	8.61E-1	2.30E-1	8.28E-3	2.87E-5
t-test|SMM^{PMBEC}				NA	2.61E-1	3.50E-3	8.49E-6
t-test|NetMHC					NA	8.57E-3	7.74E-5
t-test|EpicCapo						NA	1.95E-5

For each dataset, AUCs were evaluated based on 5-fold cross validation. In the lower part, p-values of average AUCs were calculated using paired *t*-tests (two-tailed).

Means and standard deviations were calculated by 20 iterations of 5-fold cross validation for EpicCapo and EpicCapo^{+}.

Underlined values represent the highest performance among ARB, SMM, SMM^{PMBEC}, and NetMHC.

Values in bold represent significant improvements of EpicCapo or EpicCapo^{+} AUCs from 20 iterations of 5-fold cross validation over the underlined values according to *t*-tests (one-tailed, significance level = 0.01).

3.3.2 Features selected by EpicCapo^{+REF}

The performance and number of features selected by EpicCapo^{+REF} are shown in Table 3-2. The overall performance of EpicCapo^{+REF} is higher than other methods.

Table 3-2 Number of selected features by EpicCapo^{+REF} using 14 HLA-A allele datasets.

Allele	**AUC of EpicCapo^{+REF}**	**# of features selected**
A*01:01	0.980	72
A*02:01	0.958	62
A*02:02	0.913	18
A*02:03	0.925	104
A*02:06	0.926	141
A*03:01	0.946	58
A*11:01	0.956	35
A*24:02	0.877	31
A*26:01	0.960	18
A*29:02	0.955	23
A*31:01	0.940	46
A*33:01	0.940	17
A*68:01	0.904	40
A*68:02	0.913	79
Average	0.935	

3.3.3 Candidates of promiscuous epitopes for the development of influenza A viral vaccines

The total number of promiscuous epitopes predicted by EpicCapo^{+REF} is 76. 51 peptides (67.1%) were immunologically validated as positive, whereas 9 peptides (11.8%) were validated as negative. No evidence of immunological validation could be obtained for 16 peptides (21.1%). These results indicate that our newly developed method provides a markedly high accuracy in epitope identification, given the fact that most of the identified epitopes could be correlated with immunological evidence. However, even without such evidence, those epitopes identified by our computational approach might be considered as candidates for the new vaccine development.

3.4 Conclusions

In this chapter, we have developed a novel method for epitope prediction. Our method achieved high performance in testing with the benchmark datasets. In addition, our study identified a number of candidates of promiscuous CTL epitopes from four influenza A viral strains, consistent with previously reported immunological experiments. This consistency in results strongly supports the accuracy of our method. We speculate that our techniques may be useful in the development of new vaccines.

Chapter 4 PAAQD

4.1 Introduction

Recent studies revealed that predicted peptides with high binding affinity to MHC-I molecules did not always result in T-cell responses [21, 54]. In addition, other factors were more strongly correlated to T-cell responses than MHC binding affinities [22]. Therefore, immunogenicity could not be accurately determined by existing epitope prediction methods.

In this chapter, we would like to introduce our novel T-cell reactivity predictor named PAAQD.

4.2 Methods

4.2.1 Datasets

Two datasets were used in this study. The first dataset is called the IMMA2 dataset, collected by Tung et al. (2011) [25]. The second dataset was collected from IEDB database [14] by selecting nonapeptides that were specific to the HLA-A2 supertype. All of these nonapeptides are not included in the IMMA2 dataset. We called the latter dataset as the validation dataset. The sequence preference of the validation dataset is different from the IMMA2 dataset.

4.2.2 Peptide encoding

As mentioned in the chapter 3, nonapeptides were encoded numerically using our peptide encoding scheme before input to the classifier. However, in this chapter, the allele-specific positional scoring matrices were not included in the encoding scheme. Therefore, we define a score $S_{k,i}^{(n)}$ for the i^{th} amino acid of the nonapeptide n under a k^{th} type of AAPPs as follows:

$$S_{k,i}^{(n)} = \sum_{j=1}^{L} \delta_{ij} \mathrm{E}_k\left(u_i^{(n)}, v_j\right) \Bigg/ \sum_{j=1}^{L} \delta_{ij}$$

Each encoded peptide was combined with the corresponding feature vector constructed by using QTMS descriptors [55]. There are four types of QTMS descriptors used in this study (see Table 4-1).

Table 4-1 QTMS descriptors used in this study.

Descriptor	Description	# of vector
CBFQ	Common bonds factor analysis of QTMS	6
CDFQ	Common bonds descriptor-based factor analysis of QTMS	3
CUFQ	Common bonds unfolded-data-based factor analysis of QTMS	5
ADFQ	All bonds descriptor-based factor analysis of QTMS descriptors	7

4.2.3 Prediction of peptide immunogenicity using the IMMA 2 dataset

The proposed peptide-encoding scheme was applied to the IMMA2 dataset and input to the random forest implemented in Weka [56]. The number of trees generated and the number of features randomly sampled as candidates at each split were set to 200 and 10, respectively. The predictive performance is evaluated using three measures; AUC, overall accuracy (ACC), and Matthew’s correlation coefficient (MCC). We compared our method with POPI [26] and POPISK [25].

4.2.4 Prediction of peptide immunogenicity using the validation dataset

The final model for T-cell reactivity prediction was constructed based on the IMMA 2 dataset. This model was used to predict immunogenicity of peptides in the validation dataset. The evaluated performance indicates the predictive stability when peptides with different sequence preferences from training data were input to the model. The PAAQD performance was compared with POPISK.

4.3 Results and discussion

4.3.1 The predictive performance of PAAQD on the IMMA 2 dataset

Figure 4.1 shows the performance of the concerned methods based on the IMMA 2 dataset. This result indicates that PAAQD outperformed POPI-modified and provided comparable performance with POPISK.

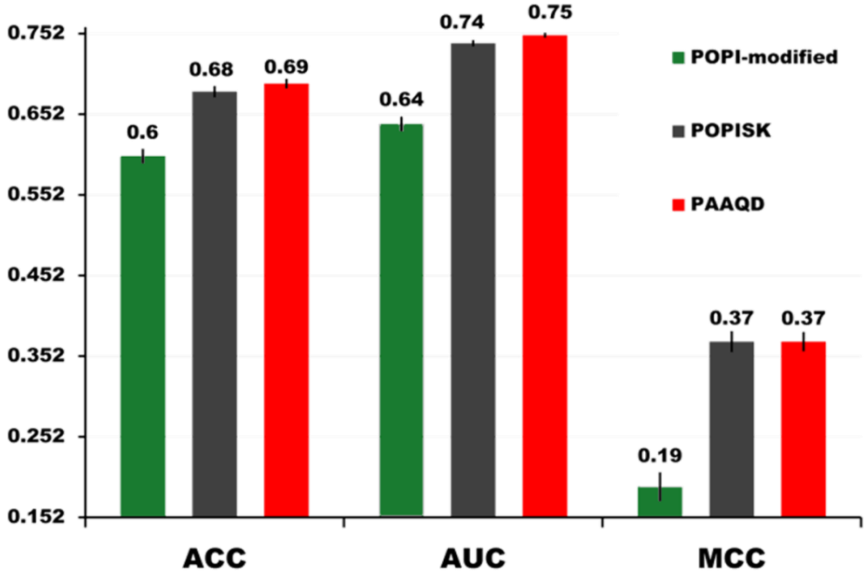

Figure 4.1 Comparison of 20 independent iterations of the 10-fold cross validation performance of POPI, POPISK, and PAAQD.

4.3.2 Result of peptide immunogenicity prediction using the validation dataset

The result of peptide immunogenicity prediction using validation dataset is shown in Figure 4.2. ACC and MCC of PAAQD were 0.72 and 0.37, respectively. ACC and MCC of POPISK were 0.68 and 0.28, respectively. PAAQD significantly outperformed POPISK 4% and 9% in ACC and MCC, respectively. This result indicated that PAAQD outperformed POPISK. Consequently, PAAQD provided more predictive stability than POPISK when using the test data with sequence preferences different from the training data.

We examined the over- and underrepresented amino acids in corresponding positions of the IMMA 2 dataset and the validation dataset using the two-sample logos [57]. In the two-sample logos, differences among amino acids were statistically significant with the level of 0.01 when using the two-sample *t*-test. The two-sample logos of the IMMA 2 and validation datasets are shown in Figures 4.3 and 4.4, respectively.

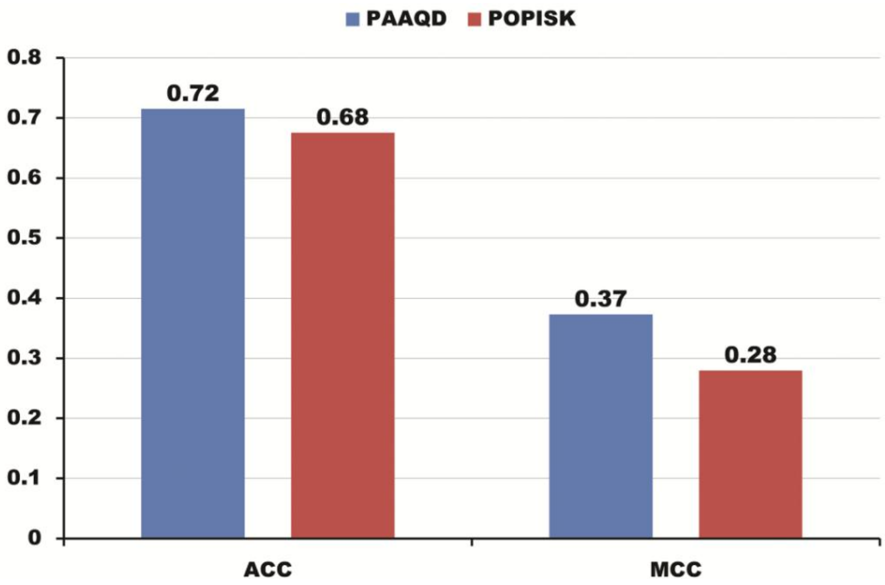

Figure 4.2 The result of peptide immunogenicity prediction evaluated on the validation dataset.

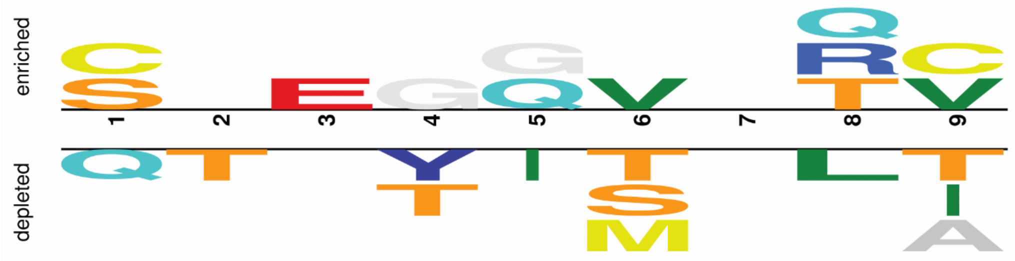

Figure 4.3 Two-sample logo that represents over- and underrepresented amino acids in the IMMA 2 dataset.

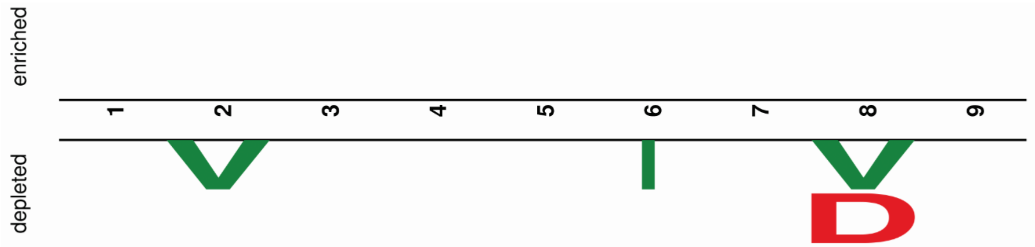

Figure 4.4 Two-sample logo that represents over- and underrepresented amino acids in the validation dataset.

4.4 Conclusion

We developed a novel method for T-cell reactivity prediction which we call PAAQD. PAAQD achieved the comparable performance with POPISK which is a high-performance T-cell reactivity predictor when testing with the IMMA 2 dataset. Additionally, PAAQD outperformed POPISK when testing with the validation dataset. This indicated that PAAQD provided more predictive stability when peptides with different sequence preferences from training data were input to the model. We speculate that PAAQD may be useful in identifying immunogenic peptides for the development of new vaccines.

Chapter 5 Conclusion and future research

5.1 Dissertation summary

Epitope prediction methods are important tools that speed up the process of the vaccine development. In this dissertation, a new epitope prediction method named EpicCapo and its variants, EpicCapo^{+} and EpicCapo^{+REF} were developed. The performance of EpicCapo^{+REF} outperformed other state of the art methods in many datasets.

According to recent studies, results of epitope prediction methods are not always reliable. Therefore, we developed a new T-cell reactivity prediction method named PAAQD. The performance of PAAQD is at least comparable with the previous high performance T-cell reactivity prediction method. However, our method shows higher predictive stability when tested with the blinded dataset.

We hope that EpicCapo^{+REF} and PAAQD may be useful in the development of new vaccines.

5.2 Future works

As we have shown before, our methods for epitope and T-cell reactivity prediction are very promising for the new vaccine development. However, an input peptide must be a nonapeptide which is a peptide composed of 9 amino acids. In the future, we will develop the length independent prediction method by using other algorithms such as string kernel in the SVM and hidden Markov model. In addition, since we developed the peptide encoding schemes for both epitope and T-cell reactivity prediction. In the upcoming works, we will apply these schemes in other studies such as protein-ligand binding, protein-protein interaction (PPI) prediction, and drug discovery.

Bibliography

1. Abbas AK, Lichtman AHH, Pillai S: *Cellular and Molecular Immunology: with STUDENT CONSULT Online Access*. Saunders; 2011.

2. Janeway CA, Travers P, Walport M, Shlomchik MJ: **Immunobiology: the immune system in health and disease**. 2005.

3. Flower DR, Macdonald IK, Ramakrishnan K, Davies MN, Doytchinova IA: **Computer aided selection of candidate vaccine antigens**. *Immunome Res* 2010, **6 Suppl 2**:S1.

4. Abbas AK, Lichtman AHH: *Basic Immunology updated edition: Functions and disorders of the immune system*. Saunders; 2010.

5. Lundegaard C, Lamberth K, Harndahl M, Buus S, Lund O, Nielsen M: **NetMHC-3.0: accurate web accessible predictions of human, mouse and monkey MHC class I affinities for peptides of length 8-11**. *Nucleic Acids Res* 2008, **36**:W509–12.

6. Lundegaard C, Hoof I, Lund O, Nielsen M: **State of the art and challenges in sequence based T-cell epitope prediction**. *Immunome Res* 2010, **6 Suppl 2**:S3.

7. Khan AR, Baker BM, Ghosh P, Biddison WE, Wiley DC: **The structure and stability of an HLA-A*0201/octameric tax peptide complex with an empty conserved peptide-N-terminal binding site**. *J Immunol* 2000, **164**:6398–6405.

8. Stern LJ, Brown JH, Jardetzky TS, Gorga JC, Urban RG, Strominger JL, Wiley DC: **Crystal structure of the human class II MHC protein HLA-DR1 complexed with an influenza virus peptide**. *Nature* 1994, **368**:215–221.

9. Liang B, Zhu L, Liang Z, Weng X, Lu X, Zhang C, Li H, Wu X: **A simplified PCR-SSP method for HLA-A2 subtype in a population of Wuhan, China**. *Cell Mol Immunol* 2006, **3**:453–458.

10. Nielsen M, Lundegaard C, Blicher T, Lamberth K, Harndahl M, Justesen S, Roder G, Peters B, Sette A, Lund O, Buus S: **NetMHCpan, a method for quantitative predictions of peptide binding to any HLA-A and -B locus protein of known sequence**. *PLoS One* 2007, **2**:e796.

11. Larsen M V, Lundegaard C, Lamberth K, Buus S, Lund O, Nielsen M: **Large-scale validation of methods for cytotoxic T-lymphocyte epitope prediction**. *BMC Bioinformatics* 2007, **8**:424.

12. Stranzl T, Larsen M V, Lundegaard C, Nielsen M: **NetCTLpan: pan-specific MHC class I pathway epitope predictions**. *Immunogenetics* 2010, **62**:357–368.

13. Wan J, Liu W, Xu Q, Ren Y, Flower DR, Li T: **SVRMHC prediction server for MHC-binding peptides**. *BMC Bioinformatics* 2006, **7**:463.

14. Peters B, Sidney J, Bourne P, Bui HH, Buus S, Doh G, Fleri W, Kronenberg M, Kubo R, Lund O, Nemazee D, Ponomarenko J V, Sathiamurthy M, Schoenberger S, Stewart S, Surko P, Way S, Wilson S, Sette A: **The immune epitope database and analysis resource: from vision to blueprint**. *PLoS Biol* 2005, **3**:e91.

15. Rammensee H, Bachmann J, Emmerich NP, Bachor OA, Stevanovic S: **SYFPEITHI: database for MHC ligands and peptide motifs**. *Immunogenetics* 1999, **50**:213–219.

16. Schonbach C, Koh JL, Flower DR, Wong L, Brusic V: **FIMM, a database of functional molecular immunology: update 2002**. *Nucleic Acids Res* 2002, **30**:226–229.

17. Brusic V, Rudy G, Harrison LC: **MHCPEP, a database of MHC-binding peptides: update 1997**. *Nucleic Acids Res* 1998, **26**:368–371.

18. Lata S, Bhasin M, Raghava GP: **MHCBN 4.0: A database of MHC/TAP binding peptides and T-cell epitopes**. *BMC Res Notes* 2009, **2**:61.

19. Toseland CP, Clayton DJ, McSparron H, Hemsley SL, Blythe MJ, Paine K, Doytchinova IA, Guan P, Hattotuwagama CK, Flower DR: **AntiJen: a quantitative immunology database integrating functional, thermodynamic, kinetic, biophysical, and cellular data**. *Immunome Res* 2005, **1**:4.

20. Chen P, Rayner S, Hu KH: **Advances of bioinformatics tools applied in virus epitopes prediction**. *Virol Sin* 2011, **26**:1–7.

21. Wu X, Xu X, Gu R, Wang Z, Chen H, Xu K, Zhang M, Hutton J, Yang T: **Prediction of HLA class I-restricted T-cell epitopes of islet autoantigen combined with binding and dissociation assays**. *Autoimmunity* 2012, **45**:176–185.

22. Tenzer S, Wee E, Burgevin A, Stewart-Jones G, Friis L, Lamberth K, Chang CH, Harndahl M, Weimershaus M, Gerstoft J, Akkad N, Klenerman P, Fugger L, Jones EY, McMichael AJ, Buus S, Schild H, Van Endert P, Iversen AK: **Antigen processing influences HIV-specific cytotoxic T lymphocyte immunodominance**. *Nat Immunol* 2009, **10**:636–646.

23. Van Regenmortel MH: **Antigenicity and immunogenicity of synthetic peptides**. *Biologicals* 2001, **29**:209–213.

24. Kanduc D: **Peptimmunology: immunogenic peptides and sequence redundancy**. *Curr Drug Discov Technol* 2005, **2**:239–244.

25. Tung CW, Ziehm M, Kamper A, Kohlbacher O, Ho SY: **POPISK: T-cell reactivity prediction using support vector machines and string kernels**. *BMC Bioinformatics* 2011, **12**:446.

26. Tung CW, Ho SY: **POPI: predicting immunogenicity of MHC class I binding peptides by mining informative physicochemical properties**. *Bioinformatics* 2007, **23**:942–949.

27. Kawashima S, Pokarowski P, Pokarowska M, Kolinski A, Katayama T, Kanehisa M: **AAindex: amino acid index database, progress report 2008**. *Nucleic Acids Res* 2008, **36**:D202–5.

28. Tomar N, De RK: **Immunoinformatics: an integrated scenario**. *Immunology* 2010, **131**:153–168.

29. Patronov A, Doytchinova I: **T-cell epitope vaccine design by immunoinformatics**. *Open Biol* 2013, **3**:120139.

30. Beale R, Jackson T: *Neural computing: an introduction*. Taylor & Francis; 1990.

31. Vapnik V: **Statistical learning theory. 1998**. 1998.

32. Schölkopf B, Burges CJC: *Advances in kernel methods: support vector learning*. The MIT press; 1999.

33. Bhasin M, Raghava GPS: **Analysis and prediction of affinity of TAP binding peptides using cascade SVM**. *Protein Science* 2004, **13**:596–607.

34. Bhasin M, Raghava GPS: **Pcleavage: an SVM based method for prediction of constitutive proteasome and immunoproteasome cleavage sites in antigenic sequences**. *Nucleic Acids Research* 2005, **33**:W202–W207.

35. Sweredoski MJ, Baldi P: **COBEpro: a novel system for predicting continuous B-cell epitopes**. *Protein Engineering Design & Selection* 2009, **22**:113–120.

36. Baum LE, Petrie T, Soules G, Weiss N: **A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains**. *The annals of mathematical statistics* 1970, **41**:164–171.

37. Zhang GL, Petrovsky N, Kwoh CK, August JT, Brusic V: **PREDTAP: a system for prediction of peptide binding to the human transporter associated with antigen processing**. *Immunome research* 2006, **2**:3.

38. McSparron H, Blythe MJ, Zygouri C, Doytchinova IA, Flower DR: **JenPep: a novel computational information resource for immunobiology and vaccinology**. *J Chem Inf Comput Sci* 2003, **43**:1276–1287.

39. Feldhahn M, Donnes P, Thiel P, Kohlbacher O: **FRED-a framework for T-cell epitope detection**. *Bioinformatics* 2009, **25**:2758–2759.

40. Huang J, Honda W: **CED: a conformational epitope database**. *BMC Immunology* 2006, **7**.

41. Saha S, Bhasin M, Raghava GPS: **Bcipep: A database of B-cell epitopes**. *BMC Genomics* 2005, **6**.

42. Schlessinger A, Ofran Y, Yachdav G, Rost B: **Epitome: database of structure-inferred antigenic epitopes**. *Nucleic Acids Research* 2006, **34**:D777–D780.

43. Kaas Q, Lefranc MP: **T cell receptor/peptide/MHC molecular characterization and standardized pMHC contact sites in IMGT/3Dstructure-DB**. *In Silico Biol* 2005, **5**:505–528.

44. King TP, Hoffman D, Lowenstein H, Marsh DG, Platts-Mills TA, Thomas W: **Allergen nomenclature. WHO/IUIS Allergen Nomenclature Subcommittee**. *Int Arch Allergy Immunol* 1994, **105**:224–233.

45. Kim C, Kwon S, Lee G, Lee H, Choi J, Kim Y, Hahn J: **A database for allergenic proteins and tools for allergenicity prediction**. *Bioinformation* 2009, **3**:344–345.

46. Ivanciuc O, Schein CH, Braun W: **SDAP: database and computational tools for allergenic proteins**. *Nucleic Acids Res* 2003, **31**:359–362.

47. Ortutay C, Siermala M, Vihinen M: **ImmTree: database of evolutionary relationships of genes and proteins in the human immune system**. *Immunome Res* 2007, **3**:4.

48. Ortutay C, Vihinen M: **Immunome: a reference set of genes and proteins for systems biology of the human immune system**. *Cell Immunol* 2006, **244**:87–89.

49. Rannikko K, Ortutay C, Vihinen M: **Immunity genes and their orthologs: a multi-species database**. *International Immunology* 2007, **19**:1361–1370.

50. Ortutay C, Vihinen M: **Immunome Knowledge Base (IKB): An integrated service for immunome research**. *BMC Immunology* 2009, **10**.

51. Kim Y, Sidney J, Pinilla C, Sette A, Peters B: **Derivation of an amino acid similarity matrix for peptide: MHC binding and its application as a Bayesian prior**. *BMC Bioinformatics* 2009, **10**:394.

52. Peters B, Bui HH, Frankild S, Nielson M, Lundegaard C, Kostem E, Basch D, Lamberth K, Harndahl M, Fleri W, Wilson SS, Sidney J, Lund O, Buus S, Sette A: **A community resource benchmarking predictions of peptide binding to MHC-I molecules**. *PLoS Comput Biol* 2006, **2**:e65.

53. Karatzoglou A, Smola A, Hornik K, Zeileis A: **kernlab-an S4 package for kernel methods in R**. 2004.

54. Feltkamp MC, Vierboom MP, Kast WM, Melief CJ: **Efficient MHC class I-peptide binding is required but does not ensure MHC class I-restricted immunogenicity**. *Mol Immunol* 1994, **31**:1391–1401.

55. Hemmateenejad B, Yousefinejad S, Mehdipour AR: **Novel amino acids indices based on quantum topological molecular similarity and their application to QSAR study of peptides**. *Amino Acids* 2011, **40**:1169–1183.

56. Frank E, Hall M, Trigg L, Holmes G, Witten IH: **Data mining in bioinformatics using Weka**. *Bioinformatics* 2004, **20**:2479–2481.

57. Vacic V, Iakoucheva LM, Radivojac P: **Two Sample Logo: a graphical representation of the differences between two sets of sequence alignments**. *Bioinformatics* 2006, **22**:1536–1537.

学位論文審査結果の要旨

平成２５年７月３０日に第１回学位論文審査委員会を開催、８月６日に口頭発表、その後に第２回審査委員会を開催し、慎重審議の結果、以下の通り判定した。なお、口頭発表における質疑を最終試験に代えるものとした。

感染症に対するワクチンを開発する際、抗体が認識する抗原の部分領域（エピトープ）を特定することは極めて重要であり、コンピュータによる高精度な予測手法が期待されている。本研究では、抗原と抗体の相互作用領域を特徴として表現する新しいエンコーディング手法を開発し、特徴選択手法と併せて用いることにより、従来法を上回る精度でエピトープ予測を行うことができた。一方、抗原と抗体が結合したからといって、必ずしも感染した細胞をＴ細胞が破壊するとは限らないことが近年分かってきた。これに対して本研究では、上記のエンコーディング手法を一部修正してこの問題に対応し、ランダムフォレスト学習器と組み合わせることにより、従来法を上回る精度でＴ細胞の細胞傷害活性を予測することができた。

以上の研究成果は、感染症に対するワクチン開発の加速と低コスト化に大きく貢献するものであり、本論文は博士（工学）に値するものと判定した。