Epitope and T-cell Reactivity Prediction Using Machine Learning Approaches

SAETHANG THAMMAKORN

July, 2013

Dissertation

Epitope and T-cell Reactivity Prediction Using Machine Learning Approaches

Graduate School of
Natural Science & Technology
Kanazawa University

Major subject:
Division of Electrical Engineering
and Computer Science

Course:
Intelligent Systems
and Information Mathematics

School registration No.: 1023112110

Name: SAETHANG THAMMAKORN

Chief advisor: Professor KENJI SATOU

Abstract

In the last decade, there are serious outbreaks appeared in the human populations such as SARS in 2003, avian flu (H5N1) in 2006, and swine flu (H1N1) in 2009. Among these, only the H1N1 outbreak was officially declared as pandemic since the virus spread worldwide in a few short weeks. Although the H1N1 virus is far less deadly than the H5N1, it is capable of being transmitted easily from person to person. In addition, because viruses are easily mutated, the next pandemic is unpredictable and can be a tragedy. Therefore, the vaccine development is very important to prevent human populations from future outbreaks.

Epitope identification is a non-trivial step in the vaccine development, since epitopes play an important role in the activation of the immune response. Epitopes are conventionally identified by synthesizing a large number of peptides and then conducting immunological experiments. However, these processes are time-consuming and laborious. The computational methods can be used to accelerate the process of the vaccine development by performing epitope prediction. The most successful approach for epitope prediction is the applications of machine learning techniques. Many methods were proposed but most of them tend to overlook the interpretability which respects to the binding potential. Consequently, they do not provide much insight into the binding of epitopes to major histocompatibility complex molecules (MHCs). Thus, the goal of this dissertation is to develop a novel epitope prediction method for the vaccine development without losing the interpretability.

In this study, a novel epitope prediction method named EpicCapo and its variants, EpicCapo^{+} and EpicCapo^{+REF} were developed. Nonapeptides, peptides with nine amino acids, were encoded numerically using a novel peptide-encoding scheme and then input to the support vector machine (SVM). This scheme utilizing the information of amino acid pairwise contact potentials (referred to as AAPPs throughout this dissertation) and peptide-MHC (pMHC) contact sites. We found that the predictive performance of EpicCapo^{+} and EpicCapo^{+REF} outperformed other state-of-the-art methods in many datasets. Interestingly, the most informative AAPPs estimated by our study were those developed by Micheletti and Simons while previous studies utilized two AAPPs developed by Miyazawa & Jernigan and Betancourt & Thirumalai. Additionally, we found that all amino acid positions in nonapeptides could effect on the performances of the predictive models including non-anchor positions such as the positions 5 and 8. Furthermore, EpicCapo^{+REF} was applied to identify candidates of promiscuous epitopes from four influenza strains: H1N1 (A/PR/8/34), H3N2 (A/Aichi/2/68), H1N1 (A/New York/4290/2009), and H5N1 (A/Hong Kong/483/97). As a result, 67.1% of the predicted nonapeptides epitopes were consistent with preceding studies based on immunological experiments. Some predicted promiscuous epitopes have not been tested in any experiment yet. These epitopes can be considered as potential candidates for the novel vaccine development.

Recent studies have demonstrated that predicted high affinity epitopes by epitope prediction methods not always successfully activate T-cell responses. Additionally, predicted low affinity epitopes not always result in low T-cell responses. Thus, immunogenicity of peptides cannot be accurately inferred from the result of epitope prediction. By these reasons, we developed novel T-cell reactivity predictor which we call PAAQD. Nonapeptides were encoded numerically, using combining information of AAPPs and quantum topological molecular similarity (QTMS) descriptors and then input to the random forest (RF). Our numerical experiments suggested that the predictive performance of PAAQD is at least comparable with POPISK, one of the pioneering methods for T-cell reactivity prediction. In addition, we found that the positions 1 and 8 of nonapeptides were the most important ones for T-cell responses. Interestingly, the anchor positions identified by other previous studies, the positions 2, 3, and 7, were not important in T-cell reactivity prediction. These findings support that epitope prediction and T-cell reactivity prediction are different and should not be used interchangeably. Moreover, we found that PAAQD provided more predictive stability than POPISK when using the test dataset that amino acids preference of sequences differs from the training dataset.

From the results of our researches, we speculate that our techniques may be useful in the development of new vaccines. The R implementation of EpicCapo^{+REF} is available at http://pirun.ku.ac.th/~fsciiok/EpicCapoREF.zip. Datasets are available at http://pirun.ku.ac.th/~fsciiok/Datasets.zip. The R implementation of PAAQD is available at http://pirun.ku.ac.th/~fsciiok/PAAQD.rar.

Acknowledgments

Completing a Ph.D. is similar to a long journey far beyond the seven seas. Before reaching the destination, there are many storms to pass, countless submerged rocks to avoid, and sometimes a big monster to fight. I would not have been able to complete this journey without the aid and support from many people over the past three years.

First and foremost, I owe my deepest gratitude to my supervisor, Professor Kenji Satou (Kanazawa University) for his invaluable guidance, motivation, enthusiasm, immense knowledge, and consistent encouragement. From the first day I arrive Japan, he helps and teaches me many things including education and daily life matters. I could not have imagined having a better advisor and mentor for my Ph.D. study.

I am indebted to Associate Professor Osamu Hirose for his grateful advice and hard works to help me writing journal papers and this thesis. Without his help, it is impossible for me to achieve a Ph.D. He always monitors in every step of my progress and teaches me to do the right things.

My sincere thanks to all committee: Professor Haruhiko Kimura (Kanazawa University), Professor Tu Bao Ho (JAIST), Associate Professor Yoichi Yamada (Kanazawa University), and Lecturer Hidetaka Nambo (Kanazawa University) for their encouragement, insightful comments, and hard questions which help me to improve this thesis.

I consider it an honor to work with Dr. Ingorn Kimkong (Kasetsart University, Thailand) who always gives me precious knowledge of immunology. I received many inspirations from her teaching and researches. She also supports and encourages me to continue doing my research.

My sincere thanks to Mr. Zhang Peng, my tutor, for his enthusiastic help and supports. He helped and taught me how to survive in Japan. I also thank all of my fellow PhD students: Lan Anh T. Nguyen, Tu Kien T. Le, Dang Xuan Tho, Vu Anh Tran, and Ngo Duc Luu for their unforgettable friendships and valuable helps. We all in the same destiny of this journey, we share our knowledge, we congratulate when one success, and we encourage together when facing with the hard times.

I would like to thank the government of Japan for providing me the scholarship (Monbukagakusho) and a great opportunity to study in Japan. I have learned a lot of things from this beautiful country. In addition, I would like to thank to Thai Students Association in Japan (TSAJ) for supporting and providing many useful information. I wish to thank to all Thai students in Kanazawa for their best friendships and a warm community.

My deepest gratitude to my beloved parents and older sister, who encouraged and support me all the time, I will never have this day without them. I would like to take this opportunity to thank them for all what I have in my life.

Finally, I would like to acknowledge Kanazawa University and all people who have directly or indirectly helped me to complete this dissertation. I wish them all the best.

Thank You!

Contents

List of Figures

List of Tables

Abbreviations

AAPP	=	amino acid pairwise contact potential
ACC	=	overall accuracy
APC	=	antigen presenting cell
AUC	=	area under the curve
BCR	=	B-cell receptor
CDR	=	complementarity determining region
CTL	=	cytotoxic T lymphocyte
F1	=	F-score
FA	=	factor analysis
FN	=	false negative
FP	=	false positive
HLA	=	human leukocyte antigen
MCC	=	Matthew's correlation coefficient
MHC	=	major histocompatibility complex
PCA	=	principal component analysis
PDB	=	protein data bank
PLS	=	partial least squares
pMHC	=	peptide-MHC
QSAR	=	quantitative structure-activity relationship
QTMS	=	quantum topological molecular similarity
RBF	=	radial basis function
ROC	=	receiver operating characteristic curve
sens	=	sensitivity
spec	=	specificity
SVM	=	support vector machine
SVR	=	support vector regression
TAP	=	transporter associated with antigen processing
TCR	=	T-cell receptor
TN	=	true negative
TP	=	true positive

Chapter 1 Introduction

In this chapter, we first review the basic knowledge of the human immune system and elucidate the need of the new vaccine development for human populations. Next, the problems of the conventional vaccine developments and the way to apply machine learning to solve these problems are described. Generally, machine learning was used in epitope prediction which attempts to search for candidate peptides that are potential for the novel vaccine development. However, recent works have found that results of epitope prediction were not always reliable. Therefore, we introduced the T-cell reactivity prediction which is another way to apply machine learning in the novel vaccine development. Afterwards, we proposed the research objectives to accomplish this dissertation. Finally, the main contributions of this thesis are clearly described and the thesis organization is also presented.

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 ability of the immune system to distinguish between self and non-self is necessary to protect our body from invading pathogens. In addition, the ability to detect malfunction cells is also important since cells infected by the virus and cancer cells can harm our body. In some cases, the immune system loses the ability to distinguish between self and non-self. This causes the immune system to destroy normal cells resulting in autoimmune diseases [2].

1.1.1 Innate immune system

Besides human, the innate immune system is found in all classes of plant and animal life. The innate immune system is the first line of defense against invading pathogens [1]. It recognizes and responds to pathogens in a generic way or non-specific manner. This means responses are even in each time of engagement with pathogens. The innate immune system immediately acts against infection. However, this immune system differs from the adaptive immune system since there is no improvement or long-term protection contributed by the innate immune system [3].

The innate immune system comprises of anatomical barriers, humoral components, and cellular components.

Anatomical barriers

Anatomical barriers in the innate immune response include defense mechanisms in the skin, gastrointestinal tracts, respiratory tracts, and eyes. Table 1-1 shows example anatomical barriers in the human body. There are three protective factors in anatomical barriers. First, mechanical factor such as the desquamation of skin epithelium which helps removing bacteria and other infectious agents attached to the epithelial surfaces. In addition, movement of cilia or peristalsis helps clearing respiratory and gastrointestinal tracts from pathogens. The flushing process by tears and saliva helps protect eyes and mouth from infection, respectively. Moreover, mucus in respiratory and gastrointestinal tracts is able to trap and immobilize microorganisms. Second, chemical factors such as fatty acids in the sweat are able to inhibit the growth of bacteria because of low pH. Lysozyme and phospholipase in tears, saliva, and nasal secretions deteriorate bacterial cell walls and membranes. In addition, small cysteine-rich cationic proteins called defensins found in lung and gastrointestinal tracts can destroy pathogens. Moreover, the surfactants in lung promote the activity of white blood cells to eliminate pathogen more effectively. Third, biological factors such as the normal flora resides on the skin and gastrointestinal tract can prevent the colonization of pathogenic bacteria by secreting toxin and competing for nutrients [4].

Table 1-1 Anatomical barriers in the innate immune system.

Anatomical barrier	Active cellular/biochemical component	Protective mechanisms
Skin	sweat, organic acids	desquamation, flushing
Gastrointestinal tract	gastric acid, bile acids, digestive enzyme, thiocyanate, defensins, gut flora, columnar cells	peristalsis, flushing, low pH
Respiratory tracts and lung	tracheal cilia, surfactant, defensins	mucociliary elevator
Nasopharynx and eyes	mucus, saliva, lysozyme, tears	flushing

Humoral components

If pathogens can penetrate anatomical barriers, they will encounter with another innate immune mechanism named acute inflammation. There are humoral components that work together in inflammation. These components are found in serum or formed at the place where the infection occurs.

The major humoral component of the innate immune system is the complement system. The complement system is series of chemical reactions that promote the ability of antibodies and phagocytic cells to eliminate pathogens. The complement system consists of a number of small proteins which reside in blood circulation. Generally, these proteins are inactive. Immediately after the infection, they will be stimulated by one of several triggers [5]. Table 1-2 shows the basic functions of the complement in overall immune system.

Besides the complement system, there are other humoral components such as lactoferrin, transferrin, interferon, lysozyme, and Interleukin-1 which play roles in the innate immune system [5].

Table 1-2 Basic functions of the complement.

Function	Description
Opsonization	enhancing phagocytosis
Chemotaxis	attracting macrophages and neutrophils
Cell lysis	destroy membranes of pathogens

Cellular components

Leukocytes, certain type of white blood cells, are cellular components in the innate immune system. Leukocytes are not strictly related to a specific organ or tissue and are different from other cells in the body. Similar to single-cell organism, leukocytes are independent and are able to move freely in our body. They can eliminate pathogens and capture foreign particles that they found throughout the body including the blood and lymphatic system [6].

Leukocytes in the innate immune system include natural killer (NK) cells, mast cells, eosinophils, basophils, and phagocytic cells. The phagocytic cells are macrophages, neutrophils, and dendritic cells. These cells kill pathogens by phagocytosis which is the process of engulfing pathogens by the cell membrane to form an internal phagosome. Afterwards, phagosome merges with either a lysosome or a granule and then pathogens will be degraded [7]. Figure 1.1 and 1.2 shows different types of leukocytes and phagocytosis process, respectively.

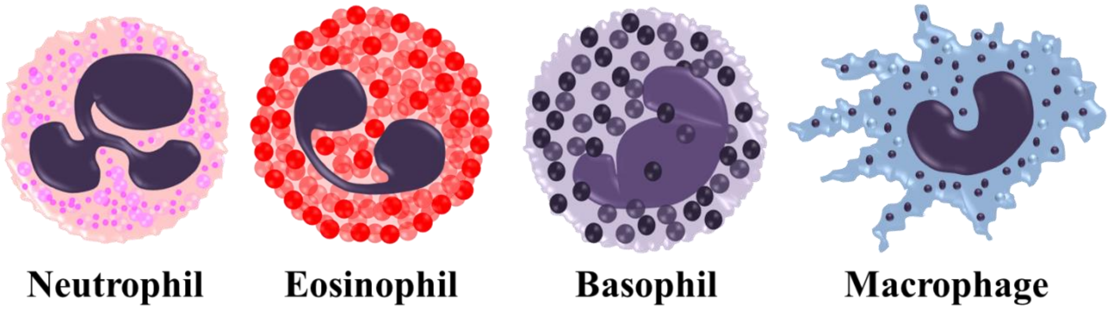


Figure 1.1 Leukocytes in the innate immune system.

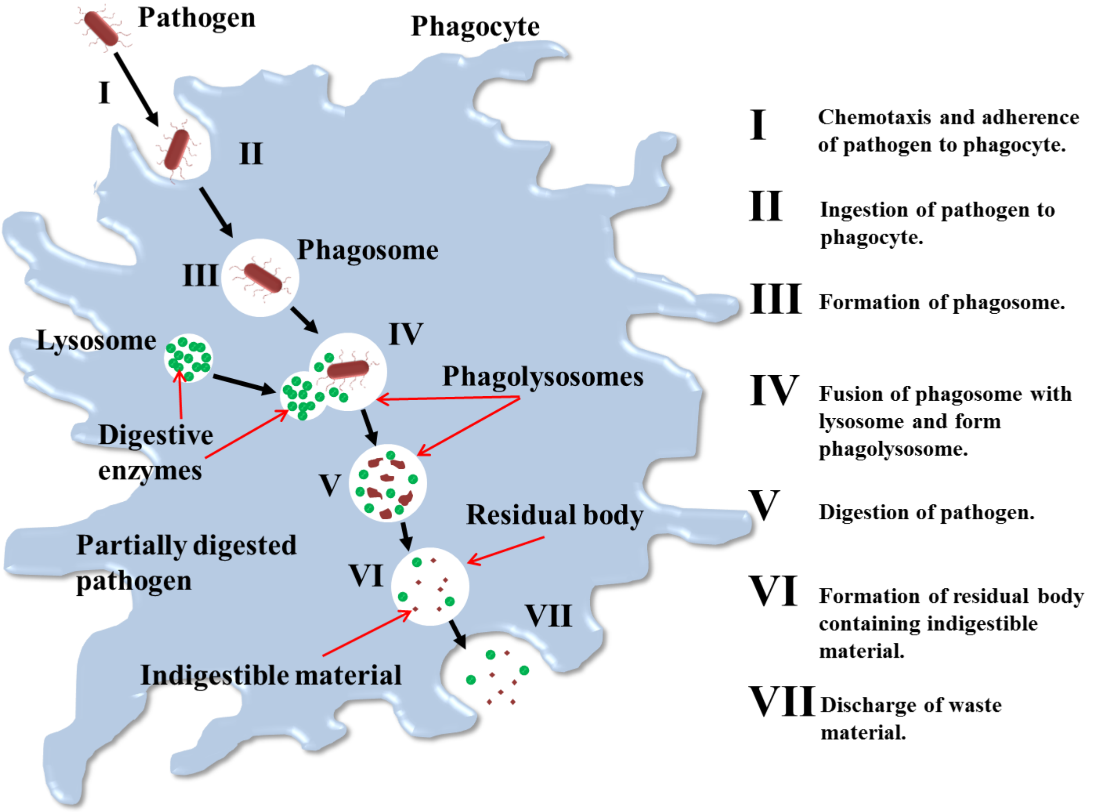

Pathogen
Phagocyte
I
II
Phagosome
III
IV
Lysosome
Phagolysosomes
Digestive enzymes
V
Residual body
Partially digested pathogen
VI
VII
Indigestible material
I Chemotaxis and adherence of pathogen to phagocyte.
II Ingestion of pathogen to phagocyte.
III Formation of phagosome.
IV Fusion of phagosome with lysosome and form phagolysosome.
V Digestion of pathogen.
VI Formation of residual body containing indigestible material.
VII Discharge of waste material.

Figure 1.2 Phagocytosis process.

1.1.2 Adaptive immune system

The adaptive or acquired immune system can learn to recognize specific types of pathogens and maintain immunogenic memory for accelerating future responses. This implies that the adaptive immune system is not able to work effectively in the first encounter or primary response to a peculiar type of pathogen. The primary response is slow and takes time up to three weeks to treat the infection. The learning from this primary response constructs the memory to a specific type of pathogen. When the memorized pathogen invades our body again, the secondary response will be faster and more efficient. This secondary response is rapid enough to eliminate pathogens before they can seriously harm our body. The immunogenic memory can confer long time protection up to our lifetime.

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 [5]. There are two subtypes of T cells: cytotoxic T-lymphocyte (CTL) and helper T-lymphocyte (Th). CTL, Th, and B cells recognize pathogens via antigen recognition.

Antigen Recognition

The term “antigen” refers to the part of a pathogen recognizable by the adaptive immune system. Generally, antigens are structural proteins such as part of bacterium cell membranes and spike proteins of viruses. Antigenic molecules are large biological polymers. These polymers introduce 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. Since a single antigen usually presents several epitopes, it can be recognized by several distinct antibodies.

Typically, T cell receptors (TCRs) of CTLs and Th cells recognize epitopes on the surface of antigen-presenting cells (APCs) whereas B cell receptors (BCRs) recognize epitopes of antigen in the extracellular fluid [1, 8]. APCs such as macrophages and dendritic cells consume pathogen by phagocytosis and digest antigen into small peptides. Some of these peptides are epitopes. These epitopes are transported to the membrane of APCs 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 professional APCs that are macrophages, B cells, and dendritic cells. In humans, MHC is referred to as human leukocyte antigen (HLA) [9].

Cytotoxic T-lymphocyte (CTL)

CTLs have a responsibility to eliminate cells infected with viruses or pathogens to stop infection processes. In addition, CTLs also detect and destroy dysfunctional and cancer cells. When CTLs are activated via antigen presentation on MHC-I, cytotoxins are released to form pores on the membrane of target cells. These pores permit ions and water to flow into the infected cells and lead to cell lysis. Moreover, CTLs also release granzymes which are serine proteases to enter cells via pores and induce programmed cell death (apoptosis) [5].

After the infection is cleared, most of CTLs are deceased. However, few will be retained as memory cells. In the future encounters with the memorized antigen, the response will be dramatically faster because of these memory cells.

Helper T-lymphocyte (Th)

Th cells are very important coordinators in the adaptive immune system. Although these cells do not possess cytotoxic or phagocytic ability, they are the center mediators which manage other immune responses. Th cells recognized epitopes via MHC-II of professional APCs. After their activation, Th cells send signals in the form of cytokines to stimulate activities of other cells such as CTLs, macrophages, and B cells [5].

There are two types of Th cells: Th1 and Th2. Th1 cells release Interferon-γ to activate the bactericidal activities of macrophages and the opsonizing of complement-fixing antibodies on B cells. Th2 cells release interleukin 4, 5, 6, 10, and 13 to activate antibody production of B cells. Antibodies are the most essential components in humoral immunity. Regularly, Th1 responses are effective against intracellular pathogens whereas Th2 responses are effective against extracellular pathogens including helminths and toxins [5].

Similar to CTLs, most of Th cells will be deceased after clearing the infection. However, few will be retained as memory cells.

B cell

B cells play the main role in antibody production. Antibodies are the major components of the humoral immunity. The term “antibody” and “immunoglobulin” (Ig) can be used interchangeably. Antibodies are characterized as a Y-shaped protein (Figure

1.3). In humans, there are five types of antibodies: IgA, IgD, IgE, IgG, and IgM. Each type of antibody has distinct biological properties and can deal with different types of antigens [10]. Antibodies function in the immune system in three ways. First, antibodies bind to pathogens to block them from entering or damaging cells. Second, pathogens coated with antibodies promote the phagocytosis activities of macrophages. Since an antibody possesses two paratopes (see Figure 1.3), two pathogens can be linked together. A number of antibodies can group many cells or particles of pathogens and cause them to be agglutinated. This helps macrophages in eliminating many cells or particles of pathogens at the same time. Third, antibodies also trigger the complement systems and other immune responses, leading to the ultimate destruction of pathogens [11].

Before antibodies production, B cells must be activated and become plasma B cells. There are two ways of B cells activation: T cell-dependent and -independent activation. For T cell-dependent activation, Th2 cells release interleukin 4, 5, 6, 10, and 13 to activate B-cells after antigen representation by professional APCs via MHC-II. For T cell-independent activation, BCR will directly bind with antigens and B-cells are then activated. Generated plasma B-cells stay for 2-3 days in our body. About 10% of these plasma cells are retained to serve as long-term antigen specific memory B cells. In the future encounters with memorized antigens, memory B cells will rapidly differentiate to plasma B-cells and then produce antibodies [5].

It is fascinating about the cooperation between components in our immune systems. The innate immune system can stimulate the adaptive immune system and vice versa. In addition, products of each immune system can promote other activities. For example, Th1 and Th2 cells send signal to activate macrophages and B cells, respectively.

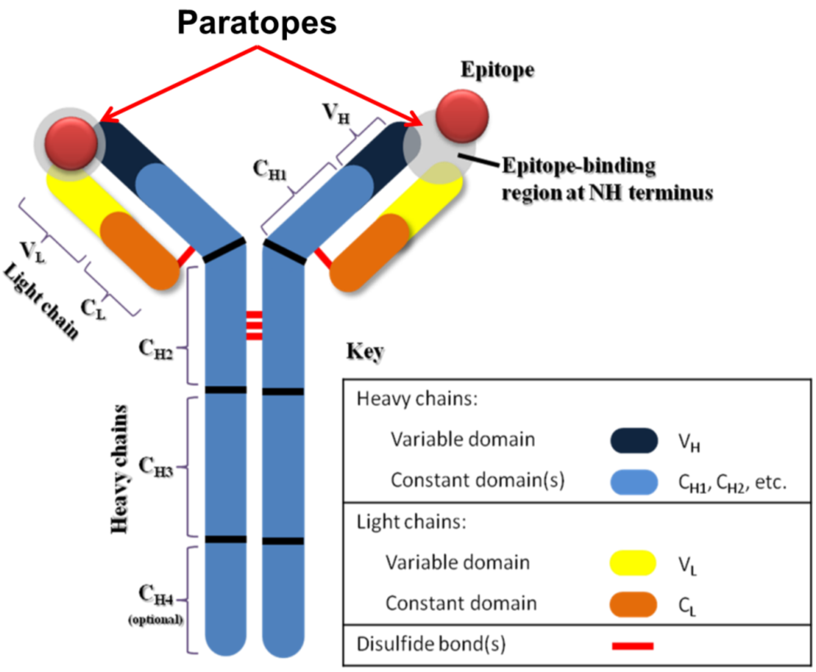


Figure 1.3 Basic structure of immunoglobulin.

The structure consists of two large heavy chains and two small light chains. Five types of antibodies are determined by difference of heavy chain. Antibodies possess two paratopes to interact with epitopes.

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 [12].

Currently, several types of vaccines have been developed. The basic vaccine technology is to use killed pathogens. Pathogens are killed by chemicals, heat, radioactivity, or antibiotics. The remains of pathogens such as cell membranes or polymers can activate immune responses. This type of vaccine has been used to prevent polio, hepatitis A, cholera, and rabies. Live attenuated-pathogens also have been used as vaccines, but they are inactivated by cultivating under conditions that disable their virulent properties. Examples of attenuated vaccines include yellow fever, measles, rubella, mumps, and typhoid vaccines. Attenuated vaccines have some advantages over killed vaccines that the stronger protection can be induced. This is because the transient growth of inactivated pathogens causes more intense immune responses. However, killed vaccines are safer since attenuated viruses may change to a virulent form and cause disease [13].

According to T and B cells, only antigen is important in the activation. Therefore, antigens should be used instead of the entire cell of pathogens. This usage is called subunit vaccines. Subunit vaccines can contain more than one antigen, and antigen can be manufactured using recombinant DNA technology. Vaccines produced by this approach are named "recombinant subunit vaccines" which was already developed for the hepatitis B virus. Subunit vaccines are safer than attenuated vaccines since only some parts of the pathogen are used [14, 15].

In addition, there are also other types of vaccines such as toxoid vaccines which used inactive bacterial toxins to stimulate immune responses. Dangerous bacterial toxins can be treated with formalin and become inactive. Inactive toxins are called toxoids. Besides, conjugate vaccines were developed for some bacteria species that have polysaccharides coat on their membranes. These polysaccharides make bacterium cells difficult to be detected by the immune system. Therefore, polysaccharides were conjugated with antigens or toxoids and cause them to be recognizable by the immune system [16, 17]. Some other types of vaccines are currently in experimental phase such as DNA vaccines, dendritic cell vaccines, recombinant vector vaccines, and T-cell receptor peptide vaccines [18, 19].

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, researches of vaccines are necessary and should be concerned by the governments as the top priority in the public health plans.

1.3 T-cell vaccine development

T-cells are considered as a center mediators in the human immune response. T-cell vaccines aim to stimulate immune responses of CTLs and Th cells via antigen presentation on MHC-I and -II, respectively. The activation of CTLs helps in terminating the infection by destroying infected cells, and also helps in the elimination of cancer cells.

The activation of Th cells is then further activate production of antibodies by B cells, and also stimulates other potential immune responses.

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 1.4. 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 [20].

To develop T-cell 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 [21]. Although epitopes identified by using machine learning are not guaranteed to be 100% correct, these predicted epitopes are promising candidates for immunological experiments.

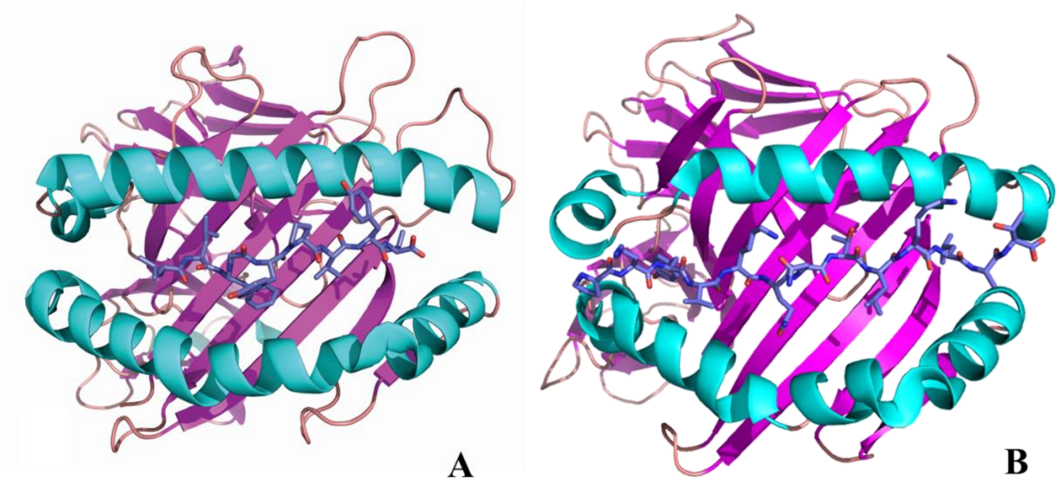


Figure 1.4 Visualization of pMHC complexes.

(A) MHC-I (PDB entry 1DUZ [22]). (B) MHC-II (PDB entry 1DLH [23]).

1.4 Applications of machine learning in CTL epitope prediction

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 [24]. In addition, peptides of length 9 known as nonapeptides have been popularly studied.

The pioneering epitope prediction methods were based on allele-specific motifs [25, 26]. The important positions of the motif were analyzed. For instance, positions 2 and 9 were the most important positions in the case of HLA-A*02:01 allele. The residues at both positions were assigned as the classical anchor residues [27]. In addition, positions 1, 3, and 7 also assigned as the secondary anchor residues [28–30]. In each anchor residue, an amino acid which frequently occurs from known epitopes was defined. New or untested peptides which comprised of matched amino acids with assigned anchor residues were identified as epitopes.

When more data of experimented epitopes are available, the matrix-based methods have been introduced. Matrices were calculated using statistical techniques. These matrices were used to estimate binding energy between HLA and peptides. The examples of matrix-based methods are BIMAS [31], RANKPEP [32], Gibbs sampler [33], ARB [34], SMM [35], and SMM$^{\mathrm{PMBEC}}$ [36].

Recently, using machine learning algorithms in epitope prediction shows great achievements. Examples of epitope prediction methods that based on machine learning techniques are NetMHC [20], NetMHCpan [37], NetCTL [38], NetCTLpan [39], and SVRMHC [40]. 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) [41] which is the largest one. In addition, there are also other available databases such as SYFPEITHI [42], FIMM [43], MHCPEP [44], MHCBN [45], and AntiJen [46].

The allele-specific motif methods, the matrix-based methods, and machine learning-based methods generally concern only sequence information. Detailed binding mechanisms cannot be provided by these methods. Therefore, three-dimensional (3D) structure-based methods have been developed [47–49] to unveil MHC-epitope binding mechanisms. Unfortunately, 3D structure-based methods require a number of crystal structures of MHC-peptide complexes, which are still not available in a large number. Besides, the performance of structure-based methods is currently lower than machine learning-based methods [21].

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 [50].

1.5 Epitope prediction versus T-cell reactivity prediction

Epitope prediction methods have been used to search for candidate peptides in the vaccine development. Predicted peptides with high binding affinity to the MHC were presumed to successfully activate T-cell responses. However, recent experiments show that those peptides with high predicted binding affinity to the MHC did not always activate T-cell responses [51]. In addition, other biological factors were more strongly correlated to T-cell responses than MHC binding affinities [52]. 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 [53–55]. Previous studies based on protein crystal structures reveal that residues at positions 4, 6, and 8 of nonapeptides were important in the binding of TCR to pMHC (TCR-pMHC) complex [56, 57]. These important positions are different from those defined by epitope prediction. In fact, the pMHC binding should directly contribute to TCR-pMHC binding. However, the results of important positions are in conflict. Therefore, the prediction and characterization of T-cell reactivity are very essential for more understanding in the immune system [55].

The first published method for T-cell reactivity prediction is POPI [58]. POPI used physicochemical properties from the AAindex database [59] to encode peptides to numerical vectors. These vectors are then input to the support vector machine

(SVM). Afterwards, POPISK [55] was developed. POPISK simply used the SVM with the string kernels.

1.6 Objectives

The vaccine development is very essential for mankind in order to establish an effective protection against infectious pathogens. In the last decade, serious outbreaks emerged and caused high mortality. The examples are the epidemics of severe acute respiratory syndrome (SARS) in 2003 and influenza A viruses, H1N1 and H5N1, in 2005–2009. Therefore, the development of new vaccines is necessary to prevent future outbreaks.

The conventional vaccine developments are laborious and time-consuming [21]. Epitope prediction can accelerate the process of the vaccine development by providing promising candidate peptides for further immunological experiments. The uses of machine learning techniques in epitope prediction have been actively studied and many methods were proposed [20, 31–40]. However, most of existing methods tend to overlook the interpretability which respects to the binding potential and thereby not provide much insight into pMHC binding mechanisms. Thus, this study is aimed at developing a novel epitope prediction method for the vaccine development. Substantially, importance of peptide positions for the pMHC binding was also analyzed to provide more understanding in pMHC binding mechanisms. The objectives of this dissertation are described as follows.

To develop a novel epitope prediction method. First, nonapeptides sequences were encoded to numerical data in order to input to machine learning algorithms. To perform this task, we established the new peptide encoding scheme. This peptide encoding scheme was created by combining information about the pMHC contact sites [60] with amino acid pairwise contact potentials (AAPPs) [59]. After peptides encoding, the support vector machine (SVM) was used for training and testing. Benchmark datasets [61] were used for evaluation of our method performance and then compared with other state of the art methods.

To analyze for important AAPPs in the pMHC binding mechanisms. For each allele dataset, only AAPPs that led to the highest performance were used in the further steps. Afterwards, we identified important positions of nonapeptide in pMHC binding. In each encoded peptide data, each feature corresponding to one position of nonapeptide. We employed the feature selection algorithm to rank the importance of features. From the rank, we could identify important positions from significant features. From these results, we selected only important AAPPs and positions of nonapeptide to estimate the final predictive models.

To identify promiscuous epitopes from influenza A viruses. We used the predictive models that were created in the previous step to identify ‘promiscuous epitopes’ from protein sequences of influenza A viral strains. The promiscuous epitope is an epitope that binds to many HLA alleles. The use of promiscuous epitopes in vaccine development will provide a high level of population coverage. The identified epitopes by our predictive models were validated by cross-checking with the publications of immunological experiments [62].

To develop a new T-cell reactivity prediction method. Recent studies showed that predicted high affinity epitopes did not always result in activation of T-cell responses. Therefore, we developed a novel T-cell reactivity prediction method by combining information of AAPPs, pMHC contact sites, and quantum topological molecular similarity (QTMS) descriptors [63]. The new peptide encoding scheme was proposed by combining AAPPs and QTMS descriptors. Peptides were encoded and then input to the random forest for training and testing. We compared the performance of our method with previous T-cell reactivity predictors [55, 58].

To analyze for important AAPPs, QTMS descriptors, and positions of peptide in TCR-pMHC binding mechanisms. We used our new T-cell reactivity prediction method to identify for important AAPPs, QTMS descriptors, and positions of peptide in TCR-pMHC binding mechanisms. The result of important positions in T-cell reactivity prediction was compared with those in epitope prediction.

1.7 Contributions

The purpose of this research is to apply machine learning techniques in epitope and T-cell reactivity prediction which are essential steps toward the vaccine development. The ultimate goals are to develop the novel epitope and T-cell reactivity prediction methods which are able to provide more insights in pMHC and TCR-pMHC binding mechanisms, respectively. The main contributions of this thesis are summarized as follows.

Novel epitope prediction method. We developed a new epitope prediction method which we called EpicCapo and its variants, EpicCapo^{+} and EpicCapo^{+REF}. Peptides were numerically encoded by using our proposed peptide encoding scheme. This scheme is the combination of pMHC contact sites with AAPPs. Our method achieved high performance and outperformed other methods in many datasets of HLA alleles. In some datasets, although there are small numbers of training peptides, our method still provided the high performance. Therefore, our method is a promising tool for the development of new vaccines.

Identification of important AAPPs and peptide positions in pMHC binding mechanisms. Based on our proposed method, we identified important AAPPs and peptide positions in pMHC binding mechanisms. We found that two AAPPs were very important in pMHC binding. In addition, we found that ten top-ranked features correspond to positions 9 and 2 in most datasets, followed by positions 3, 1, or 7. This finding is consistent with other studies which demonstrate that positions 9 and 2 are primary anchor residues, and positions 1, 3, and 7 are secondary anchor residues in the pMHC binding. However, when we identified for the optimal sets of features that led to the highest performance, features from all nine positions were included. Hence, we presumed that all nine positions are important in the pMHC binding and their effects to the binding affinity are not independent.

New promiscuous epitopes for the development of influenza A vaccines. Our proposed method was applied to identify promiscuous epitopes from four influenza A viral strains: H1N1 (A/PR/8/34), H3N2 (A/Aichi/2/68), H1N1 (A/New York/4290/2009), and H5N1 (A/Hong Kong/483/97). We found that many predicted promiscuous epitopes were in agreement with previous immunological experiments. This consistency indicates that our method has high accuracy in epitope prediction. Some predicted promiscuous epitopes have not been tested in any experiment yet. These epitopes can be considered as potential candidates for the novel vaccine development.

Novel T-cell reactivity prediction method. We developed a new T-cell reactivity prediction method which we called PAAQD. Peptides were numerically encoded by using our proposed peptide encoding scheme which is similar to that in EpicCapo. 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. Recent studies show that predicted binding affinities by epitope prediction methods were not strongly correlated to T-cell responses. This means that predicted epitopes are not guaranteed to activate immune responses. Therefore, T-cell reactivity prediction should be used rather than epitope prediction.

Identification of important AAPPs, QTMS descriptors, and peptide positions in TCR-pMHC binding mechanisms. Based on our new T-cell reactivity prediction method, we identified important AAPPs, QTMS descriptors, and peptide positions in TCR-pMHC binding mechanisms. We found six important AAPPs. One of these is also important in the pMHC binding. For QTMS descriptors, we found that all descriptors were important. By using our method, peptide positions 1 and 8 were the most important ones. This result is concordant with the previous study of T-cell reactivity prediction. Interestingly, we found that positions 2, 3, and 7 were less important than the others. These positions have been identified as anchor residues for epitope prediction in other studies. Therefore, these findings support that epitope prediction and T-cell reactivity prediction are considerably different.

1.8 Thesis organization

The thesis is divided into 5 chapters, including the current one. The first chapter covers introductory materials, motivations, and contributions of researches presented in this dissertation. The remaining chapters are organized as follows:

Chapter 2 reviews the uses of machine learning in immunoinformatics which is a new field that focused on *in silico* analysis and modeling of immunological data and problems. The major usages of machine learning algorithms in immunoinformatics that are artificial neural network, support vector machine, and hidden Markov models were described. Additionally, important immunoinformatics databases are also addressed in this chapter.

Chapter 3 describes a novel epitope prediction method which named EpicCapo. This method used our proposed peptide encoding scheme which is the combination of structural and physicochemical information. The SVM was used to conduct classification tasks after the data processing. The performance of our epitope prediction method also evaluated and compared with other state of the art methods. Moreover, the insights in pMHC binding are also shown in this chapter.

Chapter 4 Introduces new T-cell reactivity prediction method which named PAAQD. The peptide encoding scheme used in this method is the combination of structural, physicochemical, and quantum topological information. The random forest was used to conduct classification tasks after the data processing. The performance of our epitope prediction method also evaluated and compared with other high performance T-cell reactivity prediction methods. Furthermore, the insights in TCR-pMHC binding are also shown in this chapter.

Chapter 5 summarizes the principal tasks of this dissertation, including the achievements and contributions. Some limitations are also presented. In addition, future works and directions also discussed in this chapter.

Chapter 2 Review of machine learning in immunoinformatics

In this chapter, we describe the uses of machine learning algorithms in immunoinformatics. Immunoinformatics is a new branch of bioinformatics that focused on computational analysis and modeling of immunological data and problems. We introduce and give examples of commonly used algorithms in immunoinformatics that are artificial neural network, support vector machine, and hidden Markov models. In addition, remarkable immunoinformatics databases are also shown.

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. To date, many immunology-focused resources and tools are available to help in uncovering the properties of the whole immune system. 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 [64, 65]. Figure 2.1 shows an overview of immunoinformatics research area.

Most immunoinformatics researches are related to prediction of potential B- and T-cell epitopes. The outcomes help speeding up the new vaccine development. 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 [66]. Bioinformaticians frequently used ANN to solve many biological and physiochemical problems. In case of epitope prediction, some methods used ANN to learn input sequences of known epitopes and then generate the predictive models. The improved model of neural network for T-cell epitope prediction was described in [33]. The high performance methods, NetMHC [20] and NetMHCpan [37], are based on ANN and used position-specific scoring matrices. NetCTL [38] and NetCTLpan [39] integrated the prediction of pMHC-I binding, proteasomal cleavage, and transporter associated with antigen processing (TAP) together.

Most methods achieved high performance when predicting MHC-I epitopes. However, medium to low performance was acquired when predicting MHC-II epitopes. The prediction of MHC-II epitopes is more difficult because the lengths of input peptide are highly variable.

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 [67]. The SVM is described as a non-probabilistic binary classifier and belongs to the group of the kernel-based approaches [68]. A hyperplane or set of hyperplanes in a high- or infinite-dimensional space was generated by the SVM for using in classification or regression tasks. The hyperplane that cause the largest distance from the nearest point belonging to another class is the favorable one. Deriving such hyperplane should lead to optimal separation and the reliable predictive model [69].

The SVM has been widely used in immunoinformatics. Most of published methods focused on epitope prediction. SVRMHC [40], the epitope predictor based on support vector regression (SVR) used data from AntiJen and used LIBSVM [70] for SVR-related implementation. This method can perform prediction on both MHC-I and –II. Nanni [71] used feature extraction based on BLOSUM50 and then conducted classification tasks using the SVM. TAPPred [72], the MHC-I epitope predictor is based on the cascade SVM. Two layers of SVMs were used in this method and it achieved remarkable performance. In case of proteasomes cleavage prediction, Pcleavage [73] was developed to predict cleavage sites in antigenic proteins by using the SVM.

For B-cell epitope prediction, COBEpro [74] was developed to predict continuous B-cell epitopes. COBEpro consists of two-step. First, a fragment epitopic propensity score was assigned to protein sequence fragments using the SVM. Second, the score for each residue was calculated based on the previous score. By using the second score, B-cell epitopes were determined. COBEpro has been incorporated into the SCARTCH prediction suite [75].

2.1.3 Hidden Markov models

The hidden Markov models (HMMs) were described by Baum et al. [76]. HMMs were used in speech recognition [77, 78]. In 1980, HMMs were firstly applied in the analysis of biological sequences, especially DNA sequences [79]. To date, HMMs are widely used in the bioinformatics field such as the prediction of protein secondary structure [80], prediction of transmembrane regions [81], and protein homology analysis [82]. In addition, HMMs have been popularly used in sequence alignment [83], phylogenetic tree analysis [84], and gene identification [85].

For immunoinformatics field, PredTAP [86], the method based on HMM was developed to predict peptide binding to TAP molecules. This method used second-order HMM back propagation neural network. Mamitsuka [87] developed HMM based models for prediction of pMHC binding affinity. However, the models were restricted to HLA-A*02:01 and DR1 alleles. Afterwards, Udaka et al. [88] used Mamitsuka's approach to estimate predictive models for other MHC-I alleles. Moreover, Brusic et al. [89] developed HMM models to predict pMHC binding affinity of HLA-A2 alleles. In this method, only amino acids that interact with HLA molecules were used to derive the predictive models.

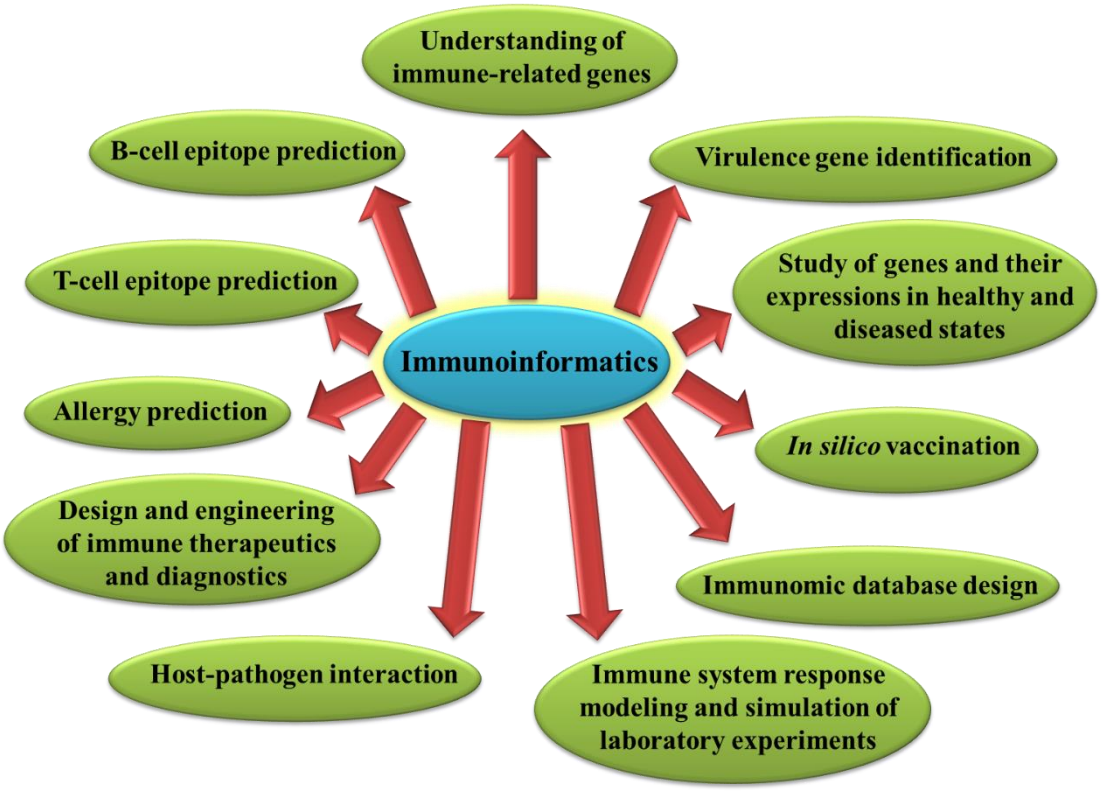

Immunoinformatics
Understanding of immune-related genes
B-cell epitope prediction
Virulence gene identification
T-cell epitope prediction
Study of genes and their expressions in healthy and diseased states
Allergy prediction
In silico vaccination
Design and engineering of immune therapeutics and diagnostics
Immunomic database design
Host-pathogen interaction
Immune system response modeling and simulation of laboratory experiments

Figure 2.1 Overview of immunoinformatics research [64].

2.2 Immunoinformatics databases

Recently, because of advancement in high throughput technology, immunological data have increased rapidly. There are many databases that store these data. 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.

In our researches, we mainly used data from IEDB [41]. This database provides vast datasets including both T- and B-cell epitopes. In addition, analysis tools are available in this database including state of the art epitope prediction methods.

Table 2-1 Available immunoinformatics databases [64].

Type	Name	URL	Ref.
T-cell epitopes	JenPep	http://www.darrenflower.info/jenpep/	[90]
	SYFPEITHI	http://www.syfpeithi.de	[42]
	FRED	http://www-bs.informatik.uni-tuebingen.de/Software/FRED	[91]
	MHCBN	http://www.imtech.res.in/raghava/mhcbn/	[45]
B-cell epitopes	CED	http://immunet.cn/ced/	[92]
	Bcipep	http://www.imtech.res.in/raghava/bcipep	[93]
	Epitome	http://cubic.bioc.columbia.edu/services/epitome/	[94]
Both T- and B- cell epitopes	IEDB	http://www.iedb.org/	[41]
	IMGT	http://www.imgt.org/	[60]
	MHCPEP	http://wehih.wehi.edu.au/mhcpep/	[44]
	AntiJen	http://www.ddg-pharmfac.net/antijen/AntiJen/antijenhomepage.htm	[46]
Allergen	Database of IUIS	http://www.allergen.org	[95]
	Allergen Pro	http://www.niab.go.kr/nabic/	[96]
	SDAP	http://fermi.utmb.edu/SDAP/	[97]
Information related to molecular evolution of immune system components	ImmTree	http://bioinf.uta.fi/ImmTree	[98]
	Immunome database	http://bioinf.uta.fi/Immunome/	[99]
	ImmunomeBase	http://bioinf.uta.fi/ImmunomeBase	[100]
	Immunome Knowledge Base	http://bioinf.uta.fi/IKB/	[101]

Chapter 3 EpicCapo: epitope prediction using combined information of amino acid pairwise contact potentials and HLA-peptide contact site information

Epitope identification is an essential step toward synthetic vaccine development since epitopes play an important role in activating immune responses. Classical experimental approaches are laborious and time consuming, and therefore computational methods for generating epitope candidates have been actively studied. Most of these methods, however, are based on sophisticated nonlinear techniques for achieving higher predictive performance. The use of these techniques tends to diminish their interpretability with respect to binding potential: that is, they do not provide much insight into binding mechanisms. We have developed a novel epitope prediction method named EpicCapo and its variants, EpicCapo^{+} and EpicCapo^{+REF}. Nonapeptides were encoded numerically using a novel peptide-encoding scheme for machine learning algorithms by utilizing 40 amino acid pairwise contact potentials. The predictive performances of EpicCapo^{+} and EpicCapo^{+REF} outperformed other state-of-the-art methods without losing interpretability. In addition, we found that all amino acid positions in nonapeptides could effect on the performances of the predictive models including non-anchor positions. Finally, EpicCapo^{+REF} was applied to identify candidates of promiscuous epitopes. As a result, 67.1% of the predicted nonapeptides epitopes were consistent with preceding studies based on immunological experiments. We speculate that our techniques may be useful in the development of new vaccines.

3.1 Introduction

CTLs play an important role in the vertebrate immune system. They recognize pathogens via peptide presentation on MHC. If the source of peptides is an infectious virus, the CTL response could be stimulated, thus leading to the elimination of virus infected cells [102]. As mentioned in the chapter 1, MHC-bound peptides are called epitopes. Epitope identification is an essential step toward synthetic vaccine development, since epitopes play an important role in the activation of the immune responses [21]. Epitopes are traditionally identified by synthesizing a large number of nonapeptides and subsequently performing affinity assays. Those peptides with high binding affinity to MHC proteins are considered as potential epitopes. However, the process of developing a new vaccine is time-consuming and laborious when performed with traditional methods. To avoid the problems of such bottlenecks, instead computational methods can be effectively applied to search for candidate peptides and identify new promising epitopes.

In human, MHC is referred to as HLA. There are three classes of HLAs: I, II, and III. Epitopes presented on HLA class I molecules are recognized by CTLs. HLA class I proteins can be categorized into three types according to their genes: HLA-A, HLA-B, and HLA-C. A majority of previous studies have focused 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 [24]. Typically, the HLA-A*02:01 epitope consists of 8–10 amino acids, and many studies have focused on nonapeptides in particular: that is, epitopes that are 9 residues long [103–105]. Figure 3.1 (A) shows the nonapeptide epitope LLFGYPVYV fitted inside the HLA-A*02:01 binding cleft, which consists of two α-helices and one β-sheet (from PDB entry 1DUZ [22]). Figure 3.1 (B) shows the conformation of the nonapeptide epitope LLFGYPVYV.

Early epitope binding prediction algorithms were based on allele-specific motifs [25, 26]. For example, for the HLA-A*02:01 allele, positions 2 and 9 of nonapeptides were the most important ones for binding. The residues at both positions were defined as classical anchor residues typically occupied by leucine, valine, and isoleucine since the MHC molecule forms hydrophobic sites for amino acids at these two positions [27]. Additionally, the residues at positions 1, 3, and 7 were identified as secondary anchor residues. Positions 1 and 3 were mainly preferred by tyrosine and phenylalanine [28, 29]. The residue at position 7 was suggested to be an amphipathic residue suitable for amino acids with small hydrophobic side-chains such as valine and alanine [30]. In this manner, unknown peptides that matched with such allele-specific motifs were determined to be epitopes.

As more data became available, statistical methods could be applied to calculate a positional scoring matrix. In the matrix, an element was defined individually for each position and a specific amino acid, resulting in an $L \times 20$ coefficient matrix where L is the length of the peptide. In general, the matrix is used under the assumption that each amino acid in a peptide sequence independently contributes a certain binding energy according to an element included in the positional scoring matrix. Overall binding energy is estimated from the summation of binding energies from all positions. There are several methods based on such a positional scoring matrix: for example, BIMAS [31], RANKPEP [32], Gibbs sampler [33], ARB [34], SMM [35], and SMM$^{\text{PMBEC}}$ [36].

Currently, the most successful approach for epitope prediction utilizes machine learning algorithms. These algorithms require large enough datasets for training in order to obtain reliable results. Fortunately, the Immune Epitope Database (IEDB) [41] provides more than 100,000 MHC binding data related to T-cell epitopes from infectious pathogens, experimental pathogens, and self-antigens (autoantigens). IEDB encompasses patent data from biotechnological and pharmaceutical companies, as well as direct submissions from research programs and partners. As reliable experimental data are provided, the volume promises a sufficient grounding for developing good predictive models. Although IEDB is not the only database that provides such information, it has more entries than other existing databases. Examples of other databases are SYFPEITHI [42], FIMM [43], MHCPEP [44], MHCBN [45], and AntiJen [46]. NetMHC [20], a predictor based on artificial neural networks, used data from both IEDB and SYFPEITHI and performed very well. SVRMHC [40], a predictor based on support vector regression (SVR) used data from AntiJen and used LIBSVM [70] for SVR-related implementation. Moreover, there also exists an epitope predictor based on a hidden Markov model [88].

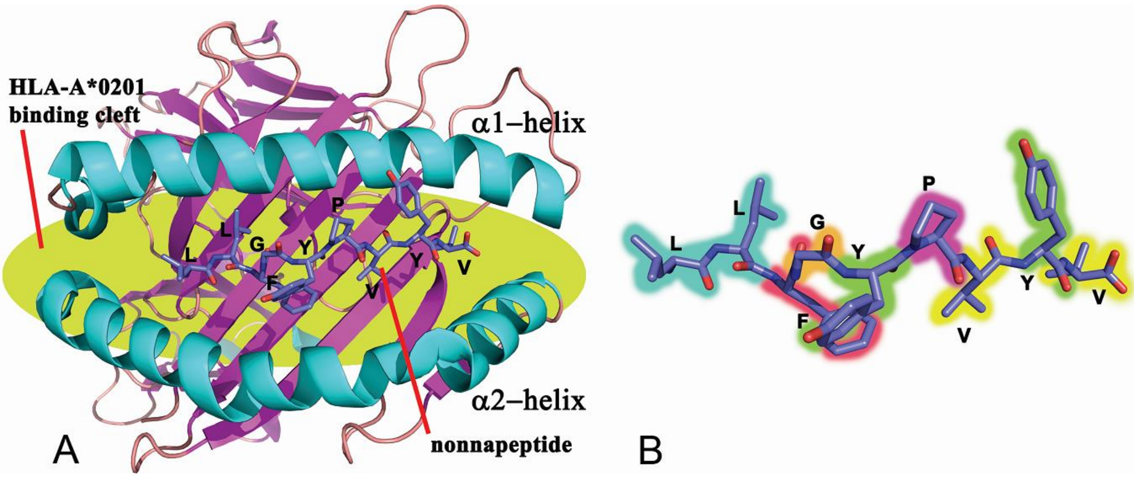


Figure 3.1 Visualization of the HLA-nonapeptide complex.

(A) Crystal structure of the LLFGYPVYV-HLA-A*02:01 complex resolved by X-ray crystal diffraction (PDB entry 1DUZ [22]). (B) Conformation of the nonapeptide extracted from the complex.

The allele-specific motif method, the positional scoring matrix method, and machine learning-based methods use only sequence information in general. Almost none of these methods can provide a clear explanation about the effects of the physicochemical properties of amino acids on binding affinity. In some cases, there are not enough peptides for training: e.g., when using data from rare alleles. Therefore, three-dimensional (3D) structure-based methods have been developed [47–49] to uncover binding mechanisms and address all forces related to binding affinity. However, such methods are currently less reliable than data-driven methods [106]. The reason is that 3D structure-based methods usually require a number of crystal structures of MHC-peptide complexes, which are still not available in large numbers.

Recently, more than 2,000 HLA alleles have been identified. Searching for epitopes that bind to a large number of those alleles would be computationally exhaustive and time-consuming. Therefore, the concept of allele supertypes was developed by clustering alleles into groups based on overlapping epitopes [107–111]. Within each supertype, most of the alleles should share the same epitopes. These epitopes are called ‘promiscuous epitopes’, which show great promise for vaccine development due to their potential for a high level of population coverage.

In this chapter, we would like to introduce our novel epitope prediction method named EpicCapo. Peptides were encoded numerically by combining information on the pMHC contact sites with AAPPs, accompanied by the SVM [112]. Our method’s performance was evaluated by using benchmark datasets and then compared with other high performance methods. In addition, identification of candidates of promiscuous CTL epitopes for influenza A viruses was demonstrated using the proposed method.

The H1N1 or H5N1 strain of influenza A virus caused a lethal flu in humans, as seen during the epidemics of 2005–2009. Although inactivated influenza vaccination is beneficial, the development of more effective vaccines is still needed, particularly in elderly adults who are more susceptible to viral infections [113]. Identification of promiscuous CTL epitopes might aid this issue by providing candidate peptides from viral proteins for vaccine development.

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 [60], the allele-specific positional scoring matrices developed by SMM^{PMBEC} [36], and the AAPPs from AAindex [59].

The reference pMHC contact sites retrieved from IMGT were modified by adding more MHC positions. The added MHC positions were determined by observing the pMHC contact sites of the selected 189 crystal structures of the HLA-nonapeptide complex collected from IMGT entries specific to the MHC-I receptor type. If there were new contact positions, the reference pMHC contact sites were modified by adding those new positions. Therefore, more HLA-nonapeptide contact positions were included in the modified pMHC contact site because the reference pMHC contact sites resulted from the use of only 74 crystal structures of the HLA-nonapeptide complex [60]. Utilizing the modified pMHC contact sites should provide more reliable results during the prediction. Table 3-1 shows the references and added pMHC contact sites positions. This information served as a binding template between the peptide and MHC. In NetMHCpan [37], the reference pMHC contact sites were used to extract a pseudo sequence representing the given MHC molecule. When performing prediction, sequence information from both peptide and MHC was taken into account. However, the pairs of amino acids between the MHC molecule and peptide were not of concern. Therefore, to generate a more informative predictive model, we used information about the pairs of amino acids at the interface between an MHC molecule and a nonapeptide, represented by AAPPs. In addition, the allele-specific positional scoring matrices developed by SMM$^{\mathrm{PMBEC}}$ were used in our study. These matrices provide information of how likely a given amino acid would be preferred or avoided in a specific residue. Like NetMHCpan, SMM$^{\mathrm{PMBEC}}$ did not use AAPPs. Consequently, we proved that a proper selection of AAPPs could lead to higher performance in the prediction. The encoded data could be further used in tasks of classification or regression using machine learning algorithms. In this study, we demonstrated the feasibility of the classification task by using the SVM implemented in the R package kernlab [112].

Here, we propose a novel scheme for encoding nonapeptides into input vectors of the SVM. Suppose $E(a_1, a_2)$ is an AAPP for the amino acids a_1 and a_2. If two or more types of AAPPs are available, we denote kth type of the AAPP by $E_k(a_1, a_2)$. Also, we denote the ith amino acid of the nonapeptide n and the jth amino acid of HLA by $u_i^{(n)}$ and v_j, respectively. In order to combine information of position-specific amino acid scores of the nonapeptides with AAPPs, we define a score $S_{k,i}^{(n)}$ for the ith amino acid of the nonapeptide n under a kth type of AAPP as follows:

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

where L is the length of the HLA protein, $T_i(a)$ is the ith 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 ith amino acid of a nonapeptide and the jth amino acid of HLA contact each other, and 0 otherwise. Here, 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. The scaling of positional scoring matrices is shown in Table 3-2. Note that $\sum_{j=1}^{L} \delta_{ij}$ is the number of contact sites for the ith amino acid of a nonapeptide (see Table 3-1). Intuitively, this score represents average pair-potential of contact sites, weighted by 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 $(M(k-1) + i)^{\text{th}}$ 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 3.2 illustrates an example of the data-encoding scheme for the first position of the nonapeptide.

Table 3-1 Reference and added pMHC contact sites for the HLA.

		Reference HLA positions	**Added HLA positions**
Nonapeptide position	1	5 59 62 63 66 163 167 171	7 9 45 58 67 164
	2	7 9 22 24 34 45 63 66 67 70	99 159
	3	97 99 152 155 156 159	9 66 67 70 160
	4	65 66 155	62 158
	5	70 73 74 97 116 155 156	65 69 72 114 147 150 151 152
	6	66 69 70 73 74 97 114 151 155	65 99 147 152 156
	7	97 114 147 150 152 155	59 63 116 133 146
	8	72 73 76 80 146	77 147
	9	77 80 81 84 95 116 123 124 143 147	26 33 55 58 97 142 146

Table 3-2 The positional scoring matrix of EpicCapo used in the experiment that peptide-encoding schemes were compared.

Amino acid	**Nonapeptide position**								
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
A	6.053	4.651	5.843	5.812	5.220	4.497	5.295	5.243	7.083
C	4.474	1.000	4.359	4.272	6.514	6.112	4.928	5.263	5.066
D	2.093	5.074	5.887	6.432	5.417	5.575	4.861	3.889	5.322
E	2.030	5.322	2.958	6.692	4.146	4.726	4.936	5.784	5.322
F	8.630	5.220	6.617	5.082	6.286	6.680	7.071	5.993	3.289
G	5.468	3.333	4.632	5.658	5.579	3.862	3.349	6.088	4.841
H	4.189	5.646	4.217	4.636	6.329	4.900	4.896	4.411	4.841
I	5.650	7.336	5.883	5.425	6.021	7.158	6.550	4.612	8.295
K	6.704	6.676	2.950	5.437	4.213	3.388	2.303	5.263	5.236
L	5.705	9.443	6.664	4.647	5.457	6.558	6.416	6.088	8.157
M	6.436	10.000	7.478	4.861	5.670	6.242	6.124	4.008	5.204
N	4.604	1.754	5.086	5.492	4.458	5.678	4.892	5.488	5.322
P	2.457	5.982	4.525	5.247	3.668	5.034	6.266	6.116	4.474
Q	5.437	6.254	5.405	4.793	5.271	6.439	5.157	4.943	4.904
R	5.239	2.891	4.418	4.813	4.261	3.451	3.211	5.200	3.436
S	5.611	4.095	5.729	5.863	4.529	5.149	5.397	6.428	5.764
T	5.425	5.611	4.486	5.492	4.193	6.155	5.101	4.861	5.512
V	6.017	6.345	5.382	5.622	5.575	6.218	5.997	3.858	9.980
W	5.871	4.497	7.020	5.168	6.621	3.487	7.351	6.246	2.271
Y	8.358	5.322	6.909	5.011	7.016	5.137	6.345	6.672	2.121

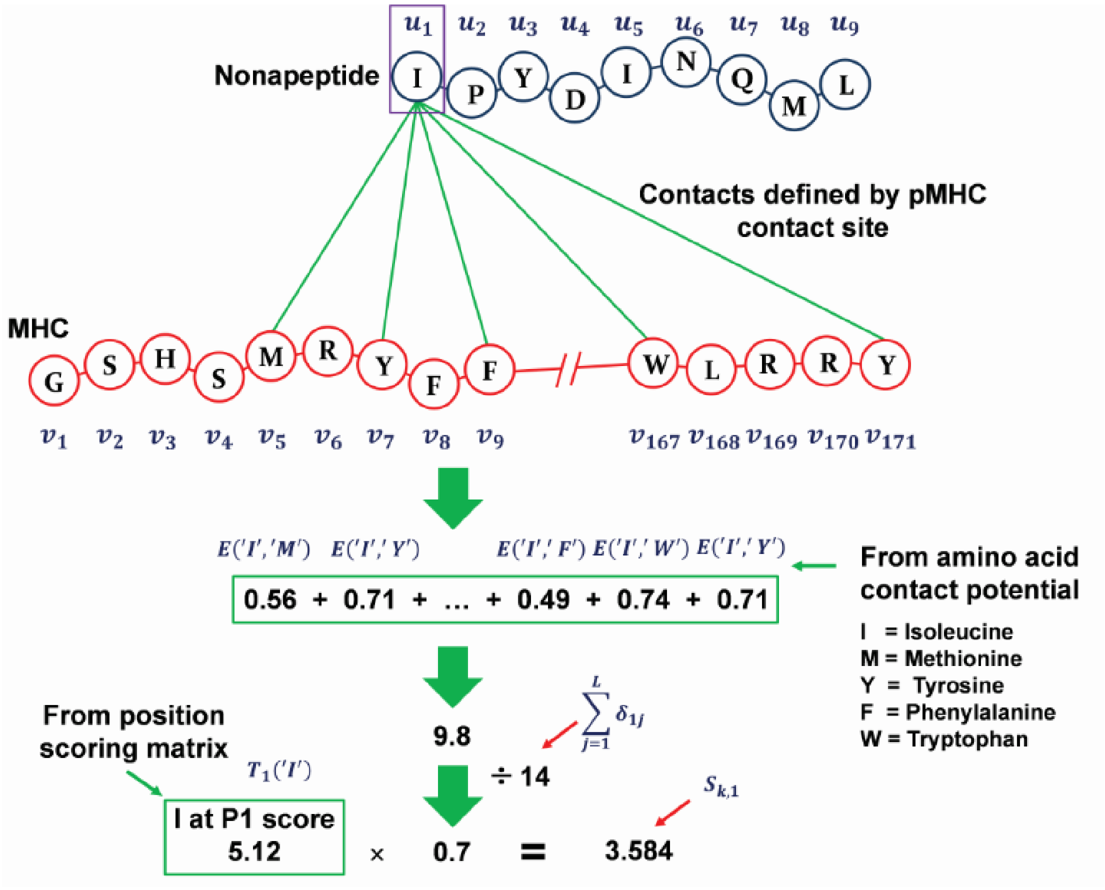

u_1 u_2 u_3 u_4 u_5 u_6 u_7 u_8 u_9
Nonapeptide
I P Y D I N Q M L
Contacts defined by pMHC contact site
MHC
G S H S M R Y F F W L R R Y
v_1 v_2 v_3 v_4 v_5 v_6 v_7 v_8 v_9 v_{167} v_{168} v_{169} v_{170} v_{171}
$E('I','M')$ $E('I','Y')$ $E('I','F')$ $E('I','W')$ $E('I','Y')$
0.56 + 0.71 + … + 0.49 +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

Figure 3.2 Our peptide data-encoding scheme, using the first position of a nonapeptide as an example.

Our peptide-encoding scheme was compared with binary peptide-encoding and with four amino acid descriptors, as shown in Table 3-3 using the dataset reported by Bi and colleagues (supplementary information for Table S2 in [114]). This dataset consists of 1,998 quantitative affinity-known HLA-A*02:01-restricted nonapeptides. The dataset was randomly partitioned into a training set containing 1,500 nonapeptides for estimating predictive models using the SVM, and a test set containing 498 nonapeptides for validating the models. For our peptide-encoding scheme, the positional scoring matrix (Table 3-2) was trained based on the external dataset downloaded from IEDB, consisting of 500 nonapeptides restricted to the HLA-A*02:01 allele. These nonapeptides were included in neither training nor test sets. For the binary peptide-encoding, each amino acid was encoded as a binary vector of length 20, resulting in a vector of length 180 for a nonapeptide. In case of using amino acid descriptors, the length of an encoded vector would be equal to *M* times larger than the length of descriptor vectors. The performances of the data-encoding schemes were evaluated in classification tasks, using a 10-fold cross validation. Throughout our experiments, the parameter C (cost of constraint violation) and the type of kernel used for the SVM were 1 and the radial basis kernel, respectively. The class for each nonapeptide was determined by using an IC_{50} affinity cutoff at 500 nM. Nonapeptides with an affinity less than 500 nM were considered to be binders and non-binders otherwise. The study by Moutaftsi et al. [115] showed that 90% of epitopes that could stimulate CTL responses bound to MHC with affinities lower than 500 nM. The predictive performance is evaluated using five measures: overall accuracy (ACC), sensitivity (sens), specificity (spec), F-score (F1), and area under receiver operating characteristic curve (AUC). ACC, sens, spec, and F1 are defined as

$$\mathrm{ACC} = \frac{\mathrm{TP+TN}}{\mathrm{TP+TN+FP+FN}},$$
$$\mathrm{sens} = \frac{\mathrm{TP}}{\mathrm{TP+FN}},$$
$$\mathrm{spec} = \frac{\mathrm{TN}}{\mathrm{FP+TN}},$$
$$\mathrm{F1} = \frac{2\times\mathrm{TP}}{\left((2\times\mathrm{TP})+\mathrm{FN}+\mathrm{FP}\right)},$$

where TP, FP, TN, and FN are the numbers of overall true positives, false positives, true negatives, and false negatives, respectively.

3.2.2 Validation of predictive models using benchmark datasets

The performance of EpicCapo was validated by using benchmark datasets of 34 MHC-I alleles provided by Peters et al. [61]. In this experiment, the positional scoring matrices were trained based on training data according to the cross validation technique. 20 iterations of 5-fold cross validation were conducted to evaluate AUCs for EpicCapo. We compared the results of our method with those of ARB, NetMHC, SMM, and $\mathrm{SMM}^{\mathrm{PMBEC}}$.

EpicCapo was further developed as $\mathrm{EpicCapo}^{+}$ by selecting AAPPs. Each encoded allele dataset was initially separated into 40 datasets according to the 40 AAPPs. The classification task was performed for each dataset to calculate AUC using the SVM and the same parameters as EpicCapo. Then, 40 datasets were ranked by AUC from highest to lowest. Next, the classification task was performed again by adding the datasets of AAPPs one by one based on their rank. Finally, the optimal subset of AAPPs that led to the highest AUC was identified for each allele. The average AUCs of all alleles as calculated from EpicCapo$^+$ were compared with those from EpicCapo and other methods using paired *t*-tests (two-tailed). For each allele, the AUCs from 20 iterations of 5-fold cross validation of EpicCapo and EpicCapo$^+$ were compared with the maximum AUC among other methods by using *t*-tests (one-tailed, significance level = 0.01).

3.2.3 Improving the performance of HLA-A-nonapeptide binding predictive models

To increase the performance of our predictive models, the positional scoring matrices used in this experiment were trained based on datasets containing larger number of nonapeptides. These matrices are available at [116]. After encoding 14 HLA-A allele datasets using the downloaded matrices, EpicCapo$^+$ was performed again to identify optimal subsets of AAPPs therein. We used the Relief-F algorithm [117] implemented in the machine learning software Weka [118] to perform the feature selection task, ranking the features according to their importance in discriminating the MHC binder peptides from the non-binder ones. The default parameters provided by Weka were used, and a 5-fold cross validation was conducted for evaluating feature importance. The best feature subsets were constructed by adding the features, one by one, from the top-ranked feature to the last one in the classification task using the SVM. The AUC gradually increased with the addition of features, until it reached the highest value. Features after this point were considered irrelevant and ignored. We named this method, accompanied with the Relief-F algorithm, EpicCapo^{+REF}.

3.2.4 Identification of candidates of promiscuous epitopes

EpicCapo^{+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). These protein sequences were downloaded from the NCBI website (http://www.ncbi.nlm.nih.gov/). The nonapeptides were generated from these sequences by using a nonamer sliding window. Next, all of the generated nonapeptides were used as inputs in EpicCapo^{+REF} predictive models. These models were estimated by using 14 HLA-A allele datasets, and each model was specific for each allele type.

The identified epitopes were validated by cross-checking with the results of immunological experiments.

3.3 Results and discussion

3.3.1 Comparison of peptide-encoding schemes

We compared our peptide-encoding scheme with binary peptide-encoding and with four amino acid descriptors (Table 3-3). The results of the comparison of the peptide-encoding schemes (Table 3-4) showed that EpicCapo performed better than others in the classification tasks. It achieved the highest average area under the curve (AUC; 0.882), followed by binary encoding (0.879), DPPS (0.878), FASGAI (0.874), z-scale (0.858), and ISA/ECI (0.796) schemes. All of standard deviations were less than 0.01. A comparison of receiver operating characteristic (ROC) curves is shown in Figure 3.3.

Although EpicCapo used the largest number of features ($M \times K = 360$)—higher than binary encoding (180), DPPS (90), FASGAI (54), z-scale (45), and ISA/ECI (18)—we confirmed that its high performance was not due to a larger number of features. In our study, the training dataset was separated into 40 datasets corresponding to 40 AAPPs. Each dataset consisted of 9 features. The classification functions were fitted to these datasets, and after that the AAPPs were ranked by AUC. The results, as shown in Table 3-4, suggested that even by using only three top-ranked AAPPs (27 features in total), the classification performance values are comparable to those obtained by using all AAPPs. These three top-ranked AAPPs were MICC010101, SIMK990101, and SIMK990105 (see Appendix B). They have been previously used in identifying native-like protein structures [119, 120], and were also identified as important AAPPs in our accompanying experiments.

Table 3-3 Amino acid descriptors acknowledged in this study.

Descriptor	Type	Technique used	# of vector	Ref.
DPPS	physicochemical	principal component	10	[103]
FASGAI	physicochemical	factor analysis (FA)	6	[121]
z-scale	physicochemical	PCA and partial least	5	[122]
ISA/ECI	quantum-chemical	-	2	[123]

Table 3-4 Classification result of peptide-encoding schemes.

Method	# of features	10-fold cross validation on training dataset only					Holdout method using training dataset and testing dataset				
		sens	spec	F1	ACC	AUC	sens	spec	F1	ACC	AUC
EpicCapo	360	**0.883** ± 0.005	0.792 ± 0.006	**0.886** ± 0.003	0.841 ± 0.004	0.915 ± 0.001	0.883	0.744	**0.831**	0.815	**0.882**
EpicCapo (3 AAPPs*)	27	0.876 ± 0.005	**0.821** ± 0.005	0.862 ± 0.003	**0.848** ± 0.003	**0.916** ± 0.001	0.855	**0.777**	0.828	**0.817**	0.878
DPPS	90	0.865 ± 0.005	0.760 ± 0.007	0.834 ± 0.004	0.816 ± 0.004	0.888 ± 0.001	0.868	0.697	0.807	0.785	0.878
FASGAI	54	0.847 ± 0.004	0.761 ± 0.004	0.825 ± 0.003	0.801 ± 0.003	0.882 ± 0.001	0.840	0.730	0.803	0.787	0.874
z-scale	45	0.847 ± 0.005	0.732 ± 0.005	0.815 ± 0.004	0.793 ± 0.004	0.873 ± 0.002	0.848	0.676	0.788	0.765	0.858
ISA/ECI	18	0.799 ± 0.005	0.652 ± 0.005	0.760 ± 0.003	0.731 ± 0.003	0.797 ± 0.001	0.829	0.643	0.766	0.739	0.796
Binary encoding	180	**0.883** ± 0.005	0.721 ± 0.006	0.831 ± 0.003	0.807 ± 0.003	0.883 ± 0.002	**0.887**	0.705	0.820	0.799	0.879

Means and standard deviations were calculated by 20 iterations of 10-fold cross validation.

Underlined values represent the highest performance.

sens = sensitivity; spec = specificity; F1 = F-score; ACC = accuracy; AUC = area under the curve.

*These three top-ranked AAPPs were MICC010101, SIMK990101, and SIMK990105 (see Appendix B)

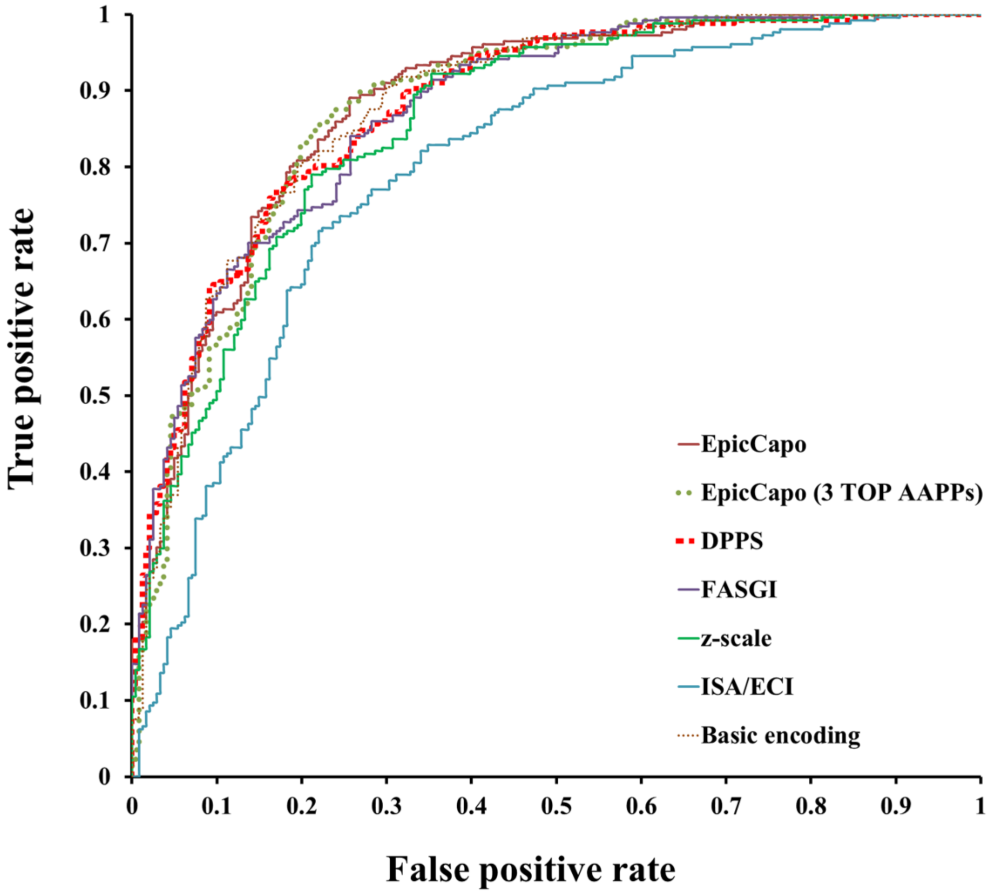


Figure 3.3 ROC curves of peptide-encoding schemes evaluated on a test set.

3.3.2 Classification results of benchmark datasets

We applied EpicCapo to benchmark datasets of 34 MHC-I alleles [61]. As shown in Table 3-5, NetMHC performed the best, ahead of ARB, SMM, and SMM^{PMBEC}. For EpicCapo, average AUCs were lower than in NetMHC (0.1%–3.4%) in 13 allele datasets and were higher than those in NetMHC (0.1%–9.3%) in 21 allele datasets when using all of the 40 AAPPs (360 features). Almost all of standard deviations were low except several alleles with results of standard deviation larger than 0.01. However, if more data are available, these standard deviations can be decreased. To improve the performance of our method, we developed EpicCapo$^+$ by selecting an appropriate subset of AAPPs. As seen in Table 3-5, 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. The IDs of AAPPs used for estimating the predictive models of EpicCapo^{+} are shown in Table 3-6.

Table 3-5 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).

Table 3-6 Optimal subsets of AAPPs identified by EpicCapo^{+} using 34 benchmark datasets.

MHC	IDs of AAPP used
HLA-A*01:01	11,14,20,24,28,33
HLA-A*02:01	9,11,14,24,26,28,31
HLA-A*02:02	14,24,28
HLA-A*02:03	3,9,11,14,19,24,25,26,28,29,31
HLA-A*02:06	9,11,13,14,19,21,22,24,25,26,28,31
HLA-A*03:01	9,11,14,20,24,26,28,33
HLA-A*11:01	11,14,26,28
HLA-A*24:02	11,14,20,24,28,31,33
HLA-A*26:01	14,28
HLA-A*29:02	5,9,11,14,19,20,22,24,26,28,33
HLA-A*31:01	1,9,11,14,20,24,26,28,31,33,38
HLA-A*33:01	1,11,14,20,24,26,28,33
HLA-A*68:01	11,14,20,26,28
HLA-A*68:02	1,2,9,11,14,19,20,22,24,26,28,33,34,39
HLA-B*07:02	1,9,11,14,20,24,26,28,33
HLA-B*08:01	4,14,18,20,40
HLA-B*15:01	14,24,26,28
HLA-B*18:01	3,14,20,24,26,28
HLA-B*27:05	9,14,20
HLA-B*35:01	14,28
HLA-B*40:02	11,14,24,28
HLA-B*44:02	9,14,20,28,32
HLA-B*44:03	13,14,20,28,33,38,39
HLA-B*51:01	6,11,14,20,24,26,33,36,38,39
HLA-B*53:01	11,14,20,24,28,33
HLA-B*54:01	1,9,11,14,20,24,26,28,33
HLA-B*57:01	5,6,8,12,22,23,24,25,27,31,37
HLA-B*58:01	14,28
H-2 Db	1,11,14,24,28
H-2 Dd	11,14,28
H-2 Kb	11,14,28
H-2 Kd	1,11,12,14,19,24,26,28,33
H-2 Kk	14,28
H-2 Ld	10,11,14,16,18,20,21,23,24,26,28,33

3.3.3 Improved HLA-A-nonapeptide binding predictive models

In this experiment, EpicCapo^{+} was further developed as EpicCapo^{+REF} to improve the predictive performance and identify important positions of nonapeptides in pMHC binding (Section 3.2.3). The IDs of AAPPs used in EpicCapo^{+REF} are shown in Table 3-7 (for more details on AAPPs, see Appendix B). The most important AAPPs identified by EpicCapo^{+} were IDs 14 (MICC010101) and 28 (SIMK990105), which were selected in 13 out of 14 alleles. IDs 11 (KESO980102) and 26 (SIMK990103) were also considered to be important, because they were selected in 9 out of 14 alleles. From previous studies that used AAPPs in MHC-I epitope prediction, AAPP IDs 19 (MIYS960102) and 2 (BETM990101) proved to be important in peptide-MHC binding prediction [104, 124, 125]. In our study, however, BETM990101 was not selected for any allele dataset, and MIYS960102 was chosen for only two alleles (A*02:03 and A*02:06). In a report by Schueler-Furman et al. [124], KESO980102 was also tested and compared with MIYS960102; however, there was no significant improvement in the predictive performance. Therefore, it is interesting that MICC010101, SIMK990105, KESO980102, and SIMK990103 were important for generating better predictive models in our study.

Table 3-7 Optimal subsets of AAPPs and numbers of selected features identified by EpicCapo^{+REF} using 14 HLA-A allele datasets.

Allele	AUC of EpicCapo^{+REF}	IDs of AAPP used	# of features selected
A*01:01	0.980	1,11,14,20,24,26,28,33	72
A*02:01	0.958	9,11,14,24,26,28,31	62
A*02:02	0.913	14,28	18
A*02:03	0.925	3,9,11,14,19,24,25,26,28,29,31,33	104
A*02:06	0.926	1,3,9,11,13,14,18,19,21,22,24,25,26,27 ,28,31,34,38,39	141
A*03:01	0.946	11,14,20,24,26,28,33	58
A*11:01	0.956	11,14,26,28	35
A*24:02	0.877	5,6,14,24,28,31	31
A*26:01	0.960	14,28	18
A*29:02	0.955	5,8,9,20,33	23
A*31:01	0.940	11,14,20,26,28,33	46
A*33:01	0.940	14,28	17
A*68:01	0.904	11,14,20,26,28,33	40
A*68:02	0.913	1,9,11,14,20,22,24,26,28,33,39	79
Average	0.935		

We further investigated the generated features according to the selected subset of AAPPs. In our peptide-encoding scheme, nine features were generated from one AAPP, corresponding to the nine amino acid positions in the nonapeptide. Previous studies have indicated that not all positions were important in pMHC binding [27–29, 103]. Therefore, some features corresponding to specific positions could be removed to improve the predictive performance.

The Relief algorithm [117] was employed in our study to rank the features according to their importance in separating the nonbinding peptides from the binding ones. The ranking results showed that the ten top-ranked features correspond to positions 9 and 2 in most of the alleles, followed by positions 3, 1, or 7 (see Appendix C). As indicated in Tables 3-5 and 3-7, the overall AUC value of EpicCapo^{+REF} was higher than that of EpicCapo^{+}; however, it was still slightly lower than that of NetMHC in the A*01:01 and A*02:06 alleles. In summary, EpicCapo^{+REF} performed better than other methods, with an average AUC of 0.935. Table 3-7 also shows the number of selected features after employing the Relief-F algorithm. These numbers were different for specific alleles. For the A*01:01, A*02:02, and A*06:01 alleles, no features were removed. However, for the A*02:06, A*24:02, A*29:02, and A*68:02 alleles, 20 or more features were removed. Interestingly, features corresponding to positions 5 and 8, which have previously been considered to not significantly contribute to HLA binding potentials, were still included in some of the selected feature subsets. Therefore, we assumed that features corresponding to different positions are not independent, and that all features from all positions should be required input to estimate the model with the highest-performance (see Appendix C).

3.3.4 Candidates of promiscuous epitopes for a development of influenza A viral vaccines

Since EpicCapo^{+REF} performed better than the other existing methods when testing with 14 HLA-A allele datasets, it was further used to find candidates of promiscuous epitopes from influenza A viral sequences. Epitopes from protein sequences of H1N1 (A/PR/8/34), H3N2 (A/Aichi/2/68), H1N1 (A/New York/4290/2009), and H5N1 (A/Hong Kong/483/97) were identified using EpicCapo^{+REF}. The prediction results of all influenza A strains categorized into specific alleles are shown in Table 3-8. All 14 alleles were assigned to supertype groups using the supertype classification defined by previous studies [107–110]. The A*01:01 and A*26:01 alleles were assigned to the

A1 group. The A*29:02 allele was assigned to an unidentified group. As shown in Table 3-8, there are a small number of predicted positive peptides in the A1 supertype. For example, in case of H1N1 (A/PR/8/34), only one peptide was identified as positive for the allele A*26:01. In contrast, there were quite high numbers of predicted positive peptides in the A2, A24, and A3 supertypes. Even the A*29:02 allele, which was assigned to an unidentified group, had a higher number of predicted positive peptides than those in the A1 group. Based on our findings, when promiscuous epitopes were identified from the overlapping epitopes of four Influenza A viral strains (Appendix D), the A1 group rarely shared peptides with other groups. As shown in Appendix D, the A*01:01 allele shared only one peptide (YSHGTGTGY) with A*29:02, and the A*26:01 allele shared the peptide DTVNRTHQY with A*29:02 and A*68:01. Moreover, the A*29:02 allele also shared peptides with the A2 and A3 groups: e.g., SMELPSFGV and QTYDWTLNR, respectively (Appendix D). Therefore, A*29:02 can be considered as a special allele that links A1, A2, and A3 together. Furthermore, Doytchinova et al. [111] assigned A*29:02 to the A3 group. However, we did not find overlapping epitopes from the four Influenza A viral strains in the A*24:02 allele assigned to the A24 group. This suggested that A*24:02 itself is different from other alleles considered here, and this might be the reason why most of the previous studies assigned it separately to the A24 group [107–110]. As shown in Appendix D, 51 peptides (67.1%) of the total 76 epitopes 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 new vaccine development.

Our results are in agreement with the study by Uchida [62], which identified promiscuous epitopes from influenza A H1N1 (A/PR/8/34), H3N2 (A/Aichi/2/68), H1N1 (A/New York/4290/2009), and H5N1 (A/Hong Kong/483/97). Uchida found experimentally confirmed CTL epitopes in the A2 group. In our results, the epitopes identified by EpicCapo^{+REF} in the A2 group were consistent with them (Table 3-9). In addition, we found promising candidates of promiscuous epitopes also for the A1 and A3 groups as shown in Appendix D.

Table 3-8 Prediction results of EpicCapo^{+REF} using four influenza A strains categorized by specific alleles.

Allele	# of predicted positive peptides				Super type
	H1N1 New York/4290/2009	H5N1 Hong Kong/483/97	H1N1 PR/8/34	H3N2 Aichi/2/68	
A*01:01	14	13	6	5	A1
A*26:01	6	9	1	5	A1
A*29:02	103	134	61	161	?
A*02:01	122	160	71	168	A2
A*02:02	302	370	162	391	A2
A*02:03	268	326	144	307	A2
A*02:06	200	250	105	264	A2
A*68:02	198	220	109	277	A2
A*24:02	90	108	50	150	A24
A*03:01	85	94	50	136	A3
A*11:01	162	176	91	229	A3
A*31:01	183	227	110	245	A3
A*33:01	96	117	62	110	A3
A*68:01	263	346	151	325	A3
Total	2092	2550	1173	2773	

Although the overall performance of EpicCapo^{+REF} was high, there are two limitations in the use of this method. The first limitation is the length of input peptides must be equal to 9. In the further study, we will improve EpicCapo^{+REF} to be applicable to peptides with the length of 8–11. The second limitation is that input amino acids must not be special or ambiguous ones. Examples of special amino acids are U (Selenocysteine) and O (Pyrrolysine). Also, examples of ambiguous amino acids are B (Asparagine or aspartic acid), Z (Glutamine or glutamic acid), and J (Leucine or Isoleucine). EpicCapo^{+REF} are not applicable with these amino acids since they are not included in AAPPs.

Table 3-9 Comparison of epitopes identified by EpicCapo^{+REF} with the broadly protective influenza A viral epitopes identified by Uchida [62].

Viral strain	**CTL epitopes identified by** [62]	**Shared alleles identified by EpicCapo^{+REF}**
H1N1 (A/PR/8/34)	GILGFVFTL	A*02:01, A*02:02, A*02:03, A*02:06
	IILKANFSV	A*02:01, A*02:02, A*02:03, A*02:06,
	GMFNMLSTV	A*02:01, A*02:02, A*02:03, A*02:06
H3N2 (A/Aichi/2/68)	GILGFVFTL	A*02:01, A*02:02, A*02:03, A*02:06
	VMLKANFSV	A*02:01, A*02:02, A*02:03, A*02:06
	GMFNMLSTV	A*02:01, A*02:02, A*02:03, A*02:06
H1N1 (A/NewYork/4290/2009)	GILGFVFTL	A*02:01, A*02:02, A*02:03, A*02:06
	IVLKANFSV	A*02:01, A*02:02, A*02:06, A*68:02
	GMFNMLSTV	A*02:01, A*02:02, A*02:03, A*02:06
H5N1 (A/Hong Kong/483/97)	GILGFVFTL	A*02:01, A*02:02, A*02:03, A*02:06
	IILKANFSV	A*02:01, A*02:02, A*02:03, A*02:06,
	GMFNMLSTV	A*02:01, A*02:02, A*02:03, A*02:06

3.4 Conclusions

In this study, we have developed a novel method for epitope prediction. Peptides were encoded numerically, combining information of pMHC contact sites and amino acid pairwise contact potentials, accompanied by an SVM for estimating the predictive model. Our method achieved high performance in testing with 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 identifying promising candidates of promiscuous epitopes for the development of new vaccines.

Chapter 4 PAAQD: Predicting immunogenicity of MHC class I binding peptides using amino acid pairwise contact potentials and quantum topological molecular similarity descriptors

Prediction of peptide immunogenicity is a promising approach for novel vaccine discovery. Conventionally, epitope prediction methods have been developed to accelerate the process of vaccine production by searching for candidate peptides from pathogenic proteins. However, recent studies revealed that peptides with high binding affinity to major histocompatibility complex molecules (MHCs) do not always result in high immunogenicity. Therefore, it is promising to predict the peptide immunogenicity rather than epitopes in order to discover new vaccines effectively. To this end, we developed a novel T-cell reactivity predictor which we call PAAQD. Nonapeptides were encoded numerically, using combining information of amino acid pairwise contact potentials (AAPPs) and quantum topological molecular similarity (QTMS) descriptors. Encoded data were used in the construction of our classification model. Our numerical experiments suggested that the predictive performance of PAAQD is at least comparable with POPISK, one of the pioneering techniques for T-cell reactivity prediction. Also, our experiment suggested that the first and eighth positions of nonapeptides are the most important for immunogenicity and most of the anchor residues in epitope prediction were not important in T-cell reactivity prediction. The R implementation of PAAQD is available at http://pirun.ku.ac.th/~fsciiok/PAAQD.rar.

4.1 Introduction

The immune system is one of the most complex mechanisms that defend an organism from infections. After antigen presenting cells (APCs) have phagocytosed pathogens, endogenous proteins from pathogens are cleaved into small peptides by a proteasome. Cleaved peptides are then transported into the endoplasmic reticulum by transporter associated with antigen processing (TAP) and selectively bound to MHCs which are HLAs in humans. At this step, pMHC complexes are translocated to the cell surface and recognized by CTLs via TCRs. Peptides are considered to be immunogenic if an immune response is successfully activated [126].

As mentioned in the chapters 1, 2, and 3, epitope prediction is extensively studied in immunoinformatics for decades [21]. Recently, most of the successful methods for the epitope prediction are applications of machine learning techniques. However, the problem of peptide immunogenicity prediction for T-cell reactivity is still not widely researched.

Prediction of peptide immunogenicity is a promising approach for the design of novel vaccines [12, 127, 128]. Traditionally, the process of developing a new vaccine is time-consuming and laborious. Computational methods for immunogenicity prediction can be effectively applied to scanning for candidate peptides; thus they have a potential to identify new promising vaccines. Conventionally, epitope prediction methods have been used to search for candidate peptides from pathogenic proteins. Predicted peptides with high binding affinity to the MHC-I were supposed to be immunogenic peptides. However, recent studies revealed that predicted peptides with high binding affinity to the MHC-I molecules did not always result in high T-cell reactivity [51, 129]. Conversely, predicted peptides with low binding affinity to the MHC-I do not necessarily result in low immunogenicity [130]. In addition, other factors such as the trimming mediated by the endoplasmic reticulum aminopeptidase (ERAAP) were more strongly correlated to T-cell immune responses than MHC binding affinities [52]. Therefore, immunogenicity could not be accurately determined by existing epitope prediction methods.

Constructing a model for predicting peptide immunogenicity is more difficult than predicting epitope. The immunogenicity does not only depend on the particular allele of HLA and the type of TCR in host immune system but is also governed by negative T-cell selection (central tolerance). The central tolerance is defined as the property of the whole proteome and cannot casually be learned by machine learning approaches [53–55].

The studies of complex protein crystal structures have been conducted to uncover TCR–pMHC binding mechanisms [56, 131, 132]. Positions 4, 6, and 8 of nonapeptides were reported to have an impact on TCR–pMHC binding. The substitution of lysine to arginine at the position 4 led to a better fit of TCR–pMHC, whereas the mutation at the position 6 increased dissociation rate of TCR–pMHC [56]. Additionally, the side chain interaction at the position 8 was crucial in TCR-peptide binding and a hydrogen bond was formed by the complementarity determining region (CDR) at this position [57]. However, precise explanations of TCR–pMHC binding mechanisms for all HLA alleles are still not concrete. This because a large number of resolved crystal structures are currently not available. Figure 4.1 shows the resolved TCR–pMHC complex crystal structure (PDB ID: 2AK4).

The first predictor for T-cell reactivity is POPI [58]. POPI used 23 informative physicochemical properties collected from the AAindex database [59] for encoding peptides and applied the SVM as a classifier. The second predictor is POPISK [55]. POPISK used the SVM with string kernels. POPISK outperformed POPI. Besides, the importance of amino acid positions of the peptides with length 9 was evaluated by removing features corresponding to each position. The positions whose deletion significantly decreased predictive performance were considered as important positions. POPISK identified six important positions (1, 4, 5, 6, 8, and 9) for T-cell reactivity.

In this chapter, we introduced our novel T-cell reactivity predictor named PAAQD. Peptides were encoded numerically, using combining information of AAPPs [59] and QTMS descriptors [63]. Previous studies have used AAPPs in the MHC-I epitope prediction [104, 124, 125]. Those studies focused on using the AAPPs of Miyazawa and Jernigan (1996) [133] and Betancourt and Thirumalai (1999) [134]. In our study, 40 AAPPs were applied, including the information from two AAPPs in their reports. The QTMS descriptors were used in constructing the quantitative structure-activity relationship (QSAR) model for predicting pMHC binding affinities. The performance of using the QTMS descriptors was comparable with other methods reported at that time [63]. The quantum chemistry methods were applied to study variations in the electrostatic field of pMHC complexes. The analyzed data provide more insights of the interactions between peptides and the MHC [135]. Therefore, the use of AAPPs and molecular quantum properties such as QTMS descriptors in T-cell reactivity prediction is a promising approach to uncover TCR–pMHC binding mechanisms. Simultaneously, new immunogenic peptides could be identified.

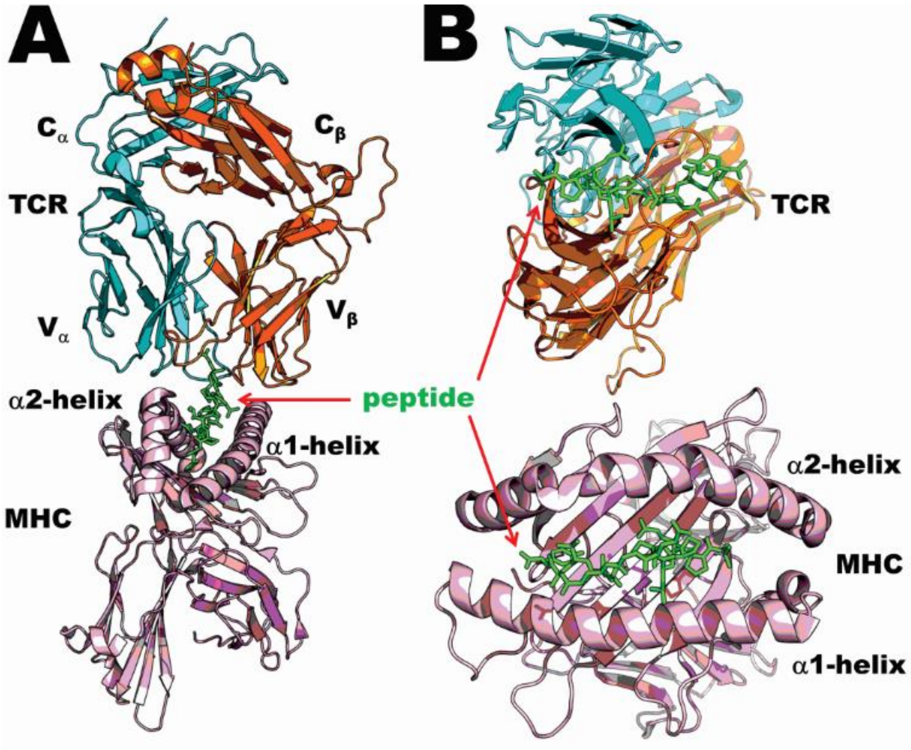


Figure 4.1 The structure of TCR–pMHC complex (PDB ID: 2AK4).

(A) The LPEP peptide is fitted inside the MHC and TCR binding clefts. (B) A closer view of a TCR-peptide and pMHC binding complex.

PAAQD's performance was evaluated by using the IMMA2 dataset published by Tung et al. (2011) [55] and compared with the two existing T-cell reactivity predictors, POPI and POPISK. We evaluated the importance of positions by removing features corresponding to the specific position of nonapeptides. The importance of AAPPs and QTMS descriptors was evaluated in the same manner by removing features corresponding to specific AAPP or QTMS descriptor. The dataset of HLA-A2 peptides collected from IEDB was used as the validation dataset to test for the predictive stability of PAAQD. This dataset consists of immunogenic and non-immunogenic peptides that have not been presented in the IMMA2 dataset. PAAQD showed comparable performance to POPISK. Positions 1 and 8 were identified as important positions in T-cell reactivity by using our method. This result was concordant with a previous study [57] and the POPISK results that the positions 1 and 8 were crucial in the response of T-cell reactivity. We speculate that our method may be useful in identifying immunogenic peptides for the development of new vaccines.

4.2 Materials and 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) [55]. This dataset consists of 558 immunogenic and 527 non-immunogenic nonapeptides associated with the HLA-A2 supertype (Appendix E). All of nonapeptides were retrieved from three databases: MHCPEP [44], SYFPEITHI [42], and IEDB [41]. The second dataset was collected from IEDB database by selecting nonapeptides that were specific to the HLA-A2 supertype. The second dataset consists of 278 immunogenic and 101 non-immunogenic nonapeptides (Appendix F). All of these nonapeptides are not included in the IMMA2 dataset. The sequence preference of the dataset we collected is different from the IMMA2 dataset. This dataset was used for evaluation of the predictive stability for the difference of datasets. We focused on IEDB since this database contains more entries than other existing immunogenic peptide databases [136]. IEDB encompasses reliable data from biotechnological and pharmaceutical companies, as well as direct submissions from research programs and partners.

4.2.2 Peptide encoding

Peptides were encoded numerically using our peptide encoding scheme. The encoding scheme was defined by utilizing the information of the pMHC contact sites retrieved from the international ImMunoGeneTics information system, IMGT [60], AAPPs from AAindex [59], and the QTMS descriptors [63].

The reference pMHC contact sites defined by Kaas and Lefranc (2005) [60] were modified by adding more MHC positions as described in chapter 3.2.1. The references and added pMHC contact sites positions were shown in Table 3-1. This information served as a binding template between the peptide and MHC. Subsequently, the AAPPs were used as a representative value for each amino acid pair, consisting of one MHC amino acid and its adjacent nonapeptide amino acid. These amino acid pairs were defined in the pMHC contact sites. In this study, 40 AAPPs were applied. The Appendix B describes details of all AAPPs used.

The used peptide encoding scheme was similar to the one used in chapter 3.2.1. However, the positional scoring matrices $T_i(a)$ are not concerned in this study. 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:

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

where L is the length of the HLA protein 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 the MHC contact each other, and 0 otherwise. Note that $\sum_{j=1}^{L} \delta_{ij}$ is the number of contact sites for the i^{th} amino acid of a nonapeptide (see Table 3-1). Intuitively, this score represents average pair-potential of contact sites. Let K be the number of pair-potential types available, and M be the length of nonapeptides, which is set to 9 throughout this study. Using this scoring scheme, we transform a nonapeptide n into a $M \times K$ -dimensional numerical vector, whose $(M(k-1)+i)^{\mathrm{th}}$ element is $S_{k,i}^{(n)}$. For example, the encoded nonapeptides consist of 9 features if the number of pair-potential types is 1 and 360 features if the number is 40. Figure 4.2 illustrates an example of the data-encoding scheme for the first position of a nonapeptide. Each encoded peptide was then combined with the corresponding feature vector constructed from using QTMS descriptors [63]. There are four types of QTMS descriptors used in this study (see Table 4-1). When these descriptors were applied, the feature vector of length 189 was produced for one nonapeptide. Therefore, the final feature vector for one nonapeptide of length 549 was generated when combining feature vectors corresponding to AAPPs and QTMS descriptors.

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

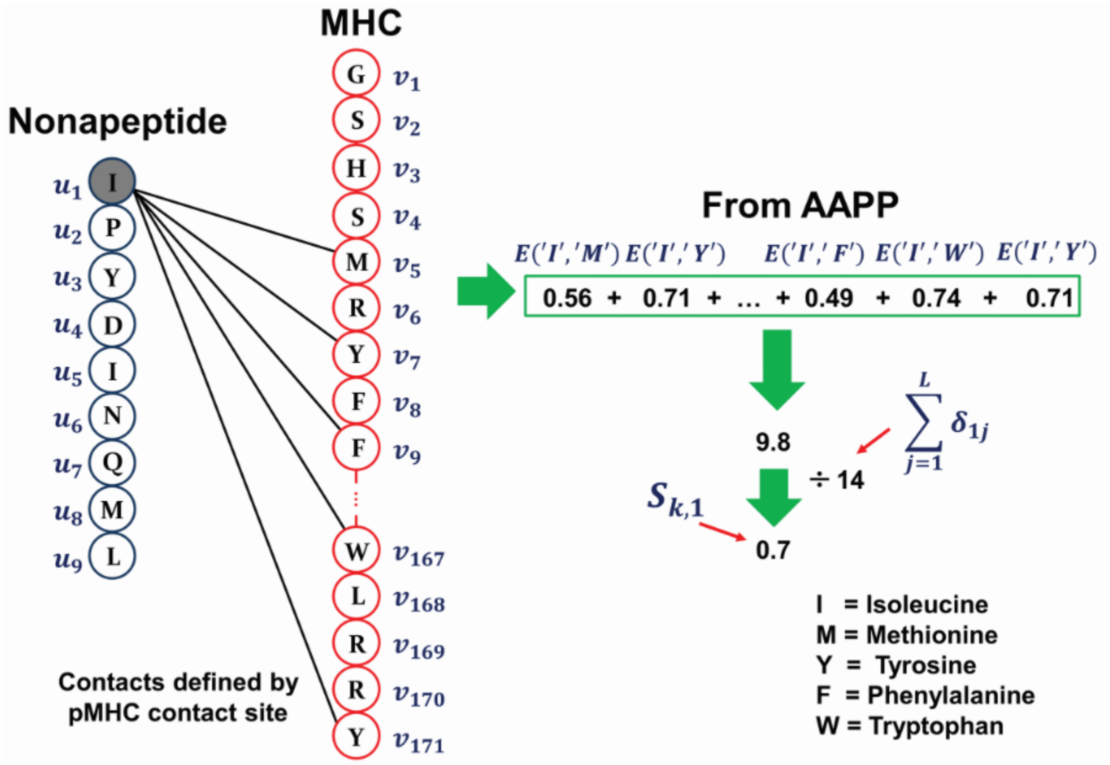


Figure 4.2 Our peptide data-encoding scheme for the first position of the nonapeptide.

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 [118]. 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; overall accuracy (ACC), Matthew's correlation coefficient (MCC), and area under receiver operating characteristic curve (AUC). ACC is defined in chapter 3.2.1 and MCC are defined as

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

where TP, FP, TN, and FN are the number of overall true positives, false positives, true negatives, and false negatives, respectively. The average and standard deviations of ACC, MCC, and AUC were evaluated by repeating 10-fold cross validation 20 times, independently. We compared our method with POPI [58] and POPISK [55]. Additionally, the encoded data were separated into two datasets. The first dataset includes 360 features corresponding to AAPPs. The second dataset includes 189 features corresponding to QTMS descriptors. Each dataset was input to the random forest to evaluate the performance using 20 iterations of 10-fold cross validation

4.2.4 Evaluation of positional importance

To uncover TCR-pMHC binding mechanisms in T-cell recognition, it is essential to identify nonapeptide positions that have a significant impact on the binding force field. Previous studies analyzed the importance of positions based on protein crystal structures of TCR-pMHC complexes. However, the discovery was specific to a small number of HLA alleles since the number of resolved crystal structures of TCR-pMHC complexes are currently not enough [56, 57, 131, 132].

In this study, we assessed the importance of each position using the method described in [55]. The decreases in the predictive performance arisen from removing features corresponding to the specific position were evaluated. The position that led to a significant decrease in the performance was considered as an important position. To evaluate the positional importance of nonapeptides, nine datasets were generated from encoded data by removing features corresponding to each position in nonapeptides. The PAAQD performance was then evaluated by using 20 independent iterations of 10-fold cross validation in the same manner as used on original encoded data. We compared our PAAQD with POPISK. To avoid the influence of the difference of classifiers, we used the SVM, the same classifier as POPISK. The SVM implementation used in this experiment is the one in the R package kernlab [112].

4.2.5 Evaluation of the importance of AAPPs and QTMS descriptors

In our peptide-encoding scheme, 40 AAPPs were used. Some AAPPs or QTMS descriptors might be redundant in TCR-pMHC binding mechanisms. Therefore, features corresponding to the specific AAPP or QTMS descriptor were removed from the encoded data. This generated 40 and 4 datasets when removing features corresponding to the specific AAPP and QTMS descriptor, respectively. Afterwards, the performance of PAAQD was evaluated on these reduced datasets using 20 independent iterations of 10-fold cross validation.

4.2.6 Prediction of peptide immunogenicity using the validation dataset

The final model for peptide immunogenic 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 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

To investigate effects of AAPPs and QTMS descriptors on the performance of PAAQD, we conducted five experiments based on the IMMA 2 dataset. The first and second experiments were conducted by using POPI-modified and POPISK respectively. In the third experiment, the performance was evaluated on the dataset that contains features corresponding to AAPPs only. In the fourth experiment, the performance was evaluated on the dataset that contains features corresponding to QTMS descriptors only. The fifth experiment was conducted by using PAAQD when the performance was evaluated on the dataset that contains features corresponding to both AAPPs and QTMS descriptors. Figure 4.3 shows the performance of five experiments based on the IMMA 2 dataset. This result indicated the comparable performance of PAAQD with POPISK. PAAQD provided 1% higher AUC than POPISK with significance level 0.01 when performing one sample *t*-test of AUCs of PAAQD against the upper bound AUCs of POPISK (0.744). Interestingly, using only encoded features from AAPPs could lead to the highest AUC of 0.75, whereas MCC was 2% lower than PAAQD. Although the performance of using encoded features from QTMS descriptors only was lower than AAPPs in all three measurements, MCC increased by 2% with significance level 0.01 compared with using the combination of both encoded features. Therefore, cooperation between physicochemical properties represented by AAPPs and quantum topological properties represented by QTMS descriptors are promising to provide more insights in TCR-pMHC binding mechanisms.

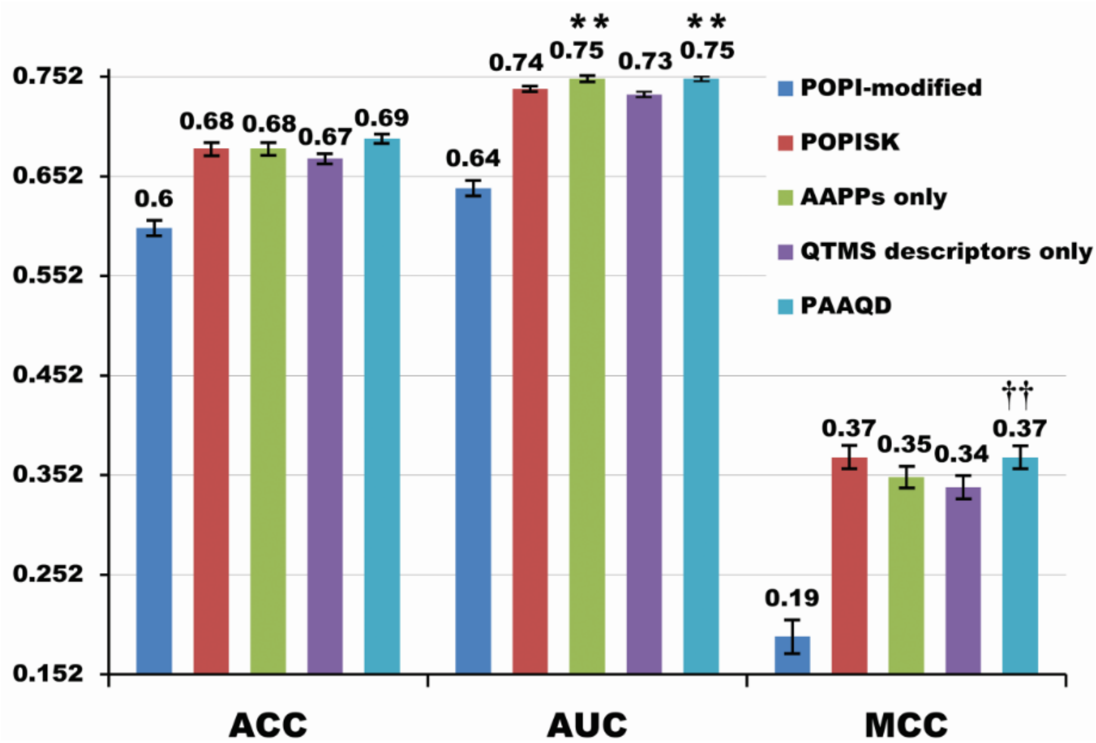


Figure 4.3 Comparison of 20 independent iterations of the 10-fold cross validation performance of POPI, POPISK, encoded features using AAPPs only, encoded features using QTMS descriptors only, and PAAQD.

The symbol ** indicates significance level 0.01 of one sample *t*-test of PAAQD AUCs with upper bound AUCs of POPISK (0.744). The symbol †† indicates significance level 0.01 of two-sample *t*-test between MCCs of PAAQD and using only AAPPs.

4.3.2 The positional importance in peptide immunogenicity

The result of important positions of nonapeptides in T-cell reactivity prediction was shown in Figure 4.4. Removing features corresponding to one of the nine positions for nonapeptides except the position 7 decreased the performance of PAAQD when compared with the use of all positions. Removing features corresponding to the position 7 reduced AUC, though this was not statistically significant. Obviously, deletions of positions 1 and 8 led to more decrease in MCC and ACC than the other seven positions. Previous studies based on the analysis of TCR-pMHC complex crystal structures identified positions 4, 6, and 8 as significant positions in TCR-pMHC binding mechanisms [56, 57]. For the position 1, there was no evidence of its importance in peptide immunogenicity. However, the position 1 was identified as an important position by POPISK [55]. For PAAQD, positions 2, 3, and 7 were less important since small decreases in the performance were observed. These findings are concordant with the result of POPISK.

The positional importance suggested by PAAQD was partially inconsistent with the result of POPISK, especially in positions 4 and 6. Therefore, we speculated that the result could be affected by the type of classifier that we used. The repeated experiments on encoded data using the SVM classifier were conducted to identify important positions. The result is consistent with the positional importance result of POPISK when positions 1, 4, 5, 6, 8, and 9 were strongly decreased the performance (Figure 4.5). Therefore, there was a high possibility that the result of feature importance was affected by the difference of the classifiers. In Figure 4.5, the performance of PPAQD with the SVM when all positions were included was 0.67, 0.73, and 0.35 for ACC, AUC, and MCC, respectively. Standard deviations of ACC, AUC, and MCC were less than 0.007. The cost parameter C used in the SVM was set to 1 and the RBF kernel was used in the training and predicting processes.

Interestingly, both results from PAAQD and POPISK indicated that the position 2, a primary anchor residue in pMHC binding [27], was the least importance in peptide immunogenicity. Similarly, positions 3 and 7, secondary anchor residues in pMHC binding [28, 29] did not strongly decrease the performance of the classification when either position was removed from the dataset. Additionally, recent studies showed that high binding affinity to MHC-I molecules does not always result in high T-cell reactivity [51, 129].

The principle of epitope prediction is based on pMHC binding mechanisms whereas T-cell reactivity prediction is based on TCR-pMHC binding mechanisms. The binding of peptide to MHC and TCR are definitely different. Therefore, both techniques cannot be used interchangeably.

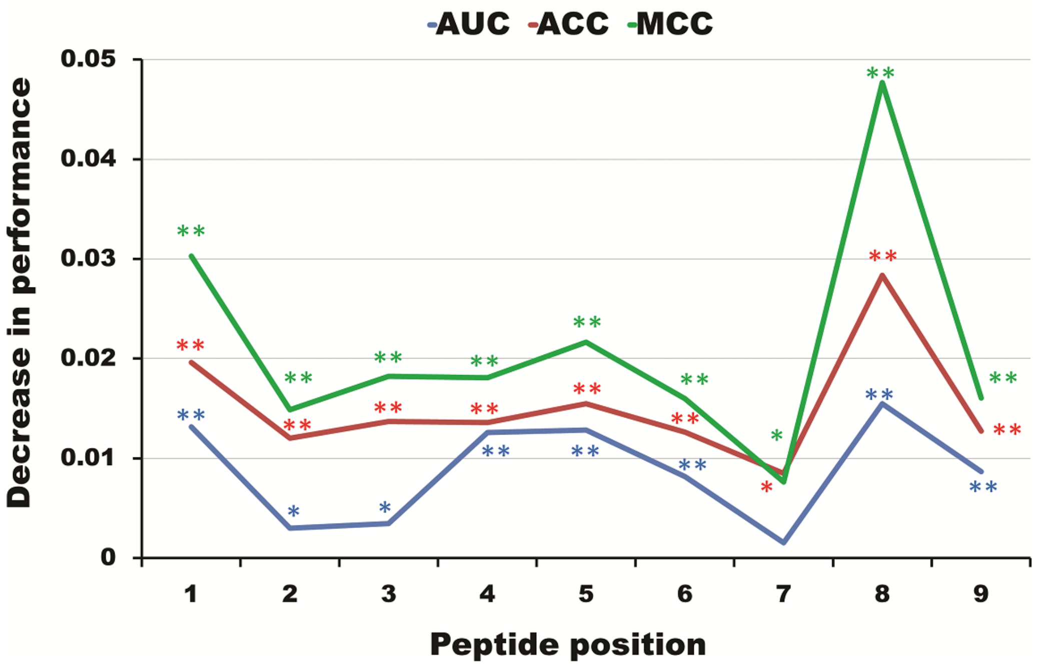


Figure 4.4 The decrease in the performance of PAAQD evaluated on datasets without features corresponding to specific positions of nonapeptides.

The symbols ** and * indicate two-sample *t*-test significance level 0.01 and 0.05, respectively, by comparing the performance between reduced datasets with the dataset including all positions.

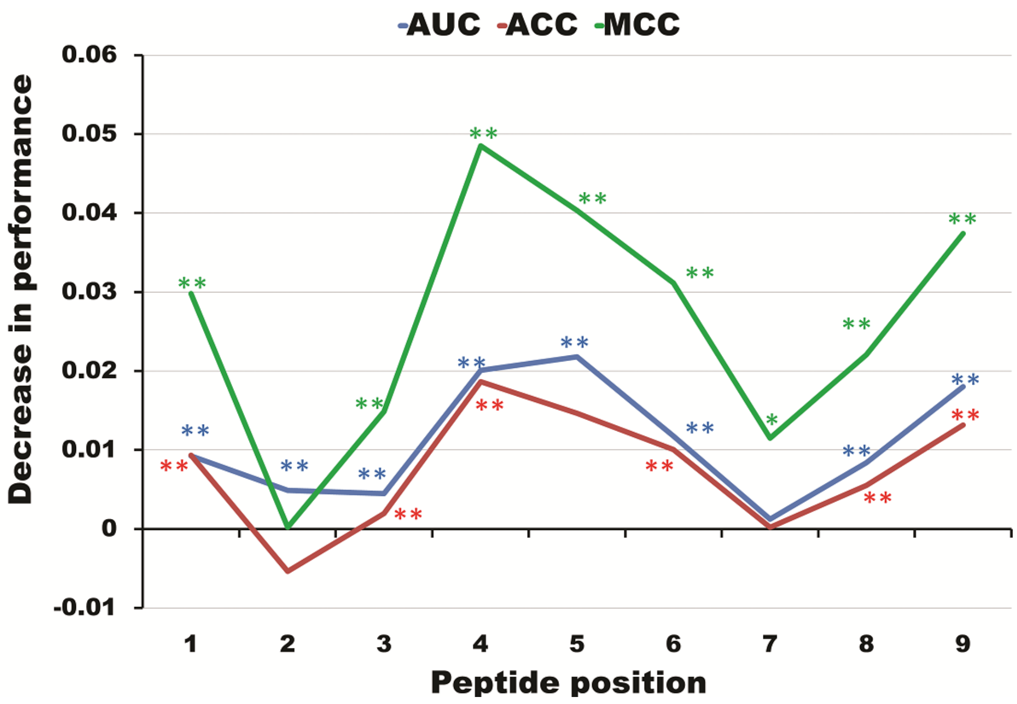


Figure 4.5 The decrease in the performance of PAAQD with the SVM evaluated on datasets without features corresponding to specific positions of nonapeptides.

The symbols ** and * indicate two-sample *t*-test significance level 0.01 and 0.05, respectively, by comparing the performance between reduced datasets with the dataset including all positions.

4.3.3 The importance of AAPP and QTMS descriptors in peptide immunogenicity

The importance of each AAPP was evaluated by removing features corresponding to the specific AAPP and was observed for the decrease in the performance. Figure 4.6 shows the result of each AAPP importance. The horizontal axis represents AAPP IDs (see Appendix B for more information) and the vertical axis represents the decrease in performance. The most important AAPP IDs were 6, 21, 26, 27, 28, and 33 with significance level 0.0001. In contrast, the least important AAPP IDs were 9, 11, 16, 24, 29, and 36. Table 4-2 shows more detail of the important AAPPs identified by PAAQD. Surprisingly, four out of the six important AAPPs were related to distance between amino acids (ID 6, 26, 27, and 28). Therefore, the distance between contacting side chains of amino acids may be an essential factor that determines the binding affinity of pMHC for TCR. This binding is a crucial step for T-cell responses.

The importance of QTMS descriptors is shown in Figure 4.7. Removing features corresponding to any QTMS descriptor decreased ACC and MCC. Removing features corresponding to the ADFQ descriptor was the least important one when compared to the other three descriptors. However, the previous study found that the ADFQ descriptor was the most important one in the HLA-peptide binding prediction [63]. Again, these findings indicate the difference between epitope prediction and T-cell reactivity prediction in influence of features to the performance of the predictive model. The QTMS descriptors were suggested to be essential in T-cell reactivity prediction since their presences improved the performance from using AAPPs alone (see Figure 4.3).

Table 4-2 Important AAPPs in T-cell reactivity prediction identified by using our method.

ID	Description	Reference
6	Distances between centers of interacting side chains in the antiparallel orientation	[137]
21	Quasichemical energy of transfer of amino acids from water to the protein environment	[138]
26	Distance-dependent statistical potential (contacts within 7.5–10 Angstroms)	[120]
27	Distance-dependent statistical potential (contacts within 10–12 Angstroms)	[120]
28	Distance-dependent statistical potential (contacts longer than 12 Angstroms)	[120]
33	Number of contacts between side chains derived from 25 X-ray protein structures	[139]

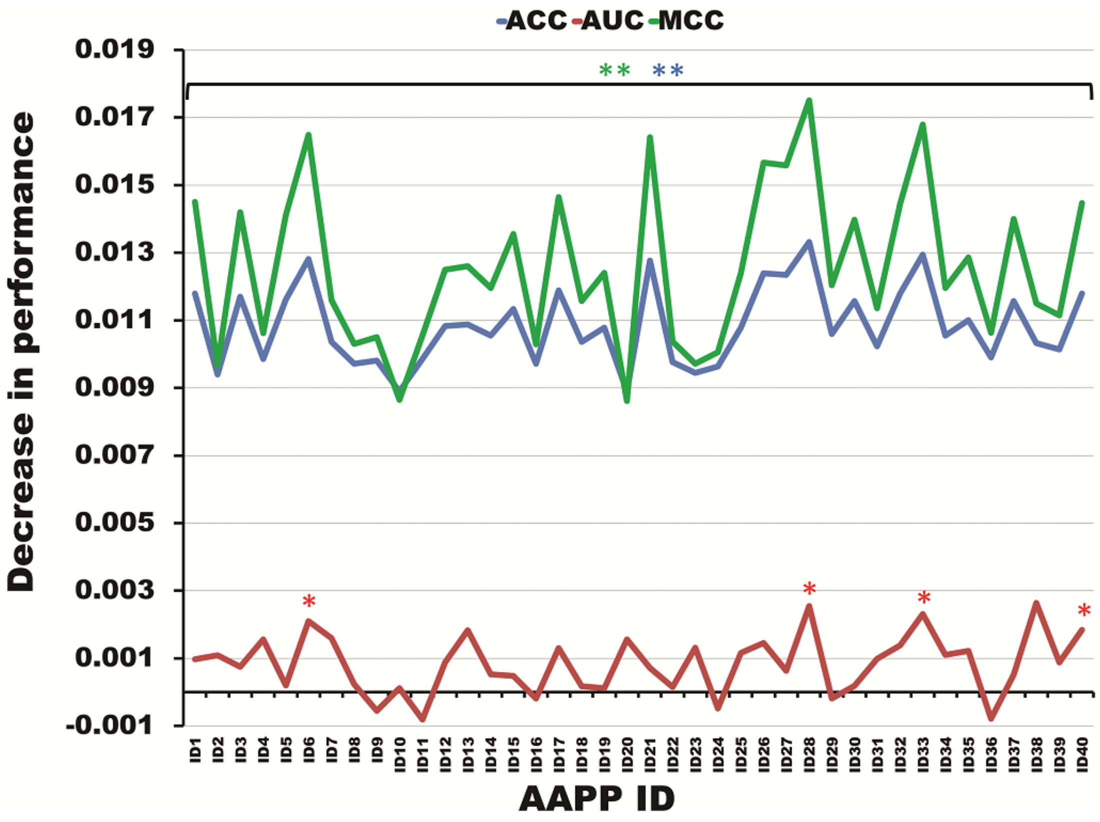


Figure 4.6 The decrease in the performance of PAAQD evaluated on datasets without features corresponding to the specific AAPP.

The symbols ** and * indicate two-sample *t*-test significance level 0.01 and 0.05, respectively, by comparing the performance between reduced datasets with the dataset including all positions.

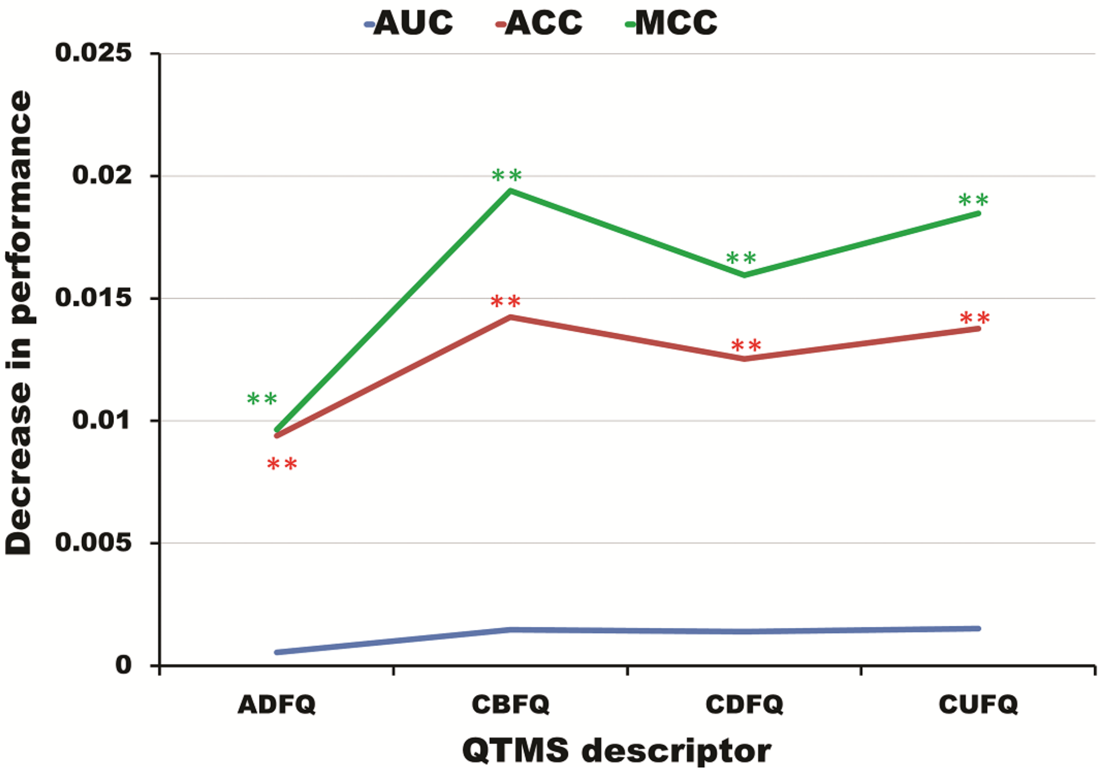


Figure 4.7 The decrease in the performance of PAAQD evaluated on datasets without features corresponding to the specific QTMS descriptor.

The symbols ** and * indicate two-sample *t*-test significance level 0.01 and 0.05, respectively, by comparing the performance between reduced datasets with the dataset including all positions.

4.3.4 Result of peptide immunogenicity prediction using the validation dataset

The result of peptide immunogenicity prediction using validation dataset is shown in Figure 4.8. 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 when peptides with different sequence preferences were input to the generated predictive model. We further examined the over- and underrepresented amino acids in corresponding positions of the IMMA 2 dataset and the validation dataset using the two-sample logos [140]. In the two-sample logos, differences among amino acids were statistically significant with level 0.01 when using the two-sample *t*-test. The two-sample logo of the IMMA 2 dataset (Figure 4.9) showed many over- and underrepresented amino acids. However, the two-sample logo of the validation dataset (Figure 4.10) showed underrepresentation of valine at the position 2, isoleucine at the position 6, and aspartic acid at the position 8. From these two-sample logos, both datasets are clearly different in preferences of amino acid sequences. This indicated that PAAQD provided more predictive stability than POPISK when using the test dataset with sequence preferences different from the training set.

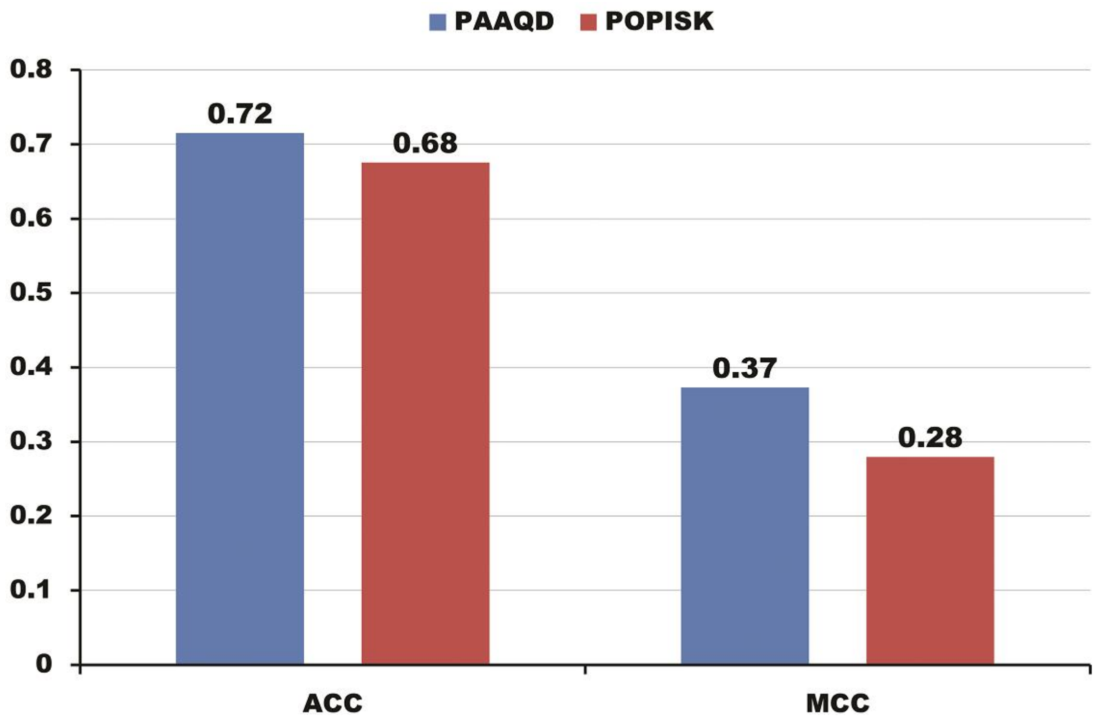


Figure 4.8 The result of peptide immunogenicity prediction evaluated on the validation dataset.

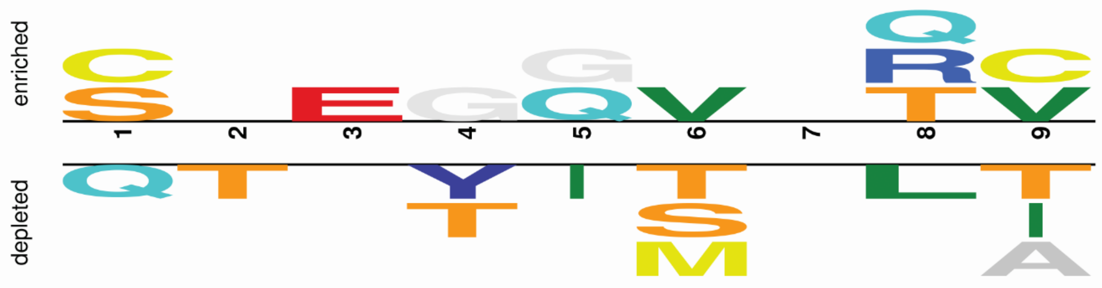


Figure 4.9 Two-sample logo that represents over- and underrepresented amino acids in the IMMA 2 dataset.

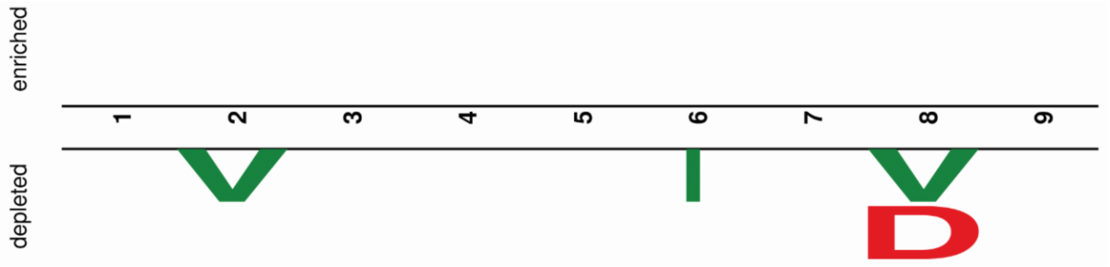


Figure 4.10 Two-sample logo that represents over- and underrepresented amino acids in the validation dataset.

4.4 Conclusion

We developed a novel method for T-cell reactivity prediction which we call PAAQD. Nonapeptides were encoded numerically, using combining information of amino acid pairwise contact potentials (AAPPs) and quantum topological molecular similarity (QTMS) descriptors. Encoded data were used in the construction of our classification model. 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 were input to the model. In this study, clear differences between epitope prediction and peptide immunogenicity prediction were demonstrated. The analysis of important positions showed that most of the anchor residues in epitope prediction were not important in T-cell reactivity prediction. Both of these two techniques are promising in vaccine development and can be used complementary. We speculate that PAAQD may be useful in identifying immunogenic peptides for the development of new vaccines.

Chapter 5 Conclusions

Previous chapters described the development of new epitope and T-cell reactivity prediction methods for advancement in the vaccine discovery. This final chapter summarizes the contributions of this thesis and presents future research directions.

5.1 Dissertation summary

Epitope is a part of an antigen recognized by the immune systems. Epitopes play the important role in activating the immune systems and are the key components in the vaccine development. The conventional vaccine development is time-consuming and laborious. Computational methods can be applied to help in epitope identification and speed up the vaccine production. From the last decade until now, many epitope prediction methods were proposed and have been used to search for new epitopes. However, recent studies found that some predicted epitopes with high binding affinities not stimulated immune responses. In addition, predicted epitopes with low binding affinities actually stimulated immune responses. Therefore, the result of epitope prediction is not always correct. Consequently, T-cell reactivity prediction has been introduced to search for immunogenic peptides instead of using epitope prediction. Hereby, the objectives of this dissertation are: (1) to develop a new epitope prediction method by using information of pMHC contact sites and AAPPs, (2) to identify important AAPPs and positions of nonapeptide in the pMHC binding, (3) to identify novel promiscuous epitopes from protein sequences of influenza A viral strains, (4) to develop a new T-cell reactivity prediction method by using information of AAPPs, pMHC contact sites, and QTMS descriptors, (5) to identify important AAPPs, QTMS descriptors, and positions of nonapeptide in the TCR-pMHC binding. The main contributions of the thesis can be summarized as follows.

Firstly, a new epitope prediction method named EpicCapo^{+REF} was developed. The combination of pMHC contact sites and AAPPs provided the better interpretability for the further analysis than other methods. Our method achieved high performance and outperformed other state of the art methods in many datasets. We speculate that our method can be applied in the development of new vaccines.

Secondly, by using our method, we are able to identify important AAPPs and positions of nonapeptides in the pMHC binding. We found that two AAPPs were very important in the pMHC binding. In addition, by ranking features in the dataset, positions 9 and 2 were the most important ones follow by positions 3, 1, or 7. Interestingly, when we remove features corresponded to one position, the performance of the method was decreased. Therefore, we suggest that all nine positions are important in the pMHC binding and their effects to the binding affinity are not independent.

Thirdly, we used EpicCapo^{+REF} to identify promiscuous epitopes from four influenza A viral strains: H1N1 (A/PR/8/34), H3N2 (A/Aichi/2/68), H1N1 (A/New York/4290/2009), and H5N1 (A/Hong Kong/483/97). 67.1% of predicted epitopes were consistent with previous immunological experiments. This consistency indicates that our method has high accuracy in epitope prediction.

Fourthly, a new T-cell reactivity prediction method named PAAQD was developed. 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.

Finally, by using PAAQD, we are able to identify important AAPPs, QTMS descriptor, and positions of nonapeptides in the TCR-pMHC binding. We found that all QTMS descriptors and six AAPPs were important. Surprisingly, positions 2, 3, and 7 were found as less important ones. However, these positions have been identified as anchor residues for epitope prediction. We suppose that epitope prediction and T-cell reactivity prediction are considerably different and should not be used interchangeably. In addition we found that position 1 and 8 were the most important ones in the TCR-pMHC binding.

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, there are limitations when using our methods. First, an input peptide must be a nonapeptide which is a peptide composed of 9 amino acids. Second, an input peptide must not contain special or ambiguous amino acids: amino acids U (Selenocysteine), O (Pyrrolysine), B (Asparagine or aspartic acid), Z (Glutamine or glutamic acid), J (Leucine or Isoleucine), and X (unknown). Our methods are not applicable with these amino acids since they are not included in AAPPs.

According to the above limitations, in our future researches, we will develop epitope and T-cell reactivity predictors that are able to be used with various lengths of peptides. However, there are small numbers of positive peptides or negative peptides in some lengths. Therefore, oversampling techniques such as SMOTE [141] can be used to generate more samples in the future study. For the problem of special or ambiguous amino acids, we still search for the practical solution. Since there are only small numbers of peptides containing these amino acids, removing them may be the best solution. In addition, there are several topics which we concerned as the future studies:

The applications of our peptide encoding schemes in other problems. We developed the peptide encoding schemes for both epitope and T-cell reactivity prediction. However, these schemes can be applied in other studies such as protein-ligand binding, protein-protein interaction (PPI) prediction, and drug discovery.

The use of data curation. We observed that many records in the databases are not reliable. For example, a peptide is reported as epitope in one record but not in another. To solve this problem, we need to look into the detail of each record and then make the decision to choose the correct one. However, this approach is time-consuming and not practical if there are large numbers of peptides. Therefore, automatic data curation needs to be developed to ease this problem.

The development of length independent epitope and T-cell reactivity predictor. Most of existing epitope or T-cell reactivity predictors including our methods are length dependent. The core algorithms required the input peptides to have the same length. However, we have considered applying other algorithms such as string kernel in the SVM and hidden Markov model. These applications will be useful in the new vaccine development.

Appendix A

The scaling of positional scoring matrices

In this study, the original and scaled positional scoring matrices are denoted by *T* and *T*'. The (*i*, *j*)th elements of *T* and *T'* represent preferences of amino acid *i* at position *j* in nonapeptides, and are denoted by $T_{i,j}$ and $T'_{i,j}$, respectively. We simply scale the original matrix *T* into *T*' as follows:

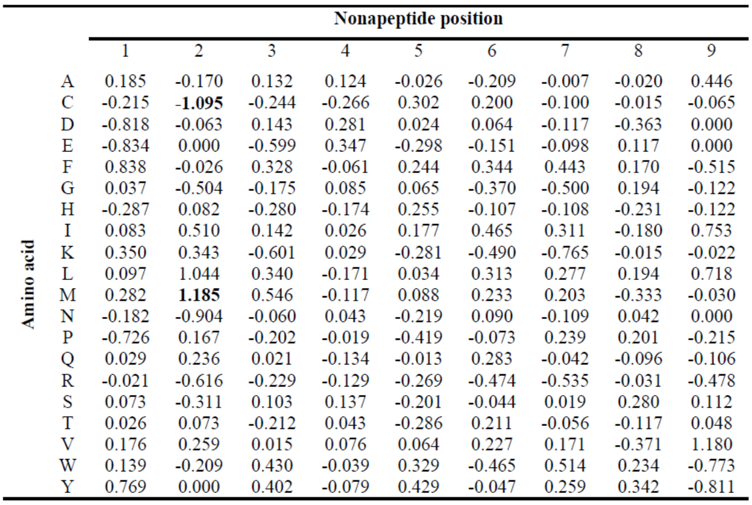
$$T'_{i,j} = 9 \times \frac{(T_{i,j} - MIN)}{(MAX - MIN)} + 1,$$

where MAX and MIN represent the maximum and minimum values in the matrix, respectively. The example of matrix scaling is shown below.

$$T_{i,j}$$

MAX = 1.185

MIN = -1.095



Amino acid	Nonapeptide position								
	1	2	3	4	5	6	7	8	9
A	0.185	-0.170	0.132	0.124	-0.026	-0.209	-0.007	-0.020	0.446
C	-0.215	**-1.095**	-0.244	-0.266	0.302	0.200	-0.100	-0.015	-0.065
D	-0.818	-0.063	0.143	0.281	0.024	0.064	-0.117	-0.363	0.000
E	-0.834	0.000	-0.599	0.347	-0.298	-0.151	-0.098	0.117	0.000
F	0.838	-0.026	0.328	-0.061	0.244	0.344	0.443	0.170	-0.515
G	0.037	-0.504	-0.175	0.085	0.065	-0.370	-0.500	0.194	-0.122
H	-0.287	0.082	-0.280	-0.174	0.255	-0.107	-0.108	-0.231	-0.122
I	0.083	0.510	0.142	0.026	0.177	0.465	0.311	-0.180	0.753
K	0.350	0.343	-0.601	0.029	-0.281	-0.490	-0.765	-0.015	-0.022
L	0.097	1.044	0.340	-0.171	0.034	0.313	0.277	0.194	0.718
M	0.282	**1.185**	0.546	-0.117	0.088	0.233	0.203	-0.333	-0.030
N	-0.182	-0.904	-0.060	0.043	-0.219	0.090	-0.109	0.042	0.000
P	-0.726	0.167	-0.202	-0.019	-0.419	-0.073	0.239	0.201	-0.215
Q	0.029	0.236	0.021	-0.134	-0.013	0.283	-0.042	-0.096	-0.106
R	-0.021	-0.616	-0.229	-0.129	-0.269	-0.474	-0.535	-0.031	-0.478
S	0.073	-0.311	0.103	0.137	-0.201	-0.044	0.019	0.280	0.112
T	0.026	0.073	-0.212	0.043	-0.286	0.211	-0.056	-0.117	0.048
V	0.176	0.259	0.015	0.076	0.064	0.227	0.171	-0.371	1.180
W	0.139	-0.209	0.430	-0.039	0.329	-0.465	0.514	0.234	-0.773
Y	0.769	0.000	0.402	-0.079	0.429	-0.047	0.259	0.342	-0.811

$$T'_{i,j}$$

$$T'_{i,j} = 9 \times \frac{(T_{i,j} - MIN)}{(MAX - MIN)} + 1$$

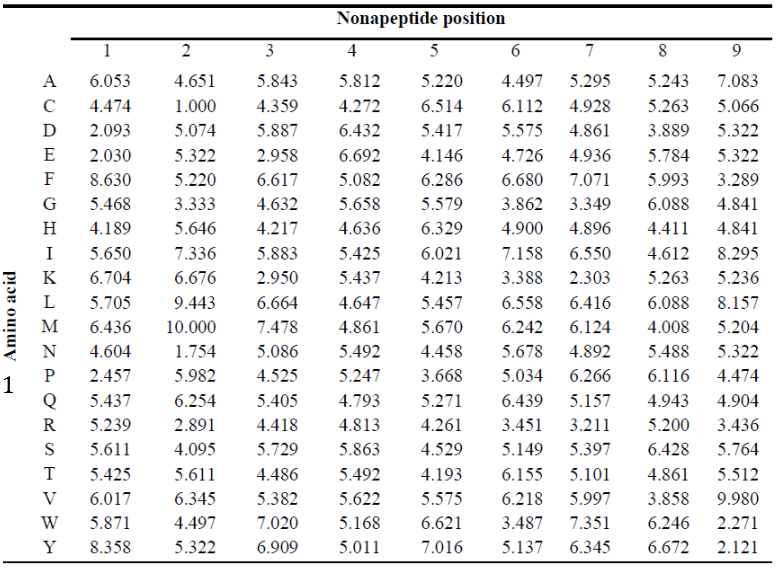
$$T'_{1,1} = 9 \times \frac{(0.185 - (-1.095))}{(1.185 - (-1.095))} + 1$$

$$= 6.053$$

$$T'_{2,1} = 9 \times \frac{((-0.215) - (-1.095))}{(1.185 - (-1.095))} + 1$$

$$= 4.474$$

$$\vdots$$



Amino acid	Nonapeptide position								
	1	2	3	4	5	6	7	8	9
A	6.053	4.651	5.843	5.812	5.220	4.497	5.295	5.243	7.083
C	4.474	1.000	4.359	4.272	6.514	6.112	4.928	5.263	5.066
D	2.093	5.074	5.887	6.432	5.417	5.575	4.861	3.889	5.322
E	2.030	5.322	2.958	6.692	4.146	4.726	4.936	5.784	5.322
F	8.630	5.220	6.617	5.082	6.286	6.680	7.071	5.993	3.289
G	5.468	3.333	4.632	5.658	5.579	3.862	3.349	6.088	4.841
H	4.189	5.646	4.217	4.636	6.329	4.900	4.896	4.411	4.841
I	5.650	7.336	5.883	5.425	6.021	7.158	6.550	4.612	8.295
K	6.704	6.676	2.950	5.437	4.213	3.388	2.303	5.263	5.236
L	5.705	9.443	6.664	4.647	5.457	6.558	6.416	6.088	8.157
M	6.436	10.000	7.478	4.861	5.670	6.242	6.124	4.008	5.204
N	4.604	1.754	5.086	5.492	4.458	5.678	4.892	5.488	5.322
P	2.457	5.982	4.525	5.247	3.668	5.034	6.266	6.116	4.474
Q	5.437	6.254	5.405	4.793	5.271	6.439	5.157	4.943	4.904
R	5.239	2.891	4.418	4.813	4.261	3.451	3.211	5.200	3.436
S	5.611	4.095	5.729	5.863	4.529	5.149	5.397	6.428	5.764
T	5.425	5.611	4.486	5.492	4.193	6.155	5.101	4.861	5.512
V	6.017	6.345	5.382	5.622	5.575	6.218	5.997	3.858	9.980
W	5.871	4.497	7.020	5.168	6.621	3.487	7.351	6.246	2.271
Y	8.358	5.322	6.909	5.011	7.016	5.137	6.345	6.672	2.121

Appendix B

Amino acid pairwise contact potentials (AAPPs) used in this study (retrieved from http://www.genome.jp/aaindex/ [59])

ID	Accession #	Description
1	BASU010101	Optimization-based potential derived by the modified perceptron criterion
2	BETM990101	Modified version of the Miyazawa-Jernigan transfer energy
3	BONM030101	Quasichemical statistical potential for the antiparallel orientation of interacting side groups
4	BONM030102	Quasichemical statistical potential for the intermediate orientation of interacting side groups
5	BONM030103	Quasichemical statistical potential for the parallel orientation of interacting side groups
6	BONM030104	Distances between centers of interacting side chains in the antiparallel orientation
7	BONM030105	Distances between centers of interacting side chains in the intermediate orientation
8	BONM030106	Distances between centers of interacting side chains in the parallel orientation
9	BRYS930101	Distance-dependent statistical potential (only energies of contacts within 0–5 Angstroms are included)
10	KESO980101	Quasichemical transfer energy derived from interfacial regions of protein-protein complexes
11	KESO980102	Quasichemical energy in an average protein environment derived from interfacial regions of protein-protein complexes
12	KOLA930101	Statistical potential derived by the quasichemical approximation
13	LIWA970101	Modified version of the Miyazawa-Jernigan transfer energy
14	MICC010101	Optimization-derived potential
15	MIRL960101	Statistical potential derived by the maximization of the harmonic mean of Z scores
16	MIYS850102	Quasichemical energy of transfer of amino acids from water to the protein environment
17	MIYS850103	Quasichemical energy of interactions in an average buried environment
18	MIYS960101	Quasichemical energy of transfer of amino acids from water to the protein environment
19	MIYS960102	Quasichemical energy of interactions in an average buried environment
20	MIYS960103	Number of contacts between side chains derived from 1168 X-ray protein structures
21	MIYS990106	Quasichemical energy of transfer of amino acids from water to the protein environment

ID	Accession #	Description
22	MIYS990107	Quasichemical energy of interactions in an average buried environment
23	MOOG990101	Quasichemical potential derived from interfacial regions of protein-protein complexes
24	SIMK990101	Distance-dependent statistical potential (contacts within 0–5 Angstroms)
25	SIMK990102	Distance-dependent statistical potential (contacts within 5–7.5 Angstroms)
26	SIMK990103	Distance-dependent statistical potential (contacts within 7.5–10 Angstroms)
27	SIMK990104	Distance-dependent statistical potential (contacts within 10–12 Angstroms)
28	SIMK990105	Distance-dependent statistical potential (contacts longer than 12 Angstroms)
29	SKOJ000101	Statistical quasichemical potential with the partially composition-corrected pair scale
30	SKOJ000102	Statistical quasichemical potential with the composition-corrected pair scale
31	SKOJ970101	Statistical potential derived by the quasichemical approximation
32	TANS760101	Statistical contact potential derived from 25 X-ray protein structures
33	TANS760102	Number of contacts between side chains derived from 25 X-ray protein structures
34	THOP960101	Mixed quasichemical and optimization-based protein contact potential
35	TOBD000101	Optimization-derived potential obtained for small set of decoys
36	TOBD000102	Optimization-derived potential obtained for large set of decoys
37	VENM980101	Statistical potential derived by the maximization of the perceptron criterion
38	ZHAC000101	Environment-dependent residue contact energies (rows = helix, cols = helix)
39	ZHAC000104	Environment-dependent residue contact energies (rows = strand, cols = strand)
40	ZHAC000106	Environment-dependent residue contact energies (rows = coil, cols = coil)

Appendix C

Features selected by EpicCapo^{+REF} separated in each allele

HLA-A datasets	# of selected features	Selected features ordered by the importance identified by Relief-F algorithm
A*01:01	72	Pos9AAPP20 ,Pos9AAPP11 ,Pos9AAPP33 ,Pos9AAPP28 ,Pos9AAPP14 ,Pos9AAPP24 ,Pos9AAPP26 ,Pos9AAPP1 ,Pos2AAPP1 ,Pos2AAPP24 ,Pos2AAPP28 ,Pos3AAPP24 ,Pos2AAPP33 ,Pos2AAPP14 ,Pos2AAPP20 ,Pos3AAPP28 ,Pos2AAPP11 ,Pos3AAPP14 ,Pos2AAPP26 ,Pos3AAPP1 ,Pos3AAPP11 ,Pos3AAPP33 ,Pos3AAPP20 ,Pos1AAPP14 ,Pos1AAPP11 ,Pos1AAPP28 ,Pos8AAPP24 ,Pos7AAPP20 ,Pos4AAPP26 ,Pos7AAPP33 ,Pos8AAPP26 ,Pos5AAPP26 ,Pos1AAPP1 ,Pos8AAPP28 ,Pos7AAPP14 ,Pos5AAPP28 ,Pos1AAPP20 ,Pos8AAPP11 ,Pos6AAPP26 ,Pos1AAPP24 ,Pos7AAPP28 ,Pos6AAPP1 ,Pos7AAPP11 ,Pos8AAPP14 ,Pos1AAPP33 ,Pos7AAPP1 ,Pos7AAPP26 ,Pos5AAPP24 ,Pos4AAPP28 ,Pos5AAPP14 ,Pos5AAPP11 ,Pos5AAPP1 ,Pos1AAPP26 ,Pos6AAPP20 ,Pos4AAPP33 ,Pos4AAPP14 ,Pos8AAPP1 ,Pos4AAPP24 ,Pos6AAPP33 ,Pos4AAPP11 ,Pos7AAPP24 ,Pos8AAPP20 ,Pos6AAPP11 ,Pos5AAPP20 ,Pos8AAPP33 ,Pos6AAPP14 ,Pos5AAPP33 ,Pos6AAPP28 ,Pos4AAPP20 ,Pos6AAPP24 ,Pos3AAPP26 ,Pos4AAPP1
A*02:01	62	Pos9AAPP26 ,Pos9AAPP28 ,Pos9AAPP14 ,Pos9AAPP11 ,Pos9AAPP24 ,Pos2AAPP9 ,Pos2AAPP31 ,Pos2AAPP14 ,Pos2AAPP26 ,Pos2AAPP28 ,Pos9AAPP9 ,Pos2AAPP11 ,Pos9AAPP31 ,Pos2AAPP24 ,Pos1AAPP28 ,Pos1AAPP26 ,Pos1AAPP14 ,Pos1AAPP9 ,Pos1AAPP11 ,Pos1AAPP24 ,Pos1AAPP31 ,Pos3AAPP11 ,Pos7AAPP24 ,Pos7AAPP26 ,Pos3AAPP14 ,Pos3AAPP9 ,Pos3AAPP31 ,Pos3AAPP24 ,Pos3AAPP28 ,Pos7AAPP28 ,Pos6AAPP28 ,Pos3AAPP26 ,Pos7AAPP14 ,Pos7AAPP11 ,Pos7AAPP31 ,Pos6AAPP14 ,Pos4AAPP14 ,Pos6AAPP24 ,Pos5AAPP9 ,Pos4AAPP31 ,Pos7AAPP9 ,Pos6AAPP26 ,Pos6AAPP11 ,Pos5AAPP14 ,Pos5AAPP28 ,Pos4AAPP11 ,Pos4AAPP26 ,Pos6AAPP31 ,Pos5AAPP11 ,Pos8AAPP14 ,Pos5AAPP31 ,Pos4AAPP9 ,Pos8AAPP26 ,Pos4AAPP24 ,Pos6AAPP9 ,Pos5AAPP24 ,Pos8AAPP28 ,Pos8AAPP31 ,Pos8AAPP24 ,Pos5AAPP26 ,Pos8AAPP11 ,Pos4AAPP28
A*02:02	18	Pos9AAPP14 ,Pos9AAPP28 ,Pos2AAPP28 ,Pos2AAPP14 ,Pos1AAPP28 ,Pos1AAPP14 ,Pos5AAPP14 ,Pos3AAPP14 ,Pos5AAPP28 ,Pos3AAPP28 ,Pos8AAPP14 ,Pos4AAPP28 ,Pos6AAPP28 ,Pos8AAPP28 ,Pos6AAPP14 ,Pos4AAPP14

HLA-A datasets	# of selected features	Selected features ordered by the importance identified by Relief-F algorithm
		,Pos7AAPP14 ,Pos7AAPP28
A*02:03	104	Pos9AAPP25 ,Pos9AAPP26 ,Pos9AAPP19 ,Pos9AAPP31 ,Pos2AAPP9 ,Pos9AAPP14 ,Pos9AAPP9 ,Pos2AAPP31 ,Pos9AAPP28 ,Pos9AAPP24 ,Pos2AAPP29 ,Pos9AAPP11 ,Pos1AAPP33 ,Pos1AAPP28 ,Pos1AAPP24 ,Pos1AAPP26 ,Pos2AAPP3 ,Pos2AAPP14 ,Pos1AAPP14 ,Pos1AAPP11 ,Pos9AAPP3 ,Pos2AAPP28 ,Pos2AAPP26 ,Pos2AAPP11 ,Pos2AAPP25 ,Pos1AAPP19 ,Pos1AAPP9 ,Pos9AAPP29 ,Pos3AAPP14 ,Pos3AAPP28 ,Pos2AAPP33 ,Pos1AAPP31 ,Pos9AAPP33 ,Pos6AAPP28 ,Pos2AAPP24 ,Pos3AAPP11 ,Pos3AAPP24 ,Pos1AAPP25 ,Pos1AAPP3 ,Pos6AAPP11 ,Pos6AAPP3 ,Pos1AAPP29 ,Pos6AAPP14 ,Pos3AAPP33 ,Pos3AAPP19 ,Pos7AAPP14 ,Pos6AAPP24 ,Pos7AAPP3 ,Pos3AAPP26 ,Pos7AAPP28 ,Pos6AAPP31 ,Pos3AAPP9 ,Pos7AAPP29 ,Pos6AAPP26 ,Pos6AAPP33 ,Pos6AAPP19 ,Pos7AAPP31 ,Pos6AAPP25 ,Pos3AAPP29 ,Pos3AAPP3 ,Pos2AAPP19 ,Pos6AAPP29 ,Pos3AAPP31 ,Pos3AAPP25 ,Pos7AAPP24 ,Pos7AAPP33 ,Pos6AAPP9 ,Pos8AAPP19 ,Pos8AAPP28 ,Pos7AAPP25 ,Pos8AAPP14 ,Pos7AAPP11 ,Pos7AAPP26 ,Pos7AAPP19 ,Pos8AAPP29 ,Pos4AAPP29 ,Pos8AAPP25 ,Pos5AAPP33 ,Pos8AAPP33 ,Pos5AAPP3 ,Pos8AAPP31 ,Pos8AAPP9 ,Pos5AAPP25 ,Pos5AAPP31 ,Pos8AAPP24 ,Pos5AAPP14 ,Pos5AAPP24 ,Pos5AAPP28 ,Pos8AAPP26 ,Pos5AAPP19 ,Pos5AAPP29 ,Pos5AAPP26 ,Pos4AAPP31 ,Pos5AAPP9 ,Pos8AAPP11 ,Pos5AAPP11 ,Pos4AAPP25 ,Pos4AAPP24 ,Pos7AAPP9 ,Pos4AAPP26 ,Pos8AAPP3 ,Pos4AAPP19 ,Pos4AAPP14 ,Pos4AAPP11
A*02:06	141	Pos9AAPP22 ,Pos1AAPP26 ,Pos9AAPP19 ,Pos9AAPP26 ,Pos1AAPP28 ,Pos9AAPP11 ,Pos1AAPP11 ,Pos1AAPP14 ,Pos9AAPP39 ,Pos9AAPP14 ,Pos9AAPP24 ,Pos9AAPP25 ,Pos9AAPP28 ,Pos9AAPP27 ,Pos1AAPP9 ,Pos9AAPP38 ,Pos1AAPP1 ,Pos1AAPP24 ,Pos9AAPP34 ,Pos9AAPP1 ,Pos3AAPP24 ,Pos1AAPP19 ,Pos9AAPP18 ,Pos9AAPP13 ,Pos1AAPP22 ,Pos9AAPP21 ,Pos1AAPP31 ,Pos1AAPP38 ,Pos9AAPP31 ,Pos1AAPP25 ,Pos1AAPP39 ,Pos1AAPP27 ,Pos9AAPP9 ,Pos6AAPP39 ,Pos6AAPP18 ,Pos1AAPP18 ,Pos1AAPP13 ,Pos1AAPP21 ,Pos3AAPP39 ,Pos8AAPP1 ,Pos1AAPP3 ,Pos1AAPP34 ,Pos6AAPP24 ,Pos6AAPP38 ,Pos3AAPP14 ,Pos6AAPP13 ,Pos3AAPP11 ,Pos3AAPP28 ,Pos3AAPP13 ,Pos4AAPP1 ,Pos6AAPP28 ,Pos3AAPP38 ,Pos3AAPP26 ,Pos3AAPP9 ,Pos8AAPP27 ,Pos7AAPP26 ,Pos6AAPP21 ,Pos6AAPP11 ,Pos8AAPP39 ,Pos6AAPP34 ,Pos8AAPP26 ,Pos3AAPP34 ,Pos3AAPP22 ,Pos7AAPP22 ,Pos3AAPP19 ,Pos3AAPP18 ,Pos6AAPP3 ,Pos8AAPP14 ,Pos6AAPP14 ,Pos7AAPP24 ,Pos8AAPP24 ,Pos8AAPP38 ,Pos8AAPP11 ,Pos7AAPP19 ,Pos2AAPP9 ,Pos6AAPP26 ,Pos8AAPP28 ,Pos3AAPP3 ,Pos3AAPP27 ,Pos6AAPP31

HLA-A datasets	# of selected features	Selected features ordered by the importance identified by Relief-F algorithm
		,Pos3AAPP21 ,Pos3AAPP1 ,Pos8AAPP18 ,Pos7AAPP27 ,Pos7AAPP38 ,Pos8AAPP21 ,Pos7AAPP39 ,Pos5AAPP9 ,Pos3AAPP31 ,Pos7AAPP1 ,Pos8AAPP31 ,Pos6AAPP1 ,Pos8AAPP22 ,Pos8AAPP25 ,Pos5AAPP28 ,Pos8AAPP13 ,Pos8AAPP19 ,Pos4AAPP31 ,Pos7AAPP3 ,Pos7AAPP13 ,Pos7AAPP28 ,Pos8AAPP34 ,Pos6AAPP22 ,Pos5AAPP24 ,Pos5AAPP38 ,Pos7AAPP14 ,Pos9AAPP3 ,Pos5AAPP13 ,Pos6AAPP9 ,Pos6AAPP19 ,Pos2AAPP25 ,Pos7AAPP25 ,Pos4AAPP14 ,Pos7AAPP11 ,Pos7AAPP9 ,Pos5AAPP3 ,Pos8AAPP3 ,Pos2AAPP22 ,Pos6AAPP25 ,Pos5AAPP39 ,Pos7AAPP21 ,Pos7AAPP18 ,Pos2AAPP39 ,Pos7AAPP31 ,Pos2AAPP26 ,Pos5AAPP14 ,Pos2AAPP27 ,Pos6AAPP27 ,Pos7AAPP34 ,Pos8AAPP9 ,Pos5AAPP34 ,Pos5AAPP18 ,Pos5AAPP27 ,Pos4AAPP11 ,Pos4AAPP25 ,Pos2AAPP14 ,Pos2AAPP19 ,Pos2AAPP11 ,Pos3AAPP25 ,Pos2AAPP18 ,Pos2AAPP21
A*03:01	58	Pos9AAPP28 ,Pos9AAPP14 ,Pos9AAPP11 ,Pos9AAPP33 ,Pos9AAPP20 ,Pos9AAPP26 ,Pos9AAPP24 ,Pos2AAPP14 ,Pos2AAPP26 ,Pos2AAPP28 ,Pos2AAPP24 ,Pos2AAPP11 ,Pos7AAPP28 ,Pos1AAPP28 ,Pos7AAPP14 ,Pos7AAPP26 ,Pos1AAPP26 ,Pos1AAPP14 ,Pos2AAPP20 ,Pos2AAPP33 ,Pos7AAPP11 ,Pos1AAPP20 ,Pos3AAPP14 ,Pos1AAPP33 ,Pos3AAPP28 ,Pos3AAPP26 ,Pos1AAPP11 ,Pos6AAPP11 ,Pos6AAPP14 ,Pos6AAPP28 ,Pos3AAPP24 ,Pos6AAPP26 ,Pos1AAPP24 ,Pos3AAPP33 ,Pos3AAPP20 ,Pos7AAPP24 ,Pos6AAPP24 ,Pos8AAPP14 ,Pos3AAPP11 ,Pos8AAPP28 ,Pos7AAPP20 ,Pos4AAPP14 ,Pos6AAPP20 ,Pos4AAPP24 ,Pos7AAPP33 ,Pos8AAPP20 ,Pos6AAPP33 ,Pos4AAPP20 ,Pos4AAPP26 ,Pos8AAPP24 ,Pos4AAPP28 ,Pos5AAPP20 ,Pos4AAPP33 ,Pos8AAPP33 ,Pos5AAPP11 ,Pos5AAPP26 ,Pos5AAPP14 ,Pos5AAPP24
A*11:01	35	Pos9AAPP11 ,Pos9AAPP28 ,Pos9AAPP14 ,Pos9AAPP26 ,Pos2AAPP26 ,Pos2AAPP14 ,Pos2AAPP28 ,Pos2AAPP11 ,Pos1AAPP28 ,Pos3AAPP14 ,Pos1AAPP14 ,Pos3AAPP28 ,Pos1AAPP26 ,Pos3AAPP26 ,Pos1AAPP11 ,Pos3AAPP11 ,Pos7AAPP28 ,Pos7AAPP26 ,Pos8AAPP14 ,Pos7AAPP11 ,Pos7AAPP14 ,Pos8AAPP28 ,Pos4AAPP14 ,Pos6AAPP28 ,Pos8AAPP26 ,Pos8AAPP11 ,Pos6AAPP14 ,Pos6AAPP26 ,Pos6AAPP11 ,Pos5AAPP11 ,Pos5AAPP28 ,Pos5AAPP14 ,Pos4AAPP28 ,Pos4AAPP11 ,Pos5AAPP26
A*24:02	31	Pos2AAPP6 ,Pos2AAPP31 ,Pos2AAPP28 ,Pos2AAPP24 ,Pos2AAPP5 ,Pos2AAPP14 ,Pos8AAPP28 ,Pos9AAPP14 ,Pos9AAPP28 ,Pos8AAPP24 ,Pos4AAPP5 ,Pos8AAPP14

HLA-A datasets	# of selected features	Selected features ordered by the importance identified by Relief-F algorithm
		,Pos9AAPP24 ,Pos9AAPP31 ,Pos1AAPP28 ,Pos5AAPP5 ,Pos1AAPP14 ,Pos4AAPP6 ,Pos7AAPP31 ,Pos1AAPP24 ,Pos3AAPP5 ,Pos5AAPP6 ,Pos8AAPP5 ,Pos7AAPP6 ,Pos8AAPP31 ,Pos4AAPP24 ,Pos3AAPP31 ,Pos9AAPP6 ,Pos5AAPP28 ,Pos5AAPP31 ,Pos6AAPP14
A*26:01	18	Pos2AAPP14 ,Pos9AAPP14 ,Pos2AAPP28 ,Pos9AAPP28 ,Pos3AAPP28 ,Pos3AAPP14 ,Pos1AAPP14 ,Pos4AAPP14 ,Pos1AAPP28 ,Pos6AAPP28 ,Pos5AAPP28 ,Pos8AAPP28 ,Pos7AAPP28 ,Pos8AAPP14 ,Pos6AAPP14 ,Pos5AAPP14 ,Pos7AAPP14 ,Pos4AAPP28
A*29:02	23	Pos9AAPP5 ,Pos9AAPP20 ,Pos9AAPP33 ,Pos9AAPP9 ,Pos9AAPP8 ,Pos2AAPP33 ,Pos5AAPP5 ,Pos2AAPP20 ,Pos3AAPP9 ,Pos2AAPP5 ,Pos3AAPP8 ,Pos2AAPP9 ,Pos2AAPP8 ,Pos1AAPP5 ,Pos7AAPP8 ,Pos1AAPP8 ,Pos3AAPP5 ,Pos5AAPP9 ,Pos1AAPP20 ,Pos1AAPP33 ,Pos7AAPP5 ,Pos5AAPP8 ,Pos8AAPP8
A*31:01	46	Pos9AAPP28 ,Pos9AAPP14 ,Pos9AAPP20 ,Pos9AAPP33 ,Pos9AAPP11 ,Pos9AAPP26 ,Pos2AAPP28 ,Pos2AAPP26 ,Pos2AAPP14 ,Pos1AAPP20 ,Pos3AAPP14 ,Pos1AAPP28 ,Pos1AAPP33 ,Pos3AAPP26 ,Pos3AAPP28 ,Pos1AAPP14 ,Pos1AAPP11 ,Pos3AAPP33 ,Pos3AAPP20 ,Pos2AAPP11 ,Pos2AAPP33 ,Pos2AAPP20 ,Pos3AAPP11 ,Pos1AAPP26 ,Pos6AAPP28 ,Pos6AAPP14 ,Pos5AAPP11 ,Pos5AAPP14 ,Pos5AAPP26 ,Pos8AAPP28 ,Pos8AAPP33 ,Pos8AAPP20 ,Pos7AAPP26 ,Pos8AAPP14 ,Pos5AAPP28 ,Pos7AAPP33 ,Pos6AAPP33 ,Pos7AAPP14 ,Pos4AAPP20 ,Pos7AAPP20 ,Pos4AAPP26 ,Pos4AAPP33 ,Pos8AAPP26 ,Pos6AAPP20 ,Pos6AAPP11 ,Pos6AAPP26
A*33:01	17	Pos9AAPP28 ,Pos9AAPP14 ,Pos1AAPP14 ,Pos8AAPP28 ,Pos1AAPP28 ,Pos3AAPP14 ,Pos7AAPP14 ,Pos7AAPP28 ,Pos3AAPP28 ,Pos4AAPP28 ,Pos8AAPP14 ,Pos5AAPP14 ,Pos4AAPP14 ,Pos5AAPP28 ,Pos6AAPP14 ,Pos6AAPP28 ,Pos2AAPP14
A*68:01	40	Pos9AAPP14 ,Pos9AAPP28 ,Pos9AAPP20 ,Pos9AAPP26 ,Pos9AAPP33 ,Pos9AAPP11 ,Pos2AAPP14 ,Pos3AAPP28 ,Pos2AAPP26 ,Pos2AAPP28 ,Pos3AAPP14 ,Pos1AAPP33 ,Pos1AAPP20 ,Pos3AAPP11 ,Pos1AAPP11 ,Pos1AAPP28

HLA-A datasets	# of selected features	Selected features ordered by the importance identified by Relief-F algorithm
		,Pos3AAPP26 ,Pos2AAPP11 ,Pos1AAPP14 ,Pos3AAPP33 ,Pos3AAPP20 ,Pos2AAPP20 ,Pos5AAPP33 ,Pos2AAPP33 ,Pos5AAPP28 ,Pos5AAPP14 ,Pos5AAPP20 ,Pos5AAPP11 ,Pos1AAPP26 ,Pos7AAPP33 ,Pos6AAPP11 ,Pos7AAPP20 ,Pos8AAPP33 ,Pos5AAPP26 ,Pos8AAPP28 ,Pos8AAPP20 ,Pos8AAPP14 ,Pos6AAPP33 ,Pos4AAPP33 ,Pos7AAPP14
A*68:02	79	Pos9AAPP26 ,Pos9AAPP22 ,Pos9AAPP39 ,Pos9AAPP24 ,Pos9AAPP11 ,Pos9AAPP14 ,Pos9AAPP28 ,Pos3AAPP14 ,Pos3AAPP28 ,Pos1AAPP33 ,Pos1AAPP14 ,Pos3AAPP24 ,Pos1AAPP11 ,Pos1AAPP20 ,Pos3AAPP1 ,Pos9AAPP9 ,Pos9AAPP1 ,Pos1AAPP24 ,Pos3AAPP39 ,Pos1AAPP28 ,Pos3AAPP22 ,Pos1AAPP1 ,Pos3AAPP11 ,Pos9AAPP33 ,Pos9AAPP20 ,Pos3AAPP9 ,Pos1AAPP9 ,Pos3AAPP26 ,Pos7AAPP1 ,Pos3AAPP20 ,Pos7AAPP14 ,Pos2AAPP22 ,Pos3AAPP33 ,Pos5AAPP24 ,Pos7AAPP20 ,Pos8AAPP24 ,Pos8AAPP1 ,Pos2AAPP20 ,Pos8AAPP26 ,Pos7AAPP28 ,Pos7AAPP33 ,Pos1AAPP22 ,Pos1AAPP39 ,Pos2AAPP33 ,Pos7AAPP9 ,Pos2AAPP26 ,Pos7AAPP24 ,Pos8AAPP28 ,Pos8AAPP39 ,Pos5AAPP14 ,Pos2AAPP39 ,Pos7AAPP11 ,Pos2AAPP24 ,Pos2AAPP14 ,Pos5AAPP33 ,Pos6AAPP14 ,Pos5AAPP28 ,Pos2AAPP9 ,Pos2AAPP28 ,Pos7AAPP39 ,Pos6AAPP11 ,Pos7AAPP26 ,Pos6AAPP28 ,Pos5AAPP39 ,Pos8AAPP11 ,Pos2AAPP11 ,Pos5AAPP26 ,Pos8AAPP14 ,Pos8AAPP22 ,Pos5AAPP1 ,Pos6AAPP39 ,Pos6AAPP24 ,Pos5AAPP22 ,Pos8AAPP33 ,Pos4AAPP14 ,Pos6AAPP20 ,Pos2AAPP1 ,Pos8AAPP20 ,Pos5AAPP20

* ‘Pos’ referred to ‘at peptide position’
‘AAPP’ referred to ‘at amino acid pairwise contact potential’

Appendix D

Candidates of promiscuous epitopes identified from overlapping epitopes of influenza A viral strains: H1N1 (A/New York/4290/2009), H5N1 (A/Hong Kong/483/97), H1N1 (A/PR/8/34), and H3N2 (A/Aichi/2/68).

Epitope	Shared alleles	T cell assay
QTYDWTLNR	A*0301, A*1101, A*2902, A*3101, A*3301, A*6801	Positive
KFFPSSSYR	A*0301, A*1101, A*2902, A*3101, A*3301, A*6801	Positive
MMMGMFNML	A*0201, A*0202, A*0203, A*0206, A*6802	Positive
FVANFSMEL	A*0201, A*0202, A*0203, A*0206, A*6802	Positive
LLTEVETYV	A*0201, A*0202, A*0203, A*0206, A*6802	Positive
ALASCMGLI	A*0201, A*0202, A*0203, A*0206, A*6802	Negative
IMFSNKMAR	A*0301, A*1101, A*3101, A*3301, A*6801	Positive
RLFFKCIYR	A*0301, A*1101, A*3101, A*3301, A*6801	Positive
AQTDCVLEA	A*0201, A*0202, A*0203, A*0206	-
RLIDFLKDV	A*0201, A*0202, A*0203, A*0206	Positive
GMFNMLSTV	A*0201, A*0202, A*0203, A*0206	Positive
NMLSTVLGV	A*0201, A*0202, A*0203, A*0206	Positive
AQMALQLFI	A*0201, A*0202, A*0203, A*0206	-
KICSTIEEL	A*0201, A*0202, A*0203, A*0206	Positive
AIVGEISPL	A*0201, A*0202, A*0203, A*0206	Positive
GILGFVFTL	A*0201, A*0202, A*0203, A*0206	Positive
SMELPSFGV	A*0201, A*0202, A*0203, A*2902	Positive
GMMMGMFNM	A*0201, A*0203, A*0206, A*2902	Negative
CVLEAMAFL	A*0201, A*0203, A*0206, A*6802	Negative
MINNDLGPA	A*0202, A*0203, A*0206, A*6802	Positive
YGFVANFSM	A*0202, A*0206, A*2902, A*6802	Positive
MSIGVTVIK	A*0301, A*1101, A*3101, A*6801	Positive
ATTHSWIPK	A*0301, A*1101, A*3101, A*6801	Positive
MVLASTTAK	A*0301, A*1101, A*3101, A*6801	Positive
FEFTSFFYR	A*0301, A*2902, A*3101, A*6801	Positive
LANTIEVFR	A*1101, A*3101, A*3301, A*6801	-
NTMTKDAER	A*1101, A*3101, A*3301, A*6801	Positive
TTHSWIPKR	A*1101, A*3101, A*3301, A*6801	Positive
KLANVVRKM	A*0201, A*0202, A*0203	Positive
VLGVSILNL	A*0201, A*0202, A*0203	Positive
VLASTTAKA	A*0201, A*0202, A*0203	Negative
LQSSDDFAL	A*0201, A*0202, A*0206	-
TALANTIEV	A*0201, A*0206, A*6802	Positive
ILSPLTKGI	A*0202, A*0203, A*0206	Positive
RMVLASTTA	A*0202, A*0203, A*0206	Positive
KTRPILSPL	A*0202, A*0203, A*3101	Negative
QLNPIDGPL	A*0202, A*0203, A*6802	Positive
SSFQVDCFL	A*0202, A*0206, A*6802	-
LTKGILGFV	A*0203, A*0206, A*6802	Negative

Epitope	Shared alleles	T cell assay
QMALQLFIK	A*0301, A*1101, A*6801	Positive
SGRLIDFLK	A*1101, A*3101, A*6801	-
RSILNTSQR	A*1101, A*3101, A*6801	Positive
DTVNRTHQY	A*2601, A*2902, A*6801	Positive
ITTHFQRKR	A*3101, A*3301, A*6801	-
IATPGMQIR	A*3101, A*3301, A*6801	-
QAGVDRFYR	A*3101, A*3301, A*6801	Positive
HSWIPKRNR	A*3101, A*3301, A*6801	-
YSHGTGTGY	A*0101, A*2902	Positive
GILHLILWI	A*0201, A*0206	Positive
MFSNKMARL	A*0202, A*0203	Positive
NMSKKKSYI	A*0202, A*0203	Positive
NLHIPEVCL	A*0202, A*0203	Positive
ILGFVFTLT	A*0202, A*0203	Positive
QMAGSSEQA	A*0202, A*0203	Negative
QSSDDFALI	A*0202, A*6802	Positive
SFFYRYGFV	A*0202, A*6802	Positive
DMSIGVTVI	A*0203, A*2902	Positive
YTMDTVNRT	A*0203, A*6802	Positive
AVATTHSWI	A*0203, A*6802	Positive
GTFEFTSFF	A*0206, A*6801	Positive
TGAPQLNPI	A*0206, A*6802	-
KMARLGKGY	A*0301, A*2902	Positive
NLYNIRNLH	A*0301, A*6801	-
NAISTTFPY	A*1101, A*6801	Negative
VSILNLGQK	A*1101, A*6801	Positive
TSFFYRYGF	A*2902, A*6802	Positive
GYTMDTVNR	A*3101, A*3301	Positive
HFQRKRRVR	A*3101, A*3301	-
VSRARIDAR	A*3101, A*3301	-
TTHFQRKRR	A*3101, A*6801	Positive
LQLFIKDYR	A*3101, A*6801	Positive
YRYTYRCHR	A*3101, A*6801	Positive
FFPSSSYRR	A*3101, A*6801	-
DAPFLDRLR	A*3301, A*6801	-
NPLIRHENR	A*3301, A*6801	Negative
TTAKAMEQM	A*6801, A*6802	Positive

Appendix E

The IMMA2 dataset

Peptide sequence	Immunogenicity
AAAGFVFTA	+
AAGIGIIQI	+
AAGIGILTV	+
AALGFVFAA	+
AALGFVFTA	+
AFHHMAREL	+
AFHHVAREL	+
AIISGDSPV	+
AILALLPAL	+
AILGFVFTA	+
AILGFVFTL	+
AIMDKNIIL	+
ALADAVKVT	+
ALAYGIDKV	+
ALCRWGLLL	+
ALGLGLLPV	+
ALGRNSFEV	+
ALHVVVIGL	+
ALIHHNTHL	+
ALINDQLIM	+
ALLKHRFEI	+
ALLNIKVKL	+
ALLVLYSFA	+
ALMDKSLHV	+
ALMEQQHYV	+
ALMPLYACI	+
ALNTPKDHI	+
ALPHIIDEV	+
ALQDSGLEV	+
ALQPGTALL	+
ALSDHHIYL	+
ALSKFPRQL	+
ALSSGLYQC	+
ALSTGLIHL	+
ALSVMGVYV	+
ALVNAVNKL	+
ALVNFLRHL	+
ALVRCIPTL	+
ALWGFFPVL	+
ALYDVVSTL	+
AMSTTDLEA	+
ATTNILEHV	+
AVAGAAILV	+
AAPTPAAPA	-
AIIGLCAYA	-
AIMEKNIML	-
AIYHPQQFV	-
ALAIPQCRL	-
ALATFTVNI	-
ALDPYNEVV	-
ALFFFDIDL	-
ALIIIRSLL	-
ALKMTMASV	-
ALLSDWLPA	-
ALLSRFFNM	-
ALMAITKNV	-
ALMRRIAVV	-
ALSLAAVLV	-
ALSPVPPVV	-
ALSTGLIHI	-
ALVALVITI	-
ALVGACITL	-
ALVLLMLPV	-
ALWIPDLFM	-
AMFTAALNI	-
AMFTTMYNI	-
AMKADIQHV	-
AMLQDMAIL	-
AMLQLDPNA	-
AMTAFFGEL	-
AVLVVMACL	-
CLLQSLQQI	-
DLVPLTVSV	-
ETDDYMFFV	-
FAAELTIGV	-
FAGKDFDTV	-
FANSKFTLV	-
FAVQTIVFI	-
FIADIGIGV	-
FIDILLFVI	-
FIDSNEYEV	-
FIDTVSVYT	-
FIHGGILYA	-
FIIEVSNCV	-
FIISVISLV	-
FILHRLHEI	-

Peptide sequence	Immunogenicity
AVGIGIAVV	+
CINGVCWTV	+
CLAAGITYV	+
CLFKDWEEL	+
CLGGLITMV	+
CLGGLLTMV	+
CLQNALDIL	+
CLTSTVQLV	+
CQWGRLWQL	+
CVNGSCFTV	+
CVNGVCWTV	+
DLIFGLNAL	+
DLMGYIPLV	+
DMWEHAFYL	+
EILGFVFTL	+
ELVSEFSRM	+
ELVSEFSRV	+
EVKEKHEFL	+
FAFRDLCIV	+
FANHKFTLV	+
FANYKFTLV	+
FIAGLIAIV	+
FIDSYICQV	+
FIFDALAEV	+
FILGIIITV	+
FIYAGSLSA	+
FLAEDLNTV	+
FLAKLNNTV	+
FLALIICNA	+
FLDEFMEGV	+
FLDPRPLTV	+
FLDQVPFSV	+
FLEESHPGI	+
FLEPGPVTA	+
FLFLRNFSL	+
FLGGTPVCL	+
FLIVSLCPT	+
FLKDVMESM	+
FLKEPVHGV	+
FLLIRYITT	+
FLLLADARV	+
FLLSLGIHL	+
FLLTRILTI	+
FLMDRHIIV	+
FLNISWFYI	+
FLNQTDETL	+
FLPATLTMV	+
FLQMNSLRV	+

Peptide sequence	Immunogenicity
FILIFNIIV	-
FISSFLLPL	-
FIVVATAAV	-
FIYLLFASM	-
FIYSIMETI	-
FLARLHAAA	-
FLAVLSPTI	-
FLGAAGSTM	-
FLGARSPSL	-
FLGGGGAGI	-
FLHNYILYA	-
FLHYCNSYA	-
FLICHNLRA	-
FLIDLAFLI	-
FLIPKGFYA	-
FLISVIVLV	-
FLKDVMVEI	-
FLLLTSIPI	-
FLLPDAQSI	-
FLLPLTSLV	-
FLLRSIIVA	-
FLLSHDAAL	-
FLPATLTMT	-
FLRYLLFGI	-
FLSNVGHYV	-
FLSRLVLYA	-
FLSYISDTV	-
FLTGTFVTA	-
FLVIAINAM	-
FLWHVRKRV	-
FLYNVYPGA	-
FMKAVCVEV	-
FMMVLPGAA	-
FMYFCEQKL	-
FMYIESIKV	-
FQQPQFQYL	-
FTGITLFLL	-
FTLIDIWFL	-
FTLNHVLAL	-
FTLVAPVSI	-
FTNSQIFNI	-
FTSAVLLLV	-
FTSSFYNYV	-
FVARVFLGL	-
FVDFVIHGL	-
FVDTMSIYI	-
FVFILTAIL	-
FVFRSPFIV	-

Peptide sequence	Immunogenicity	Peptide sequence	Immunogenicity
FLSFASLFL	+	FVFSTSFYL	-
FLSLMSLSI	+	FVILYLLAV	-
FLTSVINRV	+	FVVALIPLV	-
FLVDAIVRV	+	GFLTSMFPK	-
FLVSQLFTF	+	GIFCFRILL	-
FLWGPRALV	+	GIGILTVIL	-
FLWGPRAYA	+	GILAFVFTL	-
FLYDDNQRV	+	GILGAVFTL	-
FLYGALLLA	+	GILGFAFTL	-
FLYRLFSIL	+	GILGFKFTL	-
FMPKVNFEV	+	GILGFVATL	-
FMVELVEGA	+	GILGFVFKL	-
FMVFLQTHI	+	GILGFVKTL	-
FMYMSLLGV	+	GILTVILGV	-
FMYSDFHFI	+	GIRPYEILA	-
FQWDSNTQL	+	GLADAFILL	-
FVDYNFTIV	+	GLDDLMSGL	-
FVFLRNFSL	+	GLFIYIPGT	-
FVFYQLFVV	+	GLFLTTEAV	-
FVWLHYYSV	+	GLHCDFACL	-
GAGIGVAVL	+	GLIACLIFV	-
GAGIGVLTA	+	GLIIISIFL	-
GALGFVFTL	+	GLILFVLAL	-
GELGFVFTL	+	GLLDRLYDL	-
GFLGFVFTL	+	GLLGWSPQA	-
GGLGFVFTL	+	GLMTAVYLV	-
GIAGFVFTL	+	GLPDSLPSL	-
GIAGGLALL	+	GLTSAVIDA	-
GIGGFVFTL	+	GLVDFVKHI	-
GIGIGVLAA	+	GLVRLNAFL	-
GIKGFVFTL	+	GLYGAQYDV	-
GILGFVFAL	+	GLYLSQIAV	-
GILGFVFTA	+	GLYPGLIWL	-
GILGFVFTK	+	GLYRQWALA	-
GILGFVFTL	+	GLYYLTTEV	-
GILGFVFTM	+	GMANTTFHV	-
GILGFVFTV	+	GMGWLTIGI	-
GILGKVFTL	+	GTDGFPFKL	-
GILKFVFTL	+	GTYAVNIHV	-
GIMGFVFTL	+	HLIFSYAFL	-
GITFQVWDV	+	HLIKIPLLI	-
GIVGFVFTL	+	HLMFYTLPI	-
GKLGFVFTL	+	HLSLRGLPV	-
GLADTVVAC	+	HTICDDYFV	-
GLCTLVAML	+	IIAIVFVFI	-
GLDCARLEI	+	IILFILFFA	-
GLDSYVRSL	+	IILNGSLLT	-
GLDTYVRSL	+	IILVAIAVV	-

Peptide sequence	Immunogenicity
GLDVLTAKV	+
GLFDFVNFV	+
GLHCYEQLV	+
GLIEKNIEL	+
GLIMVLSFL	+
GLIQLVEGV	+
GLISLILQI	+
GLKAGVIAV	+
GLLGFVFTL	+
GLLGNVSTV	+
GLLGTLVQL	+
GLLHHAPSL	+
GLMKYIGEV	+
GLNDYLHSV	+
GLPVEYLQV	+
GLQDCTMLV	+
GLSGGTPSK	+
GLSRYVARL	+
GLSRYVPRL	+
GLVGLVTFL	+
GLYDGMEHL	+
GMFNMLSTV	+
GMGPSLIGL	+
GMLGFVFTL	+
GMSRIGMEV	+
GQLGFVFTL	+
GQTEPIAFV	+
GTLGFVFTL	+
GTLGIVCPI	+
GVALQTMKQ	+
GVLGFVFTL	+
GVPVDPSRV	+
HACWPAFTV	+
HIFYQLANV	+
HILLGVFML	+
HLGNVKYLV	+
HLSLRGLFV	+
HLSTAFARV	+
HLYQGCQVV	+
HMTEVVRHC	+
HMWNFISGI	+
HMWNFITGI	+
HMYFTFFDV	+
HVDGKILFV	+
IAGIGILAI	+
IISAVVGIL	+
ILAGYGAGV	+
ILAKFLHWL	+
IIMAINVFT	-
IISCTCPTV	-
IISLWDQSL	-
IISTFHLSI	-
IISYIILFI	-
ILAADLEKL	-
ILAIIFLVL	-
ILDPKTGLV	-
ILDSFDPLR	-
ILFDGHDLL	-
ILFEPVHGV	-
ILFIMFMLI	-
ILFTFLHLA	-
ILGADPLRV	-
ILKEYVHGV	-
ILLPWFVDL	-
ILLSIARVV	-
ILMYPTTLL	-
ILPVIFLSI	-
ILQYDLWNV	-
ILSCIFAFI	-
ILSDDMLNI	-
ILSPFMPLL	-
ILSPLTKGI	-
ILTAILFFM	-
ILTLDIFYL	-
ILVCYILYI	-
ILVGYMSNL	-
ILWEPVHGV	-
ILYAAFLWL	-
ILYDNVVTL	-
ILYEPVHGV	-
ILYFIAFAL	-
IMEYHLLFA	-
IMFMLIFNV	-
IMFTCMDPL	-
IMYTYFSNT	-
ITNGYLISI	-
IVFVFILTA	-
IVQENNGAV	-
KIDSTSFSV	-
KINIFMAFL	-
KLFTDNNFL	-
KLFTHDIML	-
KLFYVYYNL	-
KLGNLLLLI	-
KLHLYSHPI	-
KLIGITAIM	-

Peptide sequence	Immunogenicity	Peptide sequence	Immunogenicity
ILALVQEKI	+	KLITFLFVI	-
ILARNLVPM	+	KLLKMVTSV	-
ILDDIGHGV	+	KLLLWFNYL	-
ILDDNLYKV	+	KLLWFLTGT	-
ILDEERDKV	+	KLQAAPYIV	-
ILDQVPFSV	+	KLSCAVHLI	-
ILDSFDPLV	+	KLSDGVAVL	-
ILGFVFTLT	+	KLVSISNFI	-
ILHEPVHGV	+	KLWGLVDFV	-
ILHNGAYSL	+	KLYCSYEVA	-
ILIEHLYGL	+	KLYIALCKV	-
ILKEPVHGV	+	KLYLVDYGL	-
ILKSPVHGV	+	KLYTIVSTL	-
ILLEPVHGV	+	KMHDVIAPA	-
ILLGSLSDL	+	KMMLFYMDL	-
ILLNKHIDA	+	KMNIQFTAV	-
ILLRDAGLV	+	KMSVRETLV	-
ILPDPLKPT	+	KTLLSLALV	-
ILPSKSLEV	+	KTMAVTYEL	-
ILQDMRNTI	+	KVLSIMAFI	-
ILSDENYLL	+	KVVSLVILA	-
ILSLELMKL	+	KVYDKLFPV	-
ILSPFLPLL	+	LAALFMYYA	-
ILTVILGVL	+	LAAVLVVMA	-
IMDQVPFSV	+	LIAGIILLI	-
IMELATAGI	+	LIALSVLAV	-
IMIGHLVGV	+	LIGDDVDSV	-
IMIGVLVGV	+	LILSLTCSV	-
IMLCLIAAV	+	LIMFEQYFI	-
IMMGVLVGV	+	LIMIYFFII	-
IMNDMPIYM	+	LIMYSVIGV	-
IMSSFEFQV	+	LIPETVPYI	-
IMTSYQYLI	+	LIQEIVHEV	-
IMVLSFLFL	+	LISIFLHLV	-
ITDQVPFSV	+	LITGRLAAL	-
ITDQVPGSV	+	LIVGIIFTA	-
KASEKIFYV	+	LIVRYLIQV	-
KIDYYIPYV	+	LLAFTNPTV	-
KIFGSLAFL	+	LLAQFTSAI	-
KILGFVFTL	+	LLARNSFEV	-
KILSVFFLA	+	LLATLTMTV	-
KIMDQVQQA	+	LLDLFGPEV	-
KLAGGVAVI	+	LLFFLALSI	-
KLAGGVAVL	+	LLFILFYFA	-
KLDVGNAEV	+	LLFRFMRPL	-
KLEDENPWL	+	LLGANSFEV	-
KLEDLERDL	+	LLGLWGTAA	-
KLGEFYNQM	+	LLGRASFEV	-

Peptide sequence	Immunogenicity	Peptide sequence	Immunogenicity
KLGGYVSFV	+	LLGRDSFEV	-
KLIANNTRV	+	LLGRNAFEV	-
KLLGQINLV	+	LLGRNSAEV	-
KLLMVLMLA	+	LLGRNSFAV	-
KLLPEGYWV	+	LLGRRSFEV	-
KLLRYYTEI	+	LLIHFLLSL	-
KLMLDIHTV	+	LLILSCIFA	-
KLMPNITLL	+	LLLCGVSLV	-
KLQDCTMLV	+	LLLDYMTST	-
KLQELNYNL	+	LLLEVEQEI	-
KLQEQQSDL	+	LLLFHETGV	-
KLSEQESLL	+	LLLGLWGTA	-
KLTEAITAA	+	LLLIALWNL	-
KLTPLCVTL	+	LLLIVTPVV	-
KLTSCNTSV	+	LLLNCLWSV	-
KLTSLCNTV	+	LLLWPLYVL	-
KLVANNTRL	+	LLMMTLPSI	-
KLWCRHFCV	+	LLMTSLQYA	-
KLWESPQEI	+	LLNATAIAV	-
KMSSAVGFV	+	LLNLLLWPL	-
KMVELVHFL	+	LLNPCLINV	-
KTLPLCVTL	+	LLPENNVLS	-
KTWGQYWQV	+	LLPLGYPFV	-
KVAEIVHFL	+	LLQYWSQEL	-
KVAELVHFL	+	LLSEFCRVL	-
KVAELVWFL	+	LLSEIRFYI	-
KVDDTFYYV	+	LLSLALVGA	-
KVINYLVML	+	LLSLFSTLV	-
KVLEYVIKV	+	LLSLLVIWI	-
KVSPYLFNV	+	LLSQYLSRV	-
LAALPHSCL	+	LLTEVETYV	-
LAARAIVAI	+	LLVAPMPTA	-
LAGIGLIAA	+	LLVDLLWLL	-
LALLLLDRL	+	LLVEPCARV	-
LIEDFDIYV	+	LLVQRVTSV	-
LIFGHLPRV	+	LLWFHISCL	-
LITGRLQSL	+	LLWQDPVPA	-
LIVIGILIL	+	LLYAHINAL	-
LKLSGVVRL	+	LLYILRYIV	-
LLCLIFLLV	+	LLYPTAVDL	-
LLCPAGHAV	+	LMDCIMFDA	-
LLCPSGHVV	+	LMDMITLSL	-
LLCPTGHAV	+	LMDSIFVST	-
LLDAHIPQL	+	LMIEYNLLT	-
LLDDEAGPL	+	LMIFISSFL	-
LLDPRVRGL	+	LMLPGMNGI	-
LLDRFLATV	+	LMNNAFEWI	-
LLDVPTAAV	+	LMSTLLIYL	-

Peptide sequence	Immunogenicity	Peptide sequence	Immunogenicity
LLFGYPVYV	+	LTLDEQIFV	-
LLFLLLADA	+	LTVILGVLL	-
LLGATCMFV	+	LTYSQLMTL	-
LLGNCLPTV	+	LVITINYFL	-
LLGRNSFEV	+	LVYVNGVVV	-
LLGTFTWTL	+	MACLVPAAT	-
LLHTDFEQV	+	MALIGDSTV	-
LLIDPTSGL	+	MALLRLPLV	-
LLLAARAIV	+	MAWGGSYIA	-
LLLCLIFLL	+	MIANALDAV	-
LLLDRLNQL	+	MIFISSFLL	-
LLLGPLGPL	+	MIYGLIACL	-
LLLLTVLTV	+	MLASTLTDA	-
LLMDCSGSI	+	MLDDFSAGA	-
LLMGTLGIV	+	MLGNAPSVV	-
LLNATDIAV	+	MLLALVALV	-
LLNCAVTKL	+	MLLHVGIPL	-
LLNQLQVNL	+	MLLNVQTLI	-
LLPRRGPRL	+	MLMEVFPQL	-
LLQAEAPRL	+	MLMFIFTGI	-
LLSAWILTA	+	MLNGIMYRL	-
LLTEVETPI	+	MLQDMAILT	-
LLWAARPRL	+	MMIDDFGTA	-
LLWKGEGAV	+	MMKTYIEFV	-
LLWSYAMGV	+	MMSCSSEAT	-
LLWTLVVLL	+	MMWYWGPSL	-
LMDALKLSI	+	MQLIYDSSL	-
LMIGTAAAV	+	MTFGDIPLV	-
LMIIPLINV	+	MTSCVSEQL	-
LMNGQQIFL	+	MVNTVLITV	-
LMVLMLAAL	+	NIAEGLRAL	-
LMWAKIGPV	+	NIAEYIAGL	-
LMWDNVGLV	+	NILQKIEKI	-
LMWYELSKI	+	NISGYNFSL	-
LMYDIINSV	+	NISTILYFT	-
LQLPQGTTL	+	NLATSIYTI	-
LQTTIHDII	+	NLDDVYSYI	-
LTAGFLIFL	+	NLDLFMSHV	-
LVCGKDGVK	+	NLDTSPFFV	-
LVHFLLLKY	+	NLFDIPLLT	-
LVMAQLLRI	+	NLFPYLVSA	-
LVQENYLEY	+	NLFTFLHEI	-
LVVADLSFI	+	NLGKVIDTL	-
LVVLGLLAV	+	NLLLWPLYV	-
MIAVFLPIV	+	NLNESLIDL	-
MINAYLDKL	+	NMISDTIFV	-
MLDLQPETT	+	NMQTVKLFV	-
MLGTHTMEV	+	NVFKYLTSV	-

Peptide sequence	Immunogenicity	Peptide sequence	Immunogenicity
MLLALLYCL	+	NVIGLIVIL	-
MLLAVLYCL	+	NVLLYNRLL	-
MLNIPSINV	+	PLLPIFFCL	-
MLTNSCVKL	+	PLSSSVPSQ	-
MLWGWREHV	+	QIFEVYWYL	-
MMLVPLITV	+	QIITLTAFV	-
MTYAAPLFV	+	QLAGYILTV	-
NAHQILPKV	+	QLDHGVLLV	-
NLLKVNIHI	+	QLDPARDVL	-
NLMEQPIKV	+	QLFHLCLII	-
NLNDNAIHL	+	QLFKYVPSA	-
NLQSLTNLL	+	QLSDVIDRL	-
NLTISDVSV	+	QLTPHTKAV	-
NLVPMVATV	+	QLVWENFLA	-
NLWNGIVPT	+	QLWPEEIGV	-
NMFTPYIGV	+	QMLLALARL	-
NMLSTVLGV	+	QMWQARLTV	-
PLDGEYFTL	+	RALSLAAVL	-
PLEEELPRL	+	RIEENLEGV	-
PLKQHFQIV	+	RILPYTFKI	-
QAGIGILLA	+	RLFDFNKQA	-
QLSLLMWIT	+	RLFSYNFTT	-
RIAECILGM	+	RLGATIWQL	-
RIFAELEGV	+	RLHLWLSDM	-
RILGAVAKV	+	RLIQNSLTI	-
RLAEYQAYI	+	RLLDDTPEV	-
RLCCQLDPA	+	RLLGTFTWT	-
RLDSYVRSL	+	RLLSPTTIV	-
RLGRNSFEV	+	RLMIGTAAA	-
RLIDFLKDV	+	RLNDFLGLL	-
RLIGHISTL	+	RLNKRSYLI	-
RLLDRLVRL	+	RLRDLNQAV	-
RLLQETELV	+	RLVDFFPDI	-
RLLQTGIHV	+	RLVSGLVGA	-
RLMKQDFSV	+	RLYDLTRYA	-
RLMRTNFLI	+	RMAWGGSYI	-
RLNEVAKNL	+	RMFAANLGV	-
RLNMFTPYI	+	RMILYLESV	-
RLNQLESKV	+	RMPAVTDLV	-
RLPLVLPAV	+	RMQFSSFTV	-
RLPRIFCSC	+	RQIFIHYSV	-
RLQGISPKI	+	RVFTSAVLL	-
RLSSCVPVA	+	SIFGFQAEV	-
RLTRFLSRV	+	SIHVTVSNV	-
RLVDDFLLV	+	SIMAFILGI	-
RLVNGSLAL	+	SIVCIVAAV	-
RLVTLKDIV	+	SIYECITFL	-
RLWHYPCTA	+	SLAGFVRML	-

Peptide sequence	Immunogenicity	Peptide sequence	Immunogenicity
RLWHYPCTF	+	SLALVGACI	-
RLWHYPCTI	+	SLDVINYLI	-
RLWHYPCTL	+	SLFSLLLVI	-
RLWHYPCTV	+	SLHVGTQCA	-
RLYDYFTRV	+	SLIYYQNEV	-
RMFPNAPYL	+	SLLEIGEGV	-
RMPEAAPPV	+	SLLPATLTV	-
RTLDKVLEV	+	SLLYLILFL	-
RVIEVLQRA	+	SLSLHPLYV	-
RVYEALYYV	+	SLSVVRPMT	-
SILEGIANV	+	SLVENNFFT	-
SILLRDAGL	+	SLVRLVYIL	-
SITEVECFL	+	SLVYVNGVV	-
SLDDYNHLV	+	SLYAVSPSV	-
SLDQSVVEL	+	SMDTLLFFL	-
SLEENIVIL	+	SMIGLCACV	-
SLENFRAYV	+	SMLGIWFFT	-
SLFNTIATL	+	SMSSYDFST	-
SLFNTVATL	+	SQYYFSMLV	-
SLFPGKLEV	+	SSVVNNVAR	-
SLGGLLTMV	+	STSFYLISI	-
SLGSPVLGL	+	SVIFYFISI	-
SLLLELEEV	+	TIIALLFAL	-
SLLMWITQA	+	TLAPFNFLV	-
SLLMWITQC	+	TLARDIVLV	-
SLLMWITQS	+	TLFLLFLEI	-
SLLMWITQV	+	TLGLSAMST	-
SLLQHLIGL	+	TLIDIWFLA	-
SLLSEFCRV	+	TLLGLILFV	-
SLMAFTAAV	+	TLLVDLLWL	-
SLMAFTASI	+	TLLYATVEV	-
SLNQTVHSL	+	TLLYPLFNL	-
SLQALKVTV	+	TLSNVEVFM	-
SLQPEDFAL	+	TLSSPSPSA	-
SLQPEDFAT	+	TLTEDFFVV	-
SLRAEDTAV	+	TLVIPSWHV	-
SLREWLLRI	+	TLYDFDYYI	-
SLSAYIIRV	+	TMLSIILVI	-
SLSEKTVLL	+	TMWCLTLFV	-
SLSKILDTV	+	TTAEEAAGI	-
SLSRFSWGA	+	TVILGVLLL	-
SLVIVTTFV	+	TVLRFVPPL	-
SLYADSPSV	+	TVQEFIFSA	-
SLYFGGICV	+	VIDEILFKV	-
SLYITVATL	+	VIGDQYVKV	-
SLYITVAVL	+	VIHDYICWL	-
SLYKGVYEL	+	VISVIFYFI	-
SLYNAVATL	+	VIVIYIFTV	-

Peptide sequence	Immunogenicity	Peptide sequence	Immunogenicity
SLYNTIATL	+	VLADANETL	-
SLYNTIAVL	+	VLALYSPPL	-
SLYNTVATL	+	VLDTTLYAV	-
SLYNTVAVL	+	VLGGCRHKL	-
SMAGSSAMI	+	VLGRLDQKL	-
SMHFYGWSL	+	VLIAGIILL	-
SMIEAESSV	+	VLIALSVLA	-
SMMWMRFFV	+	VLKAAGVPV	-
SMNATLVQA	+	VLKDAIKDL	-
SMVGNWAKV	+	VLLLDVTPL	-
SPEKHHCTV	+	VLLLVVVMM	-
SSKALQRPV	+	VLLVSLGAI	-
STAPPAHGV	+	VLMTETRNL	-
STAPPVHNV	+	VLSPLPSQA	-
STPPPGTRV	+	VLVVMACLV	-
SVASTITGV	+	VLYPVIFIT	-
SVFRENLFL	+	VMKLFTISV	-
SVRDRLARL	+	VMMSCSSEA	-
SVYDFFVWL	+	VMYAFTTPL	-
TILLGIFFL	+	WAFSAIGNV	-
TIMAFRWVT	+	WIIKNSWTA	-
TINPQVSKT	+	WIVQENNGA	-
TLDSQVMSL	+	WLGAAITLV	-
TLEEFSAKL	+	WLGETFHGL	-
TLFIGSHVV	+	WLIGFDFDV	-
TLFLQMNSL	+	WLLIDTSNA	-
TLGIVCPIC	+	WLLSVLAAV	-
TLHEYMLDL	+	WLTSILLSL	-
TLLANVTAV	+	WLWYIKIFI	-
TLLNNCTRV	+	WMMAMKYPI	-
TLLNVIKSV	+	YATVEVPSL	-
TLLVYLFSL	+	YAYGWIPET	-
TLNAWVKVV	+	YFLEILWRL	-
TLNDLETDV	+	YIIDWMVDI	-
TLTSCNTSV	+	YIIKNTFNV	-
TLVCGKDGV	+	YIIRVTTEL	-
TLWVDPYEV	+	YILCNMALL	-
TLYLQMNSL	+	YILYIVFCI	-
TMYGGISLL	+	YINRALAQI	-
VCTEGKSKL	+	YIWIKNLET	-
VDGIGILTI	+	YIYYFFIRL	-
VILGVLLLI	+	YLAKLTALV	-
VIYQTMDDL	+	YLCCQLDPA	-
VIYQYMDDL	+	YLCTFMIIT	-
VLAELVKQI	+	YLDFLLLLL	-
VLAGLLGNV	+	YLDLALMSV	-
VLAGVGFFI	+	YLFGGFSTL	-
VLCLRPVGA	+	YLFNAIETM	-

Peptide sequence	Immunogenicity	Peptide sequence	Immunogenicity
VLDDLSMYL	+	YLFRIVSTV	-
VLDGLDVLL	+	YLGGCRHKL	-
VLEETSVML	+	YLGPRVCWL	-
VLFSSDFRI	+	YLILFLLFV	-
VLGPISGHV	+	YLIPAVTSL	-
VLHDDLLEA	+	YLKEYIPKA	-
VLHKRTLGL	+	YLKIGTLLV	-
VLLCESTAV	+	YLLALRYLA	-
VLLDYQGML	+	YLLAVCGCI	-
VLLPSLFLL	+	YLLCCNYKL	-
VLPDVFIRC	+	YLLDDVLYT	-
VLPFDIKKL	+	YLLFGIKCI	-
VLPHETRLL	+	YLLGDSDSV	-
VLQAGFFLL	+	YLLPGFVLT	-
VLQELNVTV	+	YLMDEEVPA	-
VLQWASLAV	+	YLMDELRYV	-
VLSDFKTWL	+	YLMDKLNLT	-
VLVGGVLAA	+	YLMKDKLNI	-
VLVKSPNHV	+	YLMPYSVYI	-
VLYDEFVTI	+	YLNMSRLFV	-
VMACLVPAA	+	YLRLYIILA	-
VVFLHVTYV	+	YLSAKITTL	-
VVLGVVFGI	+	YLSEGDMAA	-
VVQELLWFL	+	YLSIYGFGV	-
WLDQVPFSV	+	YLSKCTLAV	-
WLGNHGFEV	+	YLSSWTPVV	-
WLNEVEFKL	+	YLTAIQDFI	-
WLQYFPNPV	+	YLTVFTVYL	-
WLSDCGEAL	+	YLVSFGVWI	-
WLSLLVPFV	+	YLVSSLSEI	-
WMNRLIAFA	+	YLYALYSPL	-
YAIDLPVSV	+	YLYQPCDLL	-
YIGEVLVSV	+	YLYVHSPAL	-
YIGSGDSPV	+	YMFFVIKNL	-
YIIIGDSPV	+	YMMGIEYGL	-
YIILGDSPV	+	YMNYYTTYI	-
YIISGDLPV	+	YQLAGYILT	-
YIISGDSPL	+	YQLFVVFGL	-
YIISGDSWV	+	YQSFLFWFL	-
YIISGISPV	+	YQYVRLHEM	-
YILEETSVM	+	YTALHYYYL	-
YIYGIPLSL	+	YTFLYNFWT	-
YLAGAATMV	+	YTIERIFNA	-
YLCLRPVGA	+	YTINCLLYI	-
YLDKVRATV	+	YTQDELINV	-
YLDPAQQNL	+	YTYAFTKKV	-
YLDQVPFSV	+	YTYKWETFL	-
YLEPGPVTA	+	YVHGDTYSL	-

Peptide sequence	Immunogenicity
YLEPGPVTI	+
YLEPGPVTL	+
YLEPGPVTV	+
YLFKDWEEL	+
YLFSLVVLV	+
YLGEVIVSV	+
YLGEVLVSV	+
YLHDPEFNL	+
YLHKRTLGL	+
YLIELIDRV	+
YLIKLIEPV	+
YLISGDSPV	+
YLISIFLHL	+
YLKEPVHGV	+
YLKKIKNSL	+
YLKKIQNSL	+
YLLDGLRAQ	+
YLLDRGADI	+
YLLEMLWRL	+
YLMDTSGKV	+
YLNKIQNSL	+
YLPEVISTI	+
YLQQNWWTL	+
YLSGANLNL	+
YLVAYQATV	+
YLVSIFLHL	+
YLVTRHADV	+
YMDDVVLGA	+
YMDGTMSQV	+
YMLAHVTGL	+
YMLDLQPET	+
YMNGTMSQV	+
YSSPTLQSV	+
YVDPVITSI	+
YVNAILYQI	+
YVLLAVLFV	-
YVLLHLLVV	-
YVMTMILFL	-
YVPGYSITT	-

Appendix F

The validation dataset

Peptide sequence	Immunogenicity	Peptide sequence	Immunogenicity
AINGVMWTV	+	ALAAYCLST	-
AISANIADI	+	ALAHGVRAL	-
AITEVECFL	+	ALVVGVVCA	-
ALEGSLQKR	+	ALYDVIQKL	-
ALENNYEVL	+	ALYGVWPLL	-
ALMLLNNYV	+	ARNLVPMVA	-
ALQAIELQL	+	ATATELNNA	-
ALQTGITLV	+	AVCKVCLRL	-
ALRCASPWL	+	CLSTGCVVI	-
ALSSSLGNV	+	CLVDYPYRL	-
ALTAVAEEV	+	CMSADLEVV	-
ALWALPHAA	+	CVSGACWTV	-
ALWDSKFFT	+	DIWDWICEV	-
ALYDVVSKL	+	EVRTLQQLL	-
ALYDVVTKL	+	FISGIQYLA	-
ALYEVVSKL	+	FKDGIYFAA	-
AMAQDPHSL	+	FLARLIWWL	-
AMARDPHSL	+	FLYNFWTNV	-
AVGGAVASV	+	FVFADLRIV	-
AVNGVLWTV	+	FVSLLAPGA	-
AVNGVMWTV	+	GLLGASMDL	-
CINGLCWTV	+	GLSPAITKY	-
CINGVCWSI	+	GPGLSPGTL	-
CINGVCWSV	+	IASPKGPVI	-
CINGVMWTL	+	ILIYNGWYA	-
CISGVCWTV	+	ILSPGALVV	-
CLGGLLYMV	+	IMSGEVPST	-
CLTEYILWV	+	IMTCMSADL	-
CMLGDPVPT	+	IMVSEHFSL	-
CTNGVCWTV	+	ISVVLIFVV	-
CVNGACWTV	+	KLFNKVPIV	-
DCLVFLAPA	+	KLLKDHFDL	-
DLLEEGNTL	+	KLMPQLPGI	-
DLMGYLPLV	+	KLYAAIFGV	-
DLPPPPPLL	+	KVCGAPPCV	-
DLSPGLPAA	+	KVLVLNPSV	-
DRFYKTLRA	+	LARGSPPSV	-
EEYLQAFTY	+	LDYQGMLPV	-
ELFQDLSQL	+	LIHLHQNIV	-
ELSPLLLST	+	LLFDSNEPI	-

Peptide sequence	Immunogenicity
ERYLKDQQL	+
FFDLPLPWL	+
FFDLPLPWT	+
FIDNTDSVV	+
FISDKIKFL	+
FLAPAKAVV	+
FLDLPLPWL	+
FLDLPLPWT	+
FLGGTRVCL	+
FLHCIVFNV	+
FLLTRILTL	+
FLLVAHYAI	+
FLNTEPSQL	+
FLPRNIGNA	+
FLQDVMNIL	+
FLTKRGGQV	+
FLTKRGRQV	+
FLTKRSGQV	+
FLTKRSRQV	+
FLTRVEAQL	+
FLWEDQTLL	+
FLYALALLL	+
FLYNRPLNS	+
FMTSSWWGA	+
FMTSSWWRA	+
FTSAVLLLL	+
FTWAGKAVL	+
FTWAGQAVL	+
FVANFSMEL	+
FVEALARSI	+
FVSPSLVSA	+
GILGFVFLT	+
GIPPAPHGV	+
GIPPAPRGV	+
GLCEREDLL	+
GLCPHCINV	+
GLFPTQIQV	+
GLGTLGAAI	+
GLIYNRMGA	+
GLKGGPSTE	+
GLLNKLENI	+
GLPPPPPLL	+
GLPRYVARL	+
GLPRYVVCL	+
GLREREDLL	+
LLGCAANWI	-
LLGCIITSL	-
LLIPKSFTL	-
LLLFAGVDA	-
LLLIWFRPV	-
LLMMSVYAL	-
LLSCLTVPA	-
LLSKNTFYL	-
LVLQAGFFL	-
LVPMVATVQ	-
MCLRFLSKI	-
MLTDPSHIT	-
MMNWSPTTA	-
MVATVQGQN	-
MVGNWAKVL	-
NLPGCSFSI	-
NLSWLSLDV	-
PLIPTTAVI	-
PLLCPAGHA	-
QLLMGTCTI	-
QLLRIPQAI	-
QLRSVIRAL	-
QMWKCLIRL	-
RAYMNTPGL	-
RLCVQSTHV	-
RLLLLDEEA	-
RNLVPMVAT	-
RVGLHEYPV	-
SAIIGIYLL	-
TLFFFLLAL	-
TLHDLCQAL	-
TLHGPTPLL	-
TLKKCLNEI	-
TLPGNPAIA	-
TLQQLLMGT	-
TLREYILDL	-
TSICSLYQL	-
TVGDVMWTV	-
TVGGVIWTV	-
TVGGVTWTV	-
VLVLNPSVA	-
VLVVLLLFA	-
VPMVATVQG	-
VVLLLFAGV	-
VVTSTWVLV	-

Peptide sequence	Immunogenicity
GLTEEIDYV	+
GLWRHSPCA	+
GMWESNANV	+
GTMDCTHPL	+
GTMDCTHSL	+
GTWESNANV	+
GVFIQVYEV	+
HAVGIFRAA	+
HLAFQLSSI	+
HLFYSAVLL	+
HLWVKNMFL	+
HLWVKNVFL	+
HNFTLVASV	+
ILHSRTEFV	+
ILKSLGFKV	+
ILLMRTTWA	+
ILLNEVPYV	+
ILNPVASSL	+
ILPLHGPEA	+
ILSFLPWLV	+
ILVGRLRAA	+
ILYISFYFI	+
IMAIELAEL	+
IMIHDLCLA	+
IMIHDLCLV	+
KCQEVLAWL	+
KIQRNLRTL	+
KIQRNLWTL	+
KIYSENLKL	+
KIYSENLTL	+
KLAKLIIDL	+
KLCPVQLWV	+
KLEELHENV	+
KLQAPVQEL	+
KLQATVQEL	+
KLWEWLGYL	+
KLYSENLKL	+
KLYSENLTL	+
KMNVFDTNL	+
KTCPVQLWV	+
KTGECCLYM	+
LAGSSLNLV	+
LAGSSLNPV	+
LASEKVYAI	+
LASEKVYTI	+

Peptide sequence	Immunogenicity
WLLRDDWLL	-
WLRDIWDWI	-
WLVSNGSYL	-
WQASLALSY	-
WTVYHGQGT	-
WVLVGGVLA	-
YFDDVTAFL	-
YIEQGMMLA	-
YILDIQPQG	-
YIPLVGAPL	-
YLALDPDSV	-
YLDGQLARL	-
YLLEMIWRL	-
YLLYRMLKT	-
YMLGLKPEV	-
YMLILHPET	-

Peptide sequence	Immunogenicity	Peptide sequence	Immunogenicity
LIFDLGGGT	+		
LIILPEDCL	+		
LIISPLPRV	+		
LIQETLLFV	+		
LLDEGKQSL	+		
LLLTLLATV	+		
LLNGWRWRL	+		
LLNLPVWVL	+		
LLQEYNWEL	+		
LLQMMQICL	+		
LLQMMQVCL	+		
LLQQYCLYL	+		
LLSDEDVAL	+		
LLSDEDVEL	+		
LLSPLHCWA	+		
LLVLFIVYV	+		
LLWNGPMAV	+		
LLYGGVPEV	+		
LMDENTYAM	+		
LMLLKNGTV	+		
LNLPDKMFL	+		
LQSRGYSSL	+		
LQTHIFAEV	+		
LVMAQLLRT	+		
LVMLLVHYA	+		
LVVSQLLRI	+		
MLPSQPTLL	+		
MLVALLGAM	+		
MLVTLPVYS	+		
MMLPSQPTL	+		
MMLPSRPTL	+		
MMMGMFNML	+		
MMQICLHHL	+		
MMQVCLHHL	+		
MVWESGCTV	+		
NCLKLLESL	+		
NLLCHIYSL	+		
NLLGRFELI	+		
NLLKRWQFV	+		
NVMLVTLPV	+		
PILQERPPL	+		
PLDGGVAAA	+		
PLHCWAVLL	+		
PLHCWVVLL	+		
PLPEAPLSL	+		

Peptide sequence	Immunogenicity
QIAILVTTV	+
QLCAKVPLL	+
QLGAFLTNV	+
QLGRISLLL	+
QLPEATFMV	+
QLPPTAPPL	+
RAIEAQQHL	+
RAPPTTPAL	+
RCHELTVSL	+
RIGQRQETV	+
RLAGSSLNL	+
RLEIPAIEL	+
RLGVRATRK	+
RLLPLLALL	+
RLLPLWAAL	+
RLLSPLSPL	+
RLQREWHTL	+
RLQVPVEAV	+
RLRAPEVFL	+
RLRPLCCTA	+
RLSCPSPRA	+
RLSCSSPRA	+
RLTSTNPTM	+
RQAGDFHQV	+
RQVGDFHQV	+
RQVGDFHYV	+
RTGEVKWSV	+
SFTEVECFL	+
SILELLQFV	+
SINGVMWTV	+
SIQNYHPFA	+
SLASLLPHV	+
SLFKNVRLL	+
SLGIMAIEL	+
SLLNLPVWV	+
SLLSLPVWV	+
SLPPPGTRV	+
SLQPLALEG	+
SLQRMVQEL	+
SLQRTVQEL	+
SLQSMVQEL	+
SLQSTVQEL	+
SLTAISTTL	+
SLTTISTTL	+
SLWQLGAAV	+

Peptide sequence	Immunogenicity	Peptide sequence	Immunogenicity
SLWQLRAAV	+		
SMPQGTFPV	+		
SMSKEAVAI	+		
SMVGNMAKV	+		
STLQGLTSV	+		
SVASLLPHV	+		
SVNGVMWTV	+		
TIADFWQMV	+		
TINGVLWTV	+		
TIPTPLQPL	+		
TLGQHLPTL	+		
TLPPRPDHI	+		
TLTTGEWAV	+		
TLWGSFVDV	+		
TMLDIQPED	+		
TMLGRRAPI	+		
TMLGRRPPI	+		
TPQDLNTML	+		
TQPGPLAPL	+		
TQPGPLVPL	+		
TTYQRTRAL	+		
TVASRLGPV	+		
TVNGVLWTV	+		
TYLPTNASL	+		
VIFCHPGQL	+		
VIFDFLHCI	+		
VILAGPCIL	+		
VLASIEAEL	+		
VLASIEPEL	+		
VLATAVREL	+		
VLAWTRAFV	+		
VLDKVEETV	+		
VLDSFKTWL	+		
VLFGLLCLL	+		
VLNSLASLL	+		
VLNSVASLL	+		
VLQAGFFIL	+		
VLRDDLLEA	+		
VLSDFKSWL	+		
VLSDFRTWL	+		
VLTDFKTWL	+		
VLVEGSTRI	+		
VLYSPNVSV	+		
VVSDFKTWL	+		
YIDDVVLGA	+		

Peptide sequence	Immunogenicity	Peptide sequence	Immunogenicity
YILDLQPEN	+		
YLGSYGFRL	+		
YMDNNLFYV	+		
YQGSYGFRL	+		
YVDDVVLGA	+		
YVFDRILKV	+		
FLLRHLSSV	+		
ILQEAEQMV	+		
QLLDEGKEL	+		
QLLESLAPL	+		
SLYQLENYC	+		
YLLEEIYTV	+		
YLMQKLQNV	+		

Bibliography

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

2. Miller JF: **Self-nonself discrimination and tolerance in T and B lymphocytes**. *Immunol Res* 1993, **12**:115–130.

3. Alberts B, Johnson A, Lewis J: **Molecular Biology of the Cell. New York, NY: Garland Science; 2002**. 2011.

4. Mayer G: **Immunology-Chapter One: Innate (non-specific) Immunity**. *Microbiology and Immunology On-Line Textbook* 2006.

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

6. Maton A: *Human biology and health*. Prentice Hall; 1997.

7. Delves PJ, Martin SJ, Burton DR, Roitt IM: *Roitt's essential immunology*. Wiley-Blackwell; 2011.

8. Harvey RA, Champe PC: **Lippincott's Illustrated Reviews: Immunology**. 2008.

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

10. Mix E, Goertsches R, Zett UK: **Immunoglobulins--basic considerations**. *J Neurol* 2006, **253 Suppl** :V9–17.

11. Ravetch J V, Bolland S: **IgG Fc receptors**. *Annu Rev Immunol* 2001, **19**:275–290.

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

13. Sinha JK, Bhattacharya S: *A Text Book of Immunology*. Academic Publishers;

14. Plotkin SA, Orenstein WA, Offit PA: **Vaccines. ed, ed**. *APaWAO Stanley. Philadelphia: Saunders Elsevier* 2008, **1725**.

15. Plotkin SA: **Vaccines: past, present and future**. *Nat Med* 2005, **11**:S5–11.

16. Dagan R, Poolman J, Siegrist CA: **Glycoconjugate vaccines and immune interference: A review**. *Vaccine* 2010, **28**:5513–5523.

17. Greenberg RN, Marbury TC, Foglia G, Warny M: **Phase I dose finding studies of an adjuvanted Clostridium difficile toxoid vaccine**. *Vaccine* 2012, **30**:2245–2249.

18. Kim W, Liau LM: **Dendritic cell vaccines for brain tumors**. *Neurosurg Clin N Am* 2010, **21**:139–157.

19. Meri S, Jordens M, Jarva H: **Microbial complement inhibitors as vaccines**. *Vaccine* 2008, **26 Suppl 8**:I113–7.

20. 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.

21. 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.

22. 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.

23. 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.

24. 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.

25. Rotzschke O, Falk K, Stevanovic S, Jung G, Walden P, Rammensee HG: **Exact prediction of a natural T cell epitope**. *Eur J Immunol* 1991, **21**:2891–2894.

26. Sette A, Buus S, Appella E, Smith JA, Chesnut R, Miles C, Colon SM, Grey HM: **Prediction of major histocompatibility complex binding regions of protein antigens by sequence pattern analysis**. *Proc Natl Acad Sci U S A* 1989, **86**:3296–3300.

27. Falk K, Rotzschke O, Stevanovic S, Jung G, Rammensee HG: **Allele-specific motifs revealed by sequencing of self-peptides eluted from MHC molecules**. *Nature* 1991, **351**:290–296.

28. Ruppert J, Sidney J, Celis E, Kubo RT, Grey HM, Sette A: **Prominent role of secondary anchor residues in peptide binding to HLA-A2.1 molecules**. *Cell* 1993, **74**:929–937.

29. Madden DR, Garboczi DN, Wiley DC: **The antigenic identity of peptide-MHC complexes: a comparison of the conformations of five viral peptides presented by HLA-A2**. *Cell* 1993, **75**:693–708.

30. Saper MA, Bjorkman PJ, Wiley DC: **Refined structure of the human histocompatibility antigen HLA-A2 at 2.6 A resolution**. *J Mol Biol* 1991, **219**:277–319.

31. Parker KC, Bednarek MA, Coligan JE: **Scheme for ranking potential HLA-A2 binding peptides based on independent binding of individual peptide side-chains**. *J Immunol* 1994, **152**:163–175.

32. Reche PA, Glutting JP, Reinherz EL: **Prediction of MHC class I binding peptides using profile motifs**. *Hum Immunol* 2002, **63**:701–709.

33. Nielsen M, Lundegaard C, Worning P, Hvid CS, Lamberth K, Buus S, Brunak S, Lund O: **Improved prediction of MHC class I and class II epitopes using a novel Gibbs sampling approach**. *Bioinformatics* 2004, **20**:1388–1397.

34. Bui HH, Sidney J, Peters B, Sathiamurthy M, Sinichi A, Purton KA, Mothe BR, Chisari F V, Watkins DI, Sette A: **Automated generation and evaluation of specific MHC binding predictive tools: ARB matrix applications**. *Immunogenetics* 2005, **57**:304–314.

35. Peters B, Sette A: **Generating quantitative models describing the sequence specificity of biological processes with the stabilized matrix method**. *BMC Bioinformatics* 2005, **6**:132.

36. 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.

37. 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.

38. 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.

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

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

41. 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.

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

43. 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.

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

45. 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.

46. 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.

47. Rosenfeld R, Zheng Q, Vajda S, DeLisi C: **Flexible docking of peptides to class I major-histocompatibility-complex receptors**. *Genet Anal* 1995, **12**:1–21.

48. Bui HH, Schiewe AJ, Von Grafenstein H, Haworth IS: **Structural prediction of peptides binding to MHC class I molecules**. *Proteins* 2006, **63**:43–52.

49. Antes I, Siu SW, Lengauer T: **DynaPred: a structure and sequence based method for the prediction of MHC class I binding peptide sequences and conformations**. *Bioinformatics* 2006, **22**:e16–24.

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

51. 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.

52. 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.

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

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

55. 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.

56. Rudolph MG, Luz JG, Wilson IA: **Structural and thermodynamic correlates of T cell signaling**. *Annu Rev Biophys Biomol Struct* 2002, **31**:121–149.

57. Stewart-Jones GB, McMichael AJ, Bell JI, Stuart DI, Jones EY: **A structural basis for immunodominant human T cell receptor recognition**. *Nat Immunol* 2003, **4**:657–663.

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

59. 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.

60. 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.

61. 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.

62. Uchida T: **Development of a cytotoxic T-lymphocyte-based, broadly protective influenza vaccine**. *Microbiol Immunol* 2011, **55**:19–27.

63. 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.

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

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

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

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

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

69. Perner P: *Machine learning and data mining in pattern recognition*. Springer; 2009, **5632**.

70. Chang CC, Lin CJ: **LIBSVM: A Library for Support Vector Machines**. *Acm Transactions on Intelligent Systems and Technology* 2011, **2**.

71. Nanni L: **Machine learning algorithms for T-cell epitopes prediction**. *Neurocomputing* 2006, **69**:866–868.

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

73. 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.

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

75. Cheng J, Randall AZ, Sweredoski MJ, Baldi P: **SCRATCH: a protein structure and structural feature prediction server**. *Nucleic Acids Research* 2005, **33**:W72–W76.

76. 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.

77. Rabiner LR: **A tutorial on hidden Markov models and selected applications in speech recognition**. *Proceedings of the IEEE* 1989, **77**:257–286.

78. Huang XD, Ariki Y, Jack MA: **Hidden Markov Models for Speech Recognition, volume 7 of Edinburgh Information Technology series**. 1990.

79. Bishop MJ, Thompson EA: **Maximum likelihood alignment of DNA sequences**. *J Mol Biol* 1986, **190**:159–165.

80. Delorenzi M, Speed T: **An HMM model for coiled-coil domains and a comparison with PSSM-based predictions**. *Bioinformatics* 2002, **18**:617–625.

81. Liu Q, Zhu YS, Wang BH, Li YX: **A HMM-based method to predict the transmembrane regions of beta-barrel membrane proteins**. *Comput Biol Chem* 2003, **27**:69–76.

82. Qian B, Goldstein RA: **Performance of an iterated T-HMM for homology detection**. *Bioinformatics* 2004, **20**:2175–2180.

83. Krogh A, Brown M, Mian IS, Sjolander K, Haussler D: **Hidden Markov-Models in Computational Biology - Applications to Protein Modeling**. *Journal of Molecular Biology* 1994, **235**:1501–1531.

84. Jojic V, Jojic N, Meek C, Geiger D, Siepel A, Haussler D, Heckerman D: **Efficient approximations for learning phylogenetic HMM models from data**. *Bioinformatics* 2004, **20**:161–168.

85. Azad RK, Borodovsky M: **Probabilistic methods of identifying genes in prokaryotic genomes: connections to the HMM theory**. *Brief Bioinform* 2004, **5**:118–130.

86. 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.

87. Mamitsuka H: **Predicting peptides that bind to MHC molecules using supervised learning of hidden Markov models**. *Proteins-Structure Function and Genetics* 1998, **33**:460–474.

88. Udaka K, Mamitsuka H, Nakaseko Y, Abe N: **Empirical evaluation of a dynamic experiment design method for prediction of MHC class I-binding peptides**. *J Immunol* 2002, **169**:5744–5753.

89. Brusic V, Petrovsky N, Zhang GL, Bajic VB: **Prediction of promiscuous peptides that bind HLA class I molecules**. *Immunology and Cell Biology* 2002, **80**:280–285.

90. 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.

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

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

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

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

95. 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.

96. 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.

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

98. 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.

99. 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.

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

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

102. Shastri N, Schwab S, Serwold T: **Producing nature's gene-chips: The generation of peptides for display by MHC class I molecules**. *Annual Review of Immunology* 2002, **20**:463–493.

103. Tian F, Yang L, Lv F, Yang Q, Zhou P: **In silico quantitative prediction of peptides binding affinity to human MHC molecule: an intuitive quantitative structure-activity relationship approach**. *Amino Acids* 2009, **36**:535–554.

104. Altuvia Y, Margalit H: **A structure-based approach for prediction of MHC-binding peptides**. *Methods* 2004, **34**:454–459.

105. Du QS, Wei YT, Pang ZW, Chou KC, Huang RB: **Predicting the affinity of epitope-peptides with class I MHC molecule HLA-A*0201: an application of amino acid-based peptide prediction**. *Protein Engineering Design & Selection* 2007, **20**:417–423.

106. Lundegaard C, Lund O, Kesmir C, Brunak S, Nielsen M: **Modeling the adaptive immune system: predictions and simulations**. *Bioinformatics* 2007, **23**:3265–3275.

107. Hertz T, Yanover C: **Identifying HLA supertypes by learning distance functions**. *Bioinformatics* 2007, **23**:E148–E155.

108. Reche PA, Reinherz EL: **PEPVAC: a web server for multi-epitope vaccine development based on the prediction of supertypic MHC ligands**. *Nucleic Acids Research* 2005, **33**:W138–W142.

109. Sette A, Sidney J: **Nine major HLA class I supertypes account for the vast preponderance of HLA-A and -B polymorphism**. *Immunogenetics* 1999, **50**:201–212.

110. Lund O, Nielsen M, Kesmir C, Petersen AG, Lundegaard C, Worning P, Sylvester-Hvid C, Lamberth K, Roder G, Justesen S, Buus S, Brunak S: **Definition of supertypes for HLA molecules using clustering of specificity matrices**. *Immunogenetics* 2004, **55**:797–810.

111. Doytchinova IA, Guan PP, Flower DR: **Identifiying human MHC supertypes using bioinformatic methods**. *Journal of Immunology* 2004, **172**:4314–4323.

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

113. Treanor JD: **Influenza - The goal of control**. *New England Journal of Medicine* 2007, **357**:1439–1441.

114. Bi J, Song R, Yang H, Li B, Fan J, Liu Z, Long C: **Stepwise identification of HLA-A*0201-restricted CD8+ T-cell epitope peptides from herpes simplex virus type 1 genome boosted by a StepRank scheme**. *Biopolymers* 2011, **96**:328–339.

115. Moutaftsi M, Peters B, Pasquetto V, Tscharke DC, Sidney J, Bui HH, Grey H, Sette A: **A consensus epitope prediction approach identifies the breadth of murine T(CD8+)-cell responses to vaccinia virus**. *Nat Biotechnol* 2006, **24**:817–819.

116. **IEDB Analysis Resource** [http://tools.immuneepitope.org/analyze/html_mhcibinding20090901B/download_mhc_I_binding.html].

117. Kononenko I: **Estimating attributes: analysis and extensions of RELIEF**. In *Machine Learning: ECML-94*. Springer; 1994:171–182.

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

119. Micheletti C, Seno F, Banavar JR, Maritan A: **Learning effective amino acid interactions through iterative stochastic techniques**. *Proteins-Structure Function and Genetics* 2001, **42**:422–431.

120. Simons KT, Ruczinski I, Kooperberg C, Fox BA, Bystroff C, Baker D: **Improved recognition of native-like protein structures using a combination of sequence-dependent and sequence-independent features of proteins**. *Proteins-Structure Function and Genetics* 1999, **34**:82–95.

121. Liang GZ, Yang L, Chen ZC, Mei H, Shu M, Li ZL: **A set of new amino acid descriptors applied in prediction of MHC class I binding peptides**. *European Journal of Medicinal Chemistry* 2009, **44**:1144–1154.

122. Sandberg M, Eriksson L, Jonsson J, Sjostrom M, Wold S: **New chemical descriptors relevant for the design of biologically active peptides. A multivariate characterization of 87 amino acids**. *Journal of Medicinal Chemistry* 1998, **41**:2481–2491.

123. Collantes ER, Dunn WJ: **Amino-Acid Side-Chain Descriptors for Quantitative Structure-Activity Relationship Studies of Peptide Analogs**. *Journal of Medicinal Chemistry* 1995, **38**:2705–2713.

124. Schueler-Furman O, Altuvia Y, Sette A, Margalit H: **Structure-based prediction of binding peptides to MHC class I molecules: Application to a broad range of MHC alleles**. *Protein Science* 2000, **9**:1838–1846.

125. Singh SP, Mishra BN: **Ranking of binding and nonbinding peptides to MHC class I molecules using inverse folding approach: implications for vaccine design**. *Bioinformation* 2008, **3**:72–82.

126. Scott DW, De Groot AS: **Can we prevent immunogenicity of human protein drugs?** *Annals of the Rheumatic Diseases* 2010, **69**:72–76.

127. Toussaint NC, Donnes P, Kohlbacher O: **A Mathematical Framework for the Selection of an Optimal Set of Peptides for Epitope-Based Vaccines**. *Plos Computational Biology* 2008, **4**.

128. Sollner J, Heinzel A, Summer G, Fechete R, Stipkovits L, Szathmary S, Mayer B: **Concept and application of a computational vaccinology workflow**. *Immunome Res* 2010, **6 Suppl 2**:S7.

129. 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.

130. Rao X, Hoof I, Costa AI, Van Baarle D, Kesmir C: **HLA class I allele promiscuity revisited**. *Immunogenetics* 2011, **63**:691–701.

131. Cuendet MA, Michielin O: **Protein-protein interaction investigated by steered molecular dynamics: the TCR-pMHC complex**. *Biophys J* 2008, **95**:3575–3590.

132. Reboul CF, Meyer GR, Porebski BT, Borg NA, Buckle AM: **Epitope Flexibility and Dynamic Footprint Revealed by Molecular Dynamics of a pMHC-TCR Complex**. *Plos Computational Biology* 2012, **8**.

133. Miyazawa S, Jernigan RL: **Residue-residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading**. *J Mol Biol* 1996, **256**:623–644.

134. Betancourt MR, Thirumalai D: **Pair potentials for protein folding: choice of reference states and sensitivity of predicted native states to variations in the interaction schemes**. *Protein Sci* 1999, **8**:361–369.

135. Cardenas C, Villaveces JL, Bohorquez H, Llanos E, Suarez C, Obregon M, Patarroyo ME: **Quantum chemical analysis explains hemagglutinin peptide-MHC Class II molecule HLA-DRbeta1*0101 interactions**. *Biochem Biophys Res Commun* 2004, **323**:1265–1277.

136. Roomp K, Antes I, Lengauer T: **Predicting MHC class I epitopes in large datasets**. *BMC Bioinformatics* 2010, **11**:90.

137. Boniecki M, Rotkiewicz P, Skolnick J, Kolinski A: **Protein fragment reconstruction using various modeling techniques**. *J Comput Aided Mol Des* 2003, **17**:725–738.

138. Miyazawa S, Jernigan RL: **Self-consistent estimation of inter-residue protein contact energies based on an equilibrium mixture approximation of residues**. *Proteins* 1999, **34**:49–68.

139. Tanaka S, Scheraga HA: **Medium- and long-range interaction parameters between amino acids for predicting three-dimensional structures of proteins**. *Macromolecules* 1976, **9**:945–950.

140. 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.

141. Chawla N V, Bowyer KW, Hall LO, Kegelmeyer WP: **SMOTE: Synthetic minority over-sampling technique**. *Journal of Artificial Intelligence Research* 2002, **16**:321–357.