A novel method for HLA-peptide binding prediction and HLA supertype sub-typing

Zhao, Bing

2006

Zhao, B. (2006). Novel method for HLA-peptide binding prediction and HLA supertype sub-typing. Doctoral thesis, Nanyang Technological University, Singapore.

<https://hdl.handle.net/10356/5258>

<https://doi.org/10.32657/10356/5258>

Nanyang Technological University

Downloaded on 26 Apr 2023 20:07:18 SGT

**NANYANG
TECHNOLOGICAL
UNIVERSITY**

**A NOVEL METHOD FOR HLA-PEPTIDE
BINDING PREDICTION AND
HLA SUPERTYPE SUB-TYPING**

**ZHAO BING
SCHOOL OF MECHANICAL AND AEROSPACE
ENGINEERING
2006**

A Novel Method for HLA-peptide Binding Prediction and HLA Supertype Sub-typing

Zhao Bing

School of Mechanical and Aerospace Engineering

A thesis submitted to the Nanyang Technological University
in fulfilment of the requirement for the degree of
Doctor of Philosophy

2006

ACKNOWLEDGEMENTS

I am greatly indebted to my supervisor Professor Meena Kishore Sakharkar for providing me with the opportunity of researching in this fascinating field of bioinformatics. Her professional guidance, constant encouragement from the beginning of the work, and moral support in many ways during my PhD research, has been invaluable in my professional development.

I am also greatly indebted to Professor Pandjassarame Kanguene for his valuable suggestion and patient guidance throughout my research. The knowledge I gained under his mentorship will carry me far. His impressive diligence and efficiency also make a prefect model for me.

I would like to thank the members of the research team, present and past, Li Lei, Yu Yiting, Iti Chaturvedi, Nidhi Dandona, Bagavathi S. Perumal, and Kingshuk Ghosh. It has been a great pleasure to work with them.

I am grateful to Professor Khoo Li Pheng, Head of the Mechatronics & Design Division, for providing advice and support during the whole study.

I would like to express my sincere thanks to Mdm Aminah Bte Abas, Senior Clerical

Officer of School of Mechanical & Aerospace Engineering. The prompt and quality management of Mdm Aminah Bte Abas is highly appreciable.

I would like to thank Professor Lye Sun Who, Vice-Dean of School of Mechanical & Aerospace Engineering, for his valuable suggestion and support.

I acknowledge the help provided by CANES and NCSV laboratory staffs. Their support to my research work is incomparable and unforgettable!

I have been financially supported by the Research Student Scholarship from Nanyang Technological University for the past three years. For this assistance I am very grateful.

Finally, I would like to express my deepest love and appreciation to my mother, father, and wife for their understanding, endless patience and encouragement when it was most required. To them I dedicate this thesis.

Zhao Bing

September 2005

TABLE OF CONTENTS

CHAPTER TITLE	PAGE
TITLE PAGE	I
ACKNOWLEDGEMENTS	II
TABLE OF CONTENTS	IV
SUMMARY	VII
LIST OF FIGURES	IX
LIST OF TABLES	XI
ABBREVIATIONS	XIII
CHAPTER 1 INTRODUCTION	1
1.1 Background	1
1.2 Research Scope	4
1.3 Objectives	5
CHAPTER 2 LITERATURE REVIEW	7
2.1 HLA Polymorphism	8
2.2 Molecular Structure of HLA	11
2.3 Peptide Binding Specificities of HLA	15
2.4 Peptide Binding Pockets of HLA	18
2.5 HLA Supertypes	23
2.6 HLA Peptide Binding Prediction	27
2.6.1. Sequence Based Prediction Methods	30
2.6.1.1 Binding motif-based methods	32

2.6.1.2	Artificial Neural Network Based Methods	34
2.6.1.3	Stepwise Discriminant Analysis	38
2.6.1.4	Hidden Markov Model Based Methods	40
2.6.1.5	Support Vector Machines	43
2.6.1.6	Quantitative Matrix Based Methods	44
2.6.1.7	Positional Scanning	45
2.6.1.8	Profile Motifs	46
2.6.1.9	Additive Method	47
2.6.2.	Structure Based Prediction Methods	49
2.6.2.1	Molecular Dynamics Simulation	51
2.6.2.2	Threading with Knowledge Based Free Energy Scoring	52
2.6.2.3	Free Energy Scoring Function Based Methods	53
2.6.2.4	Virtual Matrix Based Methods	54
2.6.2.5	Computational Combinatorial Ligand Design	56
2.6.2.6	3D Quantitative Structure-affinity Relationship	57
2.6.3	Limitations of Current Methods	59
CHAPTER 3	HLA PEPTIDE BINDING PREDICTION MODEL	64
3.1	Model Development	71
3.1.1	Overview	71
3.1.2	Creation of Human HLAp X-ray Structure Dataset	73
3.1.3	Identification of Highly Essential Residue Positions	77
3.1.4	Definition of Virtual Binding Pockets	81
3.1.5	Scoring Scheme	85
3.1.6	Assessment of Model Performance	88
3.1.7	Cross Validation	89
3.2	Results and Discussion	101

3.3	Extension of the Procedure to Class II HLA	108
CHAPTER 4	HLA SUPERTYPE SUB-TYPING	120
4.1	Introduction	120
4.2	Functional Positions	125
4.3	Grouping of HLA Alleles	130
4.4	Results of Classification	132
4.5	HLA Sub-supertypes	135
4.6	Validation of Class I HLA Sub-supertypes	143
4.7	Discussion	150
CHAPTER 5	CONCLUSIONS	153
5.1	Achievements	153
5.2	Limitations and Future Improvements	154
REFERENCES		157
APPENDIX 1	Classification of pockets of HLA-A alleles based on CPFRs	170
APPENDIX 2	Classification of pockets of HLA-B alleles based on CPFRs	180
APPENDIX 3	Classification of pockets of HLA-Cw alleles based on CPFRs	194
APPENDIX 4	PUBLICATIONS	201

SUMMARY

HLA (Human Leukocyte Antigen) molecules bind short antigenic peptides (8 - 20 residues) during T-cell mediated immune response. HLA genes are highly polymorphic among different ethnic populations and more than 2,000 alleles (different variants of the gene) are known to date. Hence, the binding of peptides to HLA alleles is variable and specific. Peptides binding to HLA molecules elicit specific T cell immune responses and are useful in the development of peptide vaccines and therapeutics. Thus, prediction of HLA-binding peptide is critical. Here, a prediction model is described.

It is a difficult task to develop accurate prediction models and to undertake detailed description of the HLA peptide-binding specificities due to the HLA polymorphism. Many statistical and molecular mechanics models for HLA peptide binding prediction have been developed and tested during the last decade. However, their efficiency and HLA diversity coverage are far from satisfactory. Analysis of structural data revealed that there are polymorphic pockets in the HLA peptide binding groove that can accommodate the anchor residues of the peptide. The residues that form the pocket determine the geometry and chemical properties of these structural pockets and thus determine the antigen peptides that would be preferentially bound. This accounts for the differential ability of different alleles to bind a variety of peptides. Thus, the

peptide-binding groove of the HLA is essentially an exchange or shuffling of pockets among different allele. Based on this analysis, a novel predictive model for HLA peptide binding was developed. The model was extensively cross-validated using peptide binding data. The model produces high prediction accuracy (average 61%) with good sensitivity ($49 \pm 2.47\% - 76 \pm 12.52\%$, 95% CI) and specificity ($58 \pm 13.68\% - 67 \pm 17.74\%$, 95% CI) values. The average positive predictive value of the model is 89%, while the average negative predictive value is only 18%. The efficiency is very high in predicting binders and very low in predicting non-binders. This model is superior to many existing methods because of its potential application to any given HLA allele whose sequence is clearly defined.

The majority of alleles can be covered within a few HLA supertypes, where members bind similar peptides, yet exhibiting distinct repertoires. Since the overlap among different members of a supertype appears to be extensive, it is crucial to develop a framework for grouping alleles into supertypes just from sequence information. In this work, a strategy was developed to group HLA alleles into different categories of sub supertypes. By applying this strategy, 56% of 331 HLA-A alleles, 50% of 604 HLA-B alleles and 48% of 160 HLA-Cw alleles were clustered into just 42, 78 and 23 groups, respectively. The grouping was moderately cross-validated using available binding data, which can serve as a framework to test specific functional overlap between HLA alleles.

LIST OF FIGURES

Figure 2.1	Complete gene map of the HLA.	9
Figure 2.2	Overall structure of HLA-peptide complex.	12
Figure 2.3	Conserved structures of class I HLA-peptide complexes.	17
Figure 2.4	HLA surface and peptide binding pockets.	21
Figure 2.5	Methods for HLA peptide binding prediction.	29
Figure 2.6	The architecture of artificial neural network.	37
Figure 2.7	Stepwise discriminant analysis.	39
Figure 2.8	An example of an HMM model trained using five peptide sequences.	42
Figure 2.9	Virtual matrix based method.	55
Figure 3.1	Peptide binding pockets in HLA binding groove and the concept of pocket profile.	68
Figure 3.2	An overview of the model development for HLA peptide binding prediction and HLA supertype sub-typing.	72
Figure 3.3	Δ ASA of the first 180 residues of class I HLA α chains upon complex formation.	79
Figure 3.4	Involvement of nonamer peptide residues in binding to class I HLA.	86
Figure 3.5	Alignment of the peptides of different length bound to class II HLA based on %Burial curve.	113
Figure 3.6	Δ ASA of the first 100 residues of class II HLA α and β chains	115

upon complex formation.

Figure 4.1. Summary of grouping results of HLA-A, B and Cw. 131

LIST OF TABLES

Table 2.1	Definition of HLA Supertypes based on HLA peptide binding data.	26
Table 2.2	Summary of HLA-peptide Binding Prediction Methods.	62
Table 3.1	Class I HLA-peptide complexes in PDB.	75
Table 3.2	List of HERPs.	82
Table 3.3	Relationship between HERPs and peptide residue positions.	83
Table 3.4	Mapping between peptide residue positions and HLA virtual pockets.	84
Table 3.5	Performance of the model in dataset 1 (HLAp crystal structures).	92
Table 3.6	Performance of the model in dataset 2 (HLAp binding affinity data [IC ₅₀]).	93
Table 3.7	Summary of model performance in dataset 2 (HLAp binding affinity data [IC ₅₀]).	95
Table 3.8	Performance of the model in dataset 3 (peptides tested positive by tetramer staining).	96
Table 3.9	Summary of model performance in dataset 3 (peptides tested positive by tetramer staining).	97
Table 3.10	Summary of model performance in dataset 4 (HLA-A binding peptides from MHCBN).	98
Table 3.11	Summary of model performance in dataset 4 (HLA-B binding peptides from MHCBN).	99

Table 3.12	Class II HLA-peptide complexes in PDB.	112
Table 3.13	List of HERPs in class II HLA molecules.	116
Table 3.14	Mapping between peptide residue positions and virtual pockets in the class II HLA binding groove.	117
Table 3.15	Performance of the model in the dataset from PDB.	118
Table 4.1	Distribution of CPFRPs of the six binding pockets among 1,095 class I HLA alleles.	128
Table 4.2	HLA-A alleles (185) grouped into putative supertypes.	137
Table 4.3	HLA-B alleles (303) grouped into putative supertypes.	139
Table 4.4	HLA-C alleles (76) grouped into putative supertypes.	142
Table 4.5	Validation of HLA-A supertypes.	147
Table 4.6	Validation of HLA-B supertypes.	148
Appendix 1	Classification of pockets of HLA-A alleles based on CPFRs.	171
Appendix 2	Classification of pockets of HLA-B alleles based on CPFRs.	181
Appendix 3	Classification of pockets of HLA-Cw alleles based on CPFRs.	195

ABBREVIATIONS

AC	accuracy
ANN	artificial neural network
APC	antigen presenting cell
ASA	accessible surface area
CCLD	computational combinatorial ligand design
CI	confidence interval
CoMSIA	comparative molecular similarity indices analysis
CPFR	critical polymorphic functional residue
CPFRP	critical polymorphic functional residue position
Ct	<i>Chlamydia trachomatis</i>
CTL	cytotoxic T lymphocyte
DS	dataset
EBV	Epstein-Barr virus
ERP	essential residue position
FN	false negative
FP	false positive
HBV	hepatitis B virus
HCMV	human cytomegalovirus

HCV	hepatitis C virus
HERP	Highly Essential Residue Position
HIV	human immunodeficiency virus
HLA	human leukocyte antigen
HLAp	HLA peptide complex
HMM	hidden Markov models
IC₅₀	inhibitory concentration 50
IMGT	ImMunoGeneTics database
MDS	molecular dynamics simulation
MHC	major histocompatibility complex
MHCBN	major histocompatibility complex binding and non-binding peptides
NPV	negative predictive value
PBMC	peripheral blood mononuclear cell
PDB	protein data bank
Pf	<i>Plasmodium falciparum</i>
PPV	positive predictive value
PSSM	position specific scoring matrix
QSAR	quantitative structure activity relationship
SE	sensitivity
SEHPR	solvent exposed hydrophobic peptide residue
SP	specificity

SSS	successive state splitting
SVMHC	support vector machines
TCR	T-cell receptor
TN	true negative
TP	true positive
vdWC	<i>van der Waals</i> clash

CHAPTER 1

INTRODUCTION

1.1 Background

With the considerable progress made in immunology, functional genomics and proteomics, the antigenic peptides, which are presented by Major Histocompatibility Complex (MHC) to mark the infected cells or parent antigens as the target of T-cell digestion, became the most interesting target of vaccine design in recent years.

The MHC plays a pivotal role in regulating the immune response. The human MHC is known as Human Leukocyte Antigen (HLA), which has more than 2,000 alleles known in its polymorphic gene loci (Robinson et al. 2003). Formation of the ternary complex between T-cell receptor (TCR) and the HLA-peptide complex is the key molecular recognition event at the heart of the adaptive and memory cellular immune responses. In the process of T-cell activation, the peptides derived from intracellular fragmentation of pathogen protein antigens, are bound to HLA molecules and subsequently presented on the cell surface of antigen presenting cells (APCs), for the recognition by T-cell specific receptors (Pamer *et al.* 1998). Theoretically, only a small fragment of the whole protein antigen is necessary for eliciting the host immune

response. These antigenic peptides, called T-cell epitopes, can be developed as a new generation of high efficiency, low toxicity molecular vaccines.

However, the extreme degree of polymorphism in the HLA poses limitations for the development of a potentially global vaccines for infectious diseases because the ability to mount an effective T-cell response is partly determined by the HLA allotypes of the individual; moreover different individuals have different HLA allotypes (MacDonald *et al.* 2001). Genetically identical antigens will elicit different immune response in hosts with different HLA haplotype. Therefore, a complete analysis of the genetic diversity of HLA genes and the functional differences in peptide-binding specificities and CTL (cytotoxic T lymphocytes) responses is a prerequisite for the design of vaccines with more universal effectiveness.

It is known that people from the same racial or geographical region generally have similar HLA alleles, which are called HLA supertypes. The HLA supertypes are characterized by largely overlapping peptide-presenting specificities based on structural similarities in the antigen-binding groove. The alleles in a given HLA supertype present the same epitopes for T-cell recognition. The HLA supertype expressed by an individual or cohort of individuals could determine the epitopes of the infecting agent to which cytotoxic T lymphocytes respond. Thus, if the HLA types present in a target population are known, together with the particular epitopes that are restricted by these HLA types, the epitopes can be developed as vaccines. Therefore, a prior knowledge of the HLA supertypes is necessary.

Several HLA supertypes have been proposed based on experimental peptide binding

data. However, determining the peptide specificity of even a single allele is a non-trivial undertaking (Sette *et al.* 2002). To undertake this task for over 2,000 alleles is intensive. At the same time, experimental approaches to identify T-cell epitopes within these antigens, i.e. the synthesis and assay of overlapping peptides from proteins of interest, are not applicable to a large number of protein sequences, especially when the number of possible disease-specific antigens has been dramatically increasing with the fast-growing genome-scale DNA sequence information. A viable alternative to these problems is to seek a bioinformatics approach. The first and most important step is to describe the peptide-binding specificities of HLA allotypes and to develop prediction models for HLA peptide binding.

The development of an efficient prediction method is complicated by two factors. The first is that the same HLA allotype can bind a range of peptides. The second is HLA allelic polymorphism. The actual situation is that there are thousands different HLA allotypes within the human species, each of which can bind a wide range of peptides. Furthermore, HLA peptide recognition does not follow a simple rule. It is a complex and difficult task to undertake a detailed description of the peptide-binding specificities of the ethnic repertoire of polymorphic HLA molecules and to develop accurate prediction for their ability to bind peptides. In our view, the problem can be partially simplified by identifying sets of key HLA amino acid residues specifically involved in peptide binding.

A large number of interaction functions and scoring schemes have been developed over the last few years for HLA-based epitope prediction on the basis of rules that govern the binding of a peptide to HLA molecule, as the result of the increasing

structural and functional information that became available over the past decade. Briefly HLA binding peptide prediction methods can be divided into two groups: a) Sequence based methods, b) Structure based methods. However, their efficiency and HLA allele coverage are far from satisfactory. Due to the limitation in the quality of available training data at current stage, most of current attempts to predict HLA peptide binding have not effectively dealt with HLA polymorphism i.e. for each allele a separate matrix has to be generated or a separate set of rules has to be applied.

In view of these limitations, a novel model for HLA-peptide binding predictions, and a systematic characterization and classification of the peptide-binding specificities of HLA alleles, based on thorough analysis of available structural and sequence data, is needed.

1.2 Research Scope

The extreme polymorphism of HLA alleles is the major obstacle in the development of HLA-based peptide vaccines. As a result, understanding the nature of HLA peptide interaction and the classification of the polymorphic HLA alleles is important.

Information of crystal structures of HLA peptide complexes is gathered from Protein Data Bank, where all publicly available protein structure information determined by X-ray crystallography or NMR are deposited. The information is analyzed from various directions by multiple computational tools. The HLA peptide binding pocket is closely examined and serves as the basis of this investigation. The analysis led to an

insight into the properties and nature of the interaction between individual HLA pocket and individual anchor residue of bound peptide. The information of HLA peptide binding pockets is used to cluster the MHC alleles into supertypes, which are in significantly smaller numbers and is used as the basis of the development of HLA peptide binding prediction methods. The prediction model and the classification of HLA supertypes is cross-validated by available HLA peptide binding data collected from literatures and public databases, compared to other available prediction methods, to ensure that the real functional similarity among the same supertype of HLA alleles is reflected.

The classification of HLA supertypes provides valuable information in peptide vaccine design, combined with information of HLA distribution among different ethnic populations. The HLA peptide binding prediction model can be employed to virtual screen antigenic peptide fragments by scanning the genome or proteome of known pathogens, thus facilitates the design and development of peptide vaccines.

1.3 Objectives

The objectives of this research project are to explore the molecular mechanism of antigenic peptide formation and presentation to T-cell by HLA molecule, and to develop a computational prediction system for virtual screening T-cell epitopes from the genome or proteome of known pathogens. Ultimately, the computational prediction system can be exploited in rational design of HLA based peptide vaccine for specific diseases.

The focus of this work is outlined below:

1. To understand the characteristics and nature of HLA peptide interactions, using computational tools and structural information of HLA peptide complexes gathered from Protein Data Bank;
2. To find major factors that affect HLA function, which are appropriate to be employed as the criteria of HLA supertype classification;
3. To develop a systematic methods for the subtyping of HLA supertypes (overlapping peptide binding function), as the basis of HLA-based vaccine design;
4. To develop a universal HLA-peptide binding prediction method, based on the understanding of the nature of HLA peptide interactions and the classification of HLA supertypes.

CHAPTER 2

LITERATURE REVIEW

Over the past decade, a detailed understanding of the cellular immune system has been achieved. The knowledge of cellular immunity could help us to develop vaccines that may prevent or treat infectious diseases including AIDS, malaria, and tuberculosis, as well as to develop immunotherapy for the treatment of cancer (van Endert 2001). Understanding cellular immunity also promises better strategies for treating autoimmune disorders and transplantation rejection. This is due to the T-cell's ability to exhibit antigen specificity and memory.

HLA plays a pivotal role in regulating cellular immunity. T-cell specificity is conferred by the clonally restricted T-cell receptor, which recognizes class I and class II HLA molecules binding with short peptides. The complex of the HLA molecule and each antigenic peptide acts as the ligand recognized by HLA-restricted T-cells (Germain *et al.* 1993; Garcia *et al.* 1999). Expression of the CD8+ or CD4+ co-receptor by T-cells limits their recognition to class I or class II HLA molecules, respectively.

In this chapter, an overview of the HLA and the molecular mechanism of HLA peptide interaction are presented. Then, current methods for HLA peptide binding prediction are introduced. The advantages and disadvantages of these methods are discussed.

2.1 HLA Polymorphism

HLA is extremely polymorphic. The HLA region spans 4×10^6 nucleotides on chromosome 6 (6p21.1 to 6p21.3), with class II, class III and class I genes located from the centromeric to the telomeric end (McCluskey *et al.* 1999) (**Figure 2.1**). Most of the genes are in linkage disequilibrium and are therefore inherited as a combined block or haplotype. Recombination occurs rarely, at a frequency of 1-3%, mostly at the HLA-A or HLA-DP ends (Shiina *et al.* 1999). The HLA genes are co-dominantly expressed and follow the Mendelian pattern of inheritance. Sequencing of HLA region (Shiina *et al.* 1999) revealed that there are more than 128 expressed genes, of which at least 40% have one or more designated immune functions.

Currently, a total of 2,088 alleles in the HLA region have been defined by nucleotide sequence in the Release 2.10 of IMGT/HLA Sequence Database (URL: <http://www.ebi.ac.uk/imgt/hla/stats.html>) maintained by the European Bioinformatics Institute. The amino acid differences that account for molecular diversity in the class I region occur mainly in the α_1 and α_2 domains within any of the seven hypervariable sequences (i.e. amino acid sequences 9-12, 40-45, 62-83, 94-97, 105-116, 137-163 and 174-194). The hypervariable regions in the class II region genes are located in the α_1 and β_1 domains of α and β chains.

Figure 2.1 Complete gene map of the HLA.

Genes are shown in order from telomere to centromere. Sizes of genes and intervening DNA segments are not shown to scale. Gene loci that were discovered or located to the HLA as a direct result of the genomic sequence are indicated by filled boxes. This map also includes many genes that are of unknown or of no known immunologic function.

(Adapted from: The MHC sequencing consortium. Complete sequence and gene map of a human major histocompatibility complex. *Nature* **401**, 921-923, 1999).

2.2 Molecular Structure of HLA

The class I and class II HLA molecules have similar overall structures with subtle functional differences (**Figure 2.2**). Class I molecules are made up of one heavy chain (45 kD) encoded within the HLA and a light chain called β_2 -microglobulin (β_2m ; 12 kD) whose gene is on chromosome 15. Class II molecules consist of one α (34 kD) and one β chain (30 kD) both of which are encoded within the HLA. The class I heavy chain has three domains of which the membrane-distal first (α_1) and the second (α_2) are the polymorphic ones. Within these domains, polymorphisms concentrate on three regions: positions 62 to 83; 92 to 121; 135 to 157 (Steinmetz *et al.* 1983). Diversity in these domains is very important in that these two domains form the antigenic peptide binding groove (**Figure 2.2 a, c**). The sides of the antigen binding groove are formed by the α_1 and α_2 domains, while the floor of the groove is comprised of eight anti-parallel β sheets (Bjorkman *et al.* 1987a; 1987b). The antigenic peptides of eight to ten amino acids (typically nonamers) bind to the groove with low specificity but high stability (Rammensee *et al.* 1995). The α_3 domain contains a conserved loop (positions 223 to 229) which serves as a binding site for CD8 (Salter *et al.* 1990).

In the class II molecule, generally both α and β chains are polymorphic. In these chains α_1 and β_2 domains form the peptide binding groove (**Figure 2.2 b, d**). Therefore, diversity is located mainly in these domains (except in HLA-DRA which is not polymorphic). Antigenic peptides of 10 to 24 amino acids long bind to the groove and extend on either side (Rammensee *et al.* 1995). In a region analogous to the CD8-binding site on class I molecules, a major CD4-binding site is contained within residues 241 to 255 in β_2 domain (Konig *et al.* 1992).

a

b

Figure 2.2 Overall structure of HLA-peptide complex.

a. Surface representation and equivalent ribbon diagram of a class I HLA molecule binding with a nonamer peptide (PDB code: 1I7U). The surface of the complex is shown, colored according to the domains shown in the ribbon representation in the right part of the panel. The class I HLA molecule is a heterodimer of a membrane-spanning α chain bound noncovalently to β_2 -microglobulin, which does not span the membrane. The α chain folds into 3 domains: $\alpha 1$, $\alpha 2$, and $\alpha 3$. The $\alpha 3$ domain and show similarities in amino acid sequence to immunoglobulin C domains and have similar folded structures. The folding of the $\alpha 1$ and $\alpha 2$ domains creates a long groove, which is the site at which peptide antigens bind to the HLA molecules.

b. Surface representation and equivalent ribbon diagram of a class II HLA molecule binding with a 13-mer peptide (PDB code: 1DLH). The class II HLA molecule is composed of two transmembrane glycoprotein chains, α and β , as shown in the right part of the panel. Each chain has two domains, and the two chains together form a compact four-domain structure similar to that of the class I HLA molecule. The two domains forming the peptide binding groove are contributed by different chains and therefore not joined by a covalent bond. Another important difference is that the peptide binding groove of class II HLA molecules is open at both ends.

c. Ribbon diagram of a class I HLA molecule binding with a nonamer peptide (PDB code: 1I7U). Only the first 180 amino acid residues of the heavy chain are shown for a clear view of the peptide binding groove. β_2m is not shown. Peptide residues are colored differently. Electrostatic potential surface of the peptide is also shown.

d. Ribbon diagram of a class II HLA molecule binding with a 13-mer peptide (PDB code: 1SEB). Only the first 80 amino acid residues of each of the α and β chain are shown for a clear view of the peptide binding groove. Peptide residues are colored differently. Electrostatic potential surface of the peptide is also shown.

2.3 Peptide Binding Specificities of HLA

HLA molecules bind peptides of diverse sequence with great affinity and long half-life. Many different HLA alleles exist within the human population; however, each individual can only express a few of these alleles. HLA alleles possess a unique peptide-binding specificity: one allelic HLA product will recognize and present one part of the universe of peptides, whereas another HLA product will recognize and present a different part of that universe. Although a single HLA molecule can bind only a single peptide at any time, same kind of HLA molecules (allotypes) can bind a wide variety of HLA peptides (Sidney *et al.* 1997). Roughly speaking, each HLA specificity samples about 0.1% to 10% of all available peptides (Kast *et al.* 1994).

The HLA peptide-binding specificity is manifested in the preference of some HLA molecules for binding peptides with specific amino acids in specific positions, depending upon the residues occupied in the peptide-binding site of that allotype. The study of single amino acid substituted antigen analogs and the sequencing of endogenously bound, naturally processed peptides has identified critical residues required for allele-specific binding to HLA molecules (Garcia *et al.* 1999).

As discussed earlier, class I and II HLA molecules have a common overall structure, featuring a unique peptide-binding site situated at the outer polymorphic domains of the HLA. These outer domains form a groove. Most HLA polymorphisms are clustered in specific sites within the groove called the peptide binding pockets that accommodate the peptide side chains (Madden 1995). The amino acid residues that form the pocket determine the properties, such as size, shape, hydrophobicity, and electronic charge, of

the pocket and thus determine the antigen peptides that would be preferentially bound by the HLA allotype. This relationship accounts for the differential ability of different alleles to bind a variety of peptides.

Most of the binding affinity is provided by the hydrogen bonding at the end of the groove between the peptide and the conserved tyrosine residues. These conserved amino acids bind the invariant portions of the peptides, the amino and carboxyl termini. Through a set of hydrogen bonds to the main chain of the peptide, the termini of the peptide are oriented into specific pockets that accommodate the chemical nature of each of the termini. Thus, the orientation (amino to carboxyl) of the peptide is fixed for all HLA molecules. However, this arrangement does not limit the length of the peptide. Longer peptides may zigzag (Madden *et al.* 1993) or bulge (Guo *et al.* 1992; Collins *et al.* 1995) to allow longer peptides to maintain the relative position of the termini. In contrast, the backbone of the peptide-binding groove is highly conserved. The root mean square deviation (RMSD) of the backbone atoms is 0.44 Å for the α_1 and α_2 domains (Vasmatzis *et al.* 1996a, 1996b) (**Figure 2.3**). Thus, the peptide-binding groove of the HLA is essentially an exchange or shuffling of pockets between different allotypes.

This binding mode, involving backbone atoms common to all peptides, explains how one HLA molecule can perform high affinity ($kD = 10^{-8} - 10^{-9}$ M) binding of a large and diverse repertoire of peptides. Only a minority of the binding energy involves peptide side-chain atoms. These interactions, however, are believed to explain the specificity of the HLA (Matsumura *et al.* 1992).

Figure 2.3 Conserved structures of class I HLA-peptide complexes.

11 class I HLA alleles are superimposed by the backbone of HLA α_1 and α_2 domains. The traces of peptides are shown within the traces of the HLA structures. Only first 180 amino acid residues of the HLA heavy chain and C_α and C_β atoms of the peptide are shown for a clear view of the peptide binding groove. β_2m is not shown.

2.4 Peptide Binding Pockets of HLA

The HLA peptide binding pockets complement a small number of specific amino acids from the antigenic peptide and thus cull out many peptides from the total pool of available peptide (Bjorkman *et al.* 1987b; Garrett *et al.* 1989a; Guo *et al.* 1993). These positions in the peptides are called anchors, because they were believed to confer significant binding energy. Typically, there are two anchors at the second ($P2$) [or fifth ($P5$)] and the final peptide position ($P\Omega$). The combination of preferences at different pockets along the groove produces a sequence motif that is useful to predict whether a particular peptide will bind to a particular HLA molecule (Falk *et al.* 1990; Jardetzky *et al.* 1991; Hunt *et al.* 1992). For example, the pockets in the class I HLA molecule HLA-A2 result in a preference of I, L, or M at position 2 ($P2$) and a preference for L or V at the carboxyl terminus of the peptide ($P\Omega$). These pockets can be changed by mutagenesis to create chimerical peptide motifs (Colbert *et al.* 1993). In addition to these pockets, there are secondary pockets that can enhance affinity of a particular peptide. For example, two secondary pockets in HLA-A2 are those that bind the first ($P1$) (Colbert *et al.* 1994) (Pogue *et al.* 1995) and the third residue ($P3$) (Ruppert *et al.* 1993) of the antigenic peptide. Changes in the sequence of the antigenic peptide at these secondary anchors may either enhance or detract from peptide affinity (Ruppert *et al.* 1993). It is becoming clear that the total binding energy is the sum of the contributions of all the residues in the binding groove. Each residue, not only the primary and secondary anchors, can offer positive or negative interactions that either enhance or detract from binding affinity.

In the peptide binding grooves of class I HLA molecules, six pockets, designated A - F,

are formed by the polymorphic HLA residues, which accommodate the anchor residues of peptide (**Figure 2.4a**). Different pockets have variable effects on the binding of peptides. Among these, pockets B and F are the most crucial for binding of peptides to HLA-A and HLA-B molecules (Matsui *et al.* 1993; Matsui *et al.* 1994). These two pockets accommodate the side chains of the second residue (*P2*) and the carboxy-terminal end of the peptide, respectively, where most of the selectivity in peptide binding is exerted (Matsumura *et al.* 1992; Bouvier *et al.* 1994). Almost 60% of known class I HLA motifs have *P2* anchors. Pocket B is located between the α_1 and α_2 helix and the β sheet and is isolated from the rest of the peptide binding groove. Its composition is thus a primary determinant of several allele-specific motifs. Residues in the middle of the bound peptide are not buried in the site and hence impose little or no restriction on peptide binding. However, a few HLA molecules such as HLA-A1 and HLA-B8 appear to have anchors interacting in the centre of the peptide binding groove, in pockets C and D (DiBrino *et al.* 1994). This flexibility in the various combinations of the six pockets allows a broad spectrum of peptides to bind class I HLA molecules (Garrett *et al.* 1989b).

The peptide-binding site of class II HLA molecules is similar to that of class I, where the amino-terminal portions of the α_1 and β_1 domains fold into β -pleated sheets while the carboxy terminal portions form the helices (**Figure 2.2d, 2.4b**). Subtle changes in the helical regions produce a binding groove with open ends, which allows peptides to hang out of the groove at both ends and thus accommodate a larger peptide than the class I molecules. Unlike class I HLA molecules, the peptide in the class II HLA molecule is held in the middle through hydrogen bonds formed at regular intervals throughout its length. Allotype specific peptide binding is imposed by positions of

polymorphism within the peptide-binding site. The most prominent pocket in the class II HLA groove is a large hydrophobic pocket at one end of the binding groove (pocket 1), formed by residues of both the α and the β chain. The specificity of this pocket is mainly influenced by the first residue (*P1*) of the peptide. The presence of a Glycine in this pocket dictates binding of large aromatic or aliphatic residues. By contrast, the presence of a larger Valine residue in the same position reduces the size of the pocket and thus creates preference for smaller aliphatic residues (Newton-Nash *et al.* 1993). Therefore, the constraints on peptide binding to class II molecules are less restrictive than those for class I molecules. This lower selectivity and potential for binding longer peptides enables promiscuous peptides to bind to different class II allotypes. An understanding of these fine details of HLA-peptide interactions could assist in designing peptide-based vaccines.

a

b

2.5 HLA Supertypes

More than 2,000 HLA alleles defined by nucleotide sequences have been deposited in the IMGT/HLA Sequence Database (Robinson *et al.* 2003). Therefore, the number of theoretically possible combinations of HLA-peptide complexes is extremely large. However, the immune system maintains a homogenous balance by specific selection, degeneration, and discrimination (self/non-self) of short peptides using HLA molecules. Although HLA molecules are polymorphic in ethnic populations, a significant overlap at the functional level exists between many of the most common HLA molecules. Some HLA alleles show similar peptide binding specificities. The HLA allelic variants that bind peptides possessing a particular HLA super-motif are collectively referred to as an HLA supertype.

A number of “HLA supertypes” have already been defined using experimentally determined peptide binding data (**Table 2.1**). It was shown that the A*0301 and A*1101 repertoires overlapped with those of A*3101, A*3301, and A*6801 (Sidney *et al.* 1996a). This group of alleles, characterized by similar peptide specificity, was defined as the A3-supertype. Similarly, a significant overlap in peptide binding repertoires was demonstrated amongst several serologically distinct HLA-B alleles (B*0702, B*3501, B*5101, B*5301, and B*5401 (Sidney *et al.* 1996b). The B7-, A2- and B44-supertypes were also defined (Sette *et al.* 1998).

The A2-supertype includes, A*0201, A*0202, A*0203, A*0204, A*0205, A*0206, A*0207, A*6802, and A*6901. These alleles recognize peptides of about 9 or 10 residues in length which bear small or aliphatic hydrophobic residues (L, I, V, M, A, T,

or Q) at *P2* and L, I, V, M, A, or T at the C- terminal position. The B pocket of A2-supertype molecules is characterized by a consensus motif including residues (where the subscript indicates HLA residue position) F/Y₉, M₄₅, E/N₆₃, K/N₆₆, V₆₇, H/Q₇₀, and Y/C₉₉. Similarly, the A2-supertype F pocket is characterized by a consensus motif including residues D₇₇, T₈₀, L₈₁, and Y₁₁₆ (Sidney *et al.* 1996a).

The A3-supertype includes the A*0301, A*1101, A*3101, A*3301, and A*6801 alleles (Sidney *et al.* 1996a). A3-supertype molecules recognize a broad motif characterized by A, V, I, L, M, S, or T at *P2*, and R or K at the C-terminus (Kubo *et al.* 1994; Sidney *et al.* 1996a). Peptide lengths of 9 to 10 amino acids have been most frequently reported, although longer peptides can bind, and sometimes be recognized, in the context of A3-supertype molecules. The B pocket of A3-supertype molecules is characterized by the consensus motif of M₄₅, N/K₆₆, M/V₆₇, Q/H₆₆, and Y₉₉. This structural motif is similar to that of A2-supertype B pockets and is in good agreement with the largely overlapping B-pocket specificity of A2- and A3-supertype alleles. The F pocket of A3-supertype molecules is characterized by D₇₇, T₈₀, L₈₁, and D₁₁₆. The dominant presence of a negatively charged residue (D) in positions 77 and 116 correlates with the specificity of A3-supertype alleles for peptides with a positively charged C-terminus.

The B7-supertype included B*0702, B*3501-03, B*51, B*5301, B*5401, B*0703-05, B*1508, B*5501-02, B*5601-02, B*6701, and B*7801. B7-supertype molecules share a peptide binding specificity for P at *P2* and a hydrophobic aliphatic (A, L, I, M, or V) or aromatic (F, W, or Y) residue at the C-terminal position of their peptide ligands. Modeling and X-ray crystallographic studies of the structure of B7 and B*3501 offered

insights into the specificity of B7-supertype molecules (Smith *et al.* 1996). Structurally, the B7-supertype molecules share a B pocket consensus motif of Y9, N63, I66, F/Y67, N/Q70, and Y99. By contrast, no discrete B7-supertype F pocket consensus motif has yet been defined.

The functional overlap between different members of the supertype is intriguing. The concept of HLA supertypes is that alleles belonging to supertypes bind a highly shared set of peptides. In principle it should be possible to predict peptide binding of other members of a supertype using experimental results based on just one member of the type. However, as illustrated in **Table 2.1**, this promise does not hold in the major supertypes. It shows that several peptides bind to some members but not all members of the A2 supertypes (**Table 2.1**). Hence, this grouping of HLA supertypes based on HLA peptide binding data is inconclusive so far, given the known number of HLA alleles. Therefore, it is necessary to develop a bioinformatics method to systematically group HLA alleles into clusters of overlapping function (supertypes) based on sequence information.

Table 2.1 Definition of HLA Supertypes based on HLA peptide binding data.

Peptide	Super -type	A*0201	A*0202	A*0203	A*0206	A*6802	Reference
LLFNILGGWV	A2	b	b	b	b	b	Scognamiglio, 1999
YLVAYQATV	A2	b	b	b	b	b	Scognamiglio, 1999
KVAELVHFL	A2	b	b	b	b	b	Kawashima, 1998
FLWGPRALV	A2	b	b	b	b	b	Kawashima, 1998
FLLLADARV	A2	b	b	b	b	b	Scognamiglio, 1999
IMIGVLVGV	A2	b	b	b	b	b	Kawashima, 1998
KIFGSLAFL	A2	b	b	b	b	nb	Kawashima, 1998
CLTSTVQLV	A2	b	b	b	b	nb	Kawashima, 1998
RLIVFPDLGV	A2	b	b	b	b	nb	Scognamiglio, 1999
YLQLVFGIEV	A2	b	b	b	b	nb	Kawashima, 1998
LLTFWNPPV	A2	b	b	b	b	nb	Kawashima, 1998
VLVGGLVLA	A2	b	b	b	b	nb	Scognamiglio, 1999
WMNRLIAFA	A2	b	b	b	nb	b	Scognamiglio, 1999
DLMGYIPLV	A2	b	nb	b	b	b	Scognamiglio, 1999
ILHNGAYSL	A2	b	b	b	nb	nb	Kawashima, 1998
YLSGANLNL	A2	b	b	b	nb	nb	Kawashima, 1998
VMAGVGSPYV	A2	b	b	b	nb	nb	Kawashima, 1998
ILAGYGAGV	A2	b	b	b	nb	nb	Scognamiglio, 1999
LMTFWNPPV	A2	b	nb	b	b	nb	Kawashima, 1998
YLVTRHADV	A2	b	nb	b	b	nb	Scognamiglio, 1999
HMWNFISGI	A2	b	nb	b	b	nb	Kawashima, 1998
YLLPRRGPR	A2	b	nb	b	b	nb	Scognamiglio, 1999
LLFLLLLADA	A2	b	b	nb	nb	nb	Scognamiglio, 1999
LLTFWNPPPT	A2	b	nb	b	nb	nb	Kawashima, 1998
ALCRWGLLL	A2	b	nb	b	nb	nb	Scognamiglio, 1999
A*0301 A*1101 A*3101 A*3301 A*6801							
KTSERQPR	A3	b	b	b	nb	b	Scognamiglio, 1999
RMVVGGEHR	A3	b	b	b	nb	nb	Scognamiglio, 1999
QLFTFSPRR	A3	b	b	nb	nb	b	Scognamiglio, 1999
LGFGAYMSK	A3	b	b	nb	nb	b	Scognamiglio, 1999
LIFCHSKKK	A3	b	b	nb	nb	b	Scognamiglio, 1999
GVAGALVAFK	A3	b	b	nb	nb	b	Scognamiglio, 1999
VAGALVAFK	A3	b	b	nb	nb	b	Scognamiglio, 1999
RLGVRATRK	A3	b	b	b	?	?	Chang, 1999
B*0702 B*3501 B*5101 B*5301 B*5401							
LPGCSFSIF	B7	b	b	b	b	nb	Chang, 1999

Abbreviations: **b** = binder; **nb** = non-binder; **?** = undetermined.

2.6 HLA Peptide Binding Prediction

HLA molecules bind to certain peptide fragments generated after proteolytic cleavage of antigen. This binding can induce immune response against the whole parent antigen, which is the basis of designing subunit and synthetic peptide vaccines. Thus the problem of peptide vaccine design can be broke down to the identification of T-cell epitopes, the peptide fragments which can bind to HLA and evoke a T-cell response.

Experimental approaches to identify T-cell epitopes within these antigens, i.e. the synthesis and assay of overlapping peptides from proteins of interest for immune reactions, are time-consuming and prohibitively expensive for a large number of protein sequences. However, computer models capable of simulating and predicting the biological process of antigen presentation can be used to minimize the number of experiments, thus enabling a systematic scanning for candidate T-cell epitopes.

Advances in understanding antigen presentation at the molecular level accelerated the development of computer models capable of predicting T-cell epitopes. Although some of the molecular aspects of antigen presentation are insufficiently defined to be of value for epitope prediction models, others, such as the interaction of peptide fragments with HLA molecules, are characterized thoroughly. As a consequence, most of the currently available epitope prediction models are based on HLA peptide binding data. This approximation is supported by the observation that HLA peptide binding is a major bottleneck in the selection of epitopes, as indicated by the finding that most peptide sequences lack the capacity to interact with HLA molecules.

The HLA peptide binding prediction model should be suited for high throughput

scanning of a pathogen proteome with high sensitivity capable of covering maximum number of HLA alleles. The method that requires very few experiments in the identification of vaccine candidates is technologically advantageous.

However, the development of an efficient prediction method is complicated by two factors. The first is that the same HLA molecule can bind a range of peptides. The second is HLA allelic polymorphism. The actual situation is that there are thousands of HLA allotypes, each of which can bind to a wide range of peptides. However, the problem can be partially simplified by identifying sets of amino acids specifically binding to HLA molecules.

Several approaches have been developed for HLA-based epitope prediction on the basis of rules that govern the binding of a peptide to HLA molecule, as the results of the increasing structural and functional information that became available over the last decade. Broadly the HLA binding peptide prediction methods can be divided into two main groups: a) sequence based methods, b) Structure based methods (**Figure 2.5**).

Figure 2.3 Overview of the classification of prediction methods for HLA peptide binding.

Broadly the HLA binding peptide prediction methods can be divided into two main groups a) sequence based methods, b) Structure based methods.

2.6.1. Sequence Based Prediction Methods

Sequence-based methods rely on the primary sequence of peptides that are known to bind specific HLA allele from HLAp binding assays. The sequence information are encoded into a binding motif (D'Amaro *et al.* 1995; Meister *et al.* 1995; Rammensee *et al.* 1995; Rammensee *et al.* 1999; Reche *et al.* 2002), a position-dependent matrix (Sette *et al.* 1989; Udaka *et al.* 2000), an artificial neural network (ANN) (Adams *et al.* 1995; Brusic *et al.* 1998; Milik *et al.* 1998), stepwise discriminant analysis (Mallios 1999), hidden Markov model (HMM) (Mamitsuka 1998; Brusic *et al.* 2002), or Support Vector Machine (SVM) (Donnes *et al.* 2002). These models are developed using peptide data in large databases derived from naturally bound peptides (Rammensee *et al.* 1995; Rammensee *et al.* 1999) or synthetic peptide libraries (Parker *et al.* 1994; Stryhn *et al.* 1996).

HLA bound peptides are generally restricted at some primary positions (second or fifth and last positions) of the peptides. These residues are known as anchor residues and positions are known as anchor positions. Allele specific sequence motifs can be identified by studying the frequencies of amino acids in anchor positions. For example, a simple motif based-prediction method is based on the observation that the peptides binding to HLA A*0201 are often nonamers, and frequently have two anchor residues, a lysine in position 2 and a Valine in position 9 (Rammensee *et al.* 1995). Besides the anchor residues, there are also weaker preferences for specific amino acids in other positions. One method to include this information is to use a profile, in the form of a matrix, where each type of amino acid in each position is given a score. The scores can be calculated from observed amino-acid frequencies in each position or be set

manually. The sum of contribution by all the residues gives predicted binding value. These methods are easy to implement, and are one of the most popular methods applied to HLAp binding prediction. One frequently used profile based prediction method is SYFPEITHI (Rammensee *et al.* 1999).

These methods are based on the assumption that different peptide positions contribute in an additive manner to the overall binding affinity, the contribution of each peptide residue to the binding affinity is independent of other neighboring residues. The interactions between different positions are not taken into account. Furthermore, their real predictive power is directly dependent on the amount of experimental data used to interpolate HLA binding properties. Thus, HLA alleles for which rather few (or no) experimental data are available are unsuitable for these prediction methods.

These methods do not consider information from non-binding peptides. This information can be used by machine learning methods. In the meantime, the predictive power of machine learning methods is not affected by the interactions among peptide residue positions. Prediction of HLA-peptides has been made by using machine learning approaches such as ANN (Adams *et al.* 1995; Brusic *et al.* 1998; Milik *et al.* 1998), stepwise discriminant analysis (Mallios 1999), HMM (Mamitsuka 1998; Brusic *et al.* 2002), and SVM (Donnes *et al.* 2002). Gulukota *et al.* (1997) showed that one advantage of machine learning algorithms compared to profile methods seems to be that they have a higher specificity. This is possible due to the inclusion of non-binding data in the training. A machine learning approach extracts useful information from a large amount of data and creates a good probabilistic model. In the case of HLA-peptide prediction, a data set of known binders and known (or supposed)

non-binders is used. This set is then used to build a model that discriminates between binding peptides and non-binding peptides. This model can then be used to predict whether a novel peptide binds or not.

These methods, as well as other methods can be considered as an extension from the sequence based methods, i.e., profile motifs (Reche *et al.* 2002), additive method (Doytchinova *et al.* 2002a), etc., as they also included the influence of the structure of the HLAp complexes.

2.6.1.1 Binding motif-based methods

Binding motif-based methods determine peptide binding property by identifying general position based pattern of amino acids in favor of HLA peptide binding (Rammensee *et al.* 1995; Rammensee *et al.* 1999). The binding of a peptide to an allele is examined on the basis of occurrence of specific residues at specific position. The presence of motifs will determine whether or not a peptide will bind to specific allele. HLA bound peptides are generally restricted at two primary positions (second or fifth and last positions). These residues are known as anchor residues and positions are known as anchor positions. Based on these anchor positions (Rammensee *et al.* 1995), simple binding motifs have been defined for specific HLA alleles. Thus, position based patterns of recurrent amino acids in a known dataset are identified and generated.

However, the compliance of a peptide sequence to such a binding motif is neither sufficient nor necessary to ensure binding (Townsend *et al.* 1986; Ruppert *et al.* 1993). The usefulness of the motifs is further diminished due to the presence of the secondary

anchor residues at the non-conserved positions (Ruppert *et al.* 1993). This method has been proved to be too simple, as the binding ability of a peptide to a given HLA molecule cannot be explained exclusively in terms of the presence or absence of a few anchor residues.

Obviously, it is not sufficient to only consider the contribution of amino acid on anchor positions to the overall binding affinity. Weaker preferences for specific amino acids in other positions should be taken into account to improve the predictive power of these methods. In some profile based methods, a score is given for each type of amino acid in each position. The scores can be calculated from observed amino acid frequencies in each position or be set manually. The sum of the scores for a given peptide is then used to make predictions.

An example of this category of method was developed by Parker *et al* (1994), which predicts the relative binding strengths of all possible nonamer peptides to the HLA class I molecule based on experimental peptide binding data. The method is based on the observation that each side-chain of the peptide contributes a certain amount to the stability of the HLA complex that is independent of the sequence of the peptide. This contribution is quantified based on the binding data from a set of 154 peptides. A table containing 180 coefficients (20 amino acids x 9 positions) is generated, each of which represents the contribution of one particular amino acid residue at a specified position within the peptide to binding to the HLA. A web-based HLA peptide binding prediction service based on this work is maintained by BIMAS (BioInformatics & Molecular Analysis Section) at NIH with the URL: http://bimas.cit.nih.gov/molbio/hla_bind/, which ranks potential 8-mer, 9-mer, or 10-mer peptides for a limited

set of HLA alleles according to predicted binding affinity. Another web site for predicting HLA binding peptides is SYFPEITHI (Rammensee *et al.* 1999) (URL: <http://www.syfpeithi.de/>), which is primarily a database for HLA binding peptides and peptide motifs. It also provides motif-based HLA binding peptide prediction. The matrices in SYFPEITHI were adjusted manually, by assigning a high score (10) for frequently occurring anchor residues, a score of 8 to amino acids that occur in a significant amount and a score of 6 to rarely occurring residues. Preferred amino acids in other positions have scores that range from 1 to 6 and amino acids regarded as unfavorable have scores ranging from -3 to -1. SYFPEITHI prediction can be done for 13 different HLA class I types.

It has been shown that profile based methods are correct about 30% of the time, meaning that one third of the predicted binders actually bind (Gulukota *et al.* 1997). Prediction models based on binding motif are mostly all-or-none algorithms with very high false negative rates. Both motifs and matrix models present only one sequence pattern in the given set of data that bind to an HLA molecule. They cannot extract multiple sequence patterns hidden in a given set of data separately, even if each of them has sufficient binding ability.

2.6.1.2 Artificial Neural Network Based Methods

Artificial Neural Networks are nonlinear and self-training systems that are able to extract and retain patterns hidden in the training data and recognize them in an input dataset. Neural networks are excellent at classifying nonlinear data. In the case of HLA

binding peptide prediction, it can utilize the information from both binders and non-binders (Sette *et al.* 1989; Adams *et al.* 1995; Brusic *et al.* 1998; Milik *et al.* 1998). The training data of peptide sequences have to be properly aligned. ANN-based models have proven very effective for the prediction of class I HLA (A*0201, Kb) (Adams *et al.* 1995; Milik *et al.* 1998) and class II HLA (DRB1*0401) binding peptides (Brusic *et al.* 1998). Gulukota *et al.* (1997) reported that the performance of the back-propagation neural networks exceeds those of matrix models and motifs in discriminating peptides that bind to an HLA molecule from other peptides (**Figure 2.6**).

The application of ANN on prediction of HLA binding peptide is based on the assumption that the binding can be influenced by the amino acids on all the positions of the peptide. An ANN has the ability to simultaneously analyze the influence of all the amino acids of the peptide and thus may improve binding predictions. Adams *et al.* (1995) applied Neural Networks to predict the binding capacity of peptides to HLA-A*0201. With a large set of binding data from 552 nonamers and 486 decamers, the neural networks achieve a predictive hit rate of 0.78 for classifying peptides as good or intermediate vs. weak or non-binders. The neural networks also depict specific motifs for different binding capacities. ANN is theoretically applicable to all class I and II HLA molecules, given a suitable training dataset of known binding affinities. The trained networks can then be used to perform a systematic search through all pathogen or tumor antigen protein sequences for potential CTL epitopes.

The major drawback of ANN based methods is that a large training set with known binding data is required. In addition, promiscuous prediction is not feasible since data

for one allele cannot be extrapolated to other alleles. An interesting approach is a combination of neural networks and an evolutionary algorithm (Brusic *et al.* 1998). An evolutionary algorithm is used to evolve a scoring matrix during preprocessing and the combined approach is shown to perform better, which achieved a correct classification percentage for both binders and non-binders in excess of 80% (Brusic *et al.* 1998).

Figure 2.6 The architecture of artificial neural network.

There are three layers of neurons in a feed-forward arrangement: the input layer (IL), the hidden layer (HL) and the output layer (OL). In the work of Brusic (1998), the IL consists of 180 neurons. Each oval represents 20 neurons and is a 20-bit word. Each such 20-bit word specifies the identity of the residue at the corresponding position. Thus the 9-mer peptide is encoded as a 180-bit Word. Each neuron in the IL is connected to every neuron in the HL. There are 50 neurons in HL, each filled circle representing ten neurons. Each neuron of HL is connected to the single neuron in the OL. The connections are unidirectional and flow from the IL toward the OL.

2.6.1.3 Stepwise Discriminant Analysis

Stepwise discriminant analysis (SDS) is another data-driven algorithm. The results are data dependent and change when the datasets change. However, as the datasets grow, the sample space is better represented and the influence of individual peptides decreases. Mallios (1999) applied an iterative SDA on a large molecular database to derive quantitative motifs for peptide binding (**Figure 2.7**). From two mutually exclusive sets, stepwise discriminant analysis builds a Bayesian discriminant function that classifies each element into one of the two sets (Mallios 1999). Specifically, an element is assigned to a set if the Bayesian posterior probability of belonging to that set exceeds the probability of belonging to the complementary set. Arguments for the function are selected from a list of potential predictor variables.

In the case of HLA binding motifs determination, the two mutually exclusive sets are the binding dataset of subsequences and the non-binding data-set of subsequences. The potential predictor variables for each case (subsequence) describe the biochemistry and position of each amino acid residue. The decision rule for classification is based on the probability of set membership. The resultant model is quantitative and can be used to predict peptide binding.

Figure 2.7 Stepwise discriminant analysis.

Mallios (1999) produced four closely related models for HLA-DR1. Each model correctly classifies >90% of the peptides in the database.

2.6.1.4 Hidden Markov Model Based Methods

The major problem in the Motif, Matrix or ANN based methods is that all these methods assume that the size of peptides that bind to HLA molecules is fixed, though actually the length of peptides that bind to HLA molecules is variable. Thus, these methods cannot predict the binding ability of a peptide whose length is longer or shorter than that of the peptides used in training, and thus, available training and test data are extremely limited. Specifically, although ANN is able to learn multiple sequence patterns in a given set of data automatically, the network parameters are given only as real-valued weights attached to edges connecting nodes in the network. Consequently, the weights cannot present any understandable training results.

Mamitsuka et al (1998) proposed to apply supervised learning of a Hidden Markov Model to overcome these shortcomings. HMMs are suitable for representing time-series sequences (strings) having flexible lengths (**Figure 2.8**). They used a fully connected HMM, which can automatically divide multiple sequence patterns hidden in a given set of data into separate patterns and is able to represent more than one sequence pattern hidden in a set of given training data. Besides, a trained HMM can be presented as a comprehensible form, just like a sequence profile derived from multiple sequence alignment. Their HMMs were trained for HLA-A2, HLA-DR1 and HLA-DR4. Two experiments were performed. Compared with a back-propagation neural network, the average discrimination accuracy of the HMMs is approximately 2–15% better.

Brusic et al (2002) applied HMM for the prediction of peptide binding to the HLA-A2

supertype. Combined with a representation of peptide/HLA interactions in which the specific HLA–peptide interaction by combining each amino acid of the peptide with the variable amino acids of its positional environment, their system, called MULTIPRED, showed high accuracy of peptide-binding predictions for HLA-A*0201, A*0204, and A*0205 alleles, good accuracy for A*0206 allele, and marginal accuracy for A*0203 allele.

Noguchi et al (2002) applied HMM with successive state splitting (S-HMM) on prediction of peptides binding to a class II HLA, HLA-DRB1*0101. In the relative operating characteristic analysis (Swets *et al.* 1988), the S-HMM prediction had values of ROC>0.85, which is better than other machine learning methods. In addition, the S-HMM may be trained on positive binding data only, and the preprocessing of training data, such as peptide alignment and the selection of binding cores, is not required. Thus the method is simpler to implement.

Figure 2.8 An example of an HMM model trained using five peptide sequences.

In this example, five peptides (ACRR, ACRKR, ACRRRRRRRRR, ACKKRKK, ACK) have the same amino acid in the first position (A), and in the second position (C). But at the third position these peptides may have either R or K, and after the third position, every peptide has different representation including no sequence, a single amino acid, or a string of amino acids. Such variation can be expressed by only one state, which is state 3 in the model. In the various regions of five peptides, there exist 2 types of amino acids, R appearing 14 times and K appearing 6 times. When the total symbol numbers which state 3 generates are 20, the symbol generation probabilities can be given as $14/20=0.7$ (R) and $6/20=0.3$ (K). To express the various regions in five peptides, state 3 was repeatedly used 15 times in 20 total times. Therefore, the transition probability of $15/20=0.75$ is given to the self-loop edge, and $1.0-0.75=0.25$ to the remaining edge. The starting node and the finishing state usually do not share symbol generation probability and meaning transition probability.

2.6.1.5 Support Vector Machines

Support vector machines are supervised classifiers that try to find a linear separation between different classes of points in a high-dimensional space. In a 2D space, this separator is a line; in 3D, it is a plane. In general, this separating surface is called a hyperplane. Support vector machines have two special features. First, instead of just finding any separating hyperplane, they are guaranteed to find the optimal one, or the one whose placement yields the largest separation between the two classes. The data points nearest the frontier between the two classes are called the support vectors, which refer to the coordinates of the data points. Second, although SVMs are linear classifiers, they can classify nonlinearly separable sets of points by transforming the original data points into a higher dimensional space in which they can be separated by a linear surface.

Donnes et al (2002) presented a novel approach, called SVHLA, based on support vector machines to predict the binding of peptides to HLA class I molecules. Their result shows that the method seems to perform slightly better than profile based methods. SVHLA currently provides prediction for 26 class I HLA alleles from the HLAPEP database or alternatively 6 class I HLA alleles from the higher quality SYFPEITHI database. The method is easy to apply to a large number of class I HLA alleles if more peptide data are available.

2.6.1.6 Quantitative Matrix Based Methods

Quantitative matrices provide a detailed linear model that is easy to implement (Sette *et al.* 1989; Parker *et al.* 1994; Gulukota *et al.* 1997). In this category of methods, the contribution to binding at each peptide position within the binding groove is quantified (Parker *et al.* 1994). This method involves producing a matrix in which every entry (X, Y) represents a score associated with amino acid residue X at position Y. The position-specific amino acid values reflect the structural properties of HLA alleles, therefore representing a fingerprint for HLA binding domains. Summing the scores for every residue in a given peptide yields a predicted binding score. The method has been extensively tested for HLA-A2.1 binding predictions. A recent implementation of matrix method is ProPred1 which allows prediction for 47 class I HLA alleles using coefficient matrixes and empirical equations.

The matrix method enables prediction for a wide pool of peptides in a high-throughput manner unlike motif based approaches. However, a serious obstacle is in the generation of binding coefficient matrix for each HLA allele which requires the experimental testing of hundreds of peptides. Another clear limitation of this method is that it assumes that every amino acid residue in a certain position influences binding independently of its neighbors. Matrix methods can give fast predictions on the basis of simple patterns, but they have less capacity to encode non-linear dependencies. This limitation does not yield matrix-based methods completely useless as there is certain simple generalization about amino acid preferences at specific positions influencing binding.

2.6.1.7 Positional Scanning

Udaka et al (2000) analyzed the specificities of three mouse HLA class I molecules, Kb, Db, and Ld, positional scanning using synthetic combinatorial peptide libraries. Graded concentrations of peptides were tested for HLA binding. Peptide concentrations that stabilized a half-maximal number of peptide-receptive class I HLA molecules (SD_{50}) were calculated from the mean fluorescence intensities acquired in an arithmetic scale. Correlation between HLA-binding scores and experimentally measured SD_{50} values was analyzed. By comparing the HLA-binding capacities of sublibraries with the completely random library as reference, the impact of a given amino acid at the position was evaluated. By scanning all nine positions, the profiles of amino acid preference by Kb, Db, and Ld were obtained. Each HLA molecule exhibited a distinct profile. The library scanning yielded a quantitative measure of the impact of every amino acid on HLA binding. The result of the analysis was used to create a scoring program to predict HLA-binding peptides in proteins.

The scoring program was then tested with a number of peptides by comparing the prediction with the experimental binding. The score and the experimental binding exhibited a linear correlation but with substantial deviations of data points. Statistically, for approximately 80% of randomly chosen peptides, HLA-binding capacity could be predicted within one log concentration of peptides for a half-maximal binding. Known cytotoxic T-lymphocyte epitope peptides could be predicted, with a few exceptions. Although the positional scanning data are only informative about an additive component of the binding energy supplied from individual amino acids on peptide, they provide better information than anchor amino acids alone.

2.6.1.8 Profile Motifs

This method is an extension from the motif matrix methods (Reche *et al.* 2002). The most simple sequence patterns that are usually extracted from large numbers of existing known peptides, or from pool sequencing experiments, has proved to be too simple. Motif matrices have been developed to overcome these limitations by accounting for the preference of every amino acid type at every position in the peptide (Rammensee *et al.* 1999). Coefficients in these matrices relate to the strength of the amino acid signals in a pool sequence of peptides eluted from a given HLA molecule, or to the occurrence of an amino acid in a set of binding peptides.

However, it is well established that position specific scoring matrices (PSSM) or profiles created from a set of aligned sequences provides a better way for defining and recognizing sequence motifs (Gribskov *et al.* 1987). There are several methods to generate PSSM from aligned sequences, usually including distinct sequence weighting methods (Thompson *et al.* 1994). In all cases, profile coefficients relate to the observed frequency of every amino acid at the position column of the alignment, corrected by the expected frequency of that amino acid in the background using a reference database. Thus, in this approach the binding potential of any peptide (query) to a given HLA molecule can be obtained by comparing the query to a PSSM created from a set of aligned HLA-specific peptides.

Reche, et al (2002) have derived alignments and profiles from a collection of peptides known to bind two specific class I HLA, Kd and Db, compatible with the structural and molecular basis of the HLAp interaction. A search algorithm, RANKPEP, which

ranks all possible peptides from a test protein using PSSM coefficients, was developed to automate the screening process. It was shown that for Kb and Db molecules, profiles created from aligned peptides are very sensitive in identifying HLA-restricted epitopes. The predictive power of the method was evaluated by running RANKPEP on proteins known to bear Kb- and Db-restricted T-cell epitopes. Analysis of the results indicates that > 80% of these epitopes are among the top 2% of scoring peptides.

These profiles are guided by structural data indicating differences in binding residues involving peptides of distinct length. Prediction of HLA peptide binding using a variety of HLA-specific PSSMs is publicly available online (Reche *et al.* 2002). This method is data driven, thus heavily depends on the quality of the binding data, while available binding data are collected from different sources with different experimental conditions and with peptide sequences of biased amino acid composition, which is a limitation of all data driven prediction method.

2.6.1.9 Additive Method

This method further developed the additivity concept, developed by Free and Wilson (1964) whereby each substituent makes an additive and constant contribution to the biological activity regardless of substituent variation in the rest of the molecule. The values of the individual group contributions are calculated by multiple linear regression (MLR) analysis. The models based on the additivity concept are simple to perform and easy to interpret. Because of that they have found a wide application in molecular design over the years. However, it has been shown that the conformation of

a certain amino acid side chain at a certain position strongly depends on the neighboring amino acids (Fremont *et al.* 1995). This means that the additivity hypothesis is not sufficient to explain the binding abilities of the peptides.

This method is developed to overcome the above limitations (Doytchinova *et al.* 2002a), which is based on the assumption that the binding affinity of a peptide depends on the contributions from each amino acid as well as on the interactions between the adjacent and every second side-chain. In this method, the partial least squares (PLS) were implemented for the multiple linear regression analysis between the different residue interaction terms and the binding affinity. PLS method belongs to so-called projection methods. These methods handle data matrixes with more variables than observations, and the data can be both noisy and highly collinear. In this situation, conventional statistical methods such as multiple regression produce a formula that fits the training data but is unreliable for prediction. PLS forms new variables as linear combinations of the old ones and then uses them as predictors of the biological activity.

Doytchinova *et al.* (2002) applied this method on class I molecule HLA-A*0201, using a training set of 420 experimental IC₅₀ values. The predictive power of the method was assessed by a “leave-one-out” cross-validation with an independent test set of 89 peptides. The mean value of the residuals between the experimental and predicted pIC₅₀ values was 0.508 for this test set.

The additive method for quantitative binding affinity prediction is easy and fast to use and gives a quantitative value for the binding affinity with very good predictive powers.

It can also give a quantitative assessment of individual amino acid contributions at any position in the peptide. The additive method has been implemented in a program for rapid T-cell epitope search. The method is universal and can be applied to any peptide-protein interaction where binding data is known. However, due to its nature as a sequence-based method, the prediction power eventually depends on the quality of the binding affinity data of the training set.

2.6.2. Structure Based Prediction Methods

The second category of prediction methods uses known three-dimensional structures of HLA complexes. The current release of Protein Data Bank (Berman *et al.* 2002) contains a number of unique HLA structures. These structures enable a better understanding of the structural principles governing peptide recognition by HLA molecules (Batalia *et al.* 1997). HLA molecules bind peptides of diverse sequence with great affinity and long half-life. Most peptides selected by class I molecules are 8 to 10 residues long and conserved amino acids bind the invariant portions of the peptides, presenting anchoring backbone atoms at positions 2 and C, N termini (Madden *et al.* 1992; Guo *et al.* 1993). Auxiliary anchors at P1 and P3 usually fine tune peptide recognition (Madden *et al.* 1992; Ruppert *et al.* 1993). Each anchoring side chain interacts with one of the 6 polymorphic HLA pockets (Saper *et al.* 1991; Guo *et al.* 1993), whose structural fold is conserved in evolution with physicochemical diversity for allele specificity (Falk *et al.* 1990). Through a set of hydrogen bonds to the main chain of the peptide, the termini of the peptide are oriented into specific pockets that

are designed to accommodate the chemical nature of the peptide residues. Thus, the orientation (amino to carboxyl) of the antigenic peptide is fixed for all HLA molecules. However, this arrangement is affected by peptide length. Longer peptides may zigzag (Madden *et al.* 1993) or bulge (Guo *et al.* 1992; Collins *et al.* 1995) to allow peptides of greater length to maintain the relative position of the termini. In addition, longer peptides may bind and maintain original binding at the N terminal end with appropriate structural adjustments at C terminal to allow for peptide extension outside the groove.

Various methods calculating binding free energy of HLA-peptide complexes have been developed. Since binding free energy of HLA-peptide complex is related to the affinity of HLA-peptide binding, binders and non-binders can be discriminated and the approach produces absolute or relative peptide binding affinity. Free energy calculation is either based on statistical pair-wise potentials tables or free energy scoring functions. A recent approach based on free energy involved threading of the peptides using known templates followed by evaluation of their binding by statistical pair-wise potentials (Altuvia *et al.* 1995; Altuvia *et al.* 1997; Schueler-Furman *et al.* 1998; Schueler-Furman *et al.* 2000). However, it does not allow the direct prediction of binding affinity values unlike methods capable of calculating the absolute binding free energies from three-dimensional homology models. Moreover, it is difficult to develop a universal free energy function for HLA-peptide binding (Rognan *et al.* 1999).

The backbone conformations of bound peptides are not generally conserved in the binding groove. The bound peptides are flexible and the middle part of the peptides usually bulges out the binding groove. This bulging part allows the backbones to take different patterns at the groove. The generic peptide structure determination methods

using Monte Carlo, molecular dynamics simulations, dynamic programming, free energy mapping, or threading are suited for binding free energy calculations, but they all have difficulty in predicting the conformation of the peptide in the groove. Another method uses computational combinatorial ligand design (CCLD) (Zeng *et al.* 2001) for placing amino acids in specificity pockets. A method based on three-dimensional quantitative structure affinity relationship (3-D QSAR) of HLAp complexes (Doytchinova *et al.* 2001; 2002b) has also been developed. Most approaches, except for threading (Altuvia *et al.* 1995), are not generally suitable for systematic high-throughput genome scanning. However, the TEPITOPE software, developed by Hammer and colleagues, using HLA pocket profiles generated on the basis of structural data (Sturniolo *et al.* 1999), was shown effective for HLA-DR alleles.

2.6.2.1 Molecular Dynamics Simulation

The first approach based on molecular dynamics simulation (MDS) of HLAp complexes (Rognan *et al.* 1994) allowed a crude discrimination of binders from non-binders. The approach produced change in free energy during simulation of the HLA-B*2705 complex with six different peptides in AMBER force field. The result exhibited unexpected structure-activity relationships. Various structural and dynamical properties of the solvated protein-peptide complexes (atomic fluctuations, solvent-accessible surface areas, hydrogen bonding pattern) were found to be in qualitative agreement with the available binding data. The MDS method could be used as a complementary tool to T-cell epitope predictions, as crystal structures of HLA

proteins available. This method is not suitable for high-throughput predictions due to extensive computational requirements in capturing the simulation trajectory.

2.6.2.2 Threading with Knowledge Based Free Energy Scoring

The knowledge based scoring method based on pair-wise contacts uses solved or modeled structure for HLAp binding calculations. The physical, chemical compatibility between peptide and HLA groove is estimated using pair-wise potential matrix. The binding score is obtained by adding all pair-wise values for residues in the pocket with the corresponding peptide residues at every position. This enables the ranking of peptides for HLAp binding (Altuvia *et al.* 1995; Altuvia *et al.* 1997; Schueler-Furman *et al.* 1998; Schueler-Furman *et al.* 2000). The structure based approach is based on two main determinants: (1) The availability of appropriate peptide structural template; (2) the choice of a pair-wise potential table.

Various knowledge based pair-wise potentials have been derived from known protein structures (Jernigan *et al.* 1996; Jones *et al.* 1996; Skolnick *et al.* 1997). A basic approximation underlying these potentials is that total “free energy” of a protein can be expressed as a sum of independent pair-wise interactions. The frequencies of residue pairs in the structures are assumed to represent the interaction preference between different types of residues. This interaction preference between two amino acids is expressed by its comparison with their affinity to a “reference state.” Various matrices have been published using distinct reference states (Skolnick *et al.* 1997). Miyazawa and Jernigan used solvent as reference state and developed a matrix with emphasis on

hydrophobic interactions (Miyazawa *et al.* 1996). Betancourt and Thirumalai (1999) modified the table by changing the reference state from solvent to a defined, single solvent-like molecule. The resulting matrix represents hydrophilic interactions. Altuvia *et al.* used the potential tables of Miyazawa and Jernigan to rank modeled HLAp structures (Altuvia *et al.* 1995; Altuvia *et al.* 1997). However, the procedure failed to predict hydrophilic interactions (Altuvia *et al.* 1997). The pair-wise potential table of Betancourt and Thirumalai (1999) successfully selected hydrophilic interactions (Schueler-Furman *et al.* 2000).

2.6.2.3 Free Energy Scoring Function Based Methods

These methods aim to determine ligands capable of binding from a series of candidate ligands by calculating binding free energy. They generally do not require predetermined experimental data for model development and can produce relatively accurate binding affinity. HLAp binding or non-binding data is not enough to predict whether the peptide can induce immune response. Accurate calculation of HLAp binding free energy difference using 3D- structures by simple free energy scoring functions is CPU intensive.

In recent years, a number of free energy scoring functions have been developed for different purposes and these functions are used for HLAp binding predictions. One of the recent approaches for class I HLAp binding prediction is a tailor-made free energy scoring function (FRESNO) combined with homology modeling (Rognan *et al.* 1999; Logean *et al.* 2001; Logean *et al.* 2002). Starting from the primary sequence of the

protein antigen, individual 3D structures of all possible class I HLA-peptide (8-, 9- and 10-mers) complexes are constructed by homology modeling. The critical issue in this approach is the identification of peptide templates for structure prediction. The FRESNO scoring function is then used to calculate binding free energy of HLAp interactions. The approach allows for the prediction of absolute binding affinities in a high throughput manner. An extension to this work is EpiDock (Logean *et al.* 2002) which is (1) shown to predict potential T-cell epitopes from viral proteomes, (2) used to roughly predict still unknown peptide binding motifs for novel class I HLA alleles.

2.6.2.4 Virtual Matrix Based Methods

Virtual matrices, like quantitative matrices, provide a detailed model in which binding of each peptide residue with HLA pockets is quantified using pocket profiles (Sturniolo *et al.* 1999; Raddrizzani *et al.* 2000). Virtual matrices are derived by assigning and combining pocket-specific binding properties using structural features or homology principles from known HLA structures and extrapolating to other alleles, while quantitative matrices are obtained using known allele-specific peptide binding data (**Figure 2.9**). The advantage over quantitative matrices is that the method is generic and can be applied to any given allele. One implementation of the algorithm is the software package TEPITOPE (Sturniolo *et al.* 1999). Its capacity has been demonstrated for 11 HLA-DR alleles. Furthermore, it has been successfully applied to predict T-cell epitopes in oncology, allergy, and autoimmune diseases (Hammer *et al.* 1995; Gross *et al.* 1998; Cochlovius *et al.* 2000; Stassar *et al.* 2001).

Figure 2.9 Virtual matrix based method.

Quantitative matrix data can be split into single pocket profiles for the assembly of virtual matrices. **a.** Schematic representation of the building process of quantitative matrices. **b.** Once a substantial pocket profile database is generated, virtual matrices can be assembled via HLA pocket homology (Raddrizzani *et al.* 2000).

2.6.2.5 Computational Combinatorial Ligand Design

Computational combinatorial ligand design (CCLD) is a computational tool used to assist drug design, by clustering compounds in classes of drug-like and non-drug-like molecules. The method selects fragments that bind favorably to a macromolecular target of known three-dimensional structure. Optimal positions and orientations of functional groups on the surface of the macromolecule are exhaustively searched, and then sorted according to an approximated binding free energy. The CCLD method allows fast and automatic generation of a multitude of highly diverse compounds, by connecting in a combinatorial fashion the functional groups in their minimized positions. The fragments are linked as two atoms may be either fused, or connected by a covalent bond or a small linker unit. To avoid the combinatorial explosion problem, pruning of the growing ligand is performed according to the average value of the approximated binding free energy of its fragments.

The CCLD method uses the 3D information from the crystal structure of the molecule. It can generate both sequence and structure of predicted ligands. Zeng et al (2001) applied the method on the prediction of peptides that bind a HLA molecule with known crystal structure. Using chemical fragments as models for amino acid residues, a set of sequences for peptides predicted to bind in the HLA peptide-binding groove were produced. The probabilities for specific amino acids occurring at each position of the peptide were calculated based on these sequences. Their results show the CCLD approach is a sensitive method that can capture the important features of both sequence and structural data.

2.6.2.6 3D Quantitative Structure-affinity Relationship

Three-dimensional quantitative structure-affinity relationship (3D QSAR) studies have been applied to explore the molecular interactions between HLA and peptides. They provide easily interpretable coefficient contour maps identifying the areas of the peptides that require a particular physicochemical property to increase binding. One of the 3D QSAR methods, Comparative Molecular Similarity Indices Analysis (CoMSIA), is a very reliable method for investigating the structure-activity trends within sets of biological molecules. It has been successively applied in pharmaceutical discovery of small molecule drugs. It is a statistic approach that seeks to correlate relative differences in molecular descriptor values to a dependent property (e.g. the binding affinity). In that respect, CoMSIA is a method able to map similarities or dissimilarities between molecules. The explanatory power of CoMSIA methods is considerable, manifest not only in their ability to accurately predict binding affinities, but also in their capacity to display advantageous and disadvantageous 3-D interaction potential mapped onto the structures of molecules being investigated.

CoMSIA method have been applied on HLA peptide binding prediction with class I HLA molecule HLA-A*0201 and two hundred and sixty-six nonamer peptides (Doytchinova *et al.* 2001; 2002b). CoMSIA uses the interaction potential around aligned sets of 3D peptide structures to describe the contributions to binding. The relationship between physico-chemical properties and the affinities of peptide binding was investigated. The X-ray structure of one nonameric viral peptide was used as a starting conformation, on which the structures of the remaining peptides were built. Five types of similarity index (steric bulk, electrostatic potential, local hydrophobicity,

hydrogen-bond donor and hydrogen-bond acceptor abilities) were calculated, using a common probe atom with 1 Å radius, charge +1, hydrophobicity +1, hydrogen-bond donor and acceptor properties +1. Since only the combination of all fields provided a complete insight, only an all-fields model was analyzed further. The same parameters as for the additive method were used to assess the predictive power of the final model. Three types of cross validation were performed.

One difficulty of the application of the CoMSIA on peptide prediction is that peptides size is large compared to small molecules and the diversity of the physico-chemical properties associated with each position being examined. However, good agreement was found between the results generated by other techniques (Doytchinova *et al.* 2001).

CoMSIA can predict the binding affinity of a peptide with an amino acid not presented in the initial training set, but it cannot assess the contribution of each amino acid at each position and the interactions between them. Another advantage of the CoMSIA method is that it returns 3D representations of the analysis for visual investigation, which indicate where adding particular kinds of functionality to the peptides would contribute to activity, either positively or negatively. However, because the method is also data driven, the predictivity of this method is also dependent on the quality of binding data.

2.6.3 Limitations of Current Methods

The relevant information of the peptide binding prediction methods discussed above is summarized in **Table 2.2**. These methods generally achieve acceptable prediction accuracy, and are commonly applied to certain HLA alleles. They have some common limitations in practical use. The main limiting factor is the quantity and quality of HLA peptide binding affinity data required for model training or development.

Ideally, large-scale HLA peptide binding assays need be carried out to obtain high quality binding affinity data of systematically designed peptide library, with broad and even coverage of amino acid properties such as size, charge, hydrophobicity, etc. However, it will be economically infeasible to carry out such large scale of experiments for the current stage, as the possible number of peptide is enormous, and the HLA polymorphism is extremely large.

Currently, most HLA peptide binding affinity data for bioinformatics research is collected from immunological or biochemical research investigating HLA peptide interaction. Normally, these research works only document the affinity data of binders. Another source of binding affinity data are from epitope mapping using synthetic peptide library. The purpose of these researches is to identify peptide binders. Thus non-binders are seldom documented. Furthermore, available binding data has many sources of potential error and inconsistency. Peptides studied by various groups are physically large with extreme physical properties, being multiply charged, zwitterionic and/or exhibiting huge ranges of hydrophobicity. Importantly, their sequences are also biased in terms of amino acid composition, often favoring hydrophobic sequences.

This arises, in part, from pre-selection processes that result in self-reinforcement. Binding motifs are often used to reduce the experimental burden of epitope discovery. Very sparse sequence patterns are matched, and the corresponding subset of peptides tested, with an enormous resulting reduction in sequence diversity. Moreover, there is little consensus on the best assay systems and affinity data are of an intrinsically inferior quality: multiple measurements of the same peptide can vary by several orders of magnitude, mixtures of different standard peptides are used in radio-ligand assays, and experiments are conducted at different temperatures and over different concentration ranges.

Due to this major practical limiting factor, the HLA peptide binding prediction approaches described above have following limitations:

- a. Most of above mentioned approaches based on ANN, HMM, SVM, etc., cannot effectively deal with HLA Polymorphism i.e. for each allele a separate matrix has to be generated or a separate set of rules has to be applied.
- b. Most of these methods have been developed only for one or two alleles and are not able to predict promiscuous HLA binders that can bind to multiple alleles.
- c. Most of these methods only produce qualitative prediction results by classifying the peptide as binders and non-binders. However, the binding or not binding does not provide sufficient information to determine whether a peptide will induce immune response, since only the peptides with HLA binding affinity within a specific range are T-cell reactive. Thus, quantitative binding affinity data is crucial for the identification of real reactive T-cell epitopes. An ideal prediction model is required to produce

quantitative prediction results.

To circumvent the practical limitations, we endeavored to develop a novel prediction model based on structural information of HLA peptide complexes.

Table 2.2 Summary of HLA-peptide Binding Prediction Methods.

	Principle	Group	Year	Training Dataset			Prediction Dataset		ROC Analysis						CC		
				Allele	Crystal structure	Modeled structure	Allele	Modeled structure	Binding data	PPV	NPV	A _{roc}	SE	SP		AC	
																	Binding Data
Sequence Based Methods	Human Data																
	BIMAS	Parker	1994	HLA-A2	-	-	154	HLA-A2	-	-	-	-	-	-	-	-	-
	ANN	Adams	1995	A*0201	-	-	1,038	A*0201	-	-	-	-	0.78	-	-	-	-
	EA & ANN	Brusic	1998	DRB1*0401	-	-	338	B1*0401	-	62	-	-	-	-	-	-	-
	SDS	Mallios	1999	DR1	-	-	526	DR1	-	-	-	-	N/A	96.8%	73.5%	93.1%	-
	HMM	Brusic	2002	HLA-A2	-	-	-	4 alleles	-	-	-	-	-	-	-	-	-
	HMM + SSS	Kobayashi	2002	DRB1*0101	-	-	305	DRB1*0101	-	-	-	-	0.89	-	-	-	-
	SVMHC	Elofsson	2002	26 alleles	-	-	-	6 alleles	-	-	-	-	-	-	-	91-100%	-
	Quantitative Matrix	Sette	1989	47 alleles	-	-	-	-	-	-	-	-	-	-	-	-	-
	Mouse Data																
Positional scanning	Udaka	2000	H-K ^b , D ^b , L ^d	-	-	-	-	-	-	-	-	-	-	-	-	80%	-
Profile motifs	Reinherz	2002	H-K ^b , D ^b	-	-	71	H-K ^b , D ^b	-	-	-	-	-	-	-	-	>80%	-
Structure Based Methods	Mostly Human Data																
	MDS	Rognan	1994	B*2705	1	5	6	-	-	-	-	-	-	-	-	-	-
	Threading	Margalit	1998	9 alleles	23	-	23	-	-	-	-	-	-	-	-	70%	-
	FESF	Rognan	1999	A*0201, H-2 ^k , 5 (A*0201)	30	H-2 ^k	-	A*0204, H-2 ^k	-	24 , 16	-	-	-	-	-	-	-
	Virtual Matrix	Hammer	1999	11 HLA-DR	-	-	-	51 alleles	-	223	-	-	-	-	-	80%	-
	Threading	Margalit	2000	10 alleles	37	-	123	-	-	106	-	-	-	-	-	71%	-0.61 to -0.07

CHAPTER 2 LITERATURE REVIEW

	CCLD	Zeng	2001	A*0201	1	-	-	-	-	-	-	-	-	-
3D QSAR		Flower	2001	A*0201	1	102	84	18	A*0201	50	50	-	-	-
EpiDock		Rognan	2002	A*0201	25/63	-	-	-	A*0201, B*2705	-	-	80- 100%	76- 85%	77-85%

Abbreviations: HMM = Hidden Markov Model, SSS = Successive State Splitting, SDS = Stepwise Discriminant Analysis, ANN = Artificial Neural Network, EA = Evolution Algorithm, SVHLA = Support Vector Machines, CCLD = Computational Combinatorial Ligand Design, MDS = Molecular Dynamics Simulation, EF = Free Energy Scoring Function, HM = Homology Modeling, QSAR = Quantitative Structure Activity Relationship, ROC = Receiver Operating Characteristic, AC = Accuracy, CC = correlation coefficient.

ROC analysis is used to compare the accuracy of the prediction methods. In ROC analysis, A_{roc} (area under the ROC curve) is used as a measure of prediction power in some investigations. An $A_{\text{roc}} > 0.8$ indicates that the method is feasible. In some investigations, the sensitivity, specificity and correlation coefficient (CC) of the prediction methods are also presented. Sensitivity (SN) is the proportion of correctly predicted binders $((\text{TP} / (\text{TP} + \text{FN})) * 100)$. Specificity (SP) is the proportion of correctly predicted non-binders $((\text{TN} / (\text{TN} + \text{FP})) * 100)$. Positive Predictive Value (PPV) is the probability that a predicted binder will actually be a binder $((\text{TP} / (\text{TP} + \text{FP})) * 100)$. Negative Predictive Value (NPV) is the probability that a predicted non-binder will actually be a non-binder $((\text{TN} / (\text{TN} + \text{FP})) * 100)$. TP = True Positive; TN = True Negative; FP = False Positive; FN = False Negative. If data is not available (not applicable as in some case) for a particular feature then it is indicated by ‘-’.

CHAPTER 3

HLA PEPTIDE BINDING PREDICTION MODEL

The current International ImmunoGeneTics Information System/Human Leukocyte Antigen (IMGT/HLA) Sequence Database (Robinson *et al.* 2003) contains over 2,000 HLA alleles (till July 2005). Only a small fraction of available peptide pool binds specifically to each of the known alleles and the HLA binding is highly combinatorial. As discussed earlier, a number of mathematical models for HLA binding predictions have been proposed and tested using computer-aided calculations in the last decade. Many of these models utilize features extracted from pooled peptide sequences and peptide residue anchor position assignment procedures. The efficiency of these models has been significantly improved by the careful utilization of the extracted features in nonlinear statistical models. However, there is a bottleneck in anchor position assignment, as a large number of peptide sequences with predetermined binding affinity to each of the known HLA alleles are required. Thus, the efficiency of all data driven predictive models is mainly determined by the quality and quantity of binding data generated for each HLA allele. However, generation of such binding data for each allele is experimentally laborious and economically expensive.

Concurrent progress in computational structural biology during the last decade has

motivated many of us to assess the usefulness of molecular mechanics and modeling calculations for HLA binding predictions. All known HLA molecules have at least 70% sequence identity to one of the known HLA structures (Kangueane *et al.* 2001). Homologous proteins share a common fold, even when the overall sequence identity is less than 10% (Abagyan *et al.* 1997). Homology modeling and side chain packing calculations are commonly performed to predict the structures of all known HLA alleles (Rognan *et al.* 1994; Lee *et al.* 1995; Altuvia *et al.* 1997; Rognan *et al.* 1999; Kangueane *et al.* 2000; Schueler-Furman *et al.* 2000). The drawback in these approaches is the current limitation in backbone prediction as peptide backbones in the binding groove play a very significant role (Schueler-Furman *et al.* 1998; Adrian *et al.* 2002). This is further complicated due to the inadequate knowledge on the physicochemical characteristics of HLA molecular recognitions (Miyazawa *et al.* 1996). Also, a large number of interaction functions and scoring schemes have been proposed over the last few years, but their efficiency is far from satisfactory.

Based on the analysis of HLA molecular interactions using crystal structures in recent years, the understanding of HLA interactions in relation to the HLA polymorphism have been significantly improved (Ren *et al.* 2000; Kangueane *et al.* 2001; Adrian *et al.* 2002; Govindarajan *et al.* 2003). It is known that six binding pockets, denoted A through F, are formed by the polymorphic amino acid residues within the binding grooves of HLA molecules (**Figure 3.1**). Anchor residues of peptide are accommodated in the pockets. The binding environments of anchor residue in the pockets, or pocket profile, are determined by the HLA polymorphic side-chains delineate the geometry and chemical properties of these structural pockets. In this work, pocket profiles are defined as those residues in a fixed neighborhood of the peptide

residues in known crystal structure complexes. Pockets are the basic functional unit of HLA molecules. The amino acid residues that form the pocket determine the properties, such as size, shape, hydrophobicity, and electronic charge, etc., of the pocket and thus determine the antigen peptides that would be preferentially bound by the HLA allotype. This accounts for the differential ability of different alleles to bind a variety of peptides. HLA alleles sharing same pocket profiles will preferentially bind same kind of peptide anchor residues. Thus, the HLA peptide-binding groove is essentially an exchange or shuffling of pockets between different HLA allotypes (**Figure 3.1d**). Most of the binding affinity of the peptides bound to class I HLA molecules is provided by the hydrogen bonding at the end of the groove between the peptide and the conserved tyrosine residues. But the binding specificities of class I HLA are determined by different combination of binding pockets (**Figure 3.1d**).

By establishing the geometry and chemical compatibilities between individual pockets and peptide anchor residues, the binding strength of peptide anchor residues can be quantified. Thereby, a peptide binding motif can be identified and the accumulated binding strength between peptide and HLA molecule can be roughly determined. However, to accurately assess the binding affinity of a whole peptide, it is not enough to solely consider individual anchor residues that interact with binding pockets, because the peptides residues other than anchor residues are also contributing to the peptide binding affinity.

To solve this problem, the concept of structural pocket was expanded. In the first step, the key functional residues for peptide binding in the HLA binding grooves were identified. Then the correlation between the peptide residue positions and the key

functional HLA residue positions was established. The key functional HLA residues that directly interact with a certain peptide residue can be viewed as a virtual pocket. This means that their residue compositions are based on structural pockets, but their positional composition will change when they interact with different individual peptide residue.

The new HLA peptide binding prediction model is developed based on this virtual pocket concept. Obviously there is a prerequisite condition: the peptide length is required be fixed. This is because the virtual pockets are defined based on their relationship with peptide residue positions. This relationship will change when the peptide length are different. Since the majority of peptides in the available crystal structures of class I HLA peptide complexes are nonamers, the model will be based solely on the structural data of HLA-nonamer peptide complexes, and will be only applicable to predict the binding of nonamer peptides. However, the model can be easily extended to the peptides of different length when enough structural data are available.

a

b

Figure 3.1 Peptide binding pockets in HLA binding groove and the concept of pocket profile.

a. A side view and a cross-section of a class I molecule, HLA-A*0201, with bound peptide (PDB Code: 1I7U). Peptide is drawn in ball-and-stick format. Peptide residues are colored differently.

b. A side view and a cross-section of a class II molecule, HLA-DR1, with bound peptide (PDB Code: 1DLH). Peptide is drawn in ball-and-stick format. Peptide residues are colored differently.

In these above two panels, HLA peptide binding pockets and peptide anchor residues are shown. The peptide binding groove of the class I molecule is closed, whereas that of the class II molecule is open. As a result, class II molecules accommodate longer peptides than do class I molecules.

c. This illustration shows the peptide binding pockets. A HLA molecule binds and displays a peptide and the T-cell receptor recognizes the two polymorphic residues of the HLA molecules and one residue of the peptide. Anchor residues of peptide are accommodated in the pockets formed by the polymorphic residues in the binding groove of the HLA molecule.

b. Six binding pockets, denoted A through F, are formed by the polymorphic amino acid residues within the binding groove of HLA molecule. Pocket profiles is a detailed description of peptide binding environments in the pockets, which are determined by the HLA polymorphic side-chains delineate the geometry and chemical properties of these structural pockets. In this work, pocket profiles are defined as those residues in a fixed neighborhood of the peptide residues in known crystal structure complexes. Pockets are the basic functional unit of HLA molecules. The amino acid residues that form the pocket determine the antigen peptides that would preferentially be bound by the HLA allotype, accounting for the differential ability of different alleles to bind a variety of peptides. HLA alleles sharing same pocket profiles will preferentially bind the same kind of peptide anchor residues. The peptide-binding groove of the HLA is essentially an exchange or shuffling of pockets between different HLA allotypes.

3.1 Model Development

3.1.1 Overview

The procedure for the development of peptide binding prediction model for class I HLA is summarized in **Figure 3.2**. The comparison of known structures of HLA-peptide complexes shows that the backbone conformation of the HLA molecules are highly conserved, even in the peptide binding grooves (Vasmatzis *et al.* 1996a, 1996b). Therefore, structural information gathered from different HLA alleles can represent general HLA backbone structural characteristics. In the first step, the structural information was gathered by examining three-dimensional structures of class I HLA peptide complexes. The residues that directly interact with the binding peptides were identified based on the interface area between HLA molecules and peptides in complex state, based on the fact that only the residues showing a changed ASA value during this process are directly involved in the interaction with the bound peptide. Among these functional residue positions, those that are conserved across HLA alleles were excluded from further analysis since these positions provide no information about functional difference among HLA alleles. Amino acid residues on these polymorphic functional positions determine the properties of the peptide binding groove, which were divided to virtual pockets by determining their interaction with peptide anchor residues. The interaction between functional residues inside each virtual pocket and peptide anchor residues was quantified using the Q matrix. The performance of the model was assessed by calculating sensitivity, specificity, and accuracy based on HLA peptide binding affinity data collected from literatures and public databases.

Figure 3.2 An overview of the model development for HLA peptide binding prediction and HLA supertype sub-typing.

3.1.2 Creation of Human HLAp X-ray Structure Dataset

46 crystal structures of class I HLA peptide complexes solved by x-ray crystallography were available in Research Collaboratory for Structural Bioinformatics (RCSB) Protein Data Bank (PDB) (Berman *et al.* 2002), till March 2005 (**Table 3.1**). 33 of these structures containing nonamer peptides were selected to glean structural information of HLA peptide interaction. The crystal structures with same HLA allele and peptide were removed to make the dataset non-redundant. Thus the selected structures represented 24 unique HLAp sequences.

It should be emphasized that peptides in selected structures are 9 residues long. HLAp structures with bound peptides longer or shorter than 9 residues were removed from the dataset. The length of the peptide throughout the investigation was fixed at 9 residues to avoid possible bias due to dissimilar HLAp binding patterns among peptides of different length (Kangueane *et al.* 2001). It was demonstrated that 8, 9, and 10 residues long peptides have different binding patterns (in physical terms), especially at the C terminal of the peptides (Kangueane *et al.* 2001). The virtual binding pockets of HLA are defined based on fixed relationship with individual peptide residue positions. This relationship will change when the peptide length are different. Thus, the length of the peptide is required to be fixed.

HLAp structures with bound nonamer peptides were best represented among available crystal structures. It should also be stated that HLAp structures with peptides of length 8 and 10 residues were not as represented as 9 residues long peptides. Therefore, the length of the peptide was fixed to 9 residues. This will not affect the application of the

model in peptide vaccine design, because the length of the peptide in the design of a synthetic vaccine can be fixed. Issues related to class I HLA bound peptides of length greater than 9 residues have to be addressed in future as more structural data of its kind are available. At this stage, the analysis was applied to class I HLA specific peptides and the work was extended to class II HLA specific peptides in the later part of the work.

Table 3.1 Class I HLA-peptide complexes in PDB.

SN	PDB	HLA	CIH	CIP	PL	Redundant	Non- redundant	Peptide Source	R [Å]	Release Year
	Code	Allele				Peptide set	Peptide Set			
1	1Q94	A*0101	a	c	9	AIFQSSMTK	AIFQSSMTK	HIV-1 RT	2.40	2004
2	1JHT	A*0201	a	c	9	ALGIGILTV	ALGIGILTV	Mart-1	2.15	2001
3	1I7U	A*0201	a	c	9	ALWGFVPVL	ALWGFVPVL	Synthetic construct	1.80	2001
4	1I7T	A*0201	a	c	9	ALWGVFPVL	ALWGVFPVL	Synthetic construct	2.80	2001
5	1JF1	A*0201	a	c	10	ELAGIGILTV	ELAGIGILTV	Mart-1	1.85	2001
6	1I7R	A*0201	a	c	9	FAPGFFPYL	FAPGFFPYL	Synthetic construct	2.20	2001
7	1I1F	A*0201	a	c	9	FLKEPVHGV	FLKEPVHGV	HIV RT	2.80	2000
8	1HHH	A*0201	a	c	10	FLPSDFFPSV	FLPSDFFPSV	HBV nucleocapsid	3.00	1993
9	1HHI	A*0201	a	c	9	GILGFVFTL	GILGFVFTL	Synthetic	2.50	1993
10	1I4F	A*0201	a	c	10	GVYDGREHTV	GVYDGREHTV	MAGE-4 Antigen	1.40	2001
11	1AKJ	A*0201	a	c	9	ILKEPVHGV	ILKEPVHGV	HIV-1 RT	2.65	1997
12	1HHJ	A*0201	a	c	9	ILKEPVHGV		Synthetic	2.50	1993
13	1QRN	A*0201	a	c	9	LLFGYAVYV	LLFGYAVYV	Tax peptide P6A	2.80	1999
14	1QSE	A*0201	a	c	9	LLFGYPRYV	LLFGYPRYV	Tax peptide V7R	2.80	1999
15	1QSF	A*0201	a	c	9	LLFGYPVAV	LLFGYPVAV	Tax peptide Y8A	2.80	1999
16	1A07	A*0201	a	c	9	LLFGYPVYV	LLFGYPVYV	HTLV-1 Tax	2.60	1997
17	1BD2	A*0201	a	c	9	LLFGYPVYV		HTLV-1 Tax	2.50	1998
18	1DUZ	A*0201	a	c	9	LLFGYPVYV		HTLV-1 Tax	1.80	2000
19	1HHK	A*0201	a	c	9	LLFGYPVYV		Synthetic	2.50	1993
20	1IM3	A*0201	a	c	9	LLFGYPVYV		HTLV-1 Tax	2.20	2001
21	2CLR	A*0201	a	c	10	MLLSVPLLIG	MLLSVPLLIG	Synthetic	2.00	1998
22	1HHG	A*0201	a	c	9	TLTSCNTSV	TLTSCNTSV	HIV-1 gp120	2.60	1993
23	1I1Y	A*0201	a	c	9	YLKEPVHGV	YLKEPVHGV	HIV-1 RT	2.20	2000
24	1TMC	A*6801	a	c	10	EVAPPEYHRK	EVAPPEYHRK	Synthetic	2.30	1995
25	1AGD	B*0801	a	c	8	GGKKKYKL	GGKKKYKL	HIV-1 gag	2.05	1997
26	1AGC	B*0801	a	c	8	GGKKKYQL	GGKKKYQL	HIV-1 gag	2.10	1997
27	1AGF	B*0801	a	c	8	GGKKRYKL	GGKKRYKL	HIV-1 gag	2.20	1997
28	1AGB	B*0801	a	c	8	GGRKKYKL	GGRKKYKL	HIV-1 gag	2.20	1997
29	1AGE	B*0801	a	c	8	GGRKKYKL		HIV-1 gag	2.30	1997
30	1HSA	B*2705	a	c	9	ARAAAAAA	ARAAAAAA	N/A	2.10	1992
31	1JGE	B*2705	a	c	9	GRFAAAIAK	GRFAAAIAK	Synthetic (M9)	2.10	2002
32	1OF2	B*2705	a	c	9	RRKWRRWHL	RRKWRRWHL	Vasoactive Intestinal Polypeptide Receptor	2.20	2004
33	1OGT	B*2705	a	c	9	RRKWRRWHL		Vasoactive Intestinal Polypeptide Receptor	1.47	2004
34	1W0W	B*2705	a	c	9	RRLPIFSRL		Butyrate Response Factor 2	2.10	2004
35	1UXS	B*2705	a	c	9	RRRWRLTV	RRRWRLTV	Gene Terminal Protein Lmp-2	1.55	2004
36	1UXW	B*2705	a	c	9	RRRWRLTV		Gene Terminal Protein Lmp-2	1.71	2004
37	1K5N	B*2709	a	c	9	GRFAAAIAK	GRFAAAIAK	Synthetic (M9)	1.09	2002
38	1A9B	B*3501	a	c	9	LPPLDITPY	LPPLDITPY	EBNA-3C	3.20	1998
39	1A9E	B*3501	a	c	9	LPPLDITPY		EBV-Ebna3c	2.50	1998
40	1A1N	B*3501	a	c	8	VPLRPMTY	VPLRPMTY	HIV-1 Nef Protein	2.00	1998
41	1E27	B*5101	a	c	9	LPPVVAKEI	LPPVVAKEI	HIV-1 Km1	2.20	2000
42	1E28	B*5101	a	c	8	TAFTIPSI	TAFTIPSI	HIV-1 Km2	3.00	2000
43	1A1O	B*5301	a	c	9	KPIVQYDNF	KPIVQYDNF	HIV-1 Nef	2.30	1998
44	1A1M	B*5301	a	c	9	TPYDINQML	TPYDINQML	HIV-2 gag	2.30	1998
45	1EFX	Cw*0304	a	c	9	GAVDPLLAL	GAVDPLLAL	Importin α -2	3.00	2000
46	1IM9	Cw*0401	a	c	9	QYDDAVYKL	QYDDAVYKL	Synthetic	2.80	2001

There were 46 crystal structures of class I HLA peptide complexes available in Research Collaboratory for Structural Bioinformatics (RCSB) Protein Data Bank (PDB) (Berman *et al.* 2002), till March 2005. 33 of these structures containing nonamer peptides were used to glean structural information of HLA peptide interaction in this work.

Duplicate peptides were removed in the non-redundant peptide set.

Abbreviations: SN = serial number; HLA = Human Leukocyte antigen; CIH= chain identifier for HLA; CIP = chain identifier for peptide; PDB= protein databank; PL = peptide length; R = resolution; Release Year = the year in which the entry was released by PDB.

3.1.3 Identification of Highly Essential Residue Positions

The residues that interact with the binding peptides were identified based on the interface area between HLA molecules and peptides in complex state. Interface area for class I HLA-peptide complexes was defined as the change in solvent accessible surface area (Δ ASA) while forming the HLA-peptide complex from the monomeric HLA and the peptide. It is obvious that only the residues showing a changed ASA value during this process are directly involved in the interaction with the bound peptide.

The accessible surface of whole molecule is generally traced out by the probe sphere center as it rolls over the protein, which is a kind of expanded *van der Waals* surface. A common algorithm calculating ASA developed by Lee and Richard (1971) was applied in this analysis. Water molecule was used as probe. Later in this section, the relative ASA of peptide residues were calculated using the same method, but was set in a trimer state with flanking Alanine as a reference state. The relative ASA, or %burial, of individual peptide residue was calculated as the percentage of ASA in the HLAp complex state, out of the relative ASA in the separate peptide state.

In this work, a residue is defined as directly interacting with the binding peptide when the Δ ASA of the residue is greater than $1 \text{ \AA}^2$ upon HLAp complex formation. Only first 180 residues in each HLA alleles were included for calculation of Δ ASA because these are the residues that form the peptide binding groove. Based on this criterion, all the 180 residue positions were examined to determine if at least one amino acid residue on a particular position is directly interacting with the binding peptide amongst the 24

non-redundant crystal structures (**Table 3.1**). The result of this calculation is shown in **Figure 3.3**. 39 residue positions in the α_1 , α_2 domains with a mean ΔASA of $\geq 1 \text{ \AA}^2$ (except for position 24) upon HLAp complex formation were identified as predominantly involved in peptide binding (**Table 3.2**). These residue positions are called highly essential residue positions (HERP). It is shown that the HERPs are quite concentrated on a set of few residue positions, which means only a small part of all the residues in the binding groove of HLA molecules are directly interacting with peptide residues.

Position 24 (with a mean ΔASA of $0.71 \text{ \AA}^2$) was included in the set of HERPs because residues at this position indicate three inter-residue interactions with the peptides in the dataset (discussed later). Within this part of the work, the positions that characterized a very low ΔASA (> 0 and $< 1 \text{ \AA}^2$) were neglected because these positions have negligible involvement in HLAp binding. These residues were excluded using an arbitrary cutoff value of ΔASA ($1 \text{ \AA}^2$), because their inclusion will add noise to the model. It should be noted that the definition of HERPs stated here are slightly different from CPERP's defined in the later part of this work. CPERP refers to polymorphic positions among HLA alleles. The HERPs defined in this part include conserved residue positions, because the entire set of residue involved in peptide binding are required to calculate the compatibility between virtual pockets and peptide residues. The HERPs are also a subset of the essential residue positions (ERPs) defined in the later part of this work. The difference is in the selection of ΔASA cutoff values used for selecting HERPs and ERPs. The ERPs reveal a ΔASA of ≥ 0 , whereas the HERPs exhibit a ΔASA of $\geq 1 \text{ \AA}^2$ in the dataset.

Figure 3.3 Δ ASA of the first 180 residues of class I HLA α chains upon complex formation.

a. Distribution of Δ ASA (unit: $\text{\AA}^2$) of the first 180 HLA residues among the 36 non-redundant class I HLA α p structures. **b.** Mean Δ ASA of the first 180 HLA residues. **c.** Standard deviation (SD). HLA residues showing change in ASA upon complex formation were identified as highly essential residue positions (HERP).

3.1.4 Definition of Virtual Binding Pockets

The HLA residue positions interacting with each of the peptide residue positions were identified by counting the number of pair-wise inter-residue contacts between the HLA and the peptide (**Table 3.2**). By definition, HLA_p residue pairs are considered to be interacting if the distance between any atom in the peptide residue and any atom in the HLA residue is less than or equal to an arbitrary cutoff distance of 4 Å. This procedure established the relationship between peptide residue positions and HERPs (**Table 3.3**). The HLA peptide residue pairs at the corresponding HERPs have at least 1 inter-residue interaction (**Table 3.4**). The set of HLA residue positions interacting with each peptide position forms the virtual binding pockets. As mentioned in step 2, these HLA residues positions (except position 24) demonstrate a $\Delta\text{ASA} \geq 1 \text{ Å}^2$ in the dataset. This suggests that the residue positions forming the virtual pockets are none other than the HERPs defined in step 2 with the inclusion of position 24. Based on the mapping established between the peptide residue positions and HLA residue positions forming virtual binding pockets, the compatibility of a particular residue type in a specific peptide position to the residues forming the virtual binding pockets in the groove can be determined using a scoring scheme. It should be noted that the virtual pockets defined in this model are not strictly identical to the A–F pockets defined by Zhang *et al.* (1998), and Guan *et al.* (A–F pockets for A3 supertypes) (2003). The A–F pockets defined in these reports are certainly related (not strictly identical) to the virtual pockets defined in our method, but a one-to-one correspondence has to be established in future for further improvement. Current attempts to develop this one-to-one correspondence is hampered because of A–F pocket overlapping for relating virtual pockets defined in this work.

Table 3.2 List of HERPs.

HERP	ΔASA ($\text{\AA}^2$)		Sum	Contacts	
	Mean	SD		Mean	SD
5	3.59	1.39	18	0.62	0.49
7	18.33	2.81	59	2.00	0.27
9	6.26	3.88	29	1.00	0.60
24	0.71	1.76	3	0.14	0.35
45	4.94	3.03	19	0.66	0.48
59	4.91	3.78	13	0.45	0.51
62	8.67	16.53	11	0.38	0.73
63	21.12	4.12	56	1.93	0.26
65	5.23	8.76	4	0.21	0.49
66	67.99	11.29	88	2.97	0.98
67	5.47	3.84	8	0.28	0.45
69	16.67	12.05	7	0.24	0.44
70	34.31	14.51	45	1.55	0.99
72	1.96	4.71	3	0.10	0.31
73	50.37	10.67	62	2.14	0.95
74	2.70	5.45	7	0.24	0.58
76	16.51	8.69	18	0.62	0.49
77	35.45	7.29	61	2.10	0.31
80	15.27	4.94	30	1.03	0.42
81	7.75	2.92	17	0.59	0.50
84	3.63	1.90	29	1.00	0.00
95	3.87	5.80	7	0.24	0.51
97	22.18	10.65	28	0.97	0.82
99	25.79	4.73	34	1.17	0.38
114	2.68	3.71	6	0.21	0.41
116	11.93	7.79	24	0.83	0.66
123	7.38	5.17	5	0.17	0.38
143	13.40	1.63	8	0.28	0.45
146	21.39	7.58	8	0.28	0.53
147	44.31	5.67	51	1.76	0.44
150	7.77	7.66	4	0.14	0.35
152	26.85	7.69	18	0.62	0.56
155	46.49	18.96	40	1.38	0.86
156	21.70	9.90	25	0.86	0.69
159	46.86	5.12	39	1.34	0.77
163	11.55	6.30	3	0.10	0.31
167	24.96	8.47	11	0.38	0.49
171	4.34	1.33	1	0.03	0.19

Mean value and standard deviation of ΔASA and inter-residue contacts is given for each highly essential residue position (HERP). The inter-residue HLA-peptide contacts were estimated at a cutoff distance of 4 $\text{\AA}$ for each HERP.

Abbreviations: HERP = highly essential residue position; ASA = accessible surface area; SD = standard deviation.

Table 3.3 Relationship between HERPs and peptide residue positions.

HERP	Peptide Residue Position									Subtotal
	1	2	3	4	5	6	7	8	9	
5	19	0	0	0	0	0	0	0	0	19
7	30	30	1	0	0	0	0	0	0	61
9	0	26	1	0	1	0	1	0	0	29
24	0	3	0	0	0	0	0	0	0	3
45	0	20	0	0	0	0	0	0	0	20
59	13	0	0	0	0	0	0	0	0	13
62	5	2	0	4	0	0	0	0	0	11
63	29	29	0	0	0	0	0	0	0	58
65	0	0	0	3	0	1	0	0	0	4
66	15	25	23	24	1	3	0	0	0	91
67	0	8	0	0	0	0	0	0	0	8
69	0	0	0	1	1	5	0	0	0	7
70	0	3	22	2	7	13	2	0	0	49
72	0	0	0	0	0	0	0	3	0	3
73	0	0	0	0	3	22	23	19	0	67
74	0	0	0	0	3	0	2	0	2	7
76	0	0	0	0	0	0	0	18	0	18
77	0	0	0	0	0	0	3	30	30	63
80	0	0	0	0	0	0	0	3	28	31
81	0	0	0	0	0	0	0	0	18	18
84	0	0	0	0	0	0	0	0	30	30
95	0	0	0	0	0	0	1	0	6	7
97	0	0	8	0	3	5	12	0	2	30
99	0	3	30	0	1	0	0	0	0	34
114	0	0	3	0	0	0	3	0	0	6
116	0	0	0	0	0	0	6	0	20	26
123	0	0	0	0	0	0	0	0	5	5
143	0	0	0	0	0	0	0	0	8	8
146	0	0	0	0	0	0	1	7	0	8
147	0	0	0	0	0	0	26	29	0	55
150	0	0	0	0	0	0	4	0	0	4
152	0	0	0	0	2	1	16	0	0	19
155	0	0	14	2	23	1	4	0	0	44
156	0	0	21	0	4	2	0	0	0	27
159	24	0	16	0	1	0	0	0	0	41
163	3	0	0	0	0	0	0	0	0	3
167	11	0	0	0	0	0	0	0	0	11
171	1	0	0	0	0	0	0	0	0	1
SUM	150	149	139	36	50	53	104	109	149	939

The table shows the total number of inter-residue contacts between HERP and peptide residue position across 24 HLA_p complex structures. The inter-residue contacts were calculated using available HLA_p structures at a cutoff distance of 4 Å.

Abbreviation: HERP = highly essential residue positions.

Table 3.4 Mapping between peptide residue positions and HLA virtual pockets.

		Peptide Residue Position																	
		P1		P2		P3		P4		P5		P6		P7		P8		P9	
P		5	19	7	30	7	1	62	4	9	1	65	1	9	1	72	3	74	
		7	30	9	26	9	1	65	3	66	1	66	3	70	2	73	19	77	
		59	13	24	3	66	23	66	24	69	1	69	5	73	23	76	18	80	
		62	5	45	20	70	22	69	1	70	7	70	13	74	2	77	30	81	
		63	29	62	2	97	8	70	2	73	3	73	22	77	3	80	3	84	
R		66	15	63	29	99	30	155	2	74	3	97	5	95	1	146	7	95	
		159	24	66	25	114	3			97	3	152	1	97	12	147	29	97	
E		163	3	67	8	155	14			99	1	155	1	114	3			116	
		167	11	70	3	156	21			152	2	156	2	116	6			123	
H		171	1	99	3	159	16			155	23			146	1			143	
										156	4			147	26				
										159	1			150	4				
														152	16				
														155	4				
SUM		150		149		139		36		45		53		54		109			

The value in the cell right of each HERP indicates the total number of inter-residue contacts between a HERP and a peptide residue position across 24 HLA_p complex structures. The inter-residue contacts were calculated using available HLA_p structures at a cutoff distance of 4 Å.

Abbreviations: HERP = highly essential residue positions; P1 to P9 = residue positions in a 9-mer peptide.

3.1.5 Scoring Scheme

The scoring scheme is a relatively important step in HLA_p binding calculations. For a given HLA allele, the virtual binding pockets (1–9) can be defined using procedure described in step 3. The binding of a given peptide to the HLA allele has to be evaluated using a scoring metric system. The overall binding affinity (predicted) is the sum of compatibility measures for peptide residues at each peptide residue position and its corresponding virtual binding pocket. In this work, the compatibility of each peptide residue with the corresponding virtual pocket made of HERPs was quantified by the Q matrix. The Q matrix provides a measure to cluster the 20 amino acids based on 237 physicochemical properties (Venkatarajan *et al.* 2001). The overall ability of a particular peptide to bind to a particular HLA is a weighted sum of the amino acid preferences of the individual binding pockets. The binding affinity between the peptide and the HLA allele was estimated as the cumulative sum of their amino acid property compatibility values in the Q matrix. A positive predicted value for a peptide indicates a binder, and a negative value indicates a non-binder.

In this calculation, the effect of peptide residue at position 4 was neglected because the residue at this position is always solvent exposed in all the HLA_p crystal structures. This is demonstrated by the %Burial for each peptide residue position (**Figures 3.4**), which was calculated by the same procedure described in step 2. It is the relative ASA for each peptide residue: (ASA in the HLA_p complex state) / (ASA in the separate peptide state). The standard deviation caused by amino acids changes at this position is also minimal (**Figure 3.4c**). This suggests that property of the amino acid at this position does not significantly affect peptide binding.

Figure 3.4 Involvement of nonamer peptide residues in binding to class I HLA.

a. %Burial curves of 24 non-redundant nonamer peptides are shown. %Burial for each peptide residue position was calculated using crystal structures of class I HLA_p complexes by the same procedure described in step 2. It is the relative ASA for each peptide residue, which is calculated as: (ASA in the HLA_p complex state) / (ASA in the separate peptide state).

b. Mean values of %Burial on each peptide residue position among the 24 non-redundant nonamer peptides.

c. Standard Deviation (SD) of %Burial on each peptide residue position among the 24 non-redundant nonamer peptides.

3.1.6 Assessment of Model Performance

The performance of the model was assessed by calculating sensitivity (SE), specificity (SP), positive predictive value (PPV), negative predictive value (NPV), and accuracy (AC) using true-positive (TP), true-negative (TN), false-positive (FP), and false-negative (FN). These values were calculated as:

$$SE = \frac{TP}{TP + FN} \times 100$$

$$SP = \frac{TN}{TN + FP} \times 100$$

$$PPV = \frac{TP}{TP + FP} \times 100$$

$$NPV = \frac{TN}{TN + FN} \times 100$$

$$\%AC = \frac{TP + TN}{TP + FP + TN + FN} \times 100$$

The standard errors of specificity and sensitivity are calculated as:

$$\sigma_p = \sqrt{\frac{p(1-p)}{n}}$$

For sensitivity, p = sensitivity, n = number of binders. For specificity, p = specificity, n = number of non-binders.

3.1.7 Cross Validation

The predictive efficiency of this model was assessed using four datasets.

Dataset 1 (DS1) is the HLA_p structures collected from PDB (**Table 3.1**). The characteristic features of this dataset were described in step 1. This dataset contains sequence information represented by a non-redundant set of 24 HLA_p structures with bound nonamer peptides. There are 8 unique HLA alleles and 24 unique nonamer peptides in the dataset. The model, developed in steps 1–5, was used to predict the binding affinity of these nonamer peptides with the corresponding HLA alleles. The predicted values are presented in **Table 3.5**. All the peptides tested in this dataset are binders as their structure is solved with their corresponding HLA alleles. Using this information, the predicted peptides are identified as TP and FN. FP and TN are not available for this dataset because this dataset does not contain non-binders. Internal cross validation indicates that TP is 15 and FN is 9 for this dataset (**Table 3.5**). Therefore, the SE of the model is 63% for this dataset.

Dataset 2 (DS2) was obtained from literature (**Table 3.6**). Peptides in this dataset are a subset of peptides used in a previous investigation of our research group (Kanguene *et al.* 2000). This dataset contains only nonamer peptides and does not include octamer and decamer peptides that are used in the previous study. In this dataset, 102 binding data are present, of which 24 are tested for A*0201; 11 each tested for A*0202, A*0203, A*0206, A*6802; 6 each tested for A*0301, A*1101, A*3101, A*3301, A*6801; and 1 each tested for B*3501, B5101, B*5301, and B5401. The definition of binders and non-binders (on the basis of IC₅₀ cutoff) stated in the previous study

(Kangueane *et al.* 2000) was followed here. IC_{50} values are measured in an assay in which both the peptide of interest and a reference peptide compete for binding to HLA. The IC_{50} value of a peptide is the concentration at which it has suppressed 50% of the reference peptides from binding to HLA. Peptides with IC_{50} values of 50,000 nM or higher are always considered non-binding, since such values cannot be reliably measured. A commonly used IC_{50} cutoff of 500 nM was applied to classifying peptides into binders and non-binders: a peptide with an $IC_{50} \geq 500$ nM was defined as binder, and a peptide with an $IC_{50} < 500$ nM was defined as non-binder. The model, described in steps 1–4, was validated using this dataset and the results are presented in **Table 3.6** and summarized in **Table 3.7**. Validation reveals that TP is 39, TN is 18, FP is 9, and FN is 36 for this dataset. Therefore, the SE is 67%, SP is 52%, AC is 55%, PPV is 81%, and NPV is 33% for this dataset. For the same dataset the side-chain packing, and the *van der Waals* clashes (vdWC)/solvent exposed hydrophobic peptide residue (SEHP) scoring procedures were applied in the previous study. The SE is 70%, SP is 53%, AC is 66%, PPV is 80%, and NPV is 40% (recalculated for 9-mer peptides) for predictions based on the side-chain packing and vdWC/SEHPR scoring procedures (**Table 3.7**).

Dataset 3 (DS3) was also obtained from literature (**Table 3.8**). Peptides in this dataset were tested positive by tetramer staining in peripheral blood mononuclear cells (PBMC) obtained from infected patients. In this dataset 46 tetramer positive peptide data are present, of which 15 are tested for A*0201; 1 tested for A*6802; 14 tested for A2; 2 tested each for B*3501, B14, B27, and B53; and 1 tested each for A3, B*0702, B*0801, B*5801, B18, B35, B7, and Cw4. These peptides are also present in the MHCBN database (Bhasin *et al.* 2003). Qualitative information on binding (BI), relative binding (RB), CTL activity, and relative CTL activity (RCTL) are available for

most peptides. Unknown information are indicated by a “?” mark (this notation applies throughout this chapter). The model described in steps 1–5 was validated using this dataset and the results are presented in **Table 3.8** and summarized in **Table 3.9**. Validation indicates that TP is 35 and FN is 11 for this model. Therefore, the SE is 76% and AC is 76% for this dataset. This dataset does not contain tetramer negative peptides in PBMC, and TN and FP are not available for this dataset.

Dataset 4 (DS4) was retrieved from MHCBN (Bhasin *et al.* 2003). Peptides in this dataset are strictly 9 residues long. Qualitative binding information is available for these peptides in MHCBN. The dataset contains 1624 HLA-A specific peptides (1574 binders and 50 non-binders) and 861 HLA-B specific peptides (859 binders and 0 non-binders). The dataset is not listed here due to its large size. In this dataset, the HLA-A peptides are spread over 33 alleles and the HLA-B peptides are spread over 52 alleles. The model described in steps 1–5 was validated using this dataset and the results are summarized in **Table 3.10** and **Table 3.11**. Validation reveals that TP is 770/539, TN is 29/?, FP is 21/? and FN is 804/320 for HLA-A/HLA-B peptides, respectively. The SE is 49%/63%, SP is 58%/?, AC is 49%/63%, PPV is 97%/?, and NPV is 3%/? for HLA-A/HLA-B peptides from MHCBN, respectively. It should be noted that this dataset contains a predominant number of binders and a very small number of non-binders (only 50 non-binders).

Table 3.5 Performance of the model in dataset 1 (HLAp crystal structures).

Serial Number	PDB Code	HLA Allele	Peptide	Predicted Value	Inference	Sensitivity (%)
1	1JHT	A*0201	ALGIGILTV	-1439.95	FN	
2	1B0G	A*0201	ALWGFFPVL	-46.64	FN	
3	1I7U	A*0201	ALWGFPVPL	156.61	TP	
4	1I7T	A*0201	ALWGVFPVL	-140.4	FN	
5	1I7R	A*0201	FAPGFFPYL	-224.39	FN	
6	1I1F	A*0201	FLKEPVHGV	754.14	TP	
7	1HHI	A*0201	GILGFVFTL	685.15	TP	
8	1AKJ	A*0201	ILKEPVHGV	578.75	TP	
9	1HHJ	A*0201	ILKEPVHGV	578.75	TP	
10	1QRN	A*0201	LLFGYAVYV	1220.34	TP	
11	1QSE	A*0201	LLFGYPRYV	1012.92	TP	
12	1QSF	A*0201	LLFGYPVAV	775.34	TP	
13	1A07	A*0201	LLFGYPVYV	831.96	TP	
14	1HHG	A*0201	TLTSCNTSV	66.71	TP	
15	1I1Y	A*0201	YLKEPVHGV	1092.96	TP	
16	1HSA	B*2705	ARAAAAAAA	-361.18	FN	
17	1JGE	B*2705	GRFAAAIAK	-753.2	FN	
18	1K5N	B*2709	GRFAAAIAK	-683.8	FN	
19	1A9B	B*3501	LPPLDITPY	753.25	TP	
20	1E27	B*5101	LPPVVAKEI	1990.63	TP	
21	1A1M	B*5301	TPYDINQML	1150.85	TP	
22	1A1O	B*5301	KPIVQYDNF	225.88	TP	
23	1EFX	C*0304	GAVDPLLAL	-2708.77	FN	
24	1IM9	C*0401	QYDDAVYKL	-410.9	FN	
						63
Standard Error						9.85
95% CI						43.7–82.3

Abbreviations: TP = true-positive value; FN = false-negative value, CI = confidence interval.

Table 3.6 Performance of the model in dataset 2 (HLA binding affinity data [IC₅₀]).

SN	HLA	Peptide	Predicted	IC ₅₀	IF	Modeling Results				Reference
	Allele		Value			vdWC	SEHPR	Remark	IF	
1	A*0201	ALSTGLIHL	-1090.68	312.5	FN	0	0	G	TP	Alexander, 1998
2	A*0201	ALCRWGLLL	-457.59	165	FN	0	1	G	TP	Kawashima, 1998
3	A*0201	ILAGYGAGV	-167.45	116.3	FN	0	0	G	TP	Alexander, 1998
4	A*0201	DLMGYIPLV	-95.25	80	FN	0	1	G	TP	Alexander, 1998
5	A*0201	LLTFWNPPPT	-83.77	178.6	FN	1	1	G	TP	Kawashima, 1998
6	A*0201	VHRDDLLEA	-70.81	365	FN	4	0	P/NB	FN	Kawashima, 1998
7	A*0201	VLHDDLLEA	-67.64	30	FN	0	0	G	TP	Kawashima, 1998
8	A*0201	KIFGSLAFL	-64.74	33.3	FN	0	1	G	TP	Kawashima, 1998
9	A*0201	CLTSTVQLV	-17.97	147.1	FN	0	1	G	TP	Kawashima, 1998
10	A*0201	LLTFWNPPV	62.46	16	TP	5	2	P/NB	FN	Kawashima, 1998
11	A*0201	KVAELVHFL	66.11	68.4	TP	0	1	G	TP	Kawashima, 1998
12	A*0201	TLTSCNTSV	66.71	+	TP	0	0	G	TP	Garboczi, 1996
13	A*0201	LMTFWNPPV	94.78	4.5	TP	0	1	G	TP	Kawashima, 1998
14	A*0201	ILHNGAYSL	392.04	74.6	TP	0	0	G	TP	Kawashima, 1998
15	A*0201	YLSGANLNL	490.7	27.8	TP	0	0	G	TP	Kawashima, 1998
16	A*0201	LLFGYPVYP	546.52	+	TP	0	0	G	TP	Garboczi, 1996
17	A*0201	ILKEPVHGV	578.75	+	TP	0	0	G	TP	Gao, 1997
18	A*0201	GILGFVFTL	685.15	+	TP	0	1	G	TP	Madden, 1993
19	A*0201	FLLLADARV	785.92	18.2	TP	1	1	G	TP	Alexander, 1998
20	A*0201	HMWNFISGI	802.42	15.2	TP	2	1	P/NB	FN	Alexander, 1998
21	A*0201	LLFGYPVYV	831.96	+	TP	0	0	G	TP	Madden, 1993
22	A*0201	IMIGVLVGV	1022.36	68.5	TP	0	0	G	TP	Kawashima, 1998
23	A*0201	FLWGPRALV	1152.63	31.3	TP	0	1	G	TP	Kawashima, 1998
24	A*0201	YLVA YQATV	1554.11	20.4	TP	0	0	G	TP	Alexander, 1998
25	A*0202	KVAELVHFL	-266.54	29	FN	2	1	P/NB	FN	Kawashima, 1998
26	A*0202	CLTSTVQLV	-256.05	457	FN	3	1	P/NB	FN	Kawashima, 1998
27	A*0202	KIFGSLAFL	87.74	9	FN	2	1	P/NB	FN	Kawashima, 1998
28	A*0202	LLTFWNPPPT	124.29	1720	FP	0	1	G	FP	Kawashima, 1998
29	A*0202	LMTFWNPPV	270.97	779	FP	1	1	G	FP	Kawashima, 1998
30	A*0202	ALCRWGLLL	84.74	>10000	TN	0	1	G	FP	Kawashima, 1998
31	A*0202	LLTFWNPPV	238.65	297	TP	0	1	G	TP	Kawashima, 1998
32	A*0202	ILHNGAYSL	259.27	358	TP	7	0	P/NB	FN	Kawashima, 1998
33	A*0202	YLSGANLNL	467.9	165	TP	1	0	G	TP	Kawashima, 1998
34	A*0202	IMIGVLVGV	578.93	62	TP	1	0	G	TP	Kawashima, 1998
35	A*0202	FLWGPRALV	1443.05	43	TP	1	1	G	TP	Kawashima, 1998
36	A*0203	KVAELVHFL	-882.17	14	FN	2	1	P/NB	FN	Kawashima, 1998
37	A*0203	ALCRWGLLL	-655.4	278	FN	0	0	G	TP	Kawashima, 1998
38	A*0203	CLTSTVQLV	-385.35	6.7	FN	1	0	G	TP	Kawashima, 1998
39	A*0203	KIFGSLAFL	-347.79	9	FN	11	2	P/NB	FN	Kawashima, 1998
40	A*0203	IMIGVLVGV	-245.8	13	FN	1	0	G	TP	Kawashima, 1998
41	A*0203	ILHNGAYSL	110.75	100	TP	5	0	P/NB	FN	Kawashima, 1998
42	A*0203	LLTFWNPPPT	273.6	67	TP	5	1	P/NB	FN	Kawashima, 1998
43	A*0203	YLSGANLNL	295.31	2.4	TP	3	0	P/NB	FN	Kawashima, 1998
44	A*0203	LLTFWNPPV	419.83	26	TP	1	1	G	TP	Kawashima, 1998
45	A*0203	LMTFWNPPV	452.15	7.6	TP	3	0	P/NB	FN	Kawashima, 1998
46	A*0203	FLWGPRALV	1784.82	14	TP	15	1	P/NB	FN	Kawashima, 1998
47	A*0206	KIFGSLAFL	-386.24	23	FN	0	1	G	TP	Kawashima, 1998
48	A*0206	KVAELVHFL	-280.73	168	FN	0	1	G	TP	Kawashima, 1998
49	A*0206	CLTSTVQLV	-37.7	308	FN	1	1	G	TP	Kawashima, 1998
50	A*0206	LLTFWNPPV	-1.04	56	FN	0	1	G	TP	Kawashima, 1998
51	A*0206	YLSGANLNL	160.02	804	FP	1	0	G	FP	Kawashima, 1998
52	A*0206	ILHNGAYSL	373.76	567	FP	0	0	G	FP	Kawashima, 1998

CHAPTER 3 HLA PEPTIDE BINDING PREDICTION MODEL

53	A*0206	ALCRWGLLL	-843.22	8863	TN	2	1	P/NB	TN	Kawashima, 1998
54	A*0206	LLTFWNPPPT	-147.27	755	TN	1	1	G	FP	Kawashima, 1998
55	A*0206	LMTFWNPPV	53.23	33	TP	0	1	G	TP	Kawashima, 1998
56	A*0206	IMIGVLVGV	565.73	106	TP	0	0	G	TP	Kawashima, 1998
57	A*0206	FLWGPRALV	955.52	336	TP	15	1	P/NB	FN	Kawashima, 1998
58	A*0301	VAGALVAFK	-1180.4	45.8	FN	0	1	G	TP	Chang, 1999
59	A*0301	LGFGAYMSK	-570.56	135.8	FN	3	0	P/NB	FN	Chang, 1999
60	A*0301	RLGVRATRK	-270.22	11.7	FN	1	0	G	TP	Chang, 1999
61	A*0301	KTSEERSQPR	-208.66	68.8	FN	0	0	G	TP	Chang, 1999
62	A*0301	QLFTFSPRR	715.52	14.7	TP	0	1	G	TP	Chang, 1999
63	A*0301	LIFCHSKKK	1293.56	20.4	TP	0	0	G	TP	Chang, 1999
64	A*1101	VAGALVAFK	-1678.12	6.7	FN	0	1	G	TP	Chang, 1999
65	A*1101	LGFGAYMSK	-1192.45	20.7	FN	0	0	G	TP	Chang, 1999
66	A*1101	QLFTFSPRR	146.67	182	TP	0	0	G	TP	Chang, 1999
67	A*1101	RLGVRATRK	199.22	207	TP	0	0	G	TP	Chang, 1999
68	A*1101	KTSEERSQPR	456.85	93.8	TP	0	0	G	TP	Chang, 1999
69	A*1101	LIFCHSKKK	747.33	1.6	TP	3	0	P/NB	FN	Chang, 1999
70	A*3101	KTSEERSQPR	-798.98	66.7	FN	0	0	G	TP	Chang, 1999
71	A*3101	RLGVRATRK	-533.25	428.6	FN	0	1	G	TP	Chang, 1999
72	A*3101	LIFCHSKKK	14.02	2535.6	FP	2	0	P/NB	TN	Chang, 1999
73	A*3101	VAGALVAFK	195.19	3750	FP	1	0	G	FP	Chang, 1999
74	A*3101	LGFGAYMSK	-266.01	2950.8	TN	10	0	P/NB	TN	Chang, 1999
75	A*3101	QLFTFSPRR	-61.93	620.7	TN	0	1	G	FP	Chang, 1999
76	A*3301	KTSEERSQPR	-659.65	1812.5	TN	0	0	G	FP	Chang, 1999
77	A*3301	RLGVRATRK	-418.56	NB	TN	1	0	G	FP	Chang, 1999
78	A*3301	LGFGAYMSK	-315.45	NB	TN	3	0	P/NB	TN	Chang, 1999
79	A*3301	LIFCHSKKK	-161.54	NB	TN	14	0	P/NB	TN	Chang, 1999
80	A*3301	QLFTFSPRR	-140.78	3766.2	TN	0	0	G	FP	Chang, 1999
81	A*3301	VAGALVAFK	-69.22	NB	TN	5	1	P/NB	TN	Chang, 1999
82	A*6801	VAGALVAFK	-1537.89	258.1	FN	2	1	P/NB	FN	Chang, 1999
83	A*6801	LGFGAYMSK	-576.33	222.2	FN	4	1	P/NB	FN	Chang, 1999
84	A*6801	KTSEERSQPR	-250.71	145.5	FN	1	0	G	TP	Chang, 1999
85	A*6801	RLGVRATRK	-506.61	NB	TN	4	0	P/NB	TN	Chang, 1999
86	A*6801	QLFTFSPRR	277.71	2.6	TP	1	0	G	TP	Chang, 1999
87	A*6801	LIFCHSKKK	723.53	333.3	TP	0	0	G	TP	Chang, 1999
88	A*6802	KVAELVHFL	-1379.03	17	FN	0	1	G	TP	Kawashima, 1998
89	A*6802	IMIGVLVGV	-500.95	89	FN	2	1	P/NB	TN	Kawashima, 1998
90	A*6802	LMTFWNPPV	1.52	3448	FP	3	0	P/NB	TN	Kawashima, 1998
91	A*6802	YLSGANLNL	64.77	>10000	FP	2	2	P/NB	TN	Kawashima, 1998
92	A*6802	ILHNGAYSL	98.77	>10000	FP	0	1	G	FP	Kawashima, 1998
93	A*6802	ALCRWGLLL	-1334.68	>10000	TN	2	1	P/NB	TN	Kawashima, 1998
94	A*6802	KIFGSLAFL	-654.82	3333	TN	3	2	P/NB	TN	Kawashima, 1998
95	A*6802	CLTSTVQLV	-205.81	8000	TN	2	0	P/NB	TN	Kawashima, 1998
96	A*6802	LLTFWNPPPT	-139.57	>10000	TN	1	0	G	FP	Kawashima, 1998
97	A*6802	LMTFWNPPV	-131.91	9442	TN	3	0	P/NB	TN	Kawashima, 1998
98	A*6802	FLWGPRALV	1026.62	40	TP	18	1	P/NB	FN	Kawashima, 1998
99	B*3501	LPGCSFSIF	-643.83	90.4	FN	2	1	P/NB	FN	Chang, 1999
100	B*5101	LPGCSFSIF	172.98	100	TP	2	1	P/NB	FN	Chang, 1999
101	B*5301	LPGCSFSIF	-109.41	114	FN	0	1	G	TP	Chang, 1999
102	B*5401	LPGCSFSIF	-577.51	6666	TN	4	1	P/NB	TN	Chang, 1999

Abbreviations: SN = Serial Number; IF = inference; FN = false negative; FP = false positive; G = good binder; P/NB = poor/non-binder; SEHPR = solvent exposed hydrophobic peptide residue; vdWC = *van der Waals* clash.

Table 3.7 Summary of model performance in dataset 2 (HLA binding affinity data [IC₅₀]).

SN	HLA Allele	Number of Peptides	TP	TN	FP	FN	SP (%)	SE (%)	AC
1	A*0201	24	15	0	0	9	–	63	63
2	A*0202	11	3	1	2	3	33	63	55
3	A*0203	11	6	0	0	5	–	55	55
4	A*0206	11	2	2	2	4	50	43	45
5	A*0301	6	2	0	0	4	–	33	33
6	A*1101	6	3	0	0	2	–	67	67
7	A*3101	6	0	2	2	2	50	–	33
8	A*3301	6	0	6	0	0	100	–	100
9	A*6801	6	1	1	0	3	100	40	50
10	A*6802	11	1	5	3	2	63	33	55
11	B*3501	1	0	0	0	1	–	–	–
12	B*5101	1	1	0	0	0	–	100	–
13	B*5301	1	0	0	0	1	–	–	–
14	B*5401	1	0	1	0	0	100	–	100
Total		102	39	18	9	36	67	52	55
Standard Error							9.05	5.76	4.93
95% CI							49.33– 84.73	40.71– 63.29	45.35– 64.65

Abbreviations: SN = serial number; TP = true positive; TN = true negative; FP = false positive; FN = false negative; SP = specificity; SE = sensitivity; AC = accuracy; CI = confidence interval.

Table 3.8 Performance of the model in dataset 3 (peptides tested positive by tetramer staining).

SN	HLA Allele	Peptide	Predicted Value	IF	MHCBN Data				Pathogen	Reference
					BI	RB	CTL	RCTL		
1	A*0201	ALYDVVTKL	749.08	TP	Y	H	Y	?	HCV	Grabowska, 2001
2	A*0201	CINGVCWTV	654.95	TP	Y	H	Y	?	HCV	Grabowska, 2001
3	A*0201	FMVFLQTHI	1190.78	TP	Y	M	N	?	EBV	Catalina, 2001
4	A*0201	ILAGYGAGV	-167.45	FN	Y	H	?	?	HCV	Prezzi, 2001
5	A*0201	MLNIPSINV	141.19	TP	Y	M	?	?	HCMV	Aubert, 2001
6	A*0201	RLNMFPTYI	220.71	TP	?	?	?	?	Ct	Kim, 2000
7	A*0201	SLYNTVATL	415.61	TP	Y	H	?	?	HIV	Shankar, 2000
8	A*0201	VLGPISGHV	-741.6	FN	Y	?	?	?	HCMV	Aubert, 2001
9	A*0201	WLSLLVPFV	167.27	TP	?	?	?	?	HBV	Reignat, 2002
10	A*0201	YLKKIKNSI	1978.02	TP	?	?	?	?	Pf	Bonelo, 2000
11	A*0201	YLKKIKNSL	1801.47	TP	?	?	?	?	Pf	Bonelo, 2000
12	A*0201	YLKKIQNSL	1526.57	TP	?	?	?	?	Pf	Bonelo, 2000
13	A*0201	YLKTIKNSL	1801.47	TP	?	?	?	?	Pf	Bonelo, 2000
14	A*0201	YLKTIQNSL	1526.57	TP	Y	?	Y	?	Pf	Bonelo, 2000
15	A*0201	YLNKIQNSL	1123.41	TP	Y	?	Y	?	Pf	Bonelo, 2000
16	A*0201	YLNTIQNSL	1123.41	TP	Y	?	Y	H	Pf	Bonelo, 2000
17	A*0201	YLQKIKNSL	1654.11	TP	Y	H	?	?	Pf	Bonelo, 2000
18	A*0201	YLQKIQNSL	1379.21	TP	Y	?	Y	?	Pf	Bonelo, 2000
19	A*0201	YLQKIRNSL	1705.06	TP	Y	?	Y	?	Pf	Bonelo, 2000
20	A*0202	ILKEPVHGV	578.75	TP	Y	?	Y	?	HIV	Kostense, 2001
21	A*0202	NLVPMVATV	397.35	TP	Y	?	Y	?	HCMV	Faint, 2001
22	A*0202	NMFPTYIGV	186.48	TP	?	?	?	?	Ct	Kim, 2000
23	A*0202	VLQWASLAV	308.12	TP	Y	H	Y	?	EBV	Catalina, 2001
24	A*0203	SLVIVTTFV	178.68	TP	Y	?	Y	?	EBV	Catalina, 2001
25	A*0203	YTAFTIPSI	-276.38	FN	Y	?	Y	?	HIV	Shankar, 2000
26	A*0204	LMIIPLINV	401.59	TP	Y	?	Y	?	EBV	Catalina, 2001
27	A*0205	TLFIGSHV	295.08	TP	Y	H	Y	?	EBV	Catalina, 2001
28	A*0206	YVLDHLIVV	1343.23	TP	Y	H	Y	?	EBV	Faint, 2001
29	A*0207	GLCTLVAML	-275.63	FN	Y	M	?	?	EBV	Hislop, 2002
30	A*0301	RVRAYTYSK	244.48	TP	Y	H	?	?	EBV	Catalina, 2001
31	A*6802	ETAYFILKL	-1566.39	FN	?	?	?	?	HIV	Appay, 2000
32	B*0701	LPCVLWVPL	-1272.42	FN	?	?	?	?	EBV	Catalina, 2001
33	B*0702	TPRVTGGGA	335.27	TP	?	?	?	?	HCMV	Hassan-Walker, 2001
34	B*0801	DANDIYRIF	-1261.3	FN	Y	H	Y	H	HCMV	Hassan-Walker, 2001
35	B*1401	DLNMLNIV	-1101.96	FN	Y	?	Y	H	HIV	Kaul, 2002
36	B*1401	ERYLRDQQL	1574.94	TP	?	?	?	?	HIV	Kaul, 2002
37	B*1801	YPLTFGWCF	496.22	TP	Y	?	Y	?	HIV	Kaul, 2002
38	B*2701	YRLLLTRVL	1376.78	TP	Y	?	Y	H	Ct	Kuon, 2001
39	B*2702	NRELIQQEL	960.49	TP	?	?	?	?	Ct	Kuon, 2001
40	B*3501	DPNPQEVVL	-1735.42	FN	Y	?	Y	?	HIV	Ogg, 1998
41	B*3501	HPNIEEVAL	-1015.99	FN	Y	?	Y	?	HCV	Sobao, 2001
42	B*3501	IPSINVHHY	906.54	TP	Y	?	Y	?	HCMV	Hassan-Walker, 2001
43	B*5301	TPQDLNMLL	-458.2	FN	Y	?	Y	?	HIV	Dorrell, 2001
44	B*5302	TPYDINQML	1150.85	TP	Y	?	Y	?	HIV	Dorell, 2001
45	B*5801	IVLPEKDSW	288.89	TP	?	?	?	?	HIV	Kostense, 2001
46	Cw*0401	KYRLKHLVW	2039.97	TP	Y	?	Y	?	HIV	Appay, 2000

Abbreviations: SN = serial number; IN = inference; TP = true positive; FN = false negative; BI = binding; RB = relative binding; CTL = cytotoxic T lymphocyte; RCTL = relative CTL activity.

Table 3.9 Summary of model performance in dataset 3 (peptides tested positive by tetramer staining).

SN	HLA Allele	Number of Peptides	TP	TN	FP	FN	SE (%)	AC (%)
1	A*0201	19	17	–	–	2	89	89
2	A*0202	4	4	–	–	0	100	100
3	A*0203	2	1	–	–	1	50	50
4	A*0204	1	1	–	–	0	100	100
5	A*0205	1	1	–	–	0	100	100
6	A*0206	1	1	–	–	0	100	100
7	A*0207	1	0	–	–	1	–	–
8	A*0301	1	1	–	–	0	100	100
9	A*6802	1	0	–	–	1	–	–
10	B*0701	1	0	–	–	1	–	–
11	B*0702	1	1	–	–	0	100	100
12	B*0801	1	0	–	–	1	–	–
13	B*1401	2	1	–	–	1	50	50
14	B*1801	1	1	–	–	0	100	100
15	B*2701	1	1	–	–	0	100	100
16	B*2702	1	1	–	–	0	100	100
17	B*3501	3	1	–	–	2	33	33
18	B*5301	1	0	–	–	1	–	–
19	B*5302	1	1	–	–	0	100	100
20	B*5801	1	1	–	–	0	100	100
21	Cw*0401	1	1	–	–	0	100	100
Total		46	35	–	–	11	76	76
Standard Error							6.29	6.39
95% CI							63.67– 88.33	63.67– 88.33

Abbreviations: SN = serial number; TP = true positive; TN = true negative; FP = false positive; FN = false negative; SE = sensitivity; AC = accuracy; CI = confidence interval.

Table 3.10 Summary of model performance in dataset 4 (HLA-A binding peptides from MHCBN).

SN	HLA Allele	Number of Peptides	TP	TN	FP	FN	SP (%)	SE (%)	AC (%)
1	A*0101	9	4	0	0	5	–	44	44
2	A*0201	794	424	18	11	341	62	55	56
3	A*0202	78	35	1	2	40	33	47	46
4	A*0303	65	21	0	0	44	–	32	32
5	A*0204	38	20	0	0	18	–	53	53
6	A*0205	23	5	0	0	18	–	22	22
7	A*0206	83	50	1	0	32	100	61	61
8	A*0207	19	8	0	0	11	–	42	42
9	A*0209	5	2	0	0	3	–	40	40
10	A*0210	3	1	0	0	2	–	33	33
11	A*0211	4	2	0	0	2	–	50	50
12	A*0214	8	8	0	0	0	–	100	100
13	A*0217	1	1	0	0	0	–	100	100
14	A*0301	66	34	0	0	32	–	52	52
15	A*0302	2	1	0	0	1	–	50	50
16	A*1101	77	13	0	0	64	–	17	17
17	A*2402	118	83	0	1	34	0	71	70
18	A*2601	4	3	0	0	1	–	75	75
19	A*2602	2	1	0	0	1	–	50	50
20	A*2603	2	0	0	0	2	–	0	–
21	A*2902	20	5	0	0	15	–	25	25
22	A*3001	1	1	0	0	0	–	100	100
23	A*3002	10	7	0	0	3	–	70	70
24	A*3003	1	1	0	0	0	–	100	100
25	A*3101	19	3	0	0	16	–	16	16
26	A*3201	1	1	0	0	0	–	100	100
27	A*3301	33	9	0	0	24	–	27	27
28	A*3402	1	0	0	0	1	–	0	–
29	A*6601	10	4	0	0	6	–	40	40
30	A*6801	63	13	0	0	50	–	21	21
31	A*6802	60	10	9	7	34	56	23	32
32	A*6901	3	0	0	0	3	–	0	–
33	A*7401	1	0	0	0	1	–	0	–
Total		1624	770	29	21	804	58	49	49
Standard Error							6.98	1.26	1.24
95% CI							44.32	46.53	46.57
							–71.68	–51.47	–51.43

Abbreviations: SN = serial number; TP = true positive; TN = true negative; FP = false positive; FN = false negative; SP = specificity; SE = sensitivity; AC = accuracy; CI = confidence interval.

Table 3.11 Summary of model performance in dataset 4 (HLA-B binding peptides from MHCBN).

SN	HLA Allele	Number of Peptides	TP	TN	FP	FN	SE (%)	AC (%)
1	B*0702	72	32	0	0	40	44	44
2	B*0801	15	12	0	0	3	80	80
3	B*0802	4	2	0	0	2	50	50
4	B*1501	11	7	0	0	4	64	64
5	B*1502	4	2	0	0	2	50	50
6	B*1503	3	2	0	0	1	67	67
7	B*1508	2	1	0	0	1	50	50
8	B*1509	9	3	0	0	6	33	33
9	B*1510	9	3	0	0	6	33	33
10	B*1512	1	0	0	0	1	0	—
11	B*1513	6	4	0	0	2	67	67
12	B*1516	7	4	0	0	3	57	57
13	B*1517	4	2	0	0	2	50	50
14	B*1518	4	1	0	0	3	25	25
15	B*2701	6	2	0	0	4	33	33
16	B*2702	13	5	0	0	8	38	38
17	B*2703	32	19	0	0	13	59	59
18	B*2704	17	12	0	0	5	71	71
20	B*2705	108	77	0	0	31	71	71
21	B*2706	23	16	0	0	7	70	70
22	B*2707	4	2	0	0	2	50	50
23	B*2709	4	2	0	0	2	50	50
24	B*3501	155	91	0	0	64	59	59
25	B*3503	1	1	0	0	0	100	100
26	B*3701	4	2	0	0	2	50	50
27	B*3801	6	4	0	0	2	67	67
28	B*3901	8	8	0	0	0	100	100
29	B*3909	3	1	0	0	2	33	33
30	B*4001	8	6	0	0	2	75	75
31	B*4002	16	6	0	0	10	38	38
32	B*4006	4	3	0	0	1	75	75
33	B*4201	1	1	0	0	0	100	100
34	B*4402	8	3	0	0	5	38	38
35	B*4403	7	1	0	0	6	14	14
36	B*4501	1	0	0	0	1	0	—
37	B*4601	5	3	0	0	2	60	60
38	B*5101	48	38	0	0	10	79	79
39	B*5102	32	26	0	0	6	81	81
40	B*5103	30	25	0	0	5	83	83
41	B*5201	6	5	0	0	1	83	83
42	B*5301	59	42	0	0	17	71	71
43	B*5401	57	41	0	0	16	72	72
44	B*5501	2	0	0	0	2	0	—

CHAPTER 3 HLA PEPTIDE BINDING PREDICTION MODEL

45	B*5502	5	1	0	0	4	20	20
46	B*5601	3	0	0	0	3	0	—
47	B*5701	10	6	0	0	4	60	60
48	B*5702	4	3	0	0	1	75	75
49	B*5801	11	7	0	0	4	64	64
50	B*5802	3	3	0	0	0	100	100
51	B*7301	2	1	0	0	1	50	50
52	B*8101	2	1	0	0	1	50	50
Total		859	539	0	0	320	63	63
Standard Error							2.71	2.71
95% CI							57.69 -68.31	57.69 -68.31

Abbreviations: SN = serial number; TP = true positive; TN = true negative; FP = false positive; FN = false negative; SE = sensitivity; AC = accuracy; CI = confidence interval.

3.2 Results and Discussion

The critical issue in the design of T-cell epitope vaccines is the identification of peptides able to bind all or most known class I or class II HLA molecules in certain ethnic population. Experimental identification of such promiscuous peptides is economically not feasible. For example, a 200 residue long surface antigen can produce 192 possible overlapping 9 residues long short peptides. Testing the binding ability of these short peptides to more than 2,000 HLA molecules requires cloning, expression, and purification of these molecules followed by binding assays on a prohibitory large scale. A reasonably validated mathematical model for the prediction of these peptides binding to multiple HLA molecules is useful and many such mathematical models are currently available. As discussed earlier, each of the currently available methods has its own merits and demerits.

In this work, a novel predictive model was developed that estimates the binding affinity of peptides to HLA molecules as a sum of compatibility measures between each residue at every peptide position to the corresponding residues forming the virtual binding pocket in the receptor molecule. The measure of compatibility between the peptide residues and the virtual binding pockets was determined using a high dimensional amino acid property matrix called the Q matrix (Mathura *et al.* 2001). Other salient features of the model discussed below include: (1) identification of HERPs from 24 HLA α structures, (2) mapping of individual peptide residue positions to virtual binding pockets, (3) exclusion of the fourth residue in the peptide for HLA α binding calculations, (4) generic nature of the model and its application to any allele whose sequence is clearly defined.

The HERPs were identified using 24 HLAp structures (**Table 3.1**). The dataset contains at least one complex from A*0201, B*2705, B*2709, B*3501, B*5101, B*5301, Cw*0304, and Cw*0401. The dataset also contains 24 unique 9 residue long peptides (**Table 3.1**). Although this dataset is not exhaustive, it is reasonably representative for this analysis. The definition of HERPs can be refined as and when more unique HLAp structures are available in the PDB. It is generally agreed that homologous proteins share a common fold, even when the overall sequence identity is less than 10% (Abagyan *et al.* 1997). HLA alleles share at least 70% sequence identity within a class (Schueler-Furman *et al.* 1998). Therefore, all the HLA molecules will have identical structural folds and binding grooves. HLA alleles share a very high homology as reflected from the sequence identity measure (greater than 70%). Studies show that proteins with such high percentage sequence similarity have negligible difference in their protein structures (3D backbone trace). The HERP defined in this work was based on HLAp structures in the PDB and the HERP's determinants will not be significantly different for homologous structures (other HLA alleles). However, this could be tested and validated as and when a large number of unique HLAp structures are available in the PDB in the future.

The functional difference between HLA alleles is caused by the polymorphism inside the peptide binding groove. The amino acid residues that form the pocket determine the geometry and chemical properties of the pocket and thus determine the antigen peptides that would be preferentially bound by the HLA allotype. This relationship accounts for the differential ability of different alleles to bind a variety of peptides. Although the side chain orientations of these residues are expected to differ between HLA molecules, the backbone traces of such residues in space are largely similar

(Schueler-Furman *et al.* 1998). It is shown that HLA side-chain atoms have a significant role in HLAp binding than HLA backbone atoms (Adrian *et al.* 2002). Despite the conservation of backbone traces among HLA alleles in the binding groove, different HLA alleles bind the same peptide with different affinity because of the variation in the physicochemical contour (properties of side-chain atoms) caused by the subtle polymorphism in the groove, which is essentially an shuffling of virtual pockets between different allotypes. The HERPs identified in this study provide a methodology to define virtual pockets and localize functions in these pockets.

The 38 HERPs reveal a mean change in ASA of at least $1 \text{ \AA}^2$ during HLAp complex formation and/or a minimum of 1 inter-residue contact between HLA and peptide at a cutoff distance of $4 \text{ \AA}$ (**Table 3.2**). A caveat in this approach is that all residue positions that exhibit a mean change in accessibility of greater than $0 \text{ \AA}^2$ and less than $1 \text{ \AA}^2$ are neglected. It is believed that these residue positions may not significantly affect the HLAp binding calculations due to their negligible contact with the peptide (**Figure 3.3a**). It should also be noted that the standard deviations of ΔASA are large at the HERPs, which reflects the polymorphism of the HLA molecules (**Figure 3.3c**). This observation proves the significance of the HERPs in terms of contribution to function difference, i.e. difference of peptide binding properties, of HLA alleles. Essentially, the conserved main-chain of HLA molecules provide a fixed framework, based on which, exchange of different amino acid side-chains among different HLA allotypes provides the variance in HLA function - peptide binding properties.

The fourth residue position in the peptide shows a low %burial value (**Figures 3.4a and 3.4b**). This means that its involvement in binding is minimum compared to other

peptide positions, irrespective of the amino acid type at that position. Thus, it was neglected in the HLA_p binding calculations. The standard deviation of the Δ ASA at each peptide position is illustrated in **Figure 3.4c**. Interestingly, the standard deviations at peptide residue position 5, 6, and 7 are large. This observation suggests that these positions are more important to the function variance of HLA allotypes.

The next challenging step was to map the HERPs to the corresponding peptide residue positions so as to define peptide-residue-position-specific virtual pockets in the HLA. Inter-residue contacts were used to map the peptide residue positions with the virtual binding pockets of the HLA molecules (**Table 3.3** and **Table 3.4**). For example, the first peptide residue position has at least one inter-residue contact with positions 5, 7, 59, 62, 63, 66, 159, 163, 167, and 171 of the HLA in the dataset of HLA_p structures (**Table 3.4**). These HERPs form the virtual pockets for binding to the first peptide residue. It should be noted that some of these HERPs are also shared with other neighboring virtual pockets. Thus, nine virtual pockets corresponding to each peptide residue positions were defined for class I HLA molecules. Due to high sequence homology among HLA alleles, the HERPs were extrapolated to other alleles and virtual binding pockets were defined for each known allele.

The peptide-residue-position-specific virtual binding pockets were defined to each known allele. After this mapping was established, it was equally important to estimate the binding affinity of a 9-mer peptide with these virtual binding pockets. In this model, the binding affinity was represented as a sum of compatibility measure of each peptide residue with its corresponding virtual pocket. The compatibility was calculated using a property matrix called the Q matrix (Mathura *et al.* 2001). The Q matrix is a high

dimensional 20×20 property matrix developed using 237 physicochemical properties of amino acids. When nonamer peptides bind to HLA molecules, each residue in the peptide interacts with a virtual pocket. The overall binding between peptide and HLA is a cumulative measure of the physical and chemical compatibility between each residue in the peptide and the residues forming the virtual pockets. The predicted binding score is the sum of compatibility scores of residues at each peptide residue positions and the corresponding virtual pockets. In this model, the peptides with negative scores are considered to be non-binders and those with positive scores as binders.

The efficiency of this model was evaluated using four different binding datasets, namely DS1, DS2, DS3, and DS4. The characteristic features of the datasets have been described earlier. A large number of manually collected peptides with known binding/non-binding-related information are included in DS2 and DS4. DS4 is the largest repositories of HLAp binding data (developed from several primary repositories and disparate sources) available in the public domain. It should be noted that the data in DS4 is less homogenous and the information noise (data quality, consistency, experimental procedure to get data, annotation) is quantifiably high. However, these datasets (DS1–DS4) provide the basis to validate the proposed model in a fair and transparent manner. DS1 contains 24 HLAp structures and the model performed well with sensitivity (SE) of 63% in this dataset (**Table 3.5**). The 9 FNs observed in this validation is probably introduced by the noise in the Q matrix used for scoring as Q matrix is a similarity matrix. Improvements in the Q matrix by fine tuning the interacting pairs will remove false negatives and improve sensitivity significantly. Future work is required to reevaluate and accordingly modify the Q matrix for further

developments.

DS2 was used in a previous study of the application of 3D side-chain packing procedures for HLA_A binding predictions (Kangueane *et al.* 2000). Encouraging results were obtained by the study and the main drawback is the cumbersome modeling procedure involving energy calculations. The same dataset was used to evaluate the performance of the new predictive model. DS2 contains 102 peptides with known HLA_A binding information (mostly from competitive binding assay) (**Table 3.6**). The model produced an overall sensitivity (SE) of 67% and specificity (SP) of 52% with an average accuracy (AC) of 55% in the dataset (**Table 3.7**). The PPV is 81% and NPV is 33%. These results are comparable to those obtained from the side-chain packing procedures (only for 9-mer peptides) (Kangueane *et al.* 2000). The PPV in the new model is 81% and the NPV is only 33%. The low NPV is because of high FN (36) and low TN (18) produced by the model. The high FN is introduced by the noise in the scoring Q matrix. The low TN is because of the nature of DS2 which contains only 27 non-binders (only 26%). NPV can be improved by the reevaluation of the scoring matrix and incorporation of more non-binders in the validation set. It should be noted that the amount of known non-binders for a given allele is significantly low as knowledge on non-binders is of less importance in inferring immunologically relevant phenomena. However, nonbinding information is critical in the development and assessment of mathematical models. Despite the limitations stated, this model is superior to the previous procedure because of easy implementation.

It is known that HLA_A binding does not fully guarantee immunogenicity (Feltkamp *et al.* 1994). In order to establish a better understanding of HLA_A binding and immune

functions, the model was assessed using 46 nonamer peptides that were documented as positive to tetramer staining in PBMC. These peptides are positive to tetramer staining in PBMC obtained from peptide specific infected patients (DS3) (**Table 3.8**). It should be noted that this does not contain peptides negative to tetramer staining. The model performed with a sensitivity and accuracy of 76% in DS3 (**Table 3.9**). The PPV cannot be calculated as DS3 lacks non-binders or peptides negative to tetramer staining. FN in this dataset is 11 and, as discussed in the previous section, the FN is because of noise introduced in the scoring matrix. It should be stated that this does not provide any additional information other than they are tetramer positive and the performance of this new model in this dataset is good (SE = 76%).

The model was also evaluated using a relatively large number of peptides (known binding information) described in DS4. DS4 contains 1624 HLA-A binding peptides (**Table 3.10**) and 859 HLA-B binding peptides (**Table 3.11**) collected from MHCBN (Bhasin *et al.* 2003). The SE of the model is 49% and 63% for HLA-A and HLA-B binding peptides, respectively. A specificity of 58% is obtained for HLA-A binding peptides. The PPV is 97% for HLA-A binding peptides. The FN is 804 in this dataset and this large number of FN is because of noise in the scoring matrix. As in other datasets, the number of non-binders in DS4 is very low (only 50 HLA-A specific non-binders). The number of non-binders is negligible compared to the binders (1574 HLA-A peptides and 859 HLA-B peptides) in DS4. Therefore, the NPV calculated for DS4 (3%) is not a real measure of the model's negative predictive value. The average specificity for HLA-A peptides in DS4 is 58%. The model produced an average accuracy of 60% in all the four datasets used for validation. The average PPV is 89% for the model with good sensitivity (50%–73%) and specificity (52%–58%) values.

It has been shown that this model produced only moderate predictive accuracy at current stage. But the approach is novel, simple, and superior to other models because of its potential application to any known class I HLA molecule whose sequence is clearly defined. The predictive power of the described model can be improved in future by the careful re-examination of HERPs for the definition of virtual pockets.

3.3 Extension of the Procedure to Class II HLA

It is a more complicated problem to apply the procedure for the development of peptide binding prediction model to class II HLA. Unlike class I HLA molecules, class II HLA molecules have a relatively open peptide binding groove. The two ends of the binding grooves of class II HLA molecules do not form hydrogen network with peptide termini as strong as that in class I HLA molecules. Thus the peptides bound to class II HLA molecules have a larger variability in their length. The length of peptides bound to class I HLA range from 8 to 10 residues in length, while the length of peptides bound to class II HLA range from 11 to 16. This makes the definition of virtual binding pocket more difficult. As discussed earlier, the virtual pockets are defined based on fixed spatial relationship between HLA residues and peptide residues that directly interact. In class II HLA peptide complexes, due to the large variability of peptide length, it is difficult to find such fixed spatial relationships.

Moreover, the model describe earlier was developed only for prediction of nonamer peptides. This is because that there are only sufficient structure data available for HLA - nonamer peptide complexes. These structure data have a relatively large HLA allele coverage. Thus the information obtained from these data is less biased. In case of class

II HLA structures, no peptide of certain length is predominant. The available crystal structures for class II HLA peptide complexes are listed in **Table 3.12**. There are 22 structures available, of which 20 are non-redundant. Totally 6 class II HLA alleles are covered. The lengths of the bound peptides are evenly distributed. Thus, not enough structural information can be extracted for peptides of certain length.

By analyzing the binding properties of the peptides of different lengths, it is found that there exists a common binding pattern. The %burial of the peptide residue upon complex formation was calculated by the procedure described earlier. The %burial roughly represents the extent of the involvement of the peptide residues in the binding to HLA molecules. It is found that the %burial curves of these peptides are quite similar, when the length variability is ignored. Thus, these %burial curves were aligned by the similarity of their profile (**Figure 3.5**). These aligned curves demonstrate that the anchor residues of the peptides have a fixed relative position, although the length of the peptide varied. Thus, the binding pockets interacting with these anchor residues should also have a fixed spatial relationship with each other. Thus, it is possible to define the virtual binding pockets for class II HLA molecules.

By examining the %burial curves of the peptides, 12 peptide residue positions were identified to have fixed relationship with the binding pockets. They are referred as virtual anchor position. The virtual binding pockets were defined following the same procedure described earlier for class I HLA. Relative Δ ASA were calculated for class II HLA residue positions based on the 22 class II HLA structures (**Figure 3.6**). Only the first 80 residues of each of the α and β chains were included in the calculation as the peptide binding grooves are formed by these residues. For class II HLA, a residue

is defined as directly interacting with the binding peptide when the ΔASA of the residue is greater than $1 \text{ \AA}^2$ upon HLAp complex formation (**Figure 3.6** and **Table 3.13**). The HLA residue positions interacting with each of the peptide residue positions were identified by counting the number of pair-wise inter-residue contacts between the two chains of HLA and the peptide. By definition, HLAp residue pairs are considered to be interacting if the distance between any atom in the peptide residue and any atom in the HLA residue is less than or equal to an arbitrary cutoff distance of $4 \text{ \AA}$. This procedure established the relationship between peptide residue positions and HERPs. The residue pairs at the corresponding HLAp positions have at least 1 inter-residue interaction (**Table 3.13**). This set of HLA residue positions are the HERPs, as described earlier. They form the virtual binding pockets, and directly interact with each of the peptide positions. Finally, these HERPs were divided into 12 virtual pockets according to residue contacts (**Table 3.14**).

The prediction model for class II HLA was constructed by the virtual pockets and the Q matrix, as described earlier. The model is applicable in calculating binding strength of peptides that have at least 12 residues. The longer peptides will be divided into overlapping 12-mer peptides. This introduces another problem, as there will be several predicted binding values for a peptide that is longer than 12-mer. In our view, it is reasonable to use only the highest predicted value. The majority of the binding strength is contributed by the twelve peptide residues. Longer peptides will be accommodated in the way that mostly favors the binding to HLA molecule. If the peptide can “slide” in the HLA binding groove, only the position that has largest binding strength can stably exist. Thus, only the highest predicted value was used to represent the binding strength of a certain peptide.

Unfortunately, insufficient peptide binding data could be collected for the validation of the new model for class II HLA peptide binding prediction. Most class II peptide binding data do not include complete information of the HLA alleles, as there are two chains in class II HLA molecule, each from different HLA gene loci. The only available dataset for verification of the model is the crystal structures from PDB (**Table 3.12**). The predicted values for this dataset are shown in **Table 3.15**. The highest predicted value for each structure is shown. For the 20 non-redundant structures, there are 17 true positive (TP) value, 3 false negative (FN) value. This indicates high prediction accuracy. However, due to the limited size of dataset for cross validation, no conclusive remark can be yet made on the performance of the class II HLA peptide binding prediction model.

Table 3.12 Class II HLA-peptide complexes in PDB.

PDB Code	HLA Allele	CIH	Redundant Peptide Set	CIP	Non-redundant Peptide Set	PL	Peptide source	R [Å]	Release Year
1JK8	DQ8	a,b	LVEALYLVCGERG	c	LVEALYLVCGERG	13	Insulin B	2.40	2001
1AQD	DR1	a,b	GSDWRFLRGYHQYA	c	GSDWRFLRGYHQYA	14	Hla-A2	2.45	1998
1DLH	DR1	a,b	PKYVKQNTLKLAT	c	PKYVKQNTLKLAT	13	Influenza virus	2.80	1994
1FYT	DR1	a,b	PKYVKQNTLKLAT	c	PKYVKQNTLKLAT	13	Hemagglutinin	2.60	2000
1HQR	DR1	a,b	VHFFKNIVTPRTP	c	VHFFKNIVTPRTP	13	Myelin Basic protein	3.20	2001
1HXY	DR1	a,b	PKYVKQNTLKLAT	c	PKYVKQNTLKLAT	12	Haemagglutinin	2.60	2001
1KG0	DR1	a,b	PKYVKQNTLKLAT	d	PKYVKQNTLKLAT	13	Haemagglutinin	2.65	2002
1KLG	DR1	a,b	GELIGILNAAKVPA	c	GELIGILNAAKVPA	15	Triosephosphate Isomerase	2.40	2002
1KLU	DR1	a,b	GELIGTLNAAKVPA	c	GELIGTLNAAKVPA	15	Triosephosphate Isomerase	1.93	2002
1PYW	Dr1	a,b	XFVKQNAAXAL X	c	XFVKQNAAXAL X	11	Influenza Virus Hemagglutinin	2.10	2004
1R5I	Drb1-1	a,b	PKYVKQNTLK LAT	c	PKYVKQNTLK LAT	13	Hemagglutinin Peptide	2.60	2004
1S9V	Dqa1*0501	a,b	LQPFPPQPELP Y	c	LQPFPPQPELP Y	11	alpha-I Gliadin	2.22	2004
1SJE	Dra*0101	a,b	PEVIPMFSAL SEGATP	c	PEVIPMFSAL SEGATP	16	HIV-1 Gag Capsid Protein	2.45	2004
1SJH	Dra*0101	a,b	PEVIPMFSAL SEG	c	PEVIPMFSAL SEG	13	HIV-1 Gag Capsid Protein	2.25	2004
1T5W	Drb1*1	a,b	AAYSDQATPL LLSPR	c	AAYSDQATPL LLSPR	15	Regulatory Protein Mig1	2.40	2004
1T5X	Drb1-1	a,b	AAYSDQATPL LLSPR	c	AAYSDQATPL LLSPR	11	Regulatory Protein Mig	2.50	2004
1SEB	DR1	a,b	AAAAAAAAAAAAA	c	AAAAAAAAAAAAA	13	Endogenous Peptide Model	2.70	1996
1BX2	DR2	a,b	ENPVVHFFKNIVTP	c	ENPVVHFFKNIVTP	14	HMBP	2.60	1998
1FV1	DR2	a,b	VHFFKNIVTPRTPP	c	VHFFKNIVTPRTPP	14	Myelin basic protein	1.90	2000
1A6A	DR3	a,b	VSKMRMATPLLMQA	c	VSKMRMATPLLMQA	14	CLIP fragment	2.75	1998
1J8H	DR4	a,b	PKYVKQNTLKLAT	c	PKYVKQNTLKLAT	13	Hemagglutinin	2.40	2002
2SEB	DR4	a,b	AYMRADAAAGGA	e	AYMRADAAAGGA	12	Collagen II	2.50	1998

Duplicate peptides were removed in the non-redundant peptide set.

Abbreviations: CIH= chain identifier for HLA; CIP = chain identifier for peptide; PDB= protein databank; PL = peptide length; R = resolution; Release Year = the year in which the entry was released by PDB.

Figure 3.5 Alignment of the peptides of different length bound to class II HLA based on %Burial curve.

a. Alignment of the peptides of different length bound to class II HLA based on %Burial curve. **b.** Mean value. **c.** Standard deviation.

%Burial for each peptide residue position was calculated by the same procedure described earlier. It is the relative ASA for each peptide residue, which is calculated as: (ASA in the HLAp complex state) / (ASA in the separate peptide state).

Figure 3.6 Δ ASA of the first 100 residues of class II HLA α and β chains upon complex formation.

The distribution of Δ ASA (unit: $\text{\AA}^2$) of the HLA residues among the 22 class II HLA_p structures is shown. HLA residues with Δ ASA $\geq 1\text{\AA}^2$ upon complex formation are identified as highly essential residue positions (HERP).

Table 3.13 List of HERPs in class II HLA molecules.

ΔASA ($\text{\AA}^2$)				Contacts					
α chain		β chain		α chain			β chain		
HERP	Mean	HERP	Mean	HERP	Sum	Mean	HERP	Sum	Mean
7	0.7	9	1.00	9	209	8.7	9	22	0.9
8	1.44	11	10.02	11	35	1.5	11	135	5.6
9	8.8	13	9.86	22	23	1	13	180	7.5
11	0.58	14	0.03	24	56	2.3	14	7	0.3
22	3.03	15	0.34	31	20	0.8	15	3	0.1
24	7.45	26	2.11	32	16	0.7	26	25	1
31	4.61	28	2.85	43	30	1.3	28	11	0.5
32	3.64	30	6.22	49	1	0	30	77	3.2
43	1.81	37	0.33	51	53	2.2	37	8	0.3
49	1.27	38	0.1	52	150	6.3	47	31	1.3
50	0.96	47	1.65	53	338	14.1	56	42	1.8
51	9.06	56	7.05	54	123	5.1	57	152	6.3
52	9.65	57	10.6	55	25	1	59	1	0
53	43.43	59	0.38	58	74	3.1	60	178	7.4
54	20.57	60	21.15	59	4	0.2	61	242	10.1
55	10.46	61	15.9	61	4	0.2	67	92	3.8
58	15.88	64	0.02	62	370	15.4	70	77	3.2
59	0.04	67	11.67	64	1	0	71	164	6.8
61	4.08	70	11.52	65	157	6.5	74	81	3.4
62	30.87	71	12.65	66	56	2.3	77	29	1.2
64	0.98	73	0.05	68	46	1.9	78	230	9.6
65	35.35	74	14.69	69	332	13.8	79	2	0.1
66	2.34	77	10.7	71	5	0.2	81	186	7.8
68	17.93	78	20.6	72	70	2.9	82	313	13
69	19.53	79	0.31	73	19	0.8	85	183	7.6
71	0.5	81	21.46	75	7	0.3	86	28	1.2
72	28.91	82	20.49	76	60	2.5	88	1	0
73	1.58	84	1.9				89	17	0.7
75	1.9	85	28						
76	5.74	86	7.29						
		87	0.01						
		88	1.06						
		89	1.55						
		90	0.12						

Abbreviations: HERP = highly essential residue position.

Table 3.14 Mapping between peptide residue positions and virtual pockets in the class II HLA binding groove.

Peptide Position	HLA α chain Position	Peptide Position	HLA β chain Position
1	51, 52, 53	1	85
2	53, 55	2	81, 85
3	32, 43, 52, 53, 54	3	82, 85, 86, 89
4	24	4	74, 77, 78
5	9, 54, 58, 62	5	81, 82
6	9, 62	6	78
7	62	7	13, 70, 71, 74, 78
8	11, 62, 65, 66	8	67, 70, 71
9	69	9	60, 61, 67, 71
10	65, 69	10	61, 60
11	68, 69, 72	11	9, 57, 61
12	72, 76	12	56, 57, 60

Table 3.15 Performance of the model in the dataset from PDB.

Serial No.	PDB Code	α chain	β chain	Bound Peptide	Peptide Fragment (12-mer)	Predicted Binding Value	IF
1	1JK8	DQA1*0301	DQB1*0302	LVEALYLVCGERG	LVEALYLVCGER	1394.28	TP
2	1JK8	DQA1*0301	DQB1*0302	LVEALYLVCGERG	VEALYLVCGERG	272.18	
3	1AQD	DRA*0101	DRB1*0101	GSDWRFLRGYHQA	GSDWRFLRGYHQ	418.25	TP
4	1AQD	DRA*0101	DRB1*0101	GSDWRFLRGYHQA	SDWRFLRGYHQY	-475.21	FN
5	1AQD	DRA*0101	DRB1*0101	GSDWRFLRGYHQA	DWRFLRGYHQA	-1079.98	
6	1DLH	DRA*0101	DRB5*0101	PKYVKQNTLKLAT	PKYVKQNTLKL	118.55	
7	1DLH	DRA*0101	DRB5*0101	PKYVKQNTLKLAT	KYVKQNTLKLAT	140.11	TP
8	1FYT	DRA*0101	DRB1*0101	PKYVKQNTLKLAT	PKYVKQNTLKL	283.3	
9	1FYT	DRA*0101	DRB1*0101	PKYVKQNTLKLAT	KYVKQNTLKLAT	521.03	TP
10	1HQR	DRA*0101	DRB5*0101	VHFFKNIVTPRTP	VHFFKNIVTPRT	1130.17	TP
11	1HQR	DRA*0101	DRB5*0101	VHFFKNIVTPRTP	HFFKNIVTPRTP	527.95	
12	1HXY	DRA*0101	DRB1*0101	PKYVKQNTLKL	PKYVKQNTLKL	283.3	TP
13	1KG0	DRA*0101	DRB1*0101	PKYVKQNTLKLAT	PKYVKQNTLKL	283.3	
14	1KG0	DRA*0101	DRB1*0101	PKYVKQNTLKLAT	KYVKQNTLKLAT	521.03	TP
15	1KLG	DRA*0101	DRB1*0101	GELIGILNAAKVPAD	GELIGILNAAKV	226.25	
16	1KLG	DRA*0101	DRB1*0101	GELIGILNAAKVPAD	ELIGILNAAKVP	-214.79	
17	1KLG	DRA*0101	DRB1*0101	GELIGILNAAKVPAD	LIGILNAAKVPA	80.4	
18	1KLG	DRA*0101	DRB1*0101	GELIGILNAAKVPAD	IGILNAAKVPAD	1864.58	TP
19	1KLU	DRA*0101	DRB1*0101	GELIGTLNAAKVPAD	GELIGTLNAAKV	449.2	
20	1KLU	DRA*0101	DRB1*0101	GELIGTLNAAKVPAD	ELIGTLNAAKVP	275.54	
21	1KLU	DRA*0101	DRB1*0101	GELIGTLNAAKVPAD	LIGTLNAAKVPA	-181.09	
22	1KLU	DRA*0101	DRB1*0101	GELIGTLNAAKVPAD	IGTLNAAKVPAD	1232.47	TP
23	1R5I	DRA*0101	DRB1*0101	PKYVKQNTLKLAT	PKYVKQNTLKL	283.3	
24	1R5I	DRA*0101	DRB1*0101	PKYVKQNTLKLAT	KYVKQNTLKLAT	521.03	TP
25	1SJE	DRA*0101	DRB1*0101	PEVIPMFSALSEGATP	PEVIPMFSALSE	116.88	
26	1SJE	DRA*0101	DRB1*0101	PEVIPMFSALSEGATP	EVIPMFSALSEG	-595.47	
27	1SJE	DRA*0101	DRB1*0101	PEVIPMFSALSEGATP	VIPMFSALSEGA	-534.71	
28	1SJE	DRA*0101	DRB1*0101	PEVIPMFSALSEGATP	IPMFSALSEGAT	65.92	
29	1SJE	DRA*0101	DRB1*0101	PEVIPMFSALSEGATP	PMFSALSEGATP	639.56	TP
30	1SJH	DRA*0101	DRB1*0101	PEVIPMFSALSEG	PEVIPMFSALSE	116.88	TP
31	1SJH	DRA*0101	DRB1*0101	PEVIPMFSALSEG	EVIPMFSALSEG	-595.47	
32	1T5W	DRA*0101	DRB1*0101	AAYSDQATPLLLSPR	AAYSDQATPLLL	1017.8	TP
33	1T5W	DRA*0101	DRB1*0101	AAYSDQATPLLLSPR	AYSDQATPLLLS	-585.34	
34	1T5W	DRA*0101	DRB1*0101	AAYSDQATPLLLSPR	YSDQATPLLLSP	-460.12	
35	1T5W	DRA*0101	DRB1*0101	AAYSDQATPLLLSPR	SDQATPLLLSPR	-320.66	
36	1T5X	DRA*0101	DRB1*0101	AAYSDQATPLLLSPR	AAYSDQATPLLL	1017.8	TP
37	1T5X	DRA*0101	DRB1*0101	AAYSDQATPLLLSPR	AYSDQATPLLLS	-585.34	
38	1T5X	DRA*0101	DRB1*0101	AAYSDQATPLLLSPR	YSDQATPLLLSP	-460.12	
39	1T5X	DRA*0101	DRB1*0101	AAYSDQATPLLLSPR	SDQATPLLLSPR	-320.66	
40	1SEB	DRA*0101	DRB1*0101	AAAAAAAAAAAAA	AAAAAAAAAAAAA	-474.39	FN
41	1SEB	DRA*0101	DRB1*0101	AAAAAAAAAAAAA	AAAAAAAAAAAAA	-474.39	
42	1BX2	DRA*0101	DRB1*1501	ENPVVHFFKNIVTP	ENPVVHFFKNIV	-831.03	

CHAPTER 3 HLA PEPTIDE BINDING PREDICTION MODEL

43	1BX2	DRA*0101	DRB1*1501	ENPVVHFFKNIVTP	NPVVHFFKNIVT	1058.68	TP
44	1BX2	DRA*0101	DRB1*1501	ENPVVHFFKNIVTP	PVVHFFKNIVTP	426.8	
45	1FV1	DRA*0101	DRB5*0101	VHFFKNIVTPRTPP	VHFFKNIVTPRT	1130.17	TP
46	1FV1	DRA*0101	DRB5*0101	VHFFKNIVTPRTPP	HFFKNIVTPRTP	527.95	
47	1FV1	DRA*0101	DRB5*0101	VHFFKNIVTPRTPP	FFKNIVTPRTPP	-1308.05	
48	1A6A	DRA*0101	DRB1*0301	VSKMRMATPLLMQA	VSKMRMATPLLM	-404.4	
49	1A6A	DRA*0101	DRB1*0301	VSKMRMATPLLMQA	SKMRMATPLLMQ	88.69	TP
50	1A6A	DRA*0101	DRB1*0301	VSKMRMATPLLMQA	KMRMATPLLMQA	-1264.37	
51	1J8H	DRA*0101	DRB1*0401	PKYVKQNTLKLAT	PKYVKQNTLKL	122.95	
52	1J8H	DRA*0101	DRB1*0401	PKYVKQNTLKLAT	KYVKQNTLKLAT	456.73	TP
53	2SEB	DRA*0101	DRB1*0401	AYMRADAAAGGA	AYMRADAAAGGA	-391.84	FN

Only the highest predicted binding value for class II HLA peptide complex is used for the validation of the model.

Abbreviations: IF = inference; TP = true positive; FN = false negative.

CHAPTER 4

HLA SUPERTYPE SUB-TYPING

4.1 Introduction

A key feature of HLA molecules is their extremely high allelic polymorphism. For example, 1,325 alleles are known for the class I HLA and 763 for the class II HLA locus in Release 2.10 (released in July 2005) of IMGT/HLA Sequence Database (Robinson *et al.* 2003). The majority of polymorphic sites in class I alleles are found in the α_1 and α_2 domains of the mature HLA protein. These domains form the peptide binding groove. The diversity of the residues located in the binding groove enables a repertoire of peptides to bind to different HLA sequences with different affinities (Barber *et al.* 1995).

Analysis of X-ray structures of HLA peptide complexes identified the HLA residues within the binding groove that are critical for peptide binding and therefore for T-cell recognition. About 20 of them are polymorphic. Six binding pockets, denoted A through F, are formed within the binding site. Some of the pockets are non-polar and can form hydrophobic contacts, but others contain polar atoms that can make hydrogen bonds with peptide side chains. The peptide side chains at certain positions bind tightly to certain pockets and form primary anchors (Saper *et al.* 1991). The combination of

two or more anchors is called a motif. It has been found experimentally that certain alleles can recognize very similar motifs (Stern *et al.* 1994), and thus be grouped into HLA supertypes.

The use of HLA supertypes as functional units for peptide binding specificity comparison is significant. The variable composition of the peptide-binding pockets and their cumulative combinations in a population confer the potential to bind a broad spectrum of peptides. Despite the observed vast genetic polymorphism of HLA alleles, many of these alleles differ by a few amino acids so that certain groups of genetically related alleles can in some circumstances present the same epitopes for immune recognition. While the information about the polymorphism of HLA is expanding, examining these molecules using a system of functional groupings is necessary.

Furthermore, classification of HLA alleles into supertypes in terms of peptide-binding specificities has direct implications for the rational design of epitope-based vaccines. It is the only practical way to measure immune responses in human population studies. Different racial groups have different supertype distributions. Information on HLA polymorphism in different racial groups is important for understanding the pattern of HLA haplotype and for the designed vaccines to have global effectiveness. In recent years, there is an increasing focus on “multi-epitope” in contrast to “whole gene” approaches to vaccines designed for infectious disease and cancers. Construction of a poly-epitope, subtype-specific vaccine, including multiple copies of immunodominant CTL epitopes across the antigen proteins as restricted by the prevailing HLA alleles, appears to be a logical approach in the design of more universal vaccines. Knowledge on how these epitopes are presented by HLA supertypes with same functional

composition is critical to the application of these approaches in the human population. Therefore, a complete analysis of the genetic diversity of HLA genes and the functional differences in peptide-binding capabilities and CTL responses of the molecules they encode is imperative before assessing the efficacy of a vaccine designed and developed.

However, it is not a simple task to determine which allelic HLA products will bind specific peptides. Different polymorphic HLA alleles, of both class I and class II, have different peptide specificities: each allele binds peptides exhibiting particular sequence patterns. Even very similar HLA molecules can have different peptide specificities. A*0201 and A*0204 differ by only one residue and yet the latter shows a much stronger preference for leucine as the C-terminal peptide anchor residue (Sudo *et al.* 1995). At the same time, peptides display a degree of promiscuity. For example, several nonamer peptides have been shown to bind to B*2705 as well as to HLA-A3 and HLA-A31 (Fruci *et al.* 1993). These anomalies can be rationalized by examining the available crystal structures. Generally, each HLA molecule binds a repertoire of peptides that exhibit common sequence motifs (Barber *et al.* 1995) which serve as “principal anchors”. The other residue positions along the peptide can vary so that a single molecule can bind many thousands of different peptides (Cox *et al.* 1994). However, the secondary anchor sites in the peptide are also important for binding to HLA molecules (Ruppert *et al.* 1993; Chen *et al.* 1994).

Practically, determining the peptide specificity of even a single allele is by no means a trivial undertaking. To undertake this rigorous task for over 2,000 alleles is too intensive to be feasible. The only viable alternative is to seek a bioinformatic approach.

In this work, a bioinformatics strategy is described for classifying HLA molecules into supertypes using information drawn solely from three dimensional HLAp structures.

In our view, the classification of HLA alleles can be approached in two ways: whole sequence-based and pocket-based. Both types of methods, although based on structural knowledge, also make use of protein sequence information. The difference between these methods is the number of residues involved in the analyses. The pocket-based classifications have more direct, practical application to vaccine development, since the binding to HLA molecule is largely determined by the anchor residues in the peptide sequence, which bind in specificity pockets in the antigen binding groove.

The pocket-based classification method was employed in this work. To represent the functional unit of the HLA molecules, a detailed description of the antigen binding pockets, or the pocket profile, is a prerequisite. The pocket profiles represent the peptide binding environments, which are determined by the HLA polymorphic side-chains delineate the geometry and chemical properties of these structural pockets. Obviously, the pocket profiles can be described by the polymorphic amino acid residues that construct the pockets. In this work, pocket profiles are defined as those residues in a fixed neighborhood of the peptide residues in known crystal structure complexes (**Figure 3.1**). The pocket profiles are actually the three dimensional structural basis for supertype classification. The overall binding strength of HLA molecule and peptide is determined by the compatibility between the individual binding pockets and the anchor residue. Thus, the specificities of HLA can be described by the specificities of the pockets. Once the pocket binding specificities, the pocket profiles, have been defined, the binding specificities of the whole HLA

molecule can be established.

Residue positions in HLA molecules crucial for peptide binding were identified by analyzing crystal structures of HLA peptide complex gathered from Protein Data Bank. The length of peptides bound to HLA molecules ranges from 8 to 10 amino acid residues. Only the complexes with 9-mer peptides were included for analysis in this work. As discussed earlier, the reason is that the residues in the binding pockets will possess a more fixed spatial relationship when the peptide length is uniform. Another reason is that the majority of peptide is nonamer in the available structures of class I HLA peptide complexes. The critical positions were then classified into six pockets interacting with six anchor residues of bound antigenic peptides, as the basis for generation of pocket profiles. Then all the pockets in the HLA alleles with amino acid sequence available are clustered to groups by the amino acid type on the critical functional positions in each pocket (**Figure 3.2**). The identified pocket profiles revealed the striking observation that only few amino acids are sufficient for the classification of an allele to a particular supertype.

The results of classification proved that the functional polymorphism can be greatly reduced upon application of pocket profile. The significance of the classification is in that, by examining the antigen binding environments from HLA molecules with known anchor residues, predictions can be made as to likely binding residues for other HLA molecules which share the same environments.

The classification is presented in the form of tables and is pertinent to class I HLA-A, HLA-B, and HLA-Cw HLA sequences. The class two HLA alleles are not included.

As discussed earlier, this is because class II HLA molecules have more open peptide binding grooves. And the peptides bound to them have larger length variability. Thus the spatial relationship of the residue positions that delineate the pockets is not fixed, neither is the relationship between peptide anchor positions and binding pockets. The other reason is the available crystal structures of class II HLA peptide complexes are far less than those of class I HLA peptide complexes. This problem was also encountered in the development of peptide binding prediction model for class II HLA in the former part of the work.

The classification of class I HLA alleles was moderately validated using available peptide binding data, and was shown to correlate well with experiments. However, the validation is limited due to lack of HLA peptide binding data. Hence, the classification can serve as a framework to test specific functional overlap between HLA alleles.

4.2 Functional Positions

The first step is to gather the structural information by examining all available class I HLA crystal structures. There were 46 crystal structures of class I HLA peptide complexes available in PDB (Berman *et al.* 2002). 33 of these structures have nonamer peptides, which were used to glean structural information of HLA peptide interaction in this work. The crystal structures with same HLA allele and peptide were removed to make the dataset non-redundant (**Table 3.1**). The comparison of known structures of HLA-peptide complexes shows that the backbone conformation of the HLA molecules are highly conserved, even in the peptide binding grooves (**Figure 2.2**). Therefore,

structural information gathered from different HLA alleles can represent general HLA backbone structural characteristics.

The HLA residues that interact with the binding peptides were identified based on the change in solvent accessibility area (Δ ASA), as described earlier. Only the first 180 residues of HLA α chain were included in calculation of Δ ASA because these are the residues that form the peptide binding groove. The result of this calculation is shown in **Figure 3.3**. In this chapter, a residue is defined as directly interacting with the binding peptide when the Δ ASA of the residue is greater than 0 (which is different from the procedure described in last chapter). Based on this criterion, all the 180 residue positions were examined to determine if at least one amino acid residue amongst the 24 crystal structures on a particular position is directly interacting with the binding peptide. The positions with at least one such residue are selected and designated as functionally essential residue positions (ERPs).

As discussed earlier, the peptide binding specificities of HLA alleles are determined by the polymorphic residues. It is found that some of the ERPs are conserved. Therefore they carry little information for the characterization of binding pockets. Alignment of the HLA allele sequences finally revealed 21 residue positions out of the ERPs with at least one amino acid polymorphism. These 21 residue positions are called Critical Polymorphic Functional Residue Positions (CPFRP). Thus, the CPFRPs are described as those positions that are predominantly involved in peptide binding in known crystal structures of HLA peptide complexes and show polymorphism at least once among known HLA alleles. It is quite striking that only 21 residue positions are responsible for the diversity of HLA peptide binding specificities.

The CPFRPs were further classified into six pockets (designated as A to F) based on spatial position in the peptide binding groove, as described by Borjkmán (1987). The result of the classification is shown in **Table 4.1**. A set of discontinuous residue segments corresponding to each pocket in the groove was generated by assembling the residues on these positions with the tandem sequences corresponding to their positions on the HLA peptide chains. The extracted discontinuous residue segment patterns formed by the CPFRPs were compared and alleles having identical CPFR segments were grouped together. Thus, alleles clustered within a group share identical CPFR segments. The pockets generated based on residue compositions at CPFRPs are thereafter called functional pockets capable of binding similar peptides with identical repertoire. These discontinuous residue segments actually represent the pocket profiles. Members of one group bind similar peptides and the groups represent HLA “sub-supertypes”.

Table 4.1 Distribution of CPEFRPs of the six binding pockets among 1,095 class I HLA alleles.

	Pocket A							Pocket B							Pocket C				Pocket D				Pocket E				Pocket F									
	7	63	66	70	99	163	167	7	9	24	35	63	66	70	76	99	9	70	73	74	97	99	114	156	97	114	152	156	77	80	81	116				
ALA(A)	0	0	0	0	0	3	0	0	0	690	0	0	0	0	54	0	0	0	81	0	0	0	1	0	0	75	1	0	0	227	0					
GLY(G)	0	0	0	0	0	2	76	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	2	0	0	0				
LEU(L)	0	0	0	0	0	312	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	631	0	0	0	631	0	0	867	72					
VAL(V)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	368	0	0	0	0	0	11	0	1	0	11	1	591	0	0	0	1					
TRP(W)	0	0	0	0	0	1	962	0	0	0	0	0	0	0	0	0	0	0	0	0	49	0	0	163	49	0	6	163	0	0	0	0				
CYC(C)	3	0	0	0	13	0	0	3	0	0	0	0	0	0	0	13	0	0	0	0	0	13	0	0	0	0	0	0	1	0	0	0				
ILE(I)	0	0	571	0	0	0	0	0	0	0	0	0	571	0	0	0	0	0	21	0	78	0	0	0	78	0	0	0	0	169	0	0				
MET(M)	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	130	0	0	0	130	0	1	0	0	0	0	0	0			
PHE(F)	0	0	0	0	90	0	0	0	130	0	0	0	0	0	0	90	130	0	0	0	0	90	0	0	0	0	0	0	0	0	0	119				
PRO(P)	0	0	0	0	0	1	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0	1	0	0	0			
ASP(D)	0	1	0	0	0	0	0	0	67	0	0	1	0	0	0	0	67	0	0	585	0	0	399	83	0	399	0	83	221	0	0	238				
GLU(E)	0	703	0	0	0	181	0	0	0	0	0	703	0	0	673	0	0	0	0	0	0	0	16	1	0	16	404	1	0	0	0	0	0			
ARG(R)	0	0	0	0	0	67	2	0	0	0	0	1071	0	0	0	0	0	0	0	0	572	0	100	128	572	100	9	128	0	0	0	0	0	0		
HIS(H)	1	0	1	235	1	0	0	1	136	0	0	0	1	235	0	1	136	235	0	92	0	1	169	0	0	169	0	0	0	0	0	18				
LYS(K)	0	0	282	24	0	0	0	0	0	0	0	0	282	24	0	0	0	24	0	0	1	0	1	0	1	1	0	0	0	75	0	0	0			
ASN(N)	0	376	239	463	0	0	0	0	0	0	0	0	376	239	463	0	0	0	463	0	1	27	0	332	0	27	332	0	0	354	499	0	2			
GLN(Q)	0	15	1	351	0	0	0	0	0	0	24	15	1	351	0	0	0	351	0	0	0	0	76	86	0	76	1	86	0	0	0	0	0	0		
SER(S)	0	0	0	22	25	0	54	0	91	246	0	0	0	22	0	25	91	22	1	0	128	25	0	1	128	0	0	1	517	0	0	258				
THR(T)	0	0	1	0	0	526	0	0	32	157	0	0	1	0	0	0	32	0	992	0	99	0	0	1	99	0	8	1	0	352	0	0	0	0		
TYR(Y)	1091	0	0	0	966	0	0	1091	639	0	0	0	0	0	0	966	639	0	0	416	0	966	0	0	0	0	0	0	0	0	0	387				
Variability	3	4	6	5	5	9	5	3	6	4	2	4	6	5	3	5	6	5	4	5	9	5	9	9	9	9	8	9	5	4	3	8				
Hydrophobic	1	0	2	0	2	7	3	1	2	3	0	0	2	0	2	2	2	0	3	1	5	2	2	3	5	2	4	3	2	1	3	3				
Hydrophilic	2	4	4	5	3	2	2	2	4	1	2	4	4	5	1	3	4	5	1	4	4	3	7	6	4	7	4	6	3	3	0	5				

The numbers in second row are positions of the residues.

Variability: The number of different types of amino acid residues that occur on this position.

Hydrophilic: The number of different types of hydrophilic amino acid residues that occur on this position.

Hydrophobic: The number of different types of hydrophobic amino acid residues that occur on this position.

There are various hydrophobicity scales available for amino acid. These scales are different with respect to the placement of specific amino acids (Cornette *et al.* 1987). In this work, the hydrophobic (non-polar) amino acids are: Alanine (A), Glycine (G), Leucine (L), Valine (V), Tryptophan (W), Cysteine (C), Isoleucine (I), Methionine (M) and Phenylalanine (F). These amino acids are listed in the upper part of the table. The hydrophilic (polar) amino acids are: Aspartic acid (D), Glutamic acid (E), Arginine (R), Histidine (H), Lysine (K), Asparagine (N), Glutamine (Q), Serine (S), Threonine (T), Tyrosine (Y), Proline (P). These amino acids are listed in the lower part of the table.

4.3 Grouping of HLA Alleles

Amino acid sequences of class I HLA alleles were downloaded from IMGT/HLA Sequence Database (Robinson *et al.* 2003). A total of 396 amino acid sequences of HLA-A alleles, 699 HLA-B alleles and 198 HLA-Cw alleles were available in the Release 2.10 of the database (released in July 2005). The HLA alleles are named according to their nucleic acid sequence. There are some redundant sequences as there may be some wobble variant of nuclear acid sequence, and the protein sequence then encoded are the same. These redundant amino acid sequences were removed. Furthermore, some of the amino acid sequences were not complete. These sequences were removed as well. The final non-redundant, complete sets of HLA amino acid sequences for grouping contained 331 HLA-A alleles, 604 HLA-B alleles and 160 HLA-Cw alleles.

As discussed earlier, the binding of peptide anchor residues to HLA show some extend of combinatorial style. The binding of individual anchor residue are relatively independent to other peptide residues. Since there are six pockets in each HLA binding groove, there are 1,986 possible combinations of six pockets (331×6) in HLA-A alleles, 3,624 in HLA-B alleles and 960 in HLA-Cw alleles. Each of the six pockets was classified into groups based on the discontinuous CPFRP segments. HLA alleles with same CPFRP segments were grouped together. The results of grouping are listed in **Appendix 1**, **Appendix 2**, and **Appendix 3**. The numbers of the groups of the HLA-A, HLA-B, and HLA-Cw alleles are summarized in **Figure 4.1**.

Figure 4.1 Summary of grouping results of HLA-A, B and Cw.

The X-axis is the pocket name. The Y-axis is the number of group in each pocket. The total number of each type of HLA alleles used in grouping is placed at leftmost part for comparison.

4.4 Results of Classification

Not all the residues in the α_1 and α_2 domains involved in the constitution of the six pockets are crucial to peptide binding. Analysis of 24 accumulated crystal structures of HLA-peptide complexes has determined that only the residues on 60 positions have direct contacts with binding peptides. 21 of these positions are CPFRPs among HLA alleles. The frequencies of CPFRs for all the 1,095 non-redundant, complete amino acid sequences of class I HLA alleles (type A, B and Cw) are listed in **Table 4.1**. The positions are polymorphic as there is more than one possible kind of amino acid at each of these positions. The residues and the amino acid variability on each position have also been listed (**Table 4.1**).

The 21 residue positions are assigned to six pockets (A to F) (**Table 4.1**). The 21 CPFRPs are unevenly distributed in the six functional pockets (named A to F) in the groove. For example, pocket B contains 9 CPFRPs and pocket D contains only 3 CPFRPs. A maximum variability of 9 is observed at 3 CPFRPs in pocket E, and at 2 CPFRPs in pocket D, and 1 CPFRP in pocket A and C. The uneven distribution of CPFRPs among functional pockets contributes to specific peptide selection in a combinatorial manner. Chemical and physical properties of binding pockets have been previously utilized in peptide selection studies (Zhang *et al.* 1998). In this analysis, it is found that the residues at the CPFRPs are prevalently hydrophilic among the 1,095 HLA-A alleles and, therefore, the functional pockets are predominantly hydrophilic. Hydrophilic to hydrophilic substitution is found to be more common than hydrophobic to hydrophobic or hydrophilic to hydrophobic substitutions among HLA-A alleles (**Table 4.1**). Some CPFRPs are shared between pockets. For example, position 7 is

assigned to pocket A and pocket B, position 9 to pocket B and C. The reason is that residues at these positions constitute the pocket rims and interact simultaneously to two anchor residues of bound peptide. Due to the structure similarity among the HLA alleles, these conserved functionally essential residue positions can be extended to those HLA alleles without structural data available and be used as the basis of the classification. Understanding the chemical properties of individual functional pockets and clustering HLA alleles by common functional pockets will provide insights to define metrics for individual pocket function.

Possible number of CPFRP segments by randomly combining the functionally essential residues with original tandem relation is large. It is the product of the numbers of residue variability of the position for each pocket. For pocket A, for instance, the number is 2400 (6×20). But most of the combinations do not actually exist. Actually only 247 unique CPFRP segments were found in the 1,986 pockets from 331 HLA-A alleles. In other words, all the pockets were classified into 274 unique categories (**Figure 4.1, Appendix 1**).

1,095 complete, non-redundant amino acid sequences of class I HLA alleles were used in this work, of which, 331 were HLA-A alleles, 604 were HLA-B alleles and 160 were HLA-Cw alleles. The discontinuous CPFR residue segment patterns were extracted for each allele. Comparison of these segment patterns among HLA-A, B and C alleles grouped A alleles into 247 unique clusters, B alleles into 261 unique clusters and C alleles into 117 unique clusters. Thus, 1,095 alleles were grouped into 625 clusters. The total diversity among the 1,095 alleles is reduced to 57% with a “reduction in variability” of 43%. The diversity in A, B and C is reduced to 75%, 43%

and 73%, with a “reduction in variability” of 25%, 57% and 27%, respectively (**Appendix 1, Appendix 2, Appendix 3, Figure 4.1**).

The summary of classification for HLA-A, HLA-B and HLA-Cw alleles is illustrated in **Figure 4.1**. The grouping results of HLA-A alleles are listed in **Appendix 1**, and will be described in details. The results for HLA-B and HLA-Cw can be found in **Appendix 2** and **Appendix 3**.

Among HLA-A alleles, the grouping produced 47 clusters for pocket A; 61 for pocket B; 39 for pocket C; 39 for pocket D; 47 for pocket E; and 14 for pocket F. Pocket B has the maximum number of clusters, whereas pocket F has the minimum number of clusters. The result suggests that the currently available sequences of HLA-A alleles have only 61 different types of CPFRP segments in pocket B. Interestingly, there are 20 groups with more than 20 allele members. It should be noted that the binding of peptides to HLA molecules is neither because of any one independent functional pocket nor because of any one independent peptide residue. However, their binding is defined by the cumulative physical, chemical properties of all the residues in the peptide and all the pockets in the HLA molecule. Knowledge on the similarity and dissimilarity of individual functional pockets will enable the development of metrics for cumulative pocket effects towards specific peptide binding. There are 14 different types of pocket F which are the most conserved and 61 types of pocket B with the most highly variability.

The most efficient grouping is that 99 different HLA-A alleles are put into group #2 in pocket F with the CPFRP segment of ‘DTLD’, 203 HLA-B alleles with the CPFRP

segment of 'YNTYR' in pocket C, 160 HLA-B alleles with CPFRP segment 'YNL' in pocket D, 157 HLA-B alleles with CPFRP segment 'SNLY' in pocket F, and 122 HLA-B alleles with CPFRP segment 'YDL' in pocket D. These are the most highly conserved type of pockets. There are totally 17 groups with more than 30 HLA-A alleles and 26 groups with more than 20 HLA-A alleles. These groups are also the most highly conserved. At the same time, there are 154 pocket groups that have no more than 3 HLA-A alleles, which are most polymorphic.

4.5 HLA Sub-supertypes

When pockets A–F are rearranged using residues at CPFRPs for the 331 HLA-A alleles, 194 clusters were produced. In other words, 194 different patterns of discontinuous sequence stretches at the CPFRPs are identified among the available HLA-A alleles.

Using the grouping procedure described above, 185 HLA-A alleles (56% of all known HLA-A alleles with non-redundant, complete amino acid sequences) were clustered into 42 groups such that each group contains at least two alleles (**Table 4.2**). Members in each of these groups have identical CPFR segment patterns. Therefore the alleles in the clusters have identical A–F functional pockets. The remaining alleles are evenly distributed in the rest clusters with each cluster containing 1 allele.

This finding provides the key to define HLA subtypes with similar peptide binding functions as observed by serologic investigations. The proposed hypothesis is that alleles in a group bind to similar peptides and members in the group form

“sub-supertypes” with identical binding repertoire. If binding of a peptide is known for one representative allele in a group, its binding to other members in the group can be deduced. However, this clustering has to be validated using experimentally determined binding data for a representative set of alleles. This requires large scale generation of binding data and it is difficult to generate binding data at such scale due to limitations in the synthesis and purification of HLA alleles for binding assay. If this hypothesis is reasonably validated using binding data, it is possible to infer the binding property of a peptide to 56% of HLA-A alleles using binding values for 42 representative A alleles. The 42 groups clustered A alleles into A1, A2, A3, A11, A23, A24, A26, A29, A30, A31, A32, A33, A34, A36, A66, A68 and A74 “predictive sub-supertypes” where members of a group is proposed to bind similar peptides (**Table 4.2**). The analysis was extended to HLA-B and HLA-Cw alleles and showed that 303 HLA-B alleles (50% of all known HLA-B alleles) were clustered into 78 groups such that each group contains at least two alleles (**Table 4.3**). It is also shown that 76 HLA-Cw alleles (47% of all known HLA-C alleles) were grouped into 23 groups such that each group contains at least two alleles (**Table 4.4**). Thus, it is demonstrated that the strategy described in this work to group alleles into sub-supertypes is extremely powerful in reducing functional diversity.

Table 4.2 HLA-A alleles (185) grouped into putative supertypes.

Group	CPFR	HLA-A allele
1	YFARENHTDANTLIYRDARRG	A*0101 A*0104 A*0109
2	YFARENQTDVDTLIYRDARRG	A*0113 A*0318
3	YFAREKHTHVDTLRYHYVLTW	A*0201 A*0209 A*0224 A*0225 A*0230 A*0231 A*0240 A*0243 A*0245 A*0246 A*0258 A*0259 A*0266 A*0267 A*0268 A*0270 A*0271 A*0273 A*0274 A*0275 A*0277 A*0283 A*0285 A*0286
4	YFAREKHTHVDTLRYHYVWTW	A*0202 A*0222 A*0250 A*0263
5	YYAREKHTHVDTLRYHYVLTW	A*0206 A*0214 A*0221 A*0228 A*0251 A*0261 A*0272 A*0279 A*0284
6	YFAREKHTHVDTLRCHYVLTW	A*0207 A*0215 A*0218
7	YFAREKHIDVDTLRYHYVLTW	A*0211 A*0269
8	YFAREKHTHVDTLRYHYVQTG	A*0219 A*0237
9	YFARENQTHVDTLRYHYVLTW	A*0256 A*0262
10	YFARENQTDVDTLIYRDELTW	A*0301 A*0303 A*0304 A*0305 A*0306 A*0313 A*0314 A*0316 A*0317
11	YFARENQTDVDTLIYRDVQTW	A*0302 A*0310
12	YYARENQTDVDTLIYRDAQRW	A*1101 A*1102 A*1105 A*1107 A*1109 A*1112 A*1113 A*1115 A*1116 A*1121 A*1123
13	YYARENQTDVDTLIYRDEQRW	A*1103 A*1120
14	YSAREKHTDENIAMFHVLTG	A*2301 A*2303 A*2306 A*2307 A*2413
15	YSAREKHTDENIAMFHYVWTG	A*2302 A*2406
16	YSAREKHTDENIAMFHYVQTG	A*2402 A*2405 A*2408 A*2409 A*2411 A*2420 A*2421 A*2426 A*2427 A*2429 A*2435 A*2437 A*2438 A*2439 A*2440 A*2443 A*2447 A*2449
17	YSAREKHTDENIAMFHYVQTW	A*2403 A*2423 A*2433
18	YSAREKHTDENIAMYHYVQTG	A*2415 A*2453
19	YSAREKHTDENIAMFRDVQTG	A*2417 A*2441
20	YSAREKHTHENIAMFHYVQTG	A*2430 A*2442
21	YYARNNHTDANTLRYQDEWRW	A*2601 A*2610 A*2614 A*2615 A*2617 A*2622 A*2623 A*2624 A*2626
22	YYARNNHTHVDTLRYQDEWRW	A*2603 A*2606 A*2621
23	YYARNNHTDANTLRYQDVWRW	A*2612 A*2618
24	YTARQNQTDANTLMYRDVLTW	A*2901 A*2902 A*2906 A*2909 A*2910 A*2911 A*2912
25	YSARENQTDVDTLIYEHWLTW	A*3001 A*3011 A*3014 A*3015
26	YSARENHTDENTLIYEHRLTW	A*3002 A*3003 A*3012
27	YSARENHTDENTLIYEHVWTW	A*3004 A*3006
28	YTARENHIDVDTLMYQDVLTW	A*3101 A*3109 A*3111

29	YTAREKHIDVDTLMYQDVLWTW	A*3102	A*3308
30	YTARENHIDVDTLIYRDVLWTW	A*3103	A*3104
31	YTAREKHTDENIAMYQDVLWTW	A*3107	A*3108
32	YFARENHTDESIAMYQDVLWTW	A*3201	A*3206
33	YTARNNHIDVDTLMYQDVLWTW	A*3301	A*3303 A*3304 A*3305 A*3306 A*3307
34	YYARNKQTDVDTLRYQDEWTW	A*3401	A*3405
35	YYARNNQTDVDTLIYRDELTW	A*3402	A*3403 A*3404
36	YFARENHTDANTLIYRDARTW	A*3601	A*3602
37	YYARNNQTDVDTLRYQDEWRW	A*6601	A*6604
38	YYARNNQTDVDTLMYRDVWTW	A*6801	A*6816 A*6819 A*6821 A*6822 A*6824 A*6825
39	YYARNNQTDVDTLRYHYVWTW	A*6802	A*6827
40	YYARNNHTHVDTLMYRDVWTW	A*6805	A*6820
41	YYARENQTDVDTLMYRDVWTW	A*6810	A*6813 A*6814
42	YFARENHTDVDTLMYQDVLWTW	A*7401	A*7402 A*7403 A*7408 A*7409 A*7411

The classification of class I HLA is based on IMGT/HLA Sequence Database Release 2.10 (released in July 2005).

CPFR: critical polymorphic functional residue; the polymorphic amino acid residues that directly contact with binding peptides in the pockets, placed according to their original positions on the HLA chains, designated in single capital letter.

HLA-A Alleles: The class I HLA-A alleles that belong to the group, with sequence based nomenclature begin with “A*” and followed by a four digit number.

Table 4.3 HLA-B alleles (303) grouped into putative supertypes.

Group	CPFR	HLA-B allele
1	YYSRNIQTDESNLSYDYEREW	B*0702 B*0710 B*0721 B*0722 B*0723 B*0730 B*0733 B*0735 B*0739 B*0741 B*0742
2	YYSRNINTDESNLSYDYEREW	B*0703 B*0708
3	YYSRNIQTDESNLSYNYEREW	B*0705 B*0706 B*0740
4	YYSRNIQTDESNLRYDYEREW	B*0707 B*0712 B*0714 B*0718
5	YYSRNIQTDESNLSYDSEREW	B*0709 B*0717
6	YYSRNINTYESNLSYDYEREW	B*0716 B*0737
7	YDSRNINTDESNLSYNYVDTW	B*0801 B*0804 B*0818 B*0819 B*0822 B*0824
8	YYTREINTYENTARYNLVLEW	B*1301 B*1311
9	YYTREINTYENTATYNLVLEW	B*1302 B*1308
10	YYSRNINTDESNLWYNFELTW	B*1401 B*1402
11	YYAREINTYESNLRDSEWLW	B*1501 B*1528 B*1533 B*1534 B*1538 B*1546 B*1556 B*1560 B*1566 B*1578 B*1581 B*1582 B*1585 B*1592 B*1596 B*1597 B*9502 B*9504
12	YYARNINTYESNLRDSELLW	B*1502 B*1521 B*3511 B*3521
13	YYSREINTYESNLRDSELLW	B*1503 B*1562 B*1574 B*1598 B*9503
14	YYAREINTYESNLTVDSEWLW	B*1504 B*1535
15	YYAREINTYESNLRDYSVLLW	B*1505 B*1520 B*3528
16	YYARNINTYESNLRDSEWLW	B*1508 B*1511 B*1515 B*3514 B*3543
17	YYSRNINTYESNLRDYEYLLW	B*1510 B*1537 B*1590
18	YYAREINTYESNLRDSEWLG	B*1512 B*1519
19	YYARNINTYENIARYDSELLW	B*1513 B*5306 B*5308
20	YYARENSTYENIARYDSELLW	B*1516 B*1567 B*1595
21	YYSRNINTYESNLRDSELLW	B*1518 B*1529 B*1564 B*1572 B*1580
22	YYAREINTYESNLRDSELLW	B*1525 B*1539
23	YYARNINTYESNLRDYSVLLW	B*1531 B*3501 B*3507 B*3519 B*3520 B*3524 B*3526 B*3532 B*3540 B*3541 B*3542 B*3546 B*3547 B*3549 B*3554 B*3557
24	YYSREINTYESNLRDSEREW	B*1547 B*1549
25	YYARNINTYESNLSYDSVLLW	B*1555 B*3505 B*3517 B*3530 B*3551
26	YYARNIQTDESNLRDSEWLW	B*1576 B*5406 B*5603
27	YHSRNINTYESNLRDYSVLTW	B*1801 B*1805 B*1807 B*1811 B*1820
28	YYARNINTYESNLRDYSVLTW	B*1804 B*3535
29	YHTREIKTDEDTLNYHVDVLEW	B*2703 B*2705 B*2713 B*2717
30	YYARNINTYESNLRYNVLLW	B*3502 B*3504 B*3509 B*3512
31	YYARNINTYESNLRDVFLLW	B*3503 B*3536 B*3555

32	YYARNINTYESNLRYSVRLW	B*3508 B*3561
33	YYARNINTYESNLRIDYVLLW	B*3534 B*3539
34	YHSREINTYEDTLRSNFVDTW	B*3701 B*3704 B*3706
35	YYSRNINTYENIARYNFVLTW	B*3801 B*3805 B*3806 B*3807 B*3809 B*3811
36	YYSRNINTDESRLRYNFVLTW	B*3901 B*3904 B*3910 B*3916 B*3919 B*3926 B*3930
37	YYSREINTDESRLRYNFVLTW	B*3902 B*3922 B*3923
38	YYSRNINTDESRLSYNFVLTW	B*3903 B*3924
39	YYSRNINTDESRLTYNFVLTW	B*3906 B*3928 B*3934
40	YHTREINTYESNRLRYNYVLEW	B*4001 B*4004 B*4007 B*4011 B*4014 B*4046 B*4054 B*4055 B*4061
41	YHTREINTYESNLSYNYVLEW	B*4002 B*4029 B*4035 B*4045 B*4056 B*4057
42	YHTREINTYESNLTNYVLEW	B*4006 B*4053
43	YHTREINTYESNLSYDYVLEW	B*4009 B*4018 B*4024 B*4031
44	YYAREINTYESNRLRYNYVLEW	B*4010 B*4021
45	YYSREINTYESNRLRYNYVLEW	B*4012 B*4803
46	YHTREINTYENIASYNYVLEW	B*4013 B*4019
47	YHTREINTYESNLSYNYEREW	B*4015 B*4016
48	YHTREINTYESNLRYSVLEW	B*4020 B*4052 B*4059
49	YHTREINTYESNRLRYNYELLW	B*4026 B*4028
50	YHTREINTYESNLSYNYVLEW	B*4030 B*4034
51	YHTREINTYESNLRIDYVLEW	B*4033 B*4042
52	YHTREINTYESNRLRYNYVDTW	B*4101 B*4103 B*4106 B*4107
53	YYSRNITQDESRLSYNYVDTW	B*4201 B*4205
54	YYTREINTYENTARYDDVDLS	B*4402 B*4408 B*4422 B*4424 B*4427 B*4433 B*4441
55	YYTREINTYENTARYDDVLLS	B*4403 B*4407 B*4413 B*4426 B*4429 B*4430 B*4432 B*4436 B*4438 B*4440
56	YHTREINTYESNRLRYNLVDLS	B*4501 B*4503 B*4505 B*4507
57	YYAREKQTDVSNLRIDSEWLW	B*4601 B*4604 B*4605
58	YYTREINTYESNLRFDHVLEW	B*4702 B*4703
59	YYSREINTYESNLSYNYVLEW	B*4801 B*4804 B*4809 B*4810 B*4812 B*4813
60	YHTREINTYESNRLRYNLELLW	B*5001 B*5004
61	YYARNINTYENIATYNYELLW	B*5101 B*5102 B*5107 B*5111 B*5112 B*5117 B*5118 B*5122 B*5124 B*5126 B*5128 B*5130 B*5132 B*5133 B*5135 B*5138
62	YYARNINTYENIARYNYELLW	B*5104 B*5106
63	YYARNINTYENIATYNYVLLW	B*5109 B*5119
64	YYARNINTYENIATYNYELEW	B*5116 B*5134
65	YYAREINTYENIATYNYELLW	B*5201 B*5202 B*5204 B*5205 B*5206 B*5207

66	YYARNINTYENIARYDSVLLW	B*5301 B*5302 B*5310
67	YYARNIQTDESNLTYNLVLTW	B*5401 B*5407 B*5502 B*5507 B*5519
68	YHARNIQTDESNLTYNLVLTW	B*5402 B*5516
69	YYARNIQTDESNLTYNLVLLW	B*5403 B*5601
70	YYARNIQTDESNLTYNLELTW	B*5501 B*5505 B*5515
71	YYARNIQTDESNLRYNLVLLW	B*5602 B*5604
72	YYARNIQTDESNLTYNYELLW	B*5605 B*5606
73	YYAREKQTDESNLTYNLVLLW	B*5608 B*5614
74	YYARENSTYENIAVYDSVLLW	B*5701 B*5706 B*5708
75	YYARENSTYENIARYDSVLLW	B*5801 B*5804 B*5809 B*5811
76	YYARNINTDESNLTYNYELLW	B*7801 B*7803
77	YYSRNIQTDESNLSYNYVLEW	B*8101 B*8102
78	YYSRNIQTDESNLRFNLVDLS	B*8201 B*8202

The classification of class I HLA is based on IMGT/HLA Sequence Database Release 2.10 (released in July 2005).

CPFR: critical polymorphic functional residue; the polymorphic amino acid residues that directly contact with binding peptides in the pockets, placed according to their original positions on the HLA chains, designated in single capital letter.

HLA-B Alleles: The class I HLA-B alleles that belong to the group, with sequence based nomenclature begin with “B*” and followed by a four digit number.

Table 4.4 HLA-C alleles (76) grouped into putative supertypes.

Group	CPFR	HLA-Cw allele			
1	YFSREKQTDVSNLWCDYERTW	Cw*0102	Cw*0107	Cw*0108	Cw*0111
2	YYAREKQTDVNKLRYDSEWEW	Cw*0202	Cw*0208	Cw*0209	Cw*0210 Cw*0211
3	YYAREKQTDVSNLRYDYELLW	Cw*0303	Cw*0304	Cw*0309	Cw*0313 Cw*0318
4	YYAREKQTDVSNLRYDLELLW	Cw*0312	Cw*0319		
5	YSAREKQADVSNKLRFNFERTW	Cw*0401	Cw*0405	Cw*0407	Cw*0409 Cw*0412
6	YYAREKQADVSNKLRFNFERTW	Cw*0403	Cw*0416		
7	YSAREKQADVSNKLRYNFERTW	Cw*0415	Cw*0417		
8	YYAQEKQTDVNKLRYNFERTW	Cw*0501	Cw*0503	Cw*0505	Cw*0506 Cw*0509 Cw*0510
9	YDSREKQADVSNKLWYDSEWTW	Cw*0602	Cw*0607	Cw*0612	Cw*0613
10	YDSRENQADVSNLRYDSALTW	Cw*0701	Cw*0706	Cw*0716	Cw*0718 Cw*0721 Cw*0724
11	YDSREKQADVSNLRSDSALTW	Cw*0702	Cw*0710	Cw*0717	Cw*0723 Cw*0725
12	YDSREKQADVSNLRYDFADTW	Cw*0704	Cw*0711	Cw*0712	
13	YDSRENQADVSNKLRYDSALTW	Cw*0707	Cw*0709		
14	YDSREKQADVSNLRYDSALTW	Cw*0714	Cw*0727		
15	YYAQEKQTDVSNLRYNFTLTW	Cw*0801	Cw*0803		
16	YYAQEKQTDVSNLRYNFERTW	Cw*0802	Cw*0807		
17	YYAQEKQTDVSNLRYDSTLTW	Cw*0809	Cw*0811		
18	YYAREKQADVSNLWYDSEWTW	Cw*1203	Cw*1206	Cw*1213	
19	YSAREKQTDVSNLWFDSEWTW	Cw*1402	Cw*1403		
20	YYARENQTDVNKLRYDLELTW	Cw*1502	Cw*1510	Cw*1512	Cw*1513 Cw*1514
21	YYARENQTDVNKLRYDSELTW	Cw*1504	Cw*1509		
22	YYAREKQADVSNKLRYNFELEW	Cw*1701	Cw*1702	Cw*1703	
23	YDSREKQADVSNKLRFNFERTW	Cw*1801	Cw*1802		

The classification of class I HLA is based on IMGT/HLA Sequence Database Release 2.10 (released in July 2005).

CPFR: critical polymorphic functional residue; the polymorphic amino acid residues that directly contact with binding peptides in the pockets, placed according to their original positions on the HLA chains, designated in single capital letter.

HLA-Cw Alleles: The class I HLA-B alleles that belong to the group, with sequence based nomenclature begin with “Cw*” and followed by a four digit number.

4.6 Validation of Class I HLA Sub-supertypes

In order to validate the proposed hypothesis, HLA peptides binding data were extracted from MHCBN (Bhasin *et al.* 2003) and JENPEP (Blythe *et al.* 2002), where experimentally determined affinity data of HLA peptide binding are deposited. But limited number of peptides that bind to more than one HLA alleles were documented in these databases. Totally five peptides were identified: (1) FLWGPRALV, (2) AAGIGILTV, (3) GILGFVFTL, (4) VLYRYGSFSV and (5) YLEPGPVTA binding to A*0201 and A*0209 (**Table 4.5**). The peptides FLWGPRALV and AAGIGILTV are documented to have “high/moderate” binding affinity to A*0201 and A*0209. Two other peptides GILGFVFTL and YLEPGPVTA are documented to have similar T-cell activity in JENPEP (Blythe *et al.* 2002). It is also known that A*0201 and A*0209 are in one cluster (**Table 4.2**). This roughly suggests that the functional repertoires of HLA molecules are as a result of changes in the CPFRs and it is possible to maintain functional similarity by the conservation of CPFRs. It is also worth mentioning that no peptide could be found so far with contrasting binding affinities for alleles grouped in the same clusters to prove the proposed hypothesis otherwise (**Table 4.2**).

However, it is important to establish the binding of these peptides to other members of the group such as A*0224, A*0225, A*0231, A*0240, A*0243, A*0245, A*0246, A*0258, A*0259, A*0266, A*0267 and A*0268, etc. (**Table 4.2**). It should be noted that it is labor intensive to clone and express all of these alleles to determine binding to the above peptides. Nonetheless, if experimentally determined binding data show similar binding to all these alleles, then extrapolation of the proposed hypothesis to other predictive sub-supertype is trivial.

Another peptide GILGFVFTL was found, which binds to A*0206 and A*0214 (**Table 4.5**). These two alleles are categorized together in ‘Group #5’ (**Table 4.2**). Interestingly, GILGFVFTL binds A*0201, A*0209, A*0206 and A*0214 (**Table 4.5**). Although GILGFVFTL binds all the above four alleles, our strategy grouped these alleles into two distinct groups (Group #3 (A*0201 & A*0209) and Group #5 (A*0206 & A*0214)). Their functional overlap was further investigated by examining their CPFR segments. Examination of CPFR segments in Group #3 and Group #5 (**Table 4.2**) shows a single residue mutation (9F to 9Y) between these two groups. F (phenylalanine) and Y (tyrosine) are aromatic and the mutation is synonymous. This explains binding overlap between members of Group #3 and Group #5. By definition, the strategy groups members with similar peptide binding and members across groups (for example, Group #3 and Group #5) do not always bind similar peptides. It was also found that a number of peptides (LLFNILGGWV, YLVAYQATV, KVAELVHFL, FLWGPRALV, FLLLADARV, IMIGVLVGV) bind to A*0201 (Group #3) and A*0206 (Group #5). A*0201 and A*0206 are grouped as members of A2 supertype defined based on binding data (**Table 2.1**). The data implies that some peptides binding to A*0201 are also found to bind A*0206. However, this is not always true. Peptides (WMNRLIAFA, ILHNGAYSL, YLSGANLNL, VMAGVGSPYV, ILAGYGAGV) that bind to A*0201 do not always bind to A*0206 (**Table 2.1**). This suggests that A*0201 and A*0206 are not strict members of the A2 “sub-supertype”.

It was suggested that peptides that bind to A*0201 should strictly bind to A*0209 as they share identical CPFR segments (**Table 4.2**). Similarly, A*0206 and A*0214 should always bind similar peptides as they share identical CPFR segments.

The CPFR segments in A*0201 (Group #3) and A*0202 (Group #4) show single residue mutation (156L to 156W). The residues leucine (L) and tryptophan (W) are hydrophobic and the mutation is synonymous. Therefore, some peptides that bind to A*0201 are also found to bind to A*0202 (**Table 2.1**). This trend is not always true (**Table 2.1**). Therefore, A*0201 and A*0202 are not strict members of the A2 “sub-supertype”. In yet another case, the peptide RASGVDRYY binds A*3002 and A*3003 (**Table 4.5**). It shows that A*3002 and A*3003 are grouped together with another allele A*3012 (Group #26) and members of this group are proposed to bind similar peptides (**Table 4.2**). Therefore, the binding of this peptide to A*3012 will be of great significance to validate this hypothesis. The grouping for B alleles were further validated using binding data for 2 pairs of alleles (B*2703/B*2705 and B*5101/B*5102) and the data is given in **Table 4.6**. The pairs B*2703/B*2705 (Group #29) and B*5101/B*5102 (Group #61) strictly bind similar peptides (**Tables 4.3** and **Table 4.6**). More than 20 peptides are shown to bind each of the allele pairs clustered in Group #29 and Group #61. However, the binding of other members of the group is required for further extrapolation.

Here, it was demonstrated using binding data for 3 pairs of HLA-A alleles (A*0201/A*0209 (Group #3), A*0206/A*0214 (Group #5) and A*3002/A*3003 (Group #26) that alleles within a group (**Table 4.2**) bind similar peptides (**Table 4.5**). The validation was extended to two pairs of B alleles B*2703/B*2705 and B*5101/B*5102 (**Table 4.3** and **Table 4.6**). Thus, the methodology described here grouped alleles into strict supertypes, where members always bind similar peptides (**Tables 4.2** and **Table 4.5**). It was also noted that there may be some degree of functional overlap (**Table 4.5**) across members of different groups (A*0201 (Group #3)

/ A*0206 (Group #5)) due to synonymous residue substitutions at the CPFRP (**Table 4.2**). Several peptide binders of A*0201 are non-binders to A*0206 (**Table 2.1**). Hence, the functional overlap across groups (**Table 2.1** and **Table 4.2**) is not always true. This explained why the grouping of HLA alleles into supertypes based on binding data is seldom conclusive and comprehensive (**Table 2.1**).

The validation of the grouping strategy is limited in the current work. Further validation is required to apply this methodology to group alleles into “sub-supertypes” from sequence information on a large scale. However, the results here suggest that it is possible to quantify the functional difference between HLA alleles and develop better methodologies for HLA-peptide binding predictions. This work provides a framework to classify class I pockets based on peptide binding principles, which have more functional generality than the alleles themselves. The current approach demonstrates a several-fold reduction in pocket variability among HLA alleles and the reduction approach will help in improving our capability to predict peptide binding to a range of HLA alleles. The challenge lies in generating CPFRPs for each allele, clustering into different pockets, storing in the form of lookup tables and establishing the common relation between functional pockets for the accurate scanning of binding functions between them. The method thus demonstrates the useful application of structural information in understanding the functional consequences of sequence polymorphism in HLA molecules.

Table 4.5 Validation of HLA-A supertypes.

Group	Peptide	Allele Pair		Binding Affinity	T-cell Activity
Category #1					
3	FLWGPRALV	A*0201	A*0209	YES (HIGH/?)	YES
	AAGIGILTV	A*0201	A*0209	YES (MOD/HIGH)	YES
	GILGFVFTL	A*0201	A*0209	YES (HIGH)	YES
	VLYRYGSFSV	A*0201	A*0209	YES (HIGH)	YES
	YLEPGPVTA	A*0201	A*0209	YES	YES
Category #2					
5	GILGFVFTL	A*0206	A*0214	YES (?/LOW)	MOD
Category #3					
26	RISGVDRYY	A*3002	A*3003	YES	?

MOD: moderate;

?: Undetermined;

The group numbers indicated in this Table refer to the group numbers designated in **Table 4.2**.

Table 4.6 Validation of HLA-B supertypes.

Group	Peptide	Allele Pair		Binding Affinity	T-cell Activity
Category #1					
27	ARHGFLPRH	B*2703	B*2705	YES (MOD)	?
	ARTAHYGSL	B*2703	B*2705	YES	?
	ARYQKSTEL	B*2703	B*2705	YES	?
	FQYNGLIHR	B*2703	B*2705	YES	?
	FRYNGLIHR	B*2703	B*2705	YES	?
	GRAFVTIGA	B*2703	B*2705	YES	?
	GRAFVTIGK	B*2703	B*2705	YES	?
	GRERFEMER	B*2703	B*2705	YES (MOD)	?
	GRFFGGDRG	B*2703	B*2705	YES	?
	GRGLSLSRF	B*2703	B*2705	YES	?
	GRIDKPILA	B*2703	B*2705	YES	?
	GRIDKPILK	B*2703	B*2705	YES	?
	SRAHSSHLK	B*2703	B*2705	YES	?
	SRFSWGAEG	B*2703	B*2705	YES (LOW)	?
	SRHKKLMFK	B*2703	B*2705	YES (HIGH)	?
	SRSGSPMAR	B*2703	B*2705	YES (MOD)	?
	SRYWAITR	B*2703	B*2705	YES	?/YES
	VRRCPPHER	B*2703	B*2705	YES (LOW/MOD)	?
	VRVCACPGR	B*2703	B*2705	YES (MOD)	?
	RRYQKSTEL	B*2703	B*2705	YES	?
	QRHGSKYLA	B*2703	B*2705	YES (LOW/MOD)	?
	RRIKEIVKK	B*2703	B*2705	YES (MOD)	?
	RRTEENLR	B*2703	B*2705	YES (MOD)	?
Category #2					
56	FPISPIETV	B*5101	B*5102	YES (HIGH/?)	?
	FPVRPQVPL	B*5101	B*5102	YES	?
	FPVRPQVPL	B*5101	B*5102	YES	?
	CPKVSFEPI	B*5101	B*5102	YES	?
	CPSGHAVGI	B*5101	B*5102	YES	?
	DARAYDTEV	B*5101	B*5102	YES (HIGH/?)	NO/?
	EPLDLPQIIB	B*5101	B*5102	YES	?
	YPFKPPKVB	B*5101	B*5102	YES	?
	IPLGDAKLV	B*5101	B*5102	YES	?
	IPTSGDVVI	B*5101	B*5102	YES	?
	LPALSTGLI	B*5101	B*5102	YES	?
	LPCRIKQIIB	B*5101	B*5102	YES (LOW/?)	?
	LPEKDSWTV	B*5101	B*5102	YES	?
	LPPLERLTL	B*5101	B*5102	YES	?
	LPPTTGPIIB	B*5101	B*5102	YES	?
	LPPVVAKEI	B*5101	B*5102	YES (HIGH)	?

NALFRNLDV	B*5101	B*5102	YES	?
NANPDCKTI	B*5101	B*5102	YES	?
NPPIPVGEIB	B*5101	B*5102	YES	?
QGWKGSPIAI	B*5101	B*5102	YES	?
TAVQMAVFI	B*5101	B*5102	YES	?
TGYLNTVTV	B*5101	B*5102	YES	?
VAQRAYRAI	B*5101	B*5102	YES	?
VGCLVGLRI	B*5101	B*5102	YES	?
VPVKLKPGM	B*5101	B*5102	YES	?
YAPPIGGQI	B*5101	B*5102	YES	?
YPCTVNFTI	B*5101	B*5102	YES	?
YPLASLKSL	B*5101	B*5102	YES	?
YPLTSLRSL	B*5101	B*5102	YES	?
APTLWARM	B*5101	B*5102	YES	?

MOD: moderate;

? : Undetermined;

The group numbers indicated in this Table refer to the group numbers designated in **Table 4.3**.

4.7 Discussion

HLA molecules are highly polymorphic and this is evident from the growth in HLA sequence data at IMGT/HLA (Marsh 2003). HLA-A and HLA-B are most polymorphic with currently more 300 and 600 alleles defined by DNA sequence analysis, respectively. HLA molecules are interesting pharmaceutical targets for vaccination and immuno-therapeutics through the design of peptide ligands. The characteristic interaction between peptide and HLA provide the molecular basis for allele-specific recognition of antigenic peptides. The peptide residues that are held by a certain pocket are determined by the physical and chemical properties of the pocket. Nearly 5.12×10^{11} combinations of 9 residue long peptide stretches can be formed from the naturally occurring 20 amino acids. Only a very small fraction of the peptide pool binds to each of the known alleles and the interactions between peptide and alleles are highly combinatorial. Therefore, combinatorial reduction approaches will provide roadmaps to quantify functional variations among HLA alleles.

The application of molecular modeling in HLA-peptide binding prediction utilized hydrophobic and hydrophilic interaction functions for scoring. Estimation of hydrophilic interaction between peptide and HLA still remains a great challenge. From current analysis, it is found that the residues at the CPFRPs in the HLA-A alleles are mostly hydrophilic and a preference for hydrophilic to hydrophilic residue substitution is observed. At positions 97, 114, 156 and 163, a very high amino acid variability of 9 is observed, and therefore pockets containing these CPFRPs contribute more towards the functional difference between alleles. Although, the alleles show considerable residue identity at the CPFRPs, the subtle dissimilarity at the CPFRPs among alleles is

the cause for considerable differences in peptide binding function. Allele specific knowledge on the chemical properties and residue variability of the CPFRPs constituting the functional pockets are crucial in HLA-peptide binding quantification. Thus, a systematic approach to quantify the critical difference between alleles will help to extrapolate the binding affinities of known HLA-peptide interactions to new HLA alleles with unknown binding information.

In the current approach, residues at the 21 CPFRPs were used to differentiate HLA alleles, and thus classified the HLA alleles into different clusters based on the residue compositions in each of the functional pockets (**Table 4.1**). Each molecule has 6 pockets and therefore, 331 HLA-A alleles constitute 1,986 pockets. However, the identification and utilization of 21 CPFRPs generated only 247 unique pockets from 331 HLA-A alleles. The methodology described here produced an eight-fold reduction for functional pocket variability among HLA-A alleles. This implies that the HLA-A alleles are functionally and quantitatively similar though not identical. The results also suggest that functional pocket variations between alleles are mainly in specific pockets. The limited number of unique pockets among HLA-A alleles presents a promising solution for reducing the large possibility of peptides binding to HLA-A molecules. When the 331 HLA-A alleles are classified based on the compositions at the CPFRPs in each pocket, 47 different types of pocket A; 61 types of pocket B; 39 types of pocket C; 39 types of pocket D; 47 types of pocket E and 14 types of pocket F, were identified. The grouping of HLA-A molecules based on the individual pocket can provide a strong tool for the understanding of pocket features. For example, pocket F is the most conserved with just 14 different clusters and pocket B is the least conserved with 61 clusters. A cluster for pocket F with residue composition “D-T-L-D” occurs in

99 different alleles. This implies that many alleles share functionally relevant residues at different pockets among alleles. The classification presented in this work will enable us to compare and contrast the functional similarities and differences between functional pockets among HLA alleles in a combinatorial manner. In other words one clusters of alleles that share all the 6 pockets, only 5 pockets, only 4 pockets, only 3 pockets, only 2 pockets, and only 1 pocket, can be found. This exercise will result in large combinations of alleles with different functional pocket categories. Based on the functional pocket similarity, strategies can be designed to extrapolate properties from known cases to unknown cases. It is found that 185 HLA-A alleles are grouped into 42 clusters by performing the simplest exercise of finding clusters of alleles with common A–F functional pockets (**Table 4.2**). HLA-B (**Table 4.3**) and HLA-Cw (**Table 4.4**) alleles are also grouped into small number of clusters by the same strategy.

This implies that a given peptide will bind to alleles in the cluster with similar binding affinities. In other words, alleles in the same group will have similar binding patterns with a defined peptide. If the binding data of a given peptide is known for one allele in each cluster then the predicted binding of that peptide to other alleles can be easily and efficiently extrapolated. However, this inference needs experimental testing and validation. With the help of effective cross validation, predicted binding data of a defined peptide for nearly 50% of class I HLA alleles can be generated using binding data for that peptide to at least one allele in each of the clusters.

CHAPTER 5

CONCLUSIONS

5.1 Achievements

The current challenge in peptide vaccine design is the identification of a minimum number of short antigenic peptides capable of binding to multiple HLA alleles for efficient and broad range immune responses. The theoretically possible combinations are overwhelmingly large, due to the extreme polymorphism of HLA and the flexible HLA peptide binding specificity. It is a challenging task to develop metrics to quantify functional differences among alleles, and to develop models for HLA peptide binding prediction. Many allele specific data driven statistical models and 3D structure based procedures are available for HLAp binding prediction and calculations. Data driven statistical methods are generally available for a limited number of HLA alleles, such as A*0201, A*1101, A*0301, B*0801, DR1, and DR4. 3D structure based procedures are generally adopted for any given allele because the prediction is based on molecular energy functions. Structural procedures utilizing structure prediction by energy minimization followed by protein-ligand interaction calculations are computationally intensive. Moreover, robust protein-ligand interaction functions are not currently available.

The method described in this work circumvents these problems because the prediction model developed is simple, fast, and robust with easy implementation. To develop this prediction model that can be applied to a broad spectrum of HLA alleles, the concept of structural pocket was extended to virtual functional pockets. The amino acid residues that form the virtual pockets determine the size, shape and charge of the pockets and thus determine the antigenic peptides that would be preferentially bound by the HLA allotype. Thus, the peptide-binding groove of the HLA is essentially an exchange or shuffling of pockets among different allotypes. Virtual pockets are defined using information gleaned from eight unique HLA alleles and the mapping of virtual pocket to position specific residue anchors is done using the HLAp structures analyzed in this work. The virtual pockets defined by this procedure is extrapolated to other class I alleles because HLA molecules are highly homologous, and homologous proteins have conserved backbone trace. The peptide residue – virtual binding pocket compatibility is estimated using the Q matrix. The average PPV of the model is 89%, whereas the average NPV is only 18%. Despite the limitations set by the model, the method described here is novel, generic, and simple. The model was extensively validated for class I alleles in this work. The procedure was also extended to class II HLA alleles and the developed prediction model is slightly cross-validated due to limited availability of class II HLA peptide binding data.

5.2 Limitations and Future Improvements

An efficient model for HLA peptide binding prediction has been developed in the

current work, but there is room for further improvement. The model development in this work is based on the definition of virtual binding pockets, mapping of virtual pockets to position specific peptide residue anchors, and estimation of peptide residue – virtual pocket compatibility. When more structural data are available in the future, the virtual binding pockets will be defined more accurately. Moreover, there is a relatively high occurrence of False Negative prediction value in the cross validation of the model. This probably is introduced by the noise in the Q matrix used for scoring as Q matrix is a similarity matrix. Improvements in the Q matrix by fine tuning the interacting pairs will remove false negatives and improve sensitivity significantly. Future work is required to reevaluate and accordingly modify the Q matrix for further developments. The model also needs to be further validated with a dataset containing more non-binders. Furthermore, class II HLA peptide binding prediction should be addressed when sufficient structural data and peptide binding data are available.

The functional overlap among alleles and the grouping of alleles into “sub-supertypes” is crucial to the description and prediction of peptide binding function of HLA alleles, and is extremely powerful in understanding peptide selection and degeneration. However, current grouping of alleles into supertypes using experimental binding data is seldom conclusive and comprehensive. In this work, the HLA alleles are grouped into sub-supertypes based on residues at 21 CPFRPs, which were identified to play a crucial role in determining functional diversity among HLA molecules for a given peptide than those at other positions in the grooves. The HLA residues on CPFRPs construct functional pockets, whose geometry and chemical properties cumulatively determine the peptide that can be preferentially bound. The identification of less number of functional pockets among large number of HLA alleles suggests that many

HLA alleles share considerable number of functional pockets among them. In other words, the functional diversity of individual pockets is less than that of the primary sequences themselves. Classified by the procedure described in this work, HLA alleles demonstrated nearly a twofold functional reduction. Above all, 56% of known HLA-A alleles (331), 50% of known HLA-B alleles (604) and 47% of known HLA-Cw alleles (160) are clustered to just 42, 78 and 23 groups, respectively.

This grouping procedure is useful in that alleles in the same cluster should bind with similar affinity (if not identical) for a given peptide. If the binding strength of a given peptide to any of the allele is known in the cluster, then its binding strength to other alleles in the cluster can be easily inferred. The findings demonstrate the functional differences and similarities among HLA alleles. Based on this work, the binding of a peptide to approximately 50% of all known alleles can be inferred using a handful of binding data representing all predictive sub-supertypes.

The classification of class I HLA specificities described in this work has been moderately cross-validated by peptide binding data. However, it is different from the supertype classifications based on peptide binding data, which are different among themselves. A comprehensive cross validation is required for large scale extrapolation. It is also hoped that consensus among these procedures can be achieved in future investigations. Another concern is the classification for class II HLA. Due to the flexible peptide binding ability of the class II HLA allotypes, and limited availability of class II HLA structural data, the classification procedure is not extended to class II HLA alleles at the current stage. It is hoped that this issue can be addressed in the future when relevant data are available.

REFERENCES

- Abagyan, RA and Batalov, S (1997). Do aligned sequences share the same fold? *J Mol Biol* **273**(1): 355-68.
- Adams, HP and Koziol, JA (1995). Prediction of binding to MHC class I molecules. *J Immunol Methods* **185**(2): 181-90.
- Adrian, PE, Rajaseger, G, Mathura, VS, Sakharkar, MK and Kanguane, P (2002). Types of inter-atomic interactions at the MHC-peptide interface: identifying commonality from accumulated data. *BMC Struct Biol* **2**(1): 2.
- Alexander, J, Del Guercio, MF, Fikes, JD, Chesnut, RW, Chisari, FV, Chang, KM, Appella, E and Sette, A (1998). Recognition of a novel naturally processed, A2 restricted, HCV-NS4 epitope triggers IFN-gamma release in absence of detectable cytopathicity. *Hum Immunol* **59**(12): 776-82.
- Altuvia, Y, Schueler, O and Margalit, H (1995). Ranking potential binding peptides to MHC molecules by a computational threading approach. *J Mol Biol* **249**(2): 244-50.
- Altuvia, Y, Sette, A, Sidney, J, Southwood, S and Margalit, H (1997). A structure-based algorithm to predict potential binding peptides to MHC molecules with hydrophobic binding pockets. *Hum Immunol* **58**(1): 1-11.
- Appay, V, Nixon, DF, Donahoe, SM, Gillespie, GM, Dong, T, King, A, Ogg, GS, Spiegel, HM, Conlon, C, Spina, CA, Havlir, DV, Richman, DD, Waters, A, Easterbrook, P, McMichael, AJ and Rowland-Jones, SL (2000). HIV-specific CD8(+) T cells produce antiviral cytokines but are impaired in cytolytic function. *J Exp Med* **192**(1): 63-75.
- Aubert, G, Hassan-Walker, AF, Madrigal, JA, Emery, VC, Morte, C, Grace, S, Koh, MB, Potter, M, Prentice, HG, Dodi, IA and Travers, PJ (2001). Cytomegalovirus-specific cellular immune responses and viremia in recipients of allogeneic stem cell transplants. *J Infect Dis* **184**(8): 955-63.
- Barber, LD, Gillece-Castro, B, Percival, L, Li, X, Clayberger, C and Parham, P (1995). Overlap in the repertoires of peptides bound in vivo by a group of related class I HLA-B allotypes. *Curr Biol* **5**(2): 179-90.
- Batalia, MA and Collins, EJ (1997). Peptide binding by class I and class II MHC molecules. *Biopolymers* **43**(4): 281-302.
- Berman, HM, Battistuz, T, Bhat, TN, Bluhm, WF, Bourne, PE, Burkhardt, K, Feng, Z, Gilliland, GL, Iype, L, Jain, S, Fagan, P, Marvin, J, Padilla, D, Ravichandran, V,

- Schneider, B, Thanki, N, Weissig, H, Westbrook, JD and Zardecki, C (2002). The Protein Data Bank. *Acta Crystallogr D Biol Crystallogr* **58**(Pt 6 No 1): 899-907.
- Betancourt, MR and Thirumalai, D (1999). Pair potentials for protein folding: choice of reference states and sensitivity of predicted native states to variations in the interaction schemes. *Protein Sci* **8**(2): 361-9.
- Bhasin, M, Singh, H and Raghava, GP (2003). MHCBN: a comprehensive database of MHC binding and non-binding peptides. *Bioinformatics* **19**(5): 665-6.
- Bjorkman, PJ, Saper, MA, Samraoui, B, Bennett, WS, Strominger, JL and Wiley, DC (1987a). The foreign antigen binding site and T cell recognition regions of class I histocompatibility antigens. *Nature* **329**(6139): 512-8.
- Bjorkman, PJ, Saper, MA, Samraoui, B, Bennett, WS, Strominger, JL and Wiley, DC (1987b). Structure of the human class I histocompatibility antigen, HLA-A2. *Nature* **329**(6139): 506-12.
- Blythe, MJ, Doytchinova, IA and Flower, DR (2002). JenPep: a database of quantitative functional peptide data for immunology. *Bioinformatics* **18**(3): 434-9.
- Bonelo, A, Valmori, D, Triponez, F, Tiercy, JM, Mentha, G, Oberholzer, J, Champagne, P, Romero, JF, Esposito, F, Nebie, I, Barbey, C, Romero, P, Herrera, S, Corradin, G and Lopez, JA (2000). Generation and characterization of malaria-specific human CD8(+) lymphocyte clones: effect of natural polymorphism on T cell recognition and endogenous cognate antigen presentation by liver cells. *Eur J Immunol* **30**(11): 3079-88.
- Bouvier, M and Wiley, DC (1994). Importance of peptide amino and carboxyl termini to the stability of MHC class I molecules. *Science* **265**(5170): 398-402.
- Brusic, V, Rudy, G, Honeyman, G, Hammer, J and Harrison, L (1998). Prediction of MHC class II-binding peptides using an evolutionary algorithm and artificial neural network. *Bioinformatics* **14**(2): 121-30.
- Brusic, V, Petrovsky, N, Zhang, G and Bajic, VB (2002). Prediction of promiscuous peptides that bind HLA class I molecules. *Immunol Cell Biol* **80**(3): 280-5.
- Catalina, MD, Sullivan, JL, Bak, KR and Luzuriaga, K (2001). Differential evolution and stability of epitope-specific CD8(+) T cell responses in EBV infection. *J Immunol* **167**(8): 4450-7.
- Chang, KM, Gruener, NH, Southwood, S, Sidney, J, Pape, GR, Chisari, FV and Sette, A (1999). Identification of HLA-A3 and -B7-restricted CTL response to hepatitis C virus in patients with acute and chronic hepatitis C. *J Immunol* **162**(2): 1156-64.

- Chen, W, Khilko, S, Fecondo, J, Margulies, DH and McCluskey, J (1994). Determinant selection of major histocompatibility complex class I-restricted antigenic peptides is explained by class I-peptide affinity and is strongly influenced by nondominant anchor residues. *J Exp Med* **180**(4): 1471-83.
- Cochlovius, B, Stassar, M, Christ, O, Raddrizzani, L, Hammer, J, Mytilineos, I and Zoller, M (2000). In vitro and in vivo induction of a Th cell response toward peptides of the melanoma-associated glycoprotein 100 protein selected by the TEPITOPE program. *J Immunol* **165**(8): 4731-41.
- Colbert, RA, Rowland-Jones, SL, McMichael, AJ and Frelinger, JA (1993). Allele-specific B pocket transplant in class I major histocompatibility complex protein changes requirement for anchor residue at P2 of peptide. *Proc Natl Acad Sci U S A* **90**(14): 6879-83.
- Colbert, RA, Rowland-Jones, SL, McMichael, AJ and Frelinger, JA (1994). Differences in peptide presentation between B27 subtypes: the importance of the P1 side chain in maintaining high affinity peptide binding to B*2703. *Immunity* **1**(2): 121-30.
- Collins, EJ, Garboczi, DN, Karpusas, MN and Wiley, DC (1995). The three-dimensional structure of a class I major histocompatibility complex molecule missing the alpha 3 domain of the heavy chain. *Proc Natl Acad Sci U S A* **92**(4): 1218-21.
- Cornette, JL, Cease, KB, Margalit, H, Spouge, JL, Berzofsky, JA and DeLisi, C (1987). Hydrophobicity scales and computational techniques for detecting amphipathic structures in proteins. *J Mol Biol* **195**(3): 659-85.
- Cox, AL, Skipper, J, Chen, Y, Henderson, RA, Darrow, TL, Shabanowitz, J, Engelhard, VH, Hunt, DF and Slingluff, CL, Jr. (1994). Identification of a peptide recognized by five melanoma-specific human cytotoxic T cell lines. *Science* **264**(5159): 716-9.
- D'Amaro, J, Houbiers, JG, Drijfhout, JW, Brandt, RM, Schipper, R, Bavinck, JN, Melief, CJ and Kast, WM (1995). A computer program for predicting possible cytotoxic T lymphocyte epitopes based on HLA class I peptide-binding motifs. *Hum Immunol* **43**(1): 13-8.
- DiBrino, M, Parker, KC, Shiloach, J, Turner, RV, Tsuchida, T, Garfield, M, Biddison, WE and Coligan, JE (1994). Endogenous peptides with distinct amino acid anchor residue motifs bind to HLA-A1 and HLA-B8. *J Immunol* **152**(2): 620-31.
- Donnes, P and Elofsson, A (2002). Prediction of MHC class I binding peptides, using SVMHC. *BMC Bioinformatics* **3**: 25.
- Dorrell, L, Willcox, BE, Jones, EY, Gillespie, G, Njai, H, Sabally, S, Jaye, A, DeGleria,

- K, Rostron, T, Lepin, E, McMichael, A, Whittle, H and Rowland-Jones, S (2001). Cytotoxic T lymphocytes recognize structurally diverse, clade-specific and cross-reactive peptides in human immunodeficiency virus type-1 gag through HLA-B53. *Eur J Immunol* **31**(6): 1747-56.
- Doytchinova, IA and Flower, DR (2001). Toward the quantitative prediction of T-cell epitopes: coMFA and coMSIA studies of peptides with affinity for the class I MHC molecule HLA-A*0201. *J Med Chem* **44**(22): 3572-81.
- Doytchinova, IA, Blythe, MJ and Flower, DR (2002a). Additive method for the prediction of protein-peptide binding affinity. Application to the MHC class I molecule HLA-A*0201. *J Proteome Res* **1**(3): 263-72.
- Doytchinova, IA and Flower, DR (2002b). A comparative molecular similarity index analysis (CoMSIA) study identifies an HLA-A2 binding supermotif. *J Comput Aided Mol Des* **16**(8-9): 535-44.
- Faint, JM, Annels, NE, Curnow, SJ, Shields, P, Pilling, D, Hislop, AD, Wu, L, Akbar, AN, Buckley, CD, Moss, PA, Adams, DH, Rickinson, AB and Salmon, M (2001). Memory T cells constitute a subset of the human CD8+CD45RA+ pool with distinct phenotypic and migratory characteristics. *J Immunol* **167**(1): 212-20.
- Falk, K, Rotzschke, O and Rammensee, HG (1990). Cellular peptide composition governed by major histocompatibility complex class I molecules. *Nature* **348**(6298): 248-51.
- Feltkamp, MC, Vierboom, MP, Kast, WM and Melief, CJ (1994). Efficient MHC class I-peptide binding is required but does not ensure MHC class I-restricted immunogenicity. *Mol Immunol* **31**(18): 1391-401.
- Free, SM, Jr. and Wilson, JW (1964). A Mathematical Contribution to Structure-Activity Studies. *J Med Chem* **53**: 395-9.
- Fremont, DH, Stura, EA, Matsumura, M, Peterson, PA and Wilson, IA (1995). Crystal structure of an H-2Kb-ovalbumin peptide complex reveals the interplay of primary and secondary anchor positions in the major histocompatibility complex binding groove. *Proc Natl Acad Sci U S A* **92**(7): 2479-83.
- Fruci, D, Rovero, P, Butler, RH, Sorrentino, R, Tosi, R and Tanigaki, N (1993). HLA class I binding of synthetic nonamer peptides carrying major anchor residue motifs of HLA-B27 (B*2705)-binding peptides. *Immunogenetics* **38**(1): 41-6.
- Gao, GF, Tormo, J, Gerth, UC, Wyer, JR, McMichael, AJ, Stuart, DI, Bell, JI, Jones, EY and Jakobsen, BK (1997). Crystal structure of the complex between human CD8alpha(alpha) and HLA-A2. *Nature* **387**(6633): 630-4.
- Garboczi, DN, Ghosh, P, Utz, U, Fan, QR, Biddison, WE and Wiley, DC (1996).

- Structure of the complex between human T-cell receptor, viral peptide and HLA-A2. *Nature* **384**(6605): 134-41.
- Garcia, KC, Teyton, L and Wilson, IA (1999). Structural basis of T cell recognition. *Annu Rev Immunol* **17**: 369-97.
- Garrett, TP, Saper, MA, Bjorkman, PJ, Strominger, JL and Wiley, DC (1989a). Specificity pockets for the side chains of peptide antigens in HLA-Aw68. *Nature* **342**(6250): 692-6.
- Garrett, TP, Saper, MA, Samraoui, B, Bjorkman, PJ and Wiley, DC (1989b). Structure/function studies of major histocompatibility antigens. *Transplant Proc* **21**(1 Pt 1): 588-90.
- Germain, RN and Margulies, DH (1993). The biochemistry and cell biology of antigen processing and presentation. *Annu Rev Immunol* **11**: 403-50.
- Govindarajan, KR, Kanguene, P, Tan, TW and Ranganathan, S (2003). MPID: MHC-Peptide Interaction Database for sequence-structure-function information on peptides binding to MHC molecules. *Bioinformatics* **19**(2): 309-10.
- Grabowska, AM, Lechner, F, Klenerman, P, Tighe, PJ, Ryder, S, Ball, JK, Thomson, BJ, Irving, WL and Robins, RA (2001). Direct ex vivo comparison of the breadth and specificity of the T cells in the liver and peripheral blood of patients with chronic HCV infection. *Eur J Immunol* **31**(8): 2388-94.
- Gribskov, M, McLachlan, AD and Eisenberg, D (1987). Profile analysis: detection of distantly related proteins. *Proc Natl Acad Sci U S A* **84**(13): 4355-8.
- Gross, DM, Forsthuber, T, Tary-Lehmann, M, Etling, C, Ito, K, Nagy, ZA, Field, JA, Steere, AC and Huber, BT (1998). Identification of LFA-1 as a candidate autoantigen in treatment-resistant Lyme arthritis. *Science* **281**(5377): 703-6.
- Guan, P, Doytchinova, IA and Flower, DR (2003). HLA-A3 supermotif defined by quantitative structure-activity relationship analysis. *Protein Eng* **16**(1): 11-8.
- Gulukota, K, Sidney, J, Sette, A and DeLisi, C (1997). Two complementary methods for predicting peptides binding major histocompatibility complex molecules. *J Mol Biol* **267**(5): 1258-67.
- Guo, HC, Jardetzky, TS, Garrett, TP, Lane, WS, Strominger, JL and Wiley, DC (1992). Different length peptides bind to HLA-Aw68 similarly at their ends but bulge out in the middle. *Nature* **360**(6402): 364-6.
- Guo, HC, Madden, DR, Silver, ML, Jardetzky, TS, Gorga, JC, Strominger, JL and Wiley, DC (1993). Comparison of the P2 specificity pocket in three human histocompatibility antigens: HLA-A*6801, HLA-A*0201, and HLA-B*2705. *Proc Natl Acad Sci U S A* **90**(17): 8053-7.

- Hammer, J, Gallazzi, F, Bono, E, Karr, RW, Guenot, J, Valsasnini, P, Nagy, ZA and Sinigaglia, F (1995). Peptide binding specificity of HLA-DR4 molecules: correlation with rheumatoid arthritis association. *J Exp Med* **181**(5): 1847-55.
- Hassan-Walker, AF, Vargas Cuero, AL, Mattes, FM, Klenerman, P, Lechner, F, Burroughs, AK, Griffiths, PD, Phillips, RE and Emery, VC (2001). CD8+ cytotoxic lymphocyte responses against cytomegalovirus after liver transplantation: correlation with time from transplant to receipt of tacrolimus. *J Infect Dis* **183**(6): 835-43.
- Hislop, AD, Annels, NE, Gudgeon, NH, Leese, AM and Rickinson, AB (2002). Epitope-specific evolution of human CD8(+) T cell responses from primary to persistent phases of Epstein-Barr virus infection. *J Exp Med* **195**(7): 893-905.
- Hunt, DF, Michel, H, Dickinson, TA, Shabanowitz, J, Cox, AL, Sakaguchi, K, Appella, E, Grey, HM and Sette, A (1992). Peptides presented to the immune system by the murine class II major histocompatibility complex molecule I-Ad. *Science* **256**(5065): 1817-20.
- Jardetzky, TS, Lane, WS, Robinson, RA, Madden, DR and Wiley, DC (1991). Identification of self peptides bound to purified HLA-B27. *Nature* **353**(6342): 326-9.
- Jernigan, RL and Bahar, I (1996). Structure-derived potentials and protein simulations. *Curr Opin Struct Biol* **6**(2): 195-209.
- Jones, DT and Thornton, JM (1996). Potential energy functions for threading. *Curr Opin Struct Biol* **6**(2): 210-6.
- Kangueane, P, Sakharkar, MK, Lim, KS, Hao, H, Lin, K, Chee, RE and Kolatkar, PR (2000). Knowledge-based grouping of modeled HLA peptide complexes. *Hum Immunol* **61**(5): 460-6.
- Kangueane, P, Sakharkar, MK, Kolatkar, PR and Ren, EC (2001). Towards the MHC-peptide combinatorics. *Hum Immunol* **62**(5): 539-56.
- Kast, WM, Brandt, RM, Sidney, J, Drijfhout, JW, Kubo, RT, Grey, HM, Melief, CJ and Sette, A (1994). Role of HLA-A motifs in identification of potential CTL epitopes in human papillomavirus type 16 E6 and E7 proteins. *J Immunol* **152**(8): 3904-12.
- Kaul, R, Rowland-Jones, SL, Gillespie, G, Kimani, J, Dong, T, Kiama, P, Simonsen, JN, Bwayo, JJ, McMichael, AJ and Plummer, FA (2002). Gonococcal cervicitis is associated with reduced systemic CD8+ T cell responses in human immunodeficiency virus type 1-infected and exposed, uninfected sex workers. *J Infect Dis* **185**(10): 1525-9.
- Kawashima, I, Hudson, SJ, Tsai, V, Southwood, S, Takesako, K, Appella, E, Sette, A

- and Celis, E (1998). The multi-epitope approach for immunotherapy for cancer: identification of several CTL epitopes from various tumor-associated antigens expressed on solid epithelial tumors. *Hum Immunol* **59**(1): 1-14.
- Kim, SK, Devine, L, Angevine, M, DeMars, R and Kavathas, PB (2000). Direct detection and magnetic isolation of Chlamydia trachomatis major outer membrane protein-specific CD8⁺ CTLs with HLA class I tetramers. *J Immunol* **165**(12): 7285-92.
- Konig, R, Huang, LY and Germain, RN (1992). MHC class II interaction with CD4 mediated by a region analogous to the MHC class I binding site for CD8. *Nature* **356**(6372): 796-8.
- Kostense, S, Ogg, GS, Manting, EH, Gillespie, G, Joling, J, Vandenberghe, K, Veenhof, EZ, van Baarle, D, Jurriaans, S, Klein, MR and Miedema, F (2001). High viral burden in the presence of major HIV-specific CD8(+) T cell expansions: evidence for impaired CTL effector function. *Eur J Immunol* **31**(3): 677-86.
- Kubo, RT, Sette, A, Grey, HM, Appella, E, Sakaguchi, K, Zhu, NZ, Arnott, D, Sherman, N, Shabanowitz, J, Michel, H and et al. (1994). Definition of specific peptide motifs for four major HLA-A alleles. *J Immunol* **152**(8): 3913-24.
- Kuon, W, Holzhutter, HG, Appel, H, Grolms, M, Kollnberger, S, Traeder, A, Henklein, P, Weiss, E, Thiel, A, Lauster, R, Bowness, P, Radbruch, A, Kloetzel, PM and Sieper, J (2001). Identification of HLA-B27-restricted peptides from the Chlamydia trachomatis proteome with possible relevance to HLA-B27-associated diseases. *J Immunol* **167**(8): 4738-46.
- Lee, B and Richards, FM (1971). The interpretation of protein structures: estimation of static accessibility. *J Mol Biol* **55**(3): 379-400.
- Lee, C and McConnell, HM (1995). A general model of invariant chain association with class II major histocompatibility complex proteins. *Proc Natl Acad Sci U S A* **92**(18): 8269-73.
- Logean, A, Sette, A and Rognan, D (2001). Customized versus universal scoring functions: application to class I MHC-peptide binding free energy predictions. *Bioorg Med Chem Lett* **11**(5): 675-9.
- Logean, A and Rognan, D (2002). Recovery of known T-cell epitopes by computational scanning of a viral genome. *J Comput Aided Mol Des* **16**(4): 229-43.
- MacDonald, KS, Matukas, L, Embree, JE, Fowke, K, Kimani, J, Nagelkerke, NJ, Oyugi, J, Kiama, P, Kaul, R, Luscher, MA, Rowland-Jones, S, Ndinya-Achola, J, Ngugi, E, Bwayo, JJ and Plummer, FA (2001). Human leucocyte antigen supertypes and immune susceptibility to HIV-1, implications for vaccine design. *Immunol Lett* **79**(1-2): 151-7.

- Madden, DR, Gorga, JC, Strominger, JL and Wiley, DC (1992). The three-dimensional structure of HLA-B27 at 2.1 Å resolution suggests a general mechanism for tight peptide binding to MHC. *Cell* **70**(6): 1035-48.
- Madden, DR, Garboczi, DN and Wiley, DC (1993). The antigenic identity of peptide-MHC complexes: a comparison of the conformations of five viral peptides presented by HLA-A2. *Cell* **75**(4): 693-708.
- Madden, DR (1995). The three-dimensional structure of peptide-MHC complexes. *Annu Rev Immunol* **13**: 587-622.
- Mallios, RR (1999). Class II MHC quantitative binding motifs derived from a large molecular database with a versatile iterative stepwise discriminant analysis meta-algorithm. *Bioinformatics* **15**(6): 432-9.
- Mamitsuka, H (1998). Predicting peptides that bind to MHC molecules using supervised learning of hidden Markov models. *Proteins* **33**(4): 460-74.
- Marsh, SG (2003). HLA nomenclature and the IMGT/HLA sequence database. *Novartis Found Symp* **254**: 165-73; discussion 173-6, 216-22, 250-2.
- Matsui, M, Hioe, CE and Frelinger, JA (1993). Roles of the six peptide-binding pockets of the HLA-A2 molecule in allorecognition by human cytotoxic T-cell clones. *Proc Natl Acad Sci U S A* **90**(2): 674-8.
- Matsui, M, Moots, RJ, McMichael, AJ and Frelinger, JA (1994). Significance of the six peptide-binding pockets of HLA-A2.1 in influenza A matrix peptide-specific cytotoxic T-lymphocyte reactivity. *Hum Immunol* **41**(2): 160-6.
- Matsumura, M, Fremont, DH, Peterson, PA and Wilson, IA (1992). Emerging principles for the recognition of peptide antigens by MHC class I molecules. *Science* **257**(5072): 927-34.
- McCluskey, J and Peh, CA (1999). The human leucocyte antigens and clinical medicine: an overview. *Rev Immunogenet* **1**(1): 3-20.
- Meister, GE, Roberts, CG, Berzofsky, JA and De Groot, AS (1995). Two novel T cell epitope prediction algorithms based on MHC-binding motifs; comparison of predicted and published epitopes from Mycobacterium tuberculosis and HIV protein sequences. *Vaccine* **13**(6): 581-91.
- Milik, M, Sauer, D, Brunmark, AP, Yuan, L, Vitiello, A, Jackson, MR, Peterson, PA, Skolnick, J and Glass, CA (1998). Application of an artificial neural network to predict specific class I MHC binding peptide sequences. *Nat Biotechnol* **16**(8): 753-6.
- Miyazawa, S and Jernigan, RL (1985). Estimation of effective inter-residue contact energies from protein crystal structure, quasi-chemical approximation. *Macromolecules* **18**: 534-552.

- Miyazawa, S and Jernigan, RL (1996). Residue-residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading. *J Mol Biol* **256**(3): 623-44.
- Newton-Nash, DK and Eckels, DD (1993). Differential effect of polymorphism at HLA-DR1 beta-chain positions 85 and 86 on binding and recognition of DR1-restricted antigenic peptides. *J Immunol* **150**(5): 1813-21.
- Noguchi, H, Kato, R, Hanai, T, Matsubara, Y, Honda, H, Brusic, V and Kobayashi, T (2002). Hidden Markov model-based prediction of antigenic peptides that interact with MHC class II molecules. *J Biosci Bioeng*(94): 264-270.
- Pamer, E and Cresswell, P (1998). Mechanisms of MHC class I--restricted antigen processing. *Annu Rev Immunol* **16**: 323-58.
- Parker, KC, Bednarek, MA and Coligan, JE (1994). Scheme for ranking potential HLA-A2 binding peptides based on independent binding of individual peptide side-chains. *J Immunol* **152**(1): 163-75.
- Pogue, RR, Eron, J, Frelinger, JA and Matsui, M (1995). Amino-terminal alteration of the HLA-A*0201-restricted human immunodeficiency virus pol peptide increases complex stability and in vitro immunogenicity. *Proc Natl Acad Sci U S A* **92**(18): 8166-70.
- Prezzi, C, Casciaro, MA, Francavilla, V, Schiaffella, E, Finocchi, L, Chircu, LV, Bruno, G, Sette, A, Abrignani, S and Barnaba, V (2001). Virus-specific CD8(+) T cells with type 1 or type 2 cytokine profile are related to different disease activity in chronic hepatitis C virus infection. *Eur J Immunol* **31**(3): 894-906.
- Raddrizzani, L and Hammer, J (2000). Epitope scanning using virtual matrix-based algorithms. *Brief Bioinform* **1**(2): 179-89.
- Rammensee, H, Bachmann, J, Emmerich, NP, Bachor, OA and Stevanovic, S (1999). SYFPEITHI: database for MHC ligands and peptide motifs. *Immunogenetics* **50**(3-4): 213-9.
- Rammensee, HG, Friede, T and Stevanoviic, S (1995). MHC ligands and peptide motifs: first listing. *Immunogenetics* **41**(4): 178-228.
- Reche, PA, Glutting, JP and Reinherz, EL (2002). Prediction of MHC class I binding peptides using profile motifs. *Hum Immunol* **63**(9): 701-9.
- Reignat, S, Webster, GJ, Brown, D, Ogg, GS, King, A, Seneviratne, SL, Dusheiko, G, Williams, R, Maini, MK and Bertoletti, A (2002). Escaping high viral load exhaustion: CD8 cells with altered tetramer binding in chronic hepatitis B virus infection. *J Exp Med* **195**(9): 1089-101.
- Ren, EC, Kanguane, P, Kolatkar, P, Lin, MT, Tseng, LH and Hansen, JA (2000). Molecular modeling of the minor histocompatibility antigen HA-1 peptides

- binding to HLA-A alleles. *Tissue Antigens* **55**(1): 24-30.
- Robinson, J, Waller, MJ, Parham, P, de Groot, N, Bontrop, R, Kennedy, LJ, Stoeckl, P and Marsh, SG (2003). IMGT/HLA and IMGT/MHC: sequence databases for the study of the major histocompatibility complex. *Nucleic Acids Res* **31**(1): 311-4.
- Rognan, D, Scapozza, L, Folkers, G and Daser, A (1994). Molecular dynamics simulation of MHC-peptide complexes as a tool for predicting potential T cell epitopes. *Biochemistry* **33**(38): 11476-85.
- Rognan, D, Lauemoller, SL, Holm, A, Buus, S and Tschinke, V (1999). Predicting binding affinities of protein ligands from three-dimensional models: application to peptide binding to class I major histocompatibility proteins. *J Med Chem* **42**(22): 4650-8.
- Ruppert, J, Sidney, J, Celis, E, Kubo, RT, Grey, HM and Sette, A (1993). Prominent role of secondary anchor residues in peptide binding to HLA-A2.1 molecules. *Cell* **74**(5): 929-37.
- Salter, RD, Benjamin, RJ, Wesley, PK, Buxton, SE, Garrett, TP, Clayberger, C, Krensky, AM, Norment, AM, Littman, DR and Parham, P (1990). A binding site for the T-cell co-receptor CD8 on the alpha 3 domain of HLA-A2. *Nature* **345**(6270): 41-6.
- Saper, MA, Bjorkman, PJ and Wiley, DC (1991). Refined structure of the human histocompatibility antigen HLA-A2 at 2.6 Å resolution. *J Mol Biol* **219**(2): 277-319.
- Schueler-Furman, O, Elber, R and Margalit, H (1998). Knowledge-based structure prediction of MHC class I bound peptides: a study of 23 complexes. *Fold Des* **3**(6): 549-64.
- Schueler-Furman, O, Altuvia, Y, Sette, A and Margalit, H (2000). Structure-based prediction of binding peptides to MHC class I molecules: application to a broad range of MHC alleles. *Protein Sci* **9**(9): 1838-46.
- Scognamiglio, P, Accapezzato, D, Casciaro, MA, Cacciani, A, Artini, M, Bruno, G, Chircu, ML, Sidney, J, Southwood, S, Abrignani, S, Sette, A and Barnaba, V (1999). Presence of effector CD8⁺ T cells in hepatitis C virus-exposed healthy seronegative donors. *J Immunol* **162**(11): 6681-9.
- Sette, A, Buus, S, Appella, E, Smith, JA, Chesnut, R, Miles, C, Colon, SM and Grey, HM (1989). Prediction of major histocompatibility complex binding regions of protein antigens by sequence pattern analysis. *Proc Natl Acad Sci U S A* **86**(9): 3296-300.
- Sette, A and Sidney, J (1998). HLA supertypes and supermotifs: a functional

- perspective on HLA polymorphism. *Curr Opin Immunol* **10**(4): 478-82.
- Sette, A, Newman, M, Livingston, B, McKinney, D, Sidney, J, Ishioka, G, Tangri, S, Alexander, J, Fikes, J and Chesnut, R (2002). Optimizing vaccine design for cellular processing, MHC binding and TCR recognition. *Tissue Antigens* **59**(6): 443-51.
- Shiina, T, Tamiya, G, Oka, A, Takishima, N and Inoko, H (1999). Genome sequencing analysis of the 1.8 Mb entire human MHC class I region. *Immunol Rev* **167**: 193-9.
- Sidney, J, Grey, HM, Southwood, S, Celis, E, Wentworth, PA, del Guercio, MF, Kubo, RT, Chesnut, RW and Sette, A (1996a). Definition of an HLA-A3-like supermotif demonstrates the overlapping peptide-binding repertoires of common HLA molecules. *Hum Immunol* **45**(2): 79-93.
- Sidney, J, Southwood, S, del Guercio, MF, Grey, HM, Chesnut, RW, Kubo, RT and Sette, A (1996b). Specificity and degeneracy in peptide binding to HLA-B7-like class I molecules. *J Immunol* **157**(8): 3480-90.
- Sidney, J, del Guercio, MF, Southwood, S, Hermanson, G, Maewal, A, Appella, E and Sette, A (1997). The HLA-A*0207 peptide binding repertoire is limited to a subset of the A*0201 repertoire. *Hum Immunol* **58**(1): 12-20.
- Sigal, LJ, Crotty, S, Andino, R and Rock, KL (1999). Cytotoxic T-cell immunity to virus-infected non-haematopoietic cells requires presentation of exogenous antigen. *Nature* **398**(6722): 77-80.
- Skolnick, J, Jaroszewski, L, Kolinski, A and Godzik, A (1997). Derivation and testing of pair potentials for protein folding. When is the quasichemical approximation correct? *Protein Sci* **6**(3): 676-88.
- Smith, KJ, Reid, SW, Harlos, K, McMichael, AJ, Stuart, DI, Bell, JI and Jones, EY (1996). Bound water structure and polymorphic amino acids act together to allow the binding of different peptides to MHC class I HLA-B53. *Immunity* **4**(3): 215-28.
- Sobao, Y, Tomiyama, H, Nakamura, S, Sekihara, H, Tanaka, K and Takiguchi, M (2001). Visual demonstration of hepatitis C virus-specific memory CD8(+) T-cell expansion in patients with acute hepatitis C. *Hepatology* **33**(1): 287-94.
- Stassar, MJ, Raddrizzani, L, Hammer, J and Zoller, M (2001). T-helper cell-response to MHC class II-binding peptides of the renal cell carcinoma-associated antigen RAGE-1. *Immunobiology* **203**(5): 743-55.
- Steinmetz, M and Hood, L (1983). Genes of the major histocompatibility complex in mouse and man. *Science* **222**(4625): 727-33.
- Stern, LJ, Brown, JH, Jardetzky, TS, Gorga, JC, Urban, RG, Strominger, JL and Wiley,

- DC (1994). Crystal structure of the human class II MHC protein HLA-DR1 complexed with an influenza virus peptide. *Nature* **368**(6468): 215-21.
- Stryhn, A, Pedersen, LO, Romme, T, Holm, CB, Holm, A and Buus, S (1996). Peptide binding specificity of major histocompatibility complex class I resolved into an array of apparently independent subspecificities: quantitation by peptide libraries and improved prediction of binding. *Eur J Immunol* **26**(8): 1911-8.
- Sturniolo, T, Bono, E, Ding, J, Raddrizzani, L, Tuereci, O, Sahin, U, Braxenthaler, M, Gallazzi, F, Protti, MP, Sinigaglia, F and Hammer, J (1999). Generation of tissue-specific and promiscuous HLA ligand databases using DNA microarrays and virtual HLA class II matrices. *Nat Biotechnol* **17**(6): 555-61.
- Sudo, T, Kamikawaji, N, Kimura, A, Date, Y, Savoie, CJ, Nakashima, H, Furuichi, E, Kuhara, S and Sasazuki, T (1995). Differences in MHC class I self peptide repertoires among HLA-A2 subtypes. *J Immunol* **155**(10): 4749-56.
- Swets, JA (1988). Measuring the accuracy of diagnostic systems. *Science* **240**(4857):1285-93.
- The MHC sequencing consortium (1999). Complete sequence and gene map of a human major histocompatibility complex. *Nature* **401** (6756): 921-23.
- Thompson, JD, Higgins, DG and Gibson, TJ (1994). Improved sensitivity of profile searches through the use of sequence weights and gap excision. *Comput Appl Biosci* **10**(1): 19-29.
- Townsend, AR, Rothbard, J, Gotch, FM, Bahadur, G, Wraith, D and McMichael, AJ (1986). The epitopes of influenza nucleoprotein recognized by cytotoxic T lymphocytes can be defined with short synthetic peptides. *Cell* **44**(6): 959-68.
- Udaka, K, Wiesmuller, KH, Kienle, S, Jung, G, Tamamura, H, Yamagishi, H, Okumura, K, Walden, P, Suto, T and Kawasaki, T (2000). An automated prediction of MHC class I-binding peptides based on positional scanning with peptide libraries. *Immunogenetics* **51**(10): 816-28.
- van Endert, PM (2001). Designing peptide vaccines for cellular cross-presentation. *Biologicals* **29**(3-4): 285-8.
- Vasmatzis, G, Cornette, J, Sezerman, U and DeLisi, C (1996a). TcR recognition of the MHC-peptide dimer: structural properties of a ternary complex. *J Mol Biol* **261**(1): 72-89.
- Vasmatzis, G, Zhang, C, Cornette, JL and DeLisi, C (1996b). Computational determination of side chain specificity for pockets in class I MHC molecules. *Mol Immunol* **33**(16): 1231-9.
- Venkatarajan, MS, Braun, W (2001). New quantitative descriptor of amino acids based on multi-dimensional scaling of a large number of physical-chemical properties.

J Mol Model **7**(12): 445-453.

Zeng, J, Treutlein, HR and Rudy, GB (2001). Predicting sequences and structures of MHC-binding peptides: a computational combinatorial approach. *J Comput Aided Mol Des* **15**(6): 573-86.

Zhang, C, Anderson, A and DeLisi, C (1998). Structural principles that govern the peptide-binding motifs of class I MHC molecules. *J Mol Biol* **281**(5): 929-47.

APPENDIX 1

APPENDIX 1 Classification of pockets of HLA-A alleles based on CPGRs.**a.**

Group	Pocket	A	HLA Alleles						
1	YENHYRG	A*0101	A*0102	A*0103	A*0104	A*0106	A*0107	A*0109	A*0110
2	YENHYWG	A*0108							
3	YENHYTG	A*0112	A*0114	A*3105					
4	YENQYRG	A*0113	A*0318						
5	YENHYGG	A*0115							
6	YENHYPR	A*0116							
		A*0201	A*0202	A*0203	A*0204	A*0205	A*0206	A*0209	A*0211
		A*0212	A*0213	A*0214	A*0221	A*0222	A*0224	A*0225	A*0226
		A*0227	A*0228	A*0230	A*0231	A*0240	A*0241	A*0242	A*0243
		A*0244	A*0245	A*0246	A*0247	A*0248	A*0249	A*0250	A*0251
7	YEKHYTW	A*0252	A*0257	A*0258	A*0259	A*0260	A*0261	A*0263	A*0264
		A*0265	A*0266	A*0267	A*0268	A*0269	A*0270	A*0271	A*0272
		A*0273	A*0274	A*0275	A*0276	A*0277	A*0279	A*0280	A*0281
		A*0283	A*0284	A*0285	A*0286	A*3007	A*3102	A*3107	A*3108
		A*3205	A*3308	A*7404					
8	YEKHCTW	A*0207	A*0215	A*0218					
		A*0208	A*0220	A*0308	A*3002	A*3003	A*3004	A*3006	A*3009
		A*3012	A*3013	A*3101	A*3103	A*3104	A*3106	A*3109	A*3110
9	YENHYTW	A*3111	A*3112	A*3201	A*3202	A*3203	A*3204	A*3206	A*3207
		A*3208	A*3210	A*3601	A*3602	A*3603	A*7401	A*7402	A*7403
		A*7405	A*7406	A*7407	A*7408	A*7409	A*7410	A*7411	
10	YEKHFTW	A*0210	A*0217	A*0239	A*2304	A*2403	A*2418	A*2422	A*2423
		A*2433							
11	YEKHYEW	A*0216							
12	YEKHYTG	A*0219	A*0236	A*0237	A*0254	A*2414	A*2415	A*2451	A*2453
13	YEQHYTW	A*0229							
14	YEKHSTW	A*0233							
15	YEKQYTW	A*0234	A*0235						
16	YEKHYSW	A*0238	A*2607						
17	YNNHYTW	A*0255	A*2609	A*3301	A*3303	A*3304	A*3305	A*3306	A*3307
		A*6803	A*6804	A*6805	A*6815	A*6820			
		A*0256	A*0262	A*0278	A*0301	A*0302	A*0303	A*0304	A*0305
18	YENQYTW	A*0306	A*0307	A*0309	A*0310	A*0312	A*0313	A*0314	A*0316
		A*0317	A*1104	A*2914	A*3001	A*3008	A*3011	A*3014	A*3015
		A*6810	A*6813	A*6814					
19	YENQFTW	A*0315							
		A*1101	A*1102	A*1103	A*1105	A*1106	A*1107	A*1108	A*1109
20	YENQYRW	A*1112	A*1113	A*1114	A*1115	A*1116	A*1120	A*1121	A*1122
		A*1123	A*2619						
21	YNNQYRW	A*1110	A*2502	A*2613	A*6601	A*6604			
22	YQNQYRW	A*1111							
23	YENHYRW	A*1117	A*1118	A*1119	A*3604				
		A*2301	A*2302	A*2303	A*2306	A*2307	A*2312	A*2402	A*2404
		A*2405	A*2406	A*2408	A*2409	A*2411	A*2413	A*2417	A*2420
24	YEKHFTG	A*2421	A*2426	A*2427	A*2428	A*2429	A*2430	A*2432	A*2434
		A*2435	A*2437	A*2438	A*2439	A*2440	A*2441	A*2442	A*2443
		A*2444	A*2447	A*2449	A*2450	A*2452			
25	CEKHFTG	A*2305	A*2425						

26	YENHFTG	A*2309
27	YEKHFRG	A*2310 A*2446
28	YEKQFTG	A*2407 A*2419
29	YEKHFRW	A*2410
30	YNNQFTG	A*2424
31	YQKHFTG	A*2431
		A*2501 A*2503 A*2504 A*2601 A*2602 A*2603 A*2605 A*2606
32	YNNHYRW	A*2608 A*2610 A*2612 A*2614 A*2615 A*2616 A*2617 A*2618
		A*2620 A*2621 A*2622 A*2623 A*2624 A*2626
33	YNNHYLW	A*2604
34	YNNHYAS	A*2611
		A*2901 A*2902 A*2905 A*2906 A*2909 A*2910 A*2911 A*2912
35	YQNQYTW	A*2913
36	YQNQYTG	A*2903
37	YQHQYTW	A*2904
38	YQNQFTW	A*2907
39	YENHHTW	A*3010
40	YENHYTR	A*3209
41	YNKQYTW	A*3401 A*3405
		A*3402 A*3403 A*3404 A*3406 A*6801 A*6802 A*6806 A*6807
42	YNNQYTW	A*6808 A*6809 A*6816 A*6817 A*6819 A*6821 A*6822 A*6823
		A*6824 A*6825 A*6827 A*6828 A*6901
43	YQNHYRW	A*4301
44	YNNQYEW	A*6602
45	YNNHYEW	A*6603
46	YNNQYTG	A*6826
47	YENHYEG	A*8001

b.

Group Pocket B		HLA Allele							
1	YFARENHAY	A*0101	A*0103	A*0104	A*0106	A*0108	A*0109	A*0110	A*0112
		A*0114	A*0115	A*0116	A*3601	A*3602	A*3603	A*3604	A*7410
2	YSARENHAY	A*0102							
3	YFARENHEY	A*0107	A*3201	A*3202	A*3203	A*3204	A*3206	A*3208	A*3209
		A*3210							
		A*0113	A*0256	A*0262	A*0301	A*0302	A*0303	A*0304	A*0305
4	YFARENQVY	A*0306	A*0307	A*0309	A*0310	A*0313	A*0314	A*0316	A*0317
		A*0318							
		A*0201	A*0202	A*0203	A*0204	A*0209	A*0211	A*0212	A*0213
		A*0216	A*0219	A*0222	A*0224	A*0225	A*0226	A*0227	A*0230
		A*0231	A*0236	A*0237	A*0238	A*0240	A*0243	A*0245	A*0246
5	YFAREKHVY	A*0247	A*0248	A*0249	A*0250	A*0252	A*0258	A*0259	A*0260
		A*0263	A*0264	A*0265	A*0266	A*0267	A*0268	A*0269	A*0270
		A*0271	A*0273	A*0274	A*0275	A*0277	A*0280	A*0283	A*0285
		A*0286	A*7404						
		A*0205	A*0206	A*0214	A*0221	A*0228	A*0241	A*0244	A*0251
6	YYAREKHVY	A*0254	A*0257	A*0261	A*0272	A*0279	A*0284		
7	YFAREKHVC	A*0207	A*0215	A*0218					
8	YYARENHVY	A*0208	A*1118	A*1119					
9	YYAREKHVF	A*0210							
10	YFAREKHVF	A*0217	A*0239						

11	YFARENHVY	A*0220	A*0308	A*7401	A*7402	A*7403	A*7405	A*7406	A*7407
		A*7408	A*7409	A*7411					
12	YFAREQHVY	A*0229							
13	YFAREKHVS	A*0233							
14	YFAREKQVY	A*0234	A*0235						
15	YFSREKHVY	A*0242							
16	YFARNNHVY	A*0255							
17	YSAREKHVY	A*0276							
		A*0278	A*0312	A*1101	A*1102	A*1103	A*1104	A*1105	A*1106
18	YYARENQVY	A*1107	A*1108	A*1109	A*1112	A*1113	A*1114	A*1115	A*1116
		A*1120	A*1121	A*1122	A*1123	A*3008	A*6810	A*6813	A*6814
19	YFAREKHEY	A*0281	A*3205						
20	YFARENQVF	A*0315							
		A*1110	A*3402	A*3403	A*3404	A*3406	A*6601	A*6602	A*6604
21	YYARNNQVY	A*6801	A*6802	A*6806	A*6807	A*6808	A*6809	A*6816	A*6817
		A*6819	A*6821	A*6822	A*6823	A*6824	A*6825	A*6826	A*6827
		A*6828	A*6901						
22	YYARQNQVY	A*1111							
23	YYARENHAY	A*1117							
		A*2301	A*2302	A*2303	A*2304	A*2306	A*2307	A*2310	A*2312
		A*2402	A*2403	A*2405	A*2406	A*2408	A*2409	A*2410	A*2411
24	YSAREKHEF	A*2413	A*2417	A*2418	A*2420	A*2421	A*2422	A*2423	A*2426
		A*2427	A*2429	A*2430	A*2432	A*2433	A*2434	A*2435	A*2437
		A*2438	A*2439	A*2440	A*2441	A*2442	A*2443	A*2446	A*2447
		A*2449	A*2452						
25	CSAREKHEF	A*2305	A*2425						
26	YSARENHEF	A*2309							
27	YSAREKHAF	A*2404							
28	YSAREKQEF	A*2407							
29	YSAREKHEY	A*2414	A*2415	A*2451	A*2453	A*3007			
30	YSAREKQVF	A*2419							
31	YSARNNQEF	A*2424							
32	YSAREKHVF	A*2428	A*2444						
33	YSARQKHEF	A*2431							
34	YYAREKHEF	A*2450							
35	YYARNNHEY	A*2501	A*2504	A*2605					
36	YYARNNQEY	A*2502							
37	YFARNNHEY	A*2503							
38	YYARNNHAY	A*2601	A*2602	A*2604	A*2608	A*2609	A*2610	A*2611	A*2612
		A*2614	A*2615	A*2617	A*2618	A*2622	A*2623	A*2624	A*2626
39	YYARNNHVY	A*2603	A*2606	A*2621	A*6603	A*6803	A*6804	A*6805	A*6815
		A*6820							
40	YYAREKHAY	A*2607							
41	YYARNNQAY	A*2613							
42	YSARNNHAY	A*2616							
43	YYARENQAY	A*2619							
44	YFARNNHAY	A*2620							
45	YTARQNQAY	A*2901	A*2902	A*2903	A*2905	A*2906	A*2909	A*2910	A*2911
		A*2912							
46	YTARQHQAY	A*2904							
47	YTARQNQAF	A*2907							

48 YTARQNQEY A*2913
 49 YTARENQAY A*2914
 50 YSARENQVY A*3001 A*3011 A*3014 A*3015
 51 YSARENHEY A*3002 A*3003 A*3004 A*3006 A*3009 A*3012 A*3207
 52 YSARENHEH A*3010
 53 YSARENHVY A*3013
 54 YTARENHVY A*3101 A*3103 A*3104 A*3105 A*3106 A*3109 A*3111 A*3112
 55 YTAREKHVY A*3102 A*3308
 56 YTAREKHEY A*3107 A*3108
 57 YTARENHEY A*3110
 58 YTARNNHVY A*3301 A*3303 A*3304 A*3305 A*3306 A*3307
 59 YYARNKQVY A*3401 A*3405
 60 YYARQNHAY A*4301
 61 YFAQENHAY A*8001

C.

Group Pocket C		HLA Allele							
1	FHTDI	A*0101	A*0104	A*0106	A*0107	A*0108	A*0109	A*0110	A*0112
		A*0114	A*0115	A*0116	A*0308	A*3204	A*3601	A*3602	A*3604
2	SHTDI	A*0102	A*2451	A*3002	A*3003	A*3004	A*3006	A*3007	A*3009
		A*3010	A*3012						
3	FHTDM	A*0103	A*3201	A*3202	A*3203	A*3205	A*3206	A*3209	A*3210
		A*3603	A*7401	A*7402	A*7403	A*7404	A*7408	A*7409	A*7410
4	FQTDI	A*0113	A*0301	A*0302	A*0303	A*0304	A*0305	A*0306	A*0307
		A*0310	A*0313	A*0314	A*0315	A*0316	A*0317	A*0318	
5	FHTHR	A*0201	A*0202	A*0203	A*0207	A*0209	A*0212	A*0213	A*0215
		A*0216	A*0218	A*0219	A*0220	A*0222	A*0224	A*0225	A*0226
6	FHTHM	A*0227	A*0229	A*0230	A*0231	A*0233	A*0236	A*0237	A*0238
		A*0239	A*0240	A*0242	A*0243	A*0245	A*0246	A*0249	A*0250
7	YHTHR	A*0252	A*0255	A*0258	A*0259	A*0260	A*0263	A*0264	A*0266
		A*0267	A*0268	A*0270	A*0271	A*0273	A*0274	A*0275	A*0277
8	FHIDR	A*0280	A*0283	A*0285	A*0286				
		A*0204	A*0217	A*0265	A*3208	A*7406			
9	FQTHR	A*0205	A*0206	A*0208	A*0210	A*0214	A*0221	A*0228	A*0241
		A*0244	A*0251	A*0254	A*0261	A*0272	A*0279	A*0284	A*2603
10	FQTDR	A*2606	A*2621	A*6815					
		A*0211	A*0269						
11	FHSHR	A*0234	A*0256	A*0262					
		A*0235							
12	FHTDR	A*0247							
		A*0248	A*0281	A*2503	A*2620				
13	YHTHM	A*0257	A*6805	A*6820					
		A*0276							
14	SHTHR	A*0278	A*2502	A*2613	A*2619	A*3401	A*3405	A*3406	A*6601
		A*6602	A*6604	A*6802	A*6823	A*6827	A*6828	A*6901	
15	YQTDR	A*0309							
		A*0312	A*1101	A*1102	A*1103	A*1104	A*1105	A*1107	A*1108
16	FQTHI	A*1109	A*1110	A*1111	A*1112	A*1113	A*1114	A*1115	A*1116
		A*1120	A*1121	A*1122	A*1123	A*3008	A*3402	A*3403	A*3404
17	YQTDI	A*1106							
18	YQTHI								

19	YHTDI	A*1117 A*1119
20	YHTHI	A*1118
		A*2301 A*2302 A*2303 A*2304 A*2305 A*2306 A*2307 A*2309
		A*2310 A*2402 A*2403 A*2404 A*2405 A*2406 A*2408 A*2409
21	SHTDM	A*2410 A*2411 A*2413 A*2415 A*2417 A*2418 A*2420 A*2421
		A*2422 A*2423 A*2425 A*2426 A*2427 A*2429 A*2431 A*2432
		A*2433 A*2435 A*2437 A*2438 A*2439 A*2440 A*2441 A*2443
		A*2444 A*2446 A*2447 A*2449 A*2453 A*3207
22	SHTHM	A*2312 A*2428 A*2430 A*2442
23	SQTDM	A*2407 A*2419 A*2424
24	SHTDR	A*2414 A*2452 A*2616
25	SHIDM	A*2434
26	YHTDM	A*2450 A*6803
		A*2501 A*2504 A*2601 A*2602 A*2604 A*2605 A*2607 A*2608
27	YHTDR	A*2609 A*2610 A*2611 A*2612 A*2614 A*2615 A*2617 A*2618
		A*2622 A*2623 A*2624 A*2626 A*4301 A*6603
28	TQTDM	A*2901 A*2902 A*2903 A*2904 A*2905 A*2906 A*2907 A*2909
		A*2910 A*2911 A*2912 A*2913 A*2914
29	SQTDI	A*3001 A*3011 A*3014 A*3015
30	SHTHI	A*3013
31	THIDM	A*3101 A*3102 A*3105 A*3106 A*3109 A*3111 A*3301 A*3303
		A*3304 A*3305 A*3306 A*3307 A*3308
32	THIDI	A*3103 A*3104
33	THTDM	A*3107 A*3108 A*3110
34	THIDK	A*3112
35	YQTDM	A*6801 A*6806 A*6807 A*6808 A*6809 A*6810 A*6813 A*6814
		A*6816 A*6817 A*6819 A*6821 A*6822 A*6824 A*6825 A*6826
36	YHIDM	A*6804
37	FHADM	A*7405
38	FHIDM	A*7407
39	FHTNI	A*8001

d.

Group Pocket D		HLA Allele							
1	FRR	A*0101	A*0103	A*0104	A*0107	A*0108	A*0109	A*0110	A*0113
		A*0114	A*0115	A*0318	A*3601	A*3602	A*3603	A*3604	
2	SRR	A*0102							
3	FRL	A*0106	A*0301	A*0303	A*0304	A*0305	A*0306	A*0307	A*0308
		A*0309	A*0313	A*0314	A*0315	A*0316	A*0317	A*3204	A*8001
4	FRQ	A*0112	A*0302	A*0310					
5	FRA	A*0116							
		A*0201	A*0204	A*0207	A*0209	A*0211	A*0215	A*0216	A*0217
		A*0218	A*0220	A*0224	A*0225	A*0226	A*0229	A*0230	A*0231
6	FHL	A*0233	A*0234	A*0235	A*0236	A*0239	A*0240	A*0242	A*0243
		A*0245	A*0246	A*0248	A*0255	A*0256	A*0258	A*0259	A*0260
		A*0262	A*0264	A*0266	A*0267	A*0268	A*0269	A*0270	A*0271
		A*0273	A*0274	A*0275	A*0277	A*0281	A*0283	A*0285	A*0286
7	FHW	A*0202	A*0203	A*0222	A*0247	A*0250	A*0263		
8	YHW	A*0205	A*0208	A*6802	A*6815	A*6827			
9	YHL	A*0206	A*0210	A*0214	A*0221	A*0228	A*0251	A*0257	A*0261
		A*0272	A*0278	A*0279	A*0284	A*6901			
10	FHQ	A*0212	A*0213	A*0219	A*0227	A*0237	A*0238		

11	YQL	A*0241	A*3406						
12	YHQ	A*0244	A*0254	A*2450					
13	FHR	A*0249							
14	FEL	A*0252							
		A*0265	A*0280	A*3201	A*3203	A*3205	A*3206	A*3208	A*3209
15	FQL	A*7401	A*7402	A*7403	A*7404	A*7405	A*7406	A*7407	A*7408
		A*7409	A*7410	A*7411					
		A*0276	A*2301	A*2303	A*2304	A*2305	A*2306	A*2307	A*2309
16	SHL	A*2310	A*2312	A*2413	A*2418	A*2424			
17	YRL	A*0312	A*3402	A*3403	A*3404	A*6808			
		A*1101	A*1102	A*1103	A*1104	A*1105	A*1106	A*1107	A*1109
18	YRQ	A*1110	A*1111	A*1112	A*1113	A*1114	A*1115	A*1116	A*1117
		A*1118	A*1119	A*1120	A*1121	A*1123	A*6809	A*6826	
19	YRR	A*1108							
20	YPQ	A*1122							
21	SHW	A*2302	A*2406	A*2422					
		A*2402	A*2403	A*2404	A*2405	A*2407	A*2408	A*2409	A*2410
		A*2411	A*2414	A*2415	A*2419	A*2420	A*2421	A*2423	A*2425
22	SHQ	A*2426	A*2427	A*2428	A*2429	A*2430	A*2431	A*2432	A*2433
		A*2434	A*2435	A*2437	A*2438	A*2439	A*2440	A*2442	A*2443
		A*2444	A*2446	A*2447	A*2449	A*2451	A*2452	A*2453	
23	SRQ	A*2417	A*2441						
		A*2501	A*2502	A*2601	A*2602	A*2603	A*2604	A*2605	A*2606
		A*2607	A*2609	A*2610	A*2612	A*2613	A*2614	A*2615	A*2617
24	YQW	A*2618	A*2619	A*2621	A*2622	A*2623	A*2624	A*2626	A*3401
		A*3405	A*4301	A*6601	A*6602	A*6603	A*6604		
25	FQW	A*2503	A*2620	A*3210					
26	YQQ	A*2504	A*2608						
27	YQS	A*2611							
28	SQW	A*2616							
		A*2901	A*2902	A*2903	A*2904	A*2906	A*2907	A*2909	A*2910
29	TRL	A*2911	A*2912	A*2913	A*2914	A*3103	A*3104	A*3106	
30	TRQ	A*2905							
		A*3001	A*3002	A*3003	A*3007	A*3009	A*3010	A*3011	A*3012
31	SEL	A*3013	A*3014	A*3015					
32	SEW	A*3004	A*3006						
33	YEL	A*3008							
		A*3101	A*3102	A*3105	A*3107	A*3108	A*3109	A*3110	A*3111
34	TQL	A*3112	A*3301	A*3303	A*3304	A*3305	A*3306	A*3307	A*3308
35	FQQ	A*3202							
36	SQL	A*3207							
		A*6801	A*6803	A*6804	A*6805	A*6807	A*6810	A*6813	A*6814
37	YRW	A*6816	A*6817	A*6819	A*6820	A*6821	A*6822	A*6823	A*6824
		A*6825							
38	YEW	A*6806							
39	YHR	A*6828							

e.

Group	Pocket	E	HLA Allele						
1	IRAR		A*0101	A*0102	A*0104	A*0107	A*0108	A*0109	A*0113
			A*0115	A*0318	A*3601	A*3602	A*3604		
2	MRAR		A*0103	A*3603					

3	IRAL	A*0106							
4	IRRR	A*0110							
5	IRVQ	A*0112	A*0302	A*0310					
6	IRAA	A*0116							
		A*0201	A*0206	A*0207	A*0209	A*0210	A*0211	A*0214	A*0215
		A*0216	A*0218	A*0220	A*0221	A*0224	A*0225	A*0228	A*0229
		A*0230	A*0231	A*0233	A*0234	A*0235	A*0236	A*0239	A*0240
7	RHVL	A*0242	A*0243	A*0245	A*0246	A*0248	A*0251	A*0255	A*0256
		A*0258	A*0259	A*0260	A*0261	A*0262	A*0264	A*0266	A*0267
		A*0268	A*0269	A*0270	A*0271	A*0272	A*0273	A*0274	A*0275
		A*0276	A*0277	A*0278	A*0279	A*0281	A*0283	A*0284	A*0285
		A*0286	A*6901						
8	RHVV	A*0202	A*0205	A*0208	A*0222	A*0247	A*0250	A*0263	A*6802
		A*6815	A*6827						
9	RHEW	A*0203							
		A*0204	A*0217	A*0257	A*2301	A*2303	A*2304	A*2305	A*2306
10	MHVL	A*2307	A*2309	A*2310	A*2312	A*2413	A*2424		
11	RHVQ	A*0212	A*0219	A*0237	A*0244	A*0254	A*2414	A*2452	
12	RHEQ	A*0213	A*0238						
13	RHEL	A*0226							
14	RHAQ	A*0227							
15	RQVL	A*0241	A*0280	A*3406					
16	RHVR	A*0249	A*6828						
17	REVL	A*0252							
		A*0265	A*3101	A*3102	A*3105	A*3107	A*3108	A*3109	A*3110
		A*3111	A*3201	A*3203	A*3205	A*3206	A*3207	A*3208	A*3209
18	MQVL	A*3301	A*3303	A*3304	A*3305	A*3306	A*3307	A*3308	A*7401
		A*7402	A*7403	A*7404	A*7405	A*7406	A*7407	A*7408	A*7409
		A*7410	A*7411						
		A*0301	A*0303	A*0304	A*0305	A*0306	A*0308	A*0309	A*0312
19	IREL	A*0313	A*0314	A*0315	A*0316	A*0317	A*3204	A*3402	A*3403
		A*3404							
20	IRVL	A*0307	A*3103	A*3104					
		A*1101	A*1102	A*1104	A*1105	A*1106	A*1107	A*1109	A*1110
21	IRAQ	A*1111	A*1112	A*1113	A*1115	A*1116	A*1117	A*1118	A*1119
		A*1121	A*1123						
22	IREQ	A*1103	A*1120						
23	IRER	A*1108							
24	IRRQ	A*1114							
25	IPAQ	A*1122							
26	MHVW	A*2302	A*2406	A*2422					
		A*2402	A*2403	A*2404	A*2405	A*2407	A*2408	A*2409	A*2410
		A*2411	A*2415	A*2419	A*2420	A*2421	A*2423	A*2425	A*2426
27	MHVQ	A*2427	A*2428	A*2429	A*2430	A*2431	A*2432	A*2433	A*2434
		A*2435	A*2437	A*2438	A*2439	A*2440	A*2442	A*2443	A*2444
		A*2446	A*2447	A*2449	A*2450	A*2453			
28	MRVQ	A*2417	A*2441	A*6809	A*6826				
29	MHEL	A*2418							
30	IHVQ	A*2451							
		A*2501	A*2502	A*2503	A*2601	A*2602	A*2603	A*2604	A*2605
		A*2606	A*2607	A*2609	A*2610	A*2613	A*2614	A*2615	A*2616
31	RQEW	A*2617	A*2619	A*2620	A*2621	A*2622	A*2623	A*2624	A*2626
		A*3401	A*3405	A*4301	A*6601	A*6602	A*6603	A*6604	

32	RQEQ	A*2504 A*2608
33	RQMS	A*2611
34	RQVW	A*2612 A*2618
35	MRVL	A*2901 A*2902 A*2903 A*2904 A*2906 A*2907 A*2909 A*2910 A*2911 A*2912 A*2913 A*2914 A*3106 A*6808
36	MREQ	A*2905
37	IEWL	A*3001 A*3008 A*3011 A*3014 A*3015
38	IERL	A*3002 A*3003 A*3007 A*3010 A*3012 A*3013
39	IEVW	A*3004 A*3006
40	IEVL	A*3009
41	KQVL	A*3112
42	MQVQ	A*3202
43	MQEW	A*3210
44	MRVW	A*6801 A*6803 A*6804 A*6805 A*6807 A*6810 A*6813 A*6814 A*6816 A*6817 A*6819 A*6820 A*6821 A*6822 A*6824 A*6825
45	MEVW	A*6806
46	RRVW	A*6823
47	IRRL	A*8001

f.

Group Pocket F		HLA allele							
1	NTLD	A*0101	A*0102	A*0103	A*0104	A*0106	A*0107	A*0108	A*0109
		A*0110	A*0112	A*0114	A*0115	A*0116	A*1117	A*2601	A*2604
		A*2605	A*2607	A*2608	A*2609	A*2610	A*2611	A*2612	A*2613
		A*2614	A*2615	A*2616	A*2617	A*2618	A*2619	A*2620	A*2622
		A*2623	A*2624	A*2626	A*2901	A*2902	A*2903	A*2904	A*2905
		A*2906	A*2907	A*2909	A*2910	A*2911	A*2912	A*2914	A*3601
		A*3602	A*3603	A*3604	A*4301	A*7410	A*8001		
		A*0113	A*0265	A*0280	A*0301	A*0302	A*0303	A*0304	A*0305
		A*0306	A*0307	A*0308	A*0309	A*0310	A*0312	A*0313	A*0314
		A*0315	A*0316	A*0317	A*0318	A*1101	A*1102	A*1103	A*1104
2	DTLD	A*1105	A*1106	A*1107	A*1108	A*1109	A*1110	A*1111	A*1112
		A*1113	A*1114	A*1115	A*1116	A*1118	A*1119	A*1120	A*1121
		A*1122	A*1123	A*2603	A*2606	A*2621	A*3101	A*3102	A*3103
		A*3104	A*3105	A*3106	A*3109	A*3111	A*3112	A*3301	A*3303
		A*3304	A*3305	A*3306	A*3307	A*3308	A*3401	A*3402	A*3403
		A*3404	A*3405	A*3406	A*6601	A*6602	A*6603	A*6604	A*6801
		A*6803	A*6804	A*6805	A*6808	A*6809	A*6810	A*6813	A*6814
		A*6816	A*6819	A*6820	A*6821	A*6822	A*6823	A*6824	A*6825
		A*6826	A*7401	A*7402	A*7403	A*7404	A*7405	A*7406	A*7407
		A*7408	A*7409	A*7411					
3	DTLY	A*0201	A*0202	A*0203	A*0204	A*0205	A*0206	A*0207	A*0208
		A*0209	A*0210	A*0211	A*0212	A*0213	A*0214	A*0215	A*0216
		A*0217	A*0218	A*0219	A*0220	A*0221	A*0222	A*0224	A*0225
		A*0226	A*0227	A*0228	A*0229	A*0230	A*0231	A*0233	A*0234
		A*0235	A*0236	A*0237	A*0238	A*0239	A*0240	A*0241	A*0242
		A*0243	A*0244	A*0245	A*0246	A*0247	A*0248	A*0249	A*0250
		A*0251	A*0254	A*0255	A*0256	A*0257	A*0258	A*0259	A*0261
		A*0262	A*0263	A*0266	A*0267	A*0268	A*0269	A*0270	A*0271
		A*0272	A*0273	A*0274	A*0275	A*0276	A*0277	A*0278	A*0279
		A*0283	A*0284	A*0285	A*0286	A*2419	A*2428	A*2444	A*6802
		A*6815	A*6827	A*6828	A*6901				

4	DTLH	A*0252 A*3001 A*3008 A*3011 A*3013 A*3014 A*3015 A*6806 A*6807
5	DTLF	A*0260
6	DTLS	A*0264
7	SIAY	A*0281 A*2432 A*2301 A*2302 A*2303 A*2304 A*2305 A*2306 A*2307 A*2309 A*2310 A*2312 A*2402 A*2403 A*2405 A*2406 A*2407 A*2408 A*2409 A*2410 A*2411 A*2413 A*2414 A*2415 A*2418 A*2420
8	NIAY	A*2421 A*2422 A*2423 A*2424 A*2425 A*2426 A*2427 A*2429 A*2430 A*2431 A*2433 A*2434 A*2435 A*2437 A*2438 A*2439 A*2440 A*2442 A*2443 A*2446 A*2447 A*2449 A*2450 A*2451 A*2452 A*2453
9	NTLY	A*2404
10	NIAD	A*2417 A*2441 A*3107 A*3108 A*3110 A*3203
11	SIAD	A*2501 A*2502 A*2503 A*2504 A*2913 A*3201 A*3202 A*3204 A*3205 A*3206 A*3207 A*3208 A*3209 A*3210
12	NTLN	A*2602
13	NTLH	A*3002 A*3003 A*3004 A*3006 A*3007 A*3009 A*3010 A*3012
14	DTLV	A*6817

The classification of class I HLA is based on IMGT/HLA Sequence Database Release 2.10 (released in July 2005).

Group: The pocket group with same positional residue constitution.

CPFR: critical polymorphic functional residue, the polymorphic amino acid residues that directly contact with binding peptides in the pockets, placed according to their original positions on the HLA chains, designated in single capital letter.

HLA Alleles: The class I HLA alleles that belong to the group, with sequence based nomenclature begin with “A*” and followed by a four digit number.

APPENDIX 2

Appendix 2 Classification of pockets of HLA-B alleles based on CPMRs.**a.**

Group Pocket A		HLA Allele							
1	YNIQYEW	B*0702	B*0704	B*0705	B*0706	B*0707	B*0709	B*0710	B*0711
		B*0712	B*0714	B*0715	B*0717	B*0718	B*0721	B*0722	B*0723
		B*0725	B*0726	B*0728	B*0729	B*0730	B*0733	B*0735	B*0736
		B*0739	B*0740	B*0741	B*0742	B*5509	B*8101	B*8102	
2	YNINYEY	B*0703	B*0708	B*0716	B*0727	B*0732	B*0737	B*0738	B*1552
		B*1810	B*2723	B*3515	B*3531	B*3533	B*4008	B*4025	B*4806
		B*5116	B*5131	B*5134					
3	YEKQYEW	B*0713							
4	YNIQYTW	B*0719	B*0731	B*0734	B*0743	B*4201	B*4202	B*4204	B*4205
		B*4206	B*5401	B*5402	B*5404	B*5407	B*5501	B*5502	B*5503
		B*5504	B*5505	B*5507	B*5510	B*5511	B*5512	B*5513	B*5514
		B*5515	B*5516	B*5517	B*5519	B*5610	B*5612	B*6701	
5	YNIQYLS	B*0720	B*8301						
6	YNIQYLW	B*0724	B*1576	B*5403	B*5406	B*5508	B*5601	B*5602	B*5603
		B*5604	B*5605	B*5606	B*5607	B*5609	B*5611	B*5613	B*5615
7	YNINYTW	B*5616							
		B*0801	B*0802	B*0804	B*0806	B*0807	B*0809	B*0812	B*0813
		B*0814	B*0815	B*0816	B*0818	B*0819	B*0820	B*0822	B*0823
		B*0824	B*1401	B*1402	B*1403	B*1405	B*1406	B*1544	B*1593
		B*1801	B*1802	B*1803	B*1804	B*1805	B*1806	B*1807	B*1809
		B*1811	B*1813	B*1814	B*1815	B*1819	B*1820	B*3535	B*3560
		B*3801	B*3802	B*3805	B*3806	B*3807	B*3809	B*3810	B*3811
		B*3901	B*3903	B*3904	B*3905	B*3906	B*3910	B*3912	B*3914
		B*3915	B*3916	B*3917	B*3918	B*3919	B*3920	B*3924	B*3926
		B*3927	B*3928	B*3929	B*3930	B*3931	B*3934	B*5121	B*5136
		B*5901							
8	YNTNYTW	B*0805							
9	YNINYRG	B*0808							
10	YDINYTW	B*0810							
11	YNINYAW	B*0811							
12	YEINYTW	B*0817	B*1306	B*1542	B*1550	B*1569	B*1583	B*1586	B*1812
		B*3709	B*3803	B*3804	B*3902	B*3908	B*3913	B*3922	B*3923
		B*4039	B*4051	B*4101	B*4102	B*4103	B*4104	B*4105	B*4106
		B*4107	B*5518						
13	YNINYLW	B*0821	B*0825	B*1502	B*1508	B*1509	B*1510	B*1511	B*1513
		B*1515	B*1518	B*1521	B*1523	B*1529	B*1531	B*1537	B*1551
		B*1555	B*1564	B*1572	B*1580	B*1588	B*1589	B*1590	B*1599
		B*3501	B*3502	B*3503	B*3504	B*3505	B*3506	B*3507	B*3508
		B*3509	B*3511	B*3512	B*3514	B*3517	B*3518	B*3519	B*3520
		B*3521	B*3522	B*3524	B*3525	B*3526	B*3527	B*3529	B*3530
		B*3532	B*3534	B*3536	B*3537	B*3538	B*3539	B*3540	B*3541
		B*3542	B*3543	B*3544	B*3546	B*3547	B*3549	B*3550	B*3551
		B*3552	B*3554	B*3555	B*3556	B*3557	B*3558	B*3561	B*3932
		B*3933	B*5101	B*5102	B*5104	B*5105	B*5106	B*5107	B*5108
		B*5109	B*5111	B*5112	B*5113	B*5114	B*5115	B*5117	B*5118
		B*5119	B*5120	B*5122	B*5124	B*5126	B*5128	B*5129	B*5130
		B*5132	B*5133	B*5135	B*5137	B*5138	B*5301	B*5302	B*5303
		B*5304	B*5305	B*5306	B*5308	B*5309	B*5310	B*7801	B*7802
		B*7803	B*7804						

		B*1301	B*1302	B*1308	B*1309	B*1310	B*1311	B*1313	B*1540
		B*1547	B*1549	B*2712	B*2716	B*2718	B*2729	B*3702	B*3707
		B*4001	B*4002	B*4003	B*4004	B*4006	B*4007	B*4009	B*4010
		B*4011	B*4012	B*4013	B*4014	B*4015	B*4016	B*4018	B*4019
14	YEINYEW	B*4020	B*4021	B*4023	B*4024	B*4027	B*4029	B*4030	B*4031
		B*4032	B*4033	B*4034	B*4035	B*4036	B*4037	B*4038	B*4040
		B*4042	B*4043	B*4044	B*4045	B*4046	B*4047	B*4048	B*4049
		B*4050	B*4052	B*4053	B*4054	B*4055	B*4056	B*4057	B*4058
		B*4059	B*4061	B*4416	B*4704	B*4705	B*4801	B*4803	B*4804
		B*4805	B*4807	B*4808	B*4809	B*4810	B*4812	B*4813	B*5208
		B*1303	B*1304	B*1501	B*1503	B*1504	B*1505	B*1507	B*1520
		B*1524	B*1525	B*1528	B*1533	B*1534	B*1535	B*1536	B*1538
		B*1539	B*1543	B*1545	B*1546	B*1548	B*1553	B*1554	B*1556
		B*1557	B*1558	B*1560	B*1561	B*1562	B*1563	B*1565	B*1566
15	YEINYLW	B*1568	B*1570	B*1571	B*1573	B*1574	B*1575	B*1577	B*1578
		B*1581	B*1582	B*1585	B*1587	B*1592	B*1596	B*1597	B*1598
		B*3513	B*3516	B*3528	B*4005	B*4026	B*4028	B*4437	B*4442
		B*4504	B*4802	B*4901	B*4902	B*4903	B*4904	B*5001	B*5004
		B*5201	B*5202	B*5203	B*5204	B*5205	B*5206	B*5207	B*7805
		B*9502	B*9503	B*9504					
16	YEINYGW	B*1312							
17	HNNNYTW	B*1404							
18	YEINFLW	B*1506	B*1527	B*1584					
19	YEINYLG	B*1512	B*1519						
		B*1514	B*1591	B*4402	B*4403	B*4405	B*4407	B*4408	B*4409
		B*4411	B*4413	B*4414	B*4415	B*4417	B*4418	B*4420	B*4422
20	YEINYLS	B*4424	B*4425	B*4426	B*4427	B*4428	B*4429	B*4430	B*4432
		B*4433	B*4434	B*4435	B*4436	B*4438	B*4439	B*4440	B*4441
		B*4501	B*4502	B*4503	B*4505	B*4507	B*5002		
		B*1516	B*1517	B*1567	B*1595	B*5701	B*5702	B*5703	B*5704
21	YENSYLW	B*5705	B*5706	B*5708	B*5709	B*5801	B*5802	B*5804	B*5806
		B*5808	B*5809	B*5811					
22	YEINSLW	B*1532							
23	YNINCTW	B*1808							
24	YNINSTW	B*1818	B*3708	B*3909					
		B*2701	B*2702	B*2703	B*2704	B*2705	B*2706	B*2707	B*2708
25	YEIKYEW	B*2709	B*2710	B*2711	B*2713	B*2714	B*2717	B*2719	B*2720
		B*2721	B*2724	B*2727	B*2730				
26	YEIKYTW	B*2715	B*2728						
27	YEIKYLW	B*2725							
28	YEIQYEW	B*2726							
29	YNINFLW	B*3523							
30	YNINYLS	B*3545	B*4406	B*4412	B*5123				
31	YNINSLW	B*3548	B*5307						
32	YEINSTW	B*3701	B*3704	B*3705	B*3706				
33	YNINYMW	B*3808							
34	YEINSEW	B*4060							
35	YEINYTS	B*4404							
36	YEINFLS	B*4410							
37	YEINYES	B*4421	B*4431						
38	YEIQYLS	B*4506							
39	YEKQYLW	B*4601	B*4602	B*4604	B*4605	B*5608	B*5614		

40	YEKQYTW	B*4603	B*6702
41	YEINFEW	B*4701	B*4702 B*4703
42	YEINYEL	B*4811	
43	YNINYLG	B*5103	
44	YENSYLS	B*5707	B*5807
45	YENSYTW	B*5805	
46	YNIKYEW	B*7301	
47	YNIQFLS	B*8201	B*8202
48	YEIQYLW	B*9501	

b.

Group	Pocket	B	HLA Allele
1	YYSRNIQEY	B*0702	B*0704 B*0705 B*0706 B*0707 B*0709 B*0710 B*0711
		B*0712	B*0714 B*0717 B*0718 B*0719 B*0720 B*0721 B*0722
		B*0723	B*0724 B*0725 B*0726 B*0728 B*0730 B*0731 B*0733
		B*0734	B*0735 B*0736 B*0739 B*0740 B*0741 B*0742 B*0743
		B*4201	B*4204 B*4205 B*4206 B*5510 B*5616 B*6701 B*8101
		B*8102	B*8301
		B*0703	B*0708 B*0716 B*0727 B*0732 B*0737 B*0738 B*1401
		B*1402	B*1403 B*1405 B*1406 B*1509 B*1510 B*1518 B*1523
		B*1529	B*1537 B*1551 B*1552 B*1564 B*1572 B*1580 B*1590
		B*1593	B*1599 B*3525 B*3801 B*3802 B*3805 B*3806 B*3807
2	YYSRNINEY	B*3808	B*3809 B*3810 B*3811 B*3901 B*3903 B*3904 B*3905
		B*3906	B*3910 B*3914 B*3915 B*3916 B*3917 B*3918 B*3919
		B*3920	B*3924 B*3926 B*3928 B*3929 B*3930 B*3932 B*3933
		B*3934	B*4806
		B*0713	B*6702
		B*0715	
3	YYSREKQVY	B*0713	B*6702
4	YYSRNIQVY	B*0715	
5	YDSRNIQEY	B*0729	
6	YDSRNINEY	B*0801	B*0802 B*0804 B*0806 B*0807 B*0808 B*0809 B*0811
		B*0812	B*0813 B*0814 B*0816 B*0818 B*0819 B*0820 B*0821
		B*0822	B*0823 B*0824 B*0825 B*3912
7	YDSRNTNEY	B*0805	
8	YDSRDINEY	B*0810	
9	YDSRNINVY	B*0815	
10	YDSREINEY	B*0817	
11	YYTREINEY	B*1301	B*1302 B*1303 B*1304 B*1306 B*1308 B*1309 B*1310
		B*1311	B*1312 B*1313 B*1553 B*4040 B*4049 B*4402 B*4403
		B*4404	B*4405 B*4407 B*4408 B*4409 B*4411 B*4413 B*4414
		B*4416	B*4417 B*4420 B*4421 B*4422 B*4424 B*4425 B*4426
		B*4427	B*4428 B*4429 B*4430 B*4431 B*4432 B*4433 B*4435
		B*4436	B*4437 B*4438 B*4440 B*4441 B*4442 B*4704 B*4705
12	HYSRNNNEY	B*1404	
13	YYAREINEY	B*1501	B*1504 B*1505 B*1507 B*1512 B*1514 B*1519 B*1520
		B*1524	B*1525 B*1528 B*1533 B*1534 B*1535 B*1536 B*1538
		B*1539	B*1540 B*1542 B*1543 B*1545 B*1546 B*1548 B*1550
		B*1556	B*1558 B*1560 B*1563 B*1565 B*1566 B*1570 B*1573
		B*1575	B*1577 B*1578 B*1581 B*1582 B*1583 B*1585 B*1586
		B*1587	B*1592 B*1596 B*1597 B*3513 B*3516 B*3528 B*4010
		B*4021	B*4058 B*4434 B*5201 B*5202 B*5203 B*5204 B*5205
		B*5206	B*5207 B*5208 B*5518 B*7805 B*9502 B*9504

		B*1502	B*1508	B*1511	B*1513	B*1515	B*1521	B*1531	B*1544
		B*1555	B*1588	B*1589	B*1804	B*3501	B*3502	B*3503	B*3504
		B*3505	B*3506	B*3507	B*3508	B*3509	B*3511	B*3512	B*3514
		B*3515	B*3517	B*3518	B*3519	B*3520	B*3521	B*3522	B*3524
		B*3526	B*3527	B*3529	B*3530	B*3531	B*3532	B*3533	B*3534
		B*3535	B*3536	B*3537	B*3538	B*3539	B*3540	B*3541	B*3542
		B*3543	B*3544	B*3545	B*3546	B*3547	B*3549	B*3551	B*3552
14	YYARNINEY	B*3554	B*3555	B*3556	B*3557	B*3558	B*3560	B*3561	B*4406
		B*5101	B*5102	B*5103	B*5104	B*5105	B*5106	B*5107	B*5108
		B*5109	B*5111	B*5112	B*5113	B*5114	B*5115	B*5116	B*5117
		B*5118	B*5119	B*5120	B*5121	B*5122	B*5123	B*5124	B*5126
		B*5128	B*5129	B*5130	B*5131	B*5132	B*5133	B*5134	B*5135
		B*5136	B*5137	B*5138	B*5301	B*5302	B*5303	B*5304	B*5305
		B*5306	B*5308	B*5309	B*5310	B*5901	B*7801	B*7802	B*7803
		B*7804							
		B*1503	B*1547	B*1549	B*1554	B*1561	B*1562	B*1568	B*1569
15	YYSREINEY	B*1574	B*1591	B*1598	B*3803	B*3804	B*3902	B*3908	B*3913
		B*3922	B*3923	B*4012	B*4801	B*4802	B*4803	B*4804	B*4805
		B*4807	B*4808	B*4809	B*4810	B*4811	B*4812	B*4813	B*9503
16	YYAREINEF	B*1506	B*1527	B*1584					
		B*1516	B*1517	B*1567	B*1595	B*5701	B*5702	B*5703	B*5704
17	YYARENSEY	B*5705	B*5706	B*5707	B*5708	B*5709	B*5801	B*5802	B*5804
		B*5805	B*5806	B*5807	B*5808	B*5809	B*5811		
18	YYAREINES	B*1532							
19	YYAREINVY	B*1557							
20	YHAREINEY	B*1571	B*4903						
		B*1576	B*5401	B*5403	B*5404	B*5406	B*5407	B*5501	B*5502
		B*5504	B*5505	B*5507	B*5508	B*5509	B*5511	B*5512	B*5513
21	YYARNIQEY	B*5514	B*5515	B*5517	B*5519	B*5601	B*5602	B*5603	B*5604
		B*5605	B*5606	B*5607	B*5609	B*5610	B*5611	B*5612	B*5613
		B*5615							
22	YHSRNINEY	B*1801	B*1802	B*1803	B*1805	B*1807	B*1809	B*1810	B*1811
		B*1813	B*1814	B*1815	B*1819	B*1820	B*3931		
23	YHSRNINVY	B*1806							
24	YHSRNINEC	B*1808							
25	YHSREINEY	B*1812	B*3702	B*3707	B*3709				
26	YHSRNINES	B*1818	B*3708						
		B*2701	B*2702	B*2703	B*2704	B*2705	B*2706	B*2707	B*2708
27	YHTREIKEY	B*2709	B*2710	B*2711	B*2713	B*2714	B*2715	B*2717	B*2719
		B*2720	B*2721	B*2724	B*2725	B*2727	B*2728	B*2730	
		B*2712	B*2716	B*2718	B*2729	B*4001	B*4002	B*4003	B*4004
		B*4005	B*4006	B*4007	B*4009	B*4011	B*4013	B*4014	B*4015
		B*4016	B*4018	B*4019	B*4020	B*4023	B*4024	B*4026	B*4027
		B*4028	B*4029	B*4030	B*4031	B*4032	B*4033	B*4034	B*4035
28	YHTREINEY	B*4036	B*4037	B*4038	B*4039	B*4042	B*4043	B*4044	B*4045
		B*4046	B*4047	B*4048	B*4050	B*4051	B*4052	B*4053	B*4054
		B*4055	B*4056	B*4057	B*4059	B*4061	B*4101	B*4102	B*4103
		B*4104	B*4105	B*4106	B*4107	B*4415	B*4418	B*4501	B*4502
		B*4503	B*4504	B*4505	B*4507	B*4901	B*4902	B*4904	B*5001
		B*5002	B*5004						
29	YHTRNINEY	B*2723	B*4008	B*4025					
30	YHTREIQEY	B*2726	B*4506						
31	YYARNINEF	B*3523							
32	YYARNINES	B*3548	B*5307						

33 YHARNINEY B*3550
 34 YHSREINES B*3701 B*3704 B*3705 B*3706
 35 YYSRNINES B*3909
 36 YYSRNINVY B*3927
 37 YHTREINES B*4060
 38 YHSRNIQEY B*4202
 39 YYTREINEF B*4410 B*4701 B*4702 B*4703
 40 YYTRNINEY B*4412
 41 YYPREINEY B*4439
 42 YYAREKQVY B*4601 B*4602 B*4603 B*4604 B*4605
 43 YHARNIQEY B*5402 B*5516
 44 YYARNIQVY B*5503
 45 YYAREKQEY B*5608 B*5614
 46 YHTRNIKVY B*7301
 47 YYSRNIQEF B*8201 B*8202
 48 YYAREIQEY B*9501

C.

Group Pocket C		HLA Allele							
1	YQTDS	B*0702	B*0704	B*0705	B*0706	B*0709	B*0710	B*0711	B*0717
		B*0719	B*0720	B*0721	B*0722	B*0723	B*0724	B*0725	B*0726
		B*0728	B*0730	B*0731	B*0733	B*0734	B*0735	B*0736	B*0739
		B*0740	B*0741	B*0742	B*0743	B*4201	B*4205	B*4206	B*5504
		B*8101	B*8102						
2	YNTDS	B*0703	B*0708	B*0732	B*1405	B*3903	B*3914	B*3924	B*3929
3	YQTDR	B*0707	B*0712	B*0714	B*0718	B*1576	B*4601	B*4603	B*4604
		B*4605	B*5406	B*5508	B*5602	B*5603	B*5604	B*5609	B*5610
		B*5611	B*5612	B*6701	B*8201	B*8202	B*8301	B*9501	
4	YQADS	B*0713	B*0715						
5	YNTYS	B*0716	B*0727	B*0737	B*0738	B*1507	B*1545	B*1555	B*1568
		B*3505	B*3516	B*3517	B*3522	B*3530	B*3531	B*3551	B*3558
		B*4040	B*4058	B*4801	B*4804	B*4805	B*4806	B*4807	B*4808
		B*4809	B*4810	B*4811	B*4812	B*4813			
6	DQTDS	B*0729							
7	DNTDS	B*0801	B*0802	B*0804	B*0805	B*0806	B*0807	B*0808	B*0810
		B*0811	B*0813	B*0814	B*0815	B*0817	B*0818	B*0819	B*0820
		B*0821	B*0822	B*0824	B*0825				
8	DNTDT	B*0809							
9	DNTDR	B*0812	B*3912						
10	DNADR	B*0816							
11	DNTYS	B*0823							
12	YNTYR	B*1301	B*1306	B*1310	B*1311	B*1312	B*1313	B*1501	B*1502
		B*1503	B*1505	B*1506	B*1508	B*1509	B*1510	B*1511	B*1512
		B*1513	B*1514	B*1515	B*1518	B*1519	B*1520	B*1521	B*1523
		B*1524	B*1525	B*1527	B*1528	B*1529	B*1531	B*1532	B*1533
		B*1534	B*1536	B*1537	B*1538	B*1539	B*1540	B*1543	B*1544
		B*1546	B*1547	B*1548	B*1549	B*1550	B*1551	B*1552	B*1553
		B*1554	B*1556	B*1557	B*1558	B*1560	B*1562	B*1563	B*1564
		B*1565	B*1566	B*1569	B*1572	B*1573	B*1574	B*1575	B*1577

		B*1578	B*1580	B*1581	B*1582	B*1584	B*1585	B*1586	B*1587
		B*1588	B*1589	B*1590	B*1591	B*1592	B*1593	B*1596	B*1597
		B*1598	B*1599	B*1804	B*3501	B*3502	B*3503	B*3504	B*3506
		B*3507	B*3508	B*3509	B*3511	B*3512	B*3513	B*3514	B*3515
		B*3518	B*3519	B*3520	B*3521	B*3523	B*3524	B*3525	B*3526
		B*3527	B*3528	B*3532	B*3533	B*3534	B*3535	B*3536	B*3538
		B*3539	B*3540	B*3541	B*3542	B*3543	B*3544	B*3545	B*3546
		B*3547	B*3548	B*3549	B*3552	B*3554	B*3555	B*3556	B*3557
		B*3561	B*3801	B*3802	B*3804	B*3805	B*3806	B*3807	B*3808
		B*3809	B*3810	B*3811	B*3905	B*3908	B*3913	B*3920	B*4010
		B*4012	B*4021	B*4049	B*4402	B*4403	B*4404	B*4405	B*4406
		B*4407	B*4408	B*4409	B*4410	B*4411	B*4412	B*4413	B*4414
		B*4416	B*4417	B*4421	B*4422	B*4424	B*4425	B*4426	B*4427
		B*4428	B*4429	B*4430	B*4431	B*4432	B*4433	B*4434	B*4435
		B*4436	B*4437	B*4438	B*4439	B*4440	B*4441	B*4442	B*4701
		B*4702	B*4703	B*4802	B*4803	B*5104	B*5106	B*5301	B*5302
		B*5303	B*5304	B*5305	B*5306	B*5307	B*5308	B*5309	B*5310
		B*9502	B*9503	B*9504					
		B*1302	B*1303	B*1304	B*1308	B*1309	B*1504	B*1535	B*1542
		B*1583	B*3537	B*3560	B*4420	B*5101	B*5102	B*5103	B*5105
		B*5107	B*5108	B*5109	B*5111	B*5112	B*5113	B*5114	B*5115
13	YNTYT	B*5116	B*5117	B*5118	B*5119	B*5121	B*5122	B*5123	B*5124
		B*5126	B*5128	B*5129	B*5130	B*5131	B*5132	B*5133	B*5134
		B*5135	B*5136	B*5137	B*5138	B*5201	B*5202	B*5203	B*5204
		B*5205	B*5206	B*5207	B*5208	B*5518	B*5901	B*7802	B*7804
		B*7805							
14	YNTDW	B*1401	B*1402	B*1403	B*1404				
		B*1406	B*1561	B*1570	B*3529	B*3803	B*3901	B*3902	B*3904
15	YNTDR	B*3909	B*3910	B*3915	B*3916	B*3917	B*3918	B*3919	B*3922
		B*3923	B*3926	B*3927	B*3930	B*3932			
16	YSTYR	B*1516	B*1517	B*1567	B*1595	B*5705	B*5801	B*5804	B*5805
		B*5809	B*5811						
		B*1571	B*1801	B*1805	B*1806	B*1807	B*1808	B*1809	B*1810
		B*1811	B*1812	B*1813	B*1815	B*1818	B*1819	B*1820	B*3550
		B*3701	B*3704	B*3705	B*3706	B*3708	B*4001	B*4004	B*4007
		B*4011	B*4014	B*4020	B*4023	B*4025	B*4026	B*4028	B*4033
17	HNTYR	B*4036	B*4038	B*4042	B*4046	B*4047	B*4048	B*4051	B*4052
		B*4054	B*4055	B*4059	B*4060	B*4061	B*4101	B*4103	B*4105
		B*4106	B*4107	B*4415	B*4418	B*4501	B*4502	B*4503	B*4504
		B*4505	B*4507	B*4901	B*4902	B*4903	B*4904	B*5001	B*5002
		B*5004							
18	HNTYN	B*1802	B*2718	B*2723	B*2729	B*3702			
19	HNTDR	B*1803	B*3931	B*4043					
		B*1814	B*3707	B*3709	B*4002	B*4003	B*4005	B*4008	B*4009
20	HNTYS	B*4013	B*4015	B*4016	B*4018	B*4019	B*4024	B*4027	B*4029
		B*4031	B*4032	B*4035	B*4037	B*4039	B*4045	B*4056	B*4057
		B*4102	B*4104						
21	HKTYN	B*2701							
22	HKTDN	B*2702	B*2703	B*2704	B*2705	B*2706	B*2708	B*2709	B*2710
		B*2713	B*2715	B*2717	B*2720	B*2725	B*2727	B*2728	
23	HKTDS	B*2707	B*2711	B*2724					
24	HNTDN	B*2712	B*2716						
25	HKTDT	B*2714	B*7301						
26	HKTDR	B*2719	B*2721	B*2730					
27	HQTDN	B*2726							

28	YNTDT	B*3906	B*3928	B*3933	B*3934	B*5120	B*7801	B*7803
29	HNTYT	B*4006	B*4044	B*4053				
30	HNTYV	B*4030	B*4034					
31	HNTDS	B*4050						
32	HQTDS	B*4202						
33	YQTD	B*4204	B*5401	B*5403	B*5404	B*5407	B*5501	B*5502
		B*5503	B*5505	B*5507	B*5509	B*5510	B*5512	B*5513
		B*5515	B*5517	B*5519	B*5601	B*5605	B*5606	B*5607
		B*5608	B*5613	B*5614	B*5615	B*5616		
34	HQTDR	B*4506						
35	YQTGR	B*4602						
36	YNTYN	B*4704	B*4705					
37	HQTD	B*5402	B*5516					
38	YQTD	B*5511						
39	YQTDV	B*5514						
40	YSTYV	B*5701	B*5702	B*5703	B*5704	B*5706	B*5707	B*5708
41	YSTYW	B*5802	B*5806	B*5807				
42	YSTYT	B*5808						
43	YQADR	B*6702						

d.

Group Pocket D		HLA Allele						
1	YDR	B*0702	B*0703	B*0707	B*0708	B*0709	B*0710	B*0711
		B*0712	B*0713	B*0714	B*0715	B*0716	B*0717	B*0718
		B*0719	B*0720	B*0721	B*0722	B*0723	B*0724	B*0727
		B*0728	B*0730	B*0731	B*0733	B*0735	B*0736	B*0737
		B*0738	B*0739	B*0741	B*0742	B*1547	B*1549	B*1565
2	YDD	B*1584	B*1599	B*3508	B*3545	B*3561	B*4404	B*4428
		B*4603	B*4808	B*5704	B*0704	B*0719	B*0725	B*1551
		B*3538	B*4402	B*4405	B*4406	B*4408	B*4409	B*4411
		B*4412	B*4416	B*4417	B*4420	B*4421	B*4422	B*4424
3	YNR	B*4425	B*4427	B*4433	B*4434	B*4441	B*4442	B*8301
		B*0705	B*0706	B*0732	B*0740	B*1403	B*3518	B*3908
		B*3933	B*4805	B*5105	B*5129	B*5509	B*5702	B*5705
4	YDL	B*0726	B*0743	B*1304	B*1310	B*1502	B*1503	B*1505
		B*1506	B*1510	B*1513	B*1516	B*1518	B*1520	B*1521
		B*1523	B*1525	B*1529	B*1531	B*1536	B*1537	B*1539
		B*1540	B*1544	B*1552	B*1555	B*1561	B*1562	B*1564
		B*1567	B*1569	B*1572	B*1574	B*1580	B*1588	B*1589
		B*1590	B*1591	B*1593	B*1595	B*1598	B*1804	B*3501
		B*3503	B*3505	B*3507	B*3511	B*3513	B*3515	B*3516
		B*3517	B*3519	B*3520	B*3521	B*3523	B*3524	B*3525
		B*3526	B*3527	B*3528	B*3529	B*3530	B*3532	B*3533
		B*3534	B*3535	B*3536	B*3537	B*3539	B*3540	B*3541
		B*3542	B*3546	B*3547	B*3548	B*3549	B*3551	B*3552
		B*3554	B*3555	B*3556	B*3557	B*3558	B*3915	B*3929
		B*4403	B*4407	B*4413	B*4426	B*4429	B*4430	B*4432
		B*4436	B*4437	B*4438	B*4439	B*4440	B*4802	B*5137
		B*5301	B*5302	B*5303	B*5304	B*5305	B*5306	B*5308
5	DDR	B*5309	B*5310	B*5514	B*5609	B*5611	B*5701	B*5706
		B*5708	B*5801	B*5802	B*5804	B*5805	B*5806	B*5807
		B*5809	B*5811	B*9503	B*0729			

6	YNL	B*0734	B*1301	B*1302	B*1303	B*1306	B*1308	B*1309	B*1311
		B*1312	B*1313	B*1401	B*1402	B*1404	B*1405	B*1406	B*1509
		B*1542	B*1548	B*1586	B*3502	B*3504	B*3506	B*3509	B*3512
		B*3522	B*3531	B*3560	B*3801	B*3802	B*3803	B*3804	B*3805
		B*3806	B*3807	B*3808	B*3809	B*3810	B*3811	B*3901	B*3902
		B*3903	B*3904	B*3905	B*3906	B*3909	B*3910	B*3913	B*3914
		B*3916	B*3917	B*3919	B*3920	B*3922	B*3923	B*3924	B*3926
		B*3927	B*3928	B*3930	B*3932	B*3934	B*4010	B*4012	B*4021
		B*4040	B*4049	B*4058	B*4410	B*4431	B*4801	B*4803	B*4804
		B*4806	B*4807	B*4809	B*4810	B*4811	B*4812	B*4813	B*5101
		B*5102	B*5103	B*5104	B*5106	B*5107	B*5109	B*5111	B*5112
		B*5113	B*5115	B*5116	B*5117	B*5118	B*5119	B*5121	B*5122
		B*5123	B*5124	B*5126	B*5128	B*5130	B*5131	B*5132	B*5133
		B*5134	B*5135	B*5138	B*5201	B*5202	B*5203	B*5204	B*5205
		B*5206	B*5207	B*5208	B*5307	B*5401	B*5403	B*5404	B*5407
		B*5501	B*5502	B*5503	B*5504	B*5505	B*5507	B*5508	B*5510
		B*5511	B*5512	B*5513	B*5515	B*5517	B*5518	B*5519	B*5601
		B*5602	B*5604	B*5605	B*5606	B*5607	B*5608	B*5610	B*5612
		B*5614	B*5615	B*5616	B*5703	B*5707	B*5808	B*5901	B*6701
		B*6702	B*7801	B*7802	B*7803	B*7804	B*7805	B*8101	B*8102
7	DND	B*0801	B*0802	B*0804	B*0805	B*0809	B*0810	B*0811	B*0812
		B*0815	B*0816	B*0817	B*0818	B*0819	B*0821	B*0822	B*0823
		B*0824							
8	DNR	B*0806	B*0820						
9	DDD	B*0807							
10	DNT	B*0808							
11	DNL	B*0813	B*0825	B*3912					
12	DHD	B*0814							
13	YDW	B*1501	B*1504	B*1507	B*1508	B*1511	B*1512	B*1514	B*1515
		B*1519	B*1524	B*1527	B*1528	B*1532	B*1533	B*1534	B*1535
		B*1538	B*1543	B*1545	B*1546	B*1550	B*1553	B*1554	B*1556
		B*1557	B*1560	B*1563	B*1566	B*1568	B*1570	B*1575	B*1576
		B*1577	B*1578	B*1581	B*1582	B*1585	B*1587	B*1592	B*1596
		B*1597	B*3514	B*3543	B*4601	B*4602	B*4604	B*4605	B*5406
		B*5603	B*9501	B*9502	B*9504				
14	YHL	B*1517	B*4701	B*4702	B*4703	B*4704	B*4705		
15	YNW	B*1558	B*1573	B*3544	B*3918				
16	HDW	B*1571	B*1819						
17	YND	B*1583	B*4201	B*4204	B*4205	B*4206	B*4414	B*5108	B*5120
		B*5136	B*5613	B*5709	B*8201	B*8202			
18	HDL	B*1801	B*1802	B*1803	B*1805	B*1806	B*1807	B*1808	B*1809
		B*1810	B*1811	B*1812	B*1814	B*1815	B*1818	B*1820	B*2706
		B*2721	B*3550	B*4003	B*4009	B*4018	B*4020	B*4024	B*4031
		B*4033	B*4042	B*4044	B*4052	B*4059	B*4060	B*4904	
19	HDR	B*1813	B*3707	B*4032					
20	HHL	B*2701	B*2702	B*2703	B*2704	B*2705	B*2708	B*2709	B*2710
		B*2712	B*2713	B*2714	B*2715	B*2716	B*2717	B*2718	B*2719
		B*2723	B*2726	B*2728	B*2729	B*2730	B*3702		
21	HNL	B*2707	B*2711	B*2720	B*2724	B*2727	B*3931	B*4001	B*4002
		B*4004	B*4005	B*4006	B*4007	B*4008	B*4011	B*4013	B*4014
		B*4019	B*4025	B*4026	B*4027	B*4028	B*4029	B*4030	B*4034
		B*4035	B*4036	B*4037	B*4038	B*4039	B*4043	B*4045	B*4046
		B*4047	B*4048	B*4050	B*4051	B*4053	B*4054	B*4055	B*4056
		B*4057	B*4061	B*4901	B*4902	B*4903	B*5001	B*5002	B*5004
		B*5402	B*5516	B*7301					

22	HHW	B*2725
		B*3701 B*3704 B*3705 B*3706 B*3708 B*4101 B*4102 B*4103
23	HND	B*4105 B*4106 B*4107 B*4202 B*4415 B*4418 B*4501 B*4502
		B*4503 B*4504 B*4505 B*4506 B*4507
24	HDD	B*3709 B*4104
25	HNR	B*4015 B*4016 B*4023
26	YDE	B*4435
27	YKL	B*5114

e.

Group Pocket E		HLA Allele							
1	SDER	B*0702	B*0703	B*0708	B*0709	B*0710	B*0711	B*0713	B*0715
		B*0716	B*0717	B*0720	B*0721	B*0722	B*0723	B*0724	B*0727
		B*0728	B*0729	B*0730	B*0731	B*0733	B*0735	B*0736	B*0737
		B*0738	B*0739	B*0741	B*0742	B*3707	B*4032	B*4808	
2	SDED	B*0704	B*0719						
3	SNER	B*0705	B*0706	B*0732	B*0740	B*0806	B*0820	B*4015	B*4016
		B*4805							
4	RDER	B*0707	B*0712	B*0714	B*0718	B*1547	B*1549	B*1565	B*1599
		B*4603							
5	SDVD	B*0725	B*0807	B*3709	B*4104				
6	SDEL	B*0726	B*0743	B*3558					
7	SNEL	B*0734	B*1405	B*2724	B*4005				
		B*0801	B*0802	B*0804	B*0805	B*0810	B*0811	B*0815	B*0817
8	SNVD	B*0818	B*0819	B*0821	B*0822	B*0823	B*0824	B*4102	B*4201
		B*4202	B*4205	B*4206					
9	SNWT	B*0808							
10	TNVD	B*0809	B*1583	B*4204	B*5108	B*5120	B*5613		
		B*0812	B*0816	B*3701	B*3704	B*3705	B*3706	B*3708	B*4101
11	RNVD	B*4103	B*4105	B*4106	B*4107	B*4414	B*4415	B*4418	B*4501
		B*4502	B*4503	B*4504	B*4505	B*4506	B*4507	B*8201	B*8202
		B*0813	B*0825	B*2707	B*2711	B*3522	B*3531	B*3903	B*3914
		B*3924	B*4002	B*4008	B*4013	B*4019	B*4027	B*4029	B*4035
12	SNVL	B*4037	B*4039	B*4040	B*4045	B*4050	B*4056	B*4057	B*4058
		B*4801	B*4804	B*4806	B*4807	B*4809	B*4810	B*4811	B*4812
		B*4813	B*5504	B*8101	B*8102				
13	SHVD	B*0814							
		B*1301	B*1306	B*1311	B*1312	B*1313	B*1548	B*1586	B*3502
		B*3504	B*3506	B*3509	B*3512	B*3801	B*3802	B*3803	B*3804
		B*3805	B*3806	B*3807	B*3808	B*3809	B*3811	B*3901	B*3902
		B*3904	B*3905	B*3909	B*3910	B*3912	B*3913	B*3916	B*3917
14	RNVL	B*3919	B*3920	B*3922	B*3923	B*3926	B*3927	B*3930	B*3931
		B*3932	B*4001	B*4004	B*4007	B*4010	B*4011	B*4012	B*4014
		B*4021	B*4025	B*4036	B*4038	B*4043	B*4046	B*4047	B*4048
		B*4049	B*4054	B*4055	B*4061	B*4410	B*4431	B*4803	B*5307
		B*5508	B*5602	B*5604	B*5610	B*5612	B*6701	B*6702	
		B*1302	B*1303	B*1308	B*1309	B*3560	B*3906	B*3928	B*3934
		B*4006	B*4053	B*5109	B*5115	B*5119	B*5131	B*5203	B*5401
15	TNVL	B*5402	B*5403	B*5404	B*5407	B*5502	B*5507	B*5510	B*5512
		B*5513	B*5516	B*5518	B*5519	B*5601	B*5607	B*5608	B*5614
		B*5615	B*5616	B*5901	B*7301	B*7804			

16	TDVL	B*1304 B*3537 B*4044
		B*1310 B*1505 B*1520 B*1531 B*1552 B*1591 B*1801 B*1803
		B*1804 B*1805 B*1806 B*1807 B*1808 B*1809 B*1810 B*1811
		B*1812 B*1818 B*1820 B*3501 B*3503 B*3507 B*3513 B*3515
		B*3519 B*3520 B*3523 B*3524 B*3525 B*3526 B*3527 B*3528
		B*3529 B*3532 B*3533 B*3534 B*3535 B*3536 B*3539 B*3540
17	RDVL	B*3541 B*3542 B*3546 B*3547 B*3548 B*3549 B*3550 B*3552
		B*3554 B*3555 B*3556 B*3557 B*3915 B*4020 B*4033 B*4042
		B*4052 B*4059 B*4060 B*4403 B*4407 B*4413 B*4426 B*4429
		B*4430 B*4432 B*4436 B*4437 B*4438 B*4439 B*4440 B*4802
		B*5301 B*5302 B*5303 B*5304 B*5305 B*5309 B*5310 B*5609
		B*5611 B*5801 B*5804 B*5809 B*5811
18	WNEL	B*1401 B*1402 B*1404
19	WNER	B*1403
20	RNEL	B*1406 B*1509 B*3810 B*4026 B*4028 B*4051 B*4901 B*4902
		B*4903 B*5001 B*5002 B*5004 B*5104 B*5106
		B*1501 B*1508 B*1511 B*1512 B*1514 B*1515 B*1519 B*1524
		B*1527 B*1528 B*1532 B*1533 B*1534 B*1538 B*1543 B*1546
21	RDEW	B*1550 B*1553 B*1554 B*1556 B*1557 B*1560 B*1563 B*1566
		B*1570 B*1571 B*1576 B*1577 B*1578 B*1581 B*1582 B*1585
		B*1587 B*1592 B*1596 B*1597 B*1819 B*3514 B*3543 B*4601
		B*4602 B*4604 B*4605 B*5406 B*5603 B*9501 B*9502 B*9504
		B*1502 B*1503 B*1506 B*1510 B*1513 B*1516 B*1518 B*1521
		B*1523 B*1525 B*1529 B*1536 B*1537 B*1539 B*1540 B*1544
22	RDEL	B*1561 B*1562 B*1564 B*1567 B*1569 B*1572 B*1574 B*1580
		B*1589 B*1590 B*1593 B*1595 B*1598 B*1815 B*2721 B*3511
		B*3521 B*4904 B*5306 B*5308 B*9503
23	TDEW	B*1504 B*1535
24	SDEW	B*1507 B*1545 B*1568
25	RHEL	B*1517
		B*1542 B*5101 B*5102 B*5103 B*5107 B*5111 B*5112 B*5113
		B*5116 B*5117 B*5118 B*5121 B*5122 B*5123 B*5124 B*5126
26	TNEL	B*5128 B*5130 B*5132 B*5133 B*5134 B*5135 B*5138 B*5201
		B*5202 B*5204 B*5205 B*5206 B*5207 B*5208 B*5501 B*5503
		B*5505 B*5515 B*5517 B*5605 B*5606 B*5808 B*7801 B*7802
		B*7803 B*7805
		B*1551 B*3538 B*4402 B*4405 B*4406 B*4408 B*4409 B*4411
27	RDVD	B*4412 B*4416 B*4417 B*4421 B*4422 B*4424 B*4425 B*4427
		B*4433 B*4434 B*4441 B*4442 B*8301
28	SDVL	B*1555 B*1814 B*3505 B*3516 B*3517 B*3530 B*3551 B*3929
		B*4003 B*4009 B*4018 B*4024 B*4031
29	RNEW	B*1558 B*1573 B*3544 B*3918
30	RDQW	B*1575
31	RDVR	B*1584 B*1813 B*3508 B*3545 B*3561 B*4404 B*4428
32	RDTL	B*1588
33	NDVL	B*1802
		B*2701 B*2702 B*2703 B*2705 B*2708 B*2709 B*2712 B*2713
34	NHVL	B*2716 B*2717 B*2723 B*2726 B*2728 B*2729 B*3702 B*4704
		B*4705
35	NHEL	B*2704 B*2710 B*2715 B*2718
36	NDEL	B*2706
37	THVL	B*2714
38	RHVL	B*2719 B*2730 B*4701 B*4702 B*4703

39	NNEL	B*2720
40	NHEW	B*2725
41	NNVL	B*2727
42	RNVR	B*3518 B*3908 B*5705
43	TNVR	B*3933 B*5105
44	RNER	B*4023
45	VNVL	B*4030 B*4034 B*5703 B*5707
46	TDVD	B*4420
47	RDVE	B*4435
48	TKEL	B*5114
49	TNER	B*5129 B*5509
50	TNED	B*5136
51	TDEL	B*5137
52	MNEL	B*5511
53	VDEL	B*5514
54	VDVL	B*5701 B*5706 B*5708
55	VNVR	B*5702
56	VDVR	B*5704
57	VNED	B*5709
58	WDVL	B*5802 B*5807
59	RDAL	B*5805
60	WDEL	B*5806

f.

Group	Pocket	F	HLA Allele							
1	SNLY		B*0702	B*0703	B*0704	B*0705	B*0706	B*0707	B*0708	B*0710
			B*0712	B*0713	B*0714	B*0715	B*0716	B*0718	B*0719	B*0720
			B*0721	B*0722	B*0723	B*0724	B*0725	B*0726	B*0729	B*0730
			B*0731	B*0732	B*0733	B*0734	B*0735	B*0737	B*0739	B*0740
			B*0741	B*0742	B*0743	B*0801	B*0804	B*0805	B*0807	B*0808
			B*0809	B*0810	B*0811	B*0812	B*0813	B*0815	B*0816	B*0818
			B*0819	B*0820	B*0821	B*0822	B*0823	B*0824	B*0825	B*1509
			B*1510	B*1537	B*1545	B*1548	B*1563	B*1583	B*1590	B*1599
			B*3502	B*3504	B*3509	B*3512	B*3518	B*3522	B*3531	B*3534
			B*3539	B*3544	B*3914	B*3929	B*4001	B*4002	B*4004	B*4005
			B*4006	B*4007	B*4008	B*4009	B*4010	B*4011	B*4012	B*4014
			B*4015	B*4016	B*4018	B*4021	B*4023	B*4024	B*4025	B*4026
			B*4028	B*4029	B*4030	B*4031	B*4032	B*4033	B*4034	B*4035
			B*4039	B*4040	B*4042	B*4043	B*4044	B*4045	B*4046	B*4049
			B*4050	B*4051	B*4053	B*4054	B*4055	B*4056	B*4057	B*4058
			B*4061	B*4101	B*4102	B*4103	B*4104	B*4106	B*4107	B*4201
			B*4202	B*4204	B*4205	B*4801	B*4803	B*4804	B*4805	B*4806
			B*4808	B*4809	B*4810	B*4811	B*4812	B*4813	B*5404	B*5504
			B*5508	B*5517	B*5605	B*5606	B*5612	B*5615	B*7801	B*7802
			B*7803	B*7804	B*7805	B*8101	B*8102			
2	SNLS		B*0709	B*0717	B*1501	B*1502	B*1503	B*1504	B*1505	B*1506
			B*1507	B*1508	B*1511	B*1512	B*1514	B*1515	B*1518	B*1519
			B*1520	B*1521	B*1525	B*1527	B*1528	B*1529	B*1531	B*1532
			B*1533	B*1534	B*1535	B*1538	B*1539	B*1540	B*1544	B*1546
			B*1547	B*1549	B*1550	B*1551	B*1552	B*1553	B*1554	B*1555
			B*1556	B*1560	B*1561	B*1562	B*1564	B*1565	B*1566	B*1568

		B*1569	B*1570	B*1571	B*1572	B*1574	B*1575	B*1576	B*1578
		B*1580	B*1581	B*1582	B*1584	B*1585	B*1588	B*1591	B*1592
		B*1593	B*1596	B*1597	B*1598	B*1801	B*1802	B*1803	B*1804
		B*1805	B*1806	B*1807	B*1808	B*1810	B*1811	B*1812	B*1813
		B*1814	B*1815	B*1818	B*1819	B*1820	B*3501	B*3505	B*3507
		B*3508	B*3511	B*3514	B*3515	B*3516	B*3517	B*3519	B*3520
		B*3521	B*3523	B*3524	B*3525	B*3526	B*3528	B*3529	B*3530
		B*3532	B*3535	B*3537	B*3540	B*3541	B*3542	B*3543	B*3545
		B*3546	B*3547	B*3548	B*3549	B*3550	B*3551	B*3554	B*3557
		B*3558	B*3561	B*4003	B*4020	B*4038	B*4052	B*4059	B*4060
		B*4601	B*4602	B*4603	B*4604	B*4605	B*4802	B*5406	B*5514
		B*5603	B*5609	B*9501	B*9502	B*9503	B*9504		
3	NNLS	B*0711	B*1557	B*3527					
4	DTLY	B*0727	B*2707	B*2727	B*3707	B*3709			
5	SNLD	B*0728	B*0814	B*1577	B*2708	B*2712	B*2718	B*2726	B*4409
		B*4702	B*4703	B*8301					
		B*0736	B*0738	B*4013	B*4019	B*4425	B*5101	B*5102	B*5103
		B*5104	B*5105	B*5106	B*5107	B*5108	B*5109	B*5111	B*5112
		B*5114	B*5116	B*5117	B*5118	B*5119	B*5120	B*5121	B*5122
6	NIAY	B*5123	B*5124	B*5126	B*5128	B*5129	B*5130	B*5131	B*5132
		B*5133	B*5134	B*5135	B*5136	B*5138	B*5201	B*5202	B*5203
		B*5204	B*5205	B*5206	B*5207	B*5208	B*5702	B*5703	B*5705
		B*5707	B*5709	B*5808					
7	NTAY	B*0802	B*1313	B*4047	B*4405	B*4431	B*4442		
8	NNLY	B*0806	B*0817	B*4037					
9	NTAL	B*1301	B*1302	B*1303	B*1306	B*1308	B*1311	B*1312	B*4410
		B*4902	B*5607						
10	NTAS	B*1304	B*1310	B*1536	B*1589	B*1809	B*4417	B*5309	
		B*1309	B*1542	B*1573	B*1586	B*3560	B*3917	B*4048	B*4501
		B*4503	B*4504	B*4505	B*4506	B*4507	B*5001	B*5002	B*5004
11	SNLL	B*5401	B*5402	B*5403	B*5407	B*5501	B*5502	B*5503	B*5505
		B*5507	B*5509	B*5510	B*5511	B*5515	B*5516	B*5518	B*5519
		B*5601	B*5602	B*5604	B*5608	B*5610	B*5613	B*5614	B*5616
		B*8201	B*8202						
		B*1401	B*1402	B*1403	B*1404	B*1405	B*1406	B*1558	B*3503
		B*3506	B*3513	B*3533	B*3536	B*3538	B*3555	B*3901	B*3902
12	SNLF	B*3903	B*3904	B*3905	B*3906	B*3908	B*3909	B*3910	B*3912
		B*3913	B*3915	B*3916	B*3918	B*3919	B*3922	B*3923	B*3924
		B*3926	B*3927	B*3928	B*3930	B*3931	B*3932	B*3933	B*3934
		B*4036	B*4206	B*4502	B*4807	B*5513	B*5611	B*6701	B*6702
		B*1513	B*1516	B*1523	B*1524	B*1567	B*1595	B*4904	B*5137
13	NIAS	B*5301	B*5302	B*5306	B*5308	B*5310	B*5701	B*5706	B*5708
		B*5801	B*5802	B*5804	B*5805	B*5806	B*5807	B*5809	B*5811
14	NIAD	B*1517	B*2702	B*2730	B*4406	B*5704			
15	DTLS	B*1543	B*5303						
16	SIAS	B*1587	B*5305						
		B*2701	B*4402	B*4403	B*4404	B*4407	B*4408	B*4412	B*4413
17	NTAD	B*4414	B*4416	B*4420	B*4421	B*4422	B*4424	B*4426	B*4427
		B*4428	B*4429	B*4430	B*4432	B*4433	B*4434	B*4435	B*4436
		B*4437	B*4438	B*4439	B*4440	B*4441	B*4704		
18	DTLD	B*2703	B*2705	B*2710	B*2713	B*2714	B*2716	B*2717	B*2719
		B*2728	B*2729	B*3702	B*4701	B*4705			
19	STLD	B*2704	B*2715	B*2723	B*2725				
20	STLY	B*2706	B*2711	B*2720	B*2721	B*2724			

21	DTLH	B*2709
22	CNLS	B*3552
23	NNLF	B*3556 B*3920
24	DTLF	B*3701 B*3704 B*3706 B*3708
25	DNLF	B*3705
26	NIAF	B*3801 B*3805 B*3806 B*3807 B*3809 B*3810 B*3811 B*5113 B*5304 B*5307
27	NTAF	B*3802 B*3804 B*3808
28	STAF	B*3803
29	SNLN	B*4027
30	SKLY	B*4105
31	NTPD	B*4411
32	STAL	B*4415
33	NIAL	B*4418 B*4901 B*4903 B*5115 B*5901
34	NNLL	B*5512
35	GNLF	B*7301

The classification of class I HLA is based on IMGT/HLA Sequence Database Release 2.10 (released in July 2005).

Group: The pocket group with same positional residue constitution.

CPFR: critical polymorphic functional residue, the polymorphic amino acid residues that directly contact with binding peptides in the pockets, placed according to their original positions on the HLA chains, designated in single capital letter.

HLA Alleles: The class I HLA alleles that belong to the group, with sequence based nomenclature begin with “B*” and followed by a four digit number.

APPENDIX 3

Appendix 3 Classification of pockets of HLA-Cw alleles based on CPFRs.**a.**

Group	Pocket A	HLA Allele						
1	YEKQCTW	Cw*0102	Cw*0103	Cw*0104	Cw*0106	Cw*0107	Cw*0108	Cw*0109
		Cw*0111						
2	YEKQSTW	Cw*0105	Cw*0702	Cw*0710	Cw*0713	Cw*0717	Cw*0723	Cw*0725
		Cw*0729	Cw*1211	Cw*1606				
3	YEKQCEW	Cw*0110						
4	YEKQYEW	Cw*0202	Cw*0203	Cw*0205	Cw*0206	Cw*0207	Cw*0208	Cw*0209
		Cw*0210	Cw*0211	Cw*0212	Cw*0608	Cw*1701	Cw*1702	Cw*1703
5	CEKQYEW	Cw*0204						
6	YEKQYLW	Cw*0302	Cw*0303	Cw*0304	Cw*0305	Cw*0306	Cw*0307	Cw*0309
		Cw*0310	Cw*0311	Cw*0312	Cw*0313	Cw*0315	Cw*0317	Cw*0318
		Cw*0319						
7	YENQYLW	Cw*0308						
		Cw*0314	Cw*0316	Cw*0415	Cw*0417	Cw*0501	Cw*0503	Cw*0504
		Cw*0505	Cw*0506	Cw*0508	Cw*0509	Cw*0510	Cw*0511	Cw*0602
		Cw*0603	Cw*0604	Cw*0605	Cw*0606	Cw*0607	Cw*0609	Cw*0610
		Cw*0611	Cw*0612	Cw*0613	Cw*0704	Cw*0705	Cw*0711	Cw*0712
8	YEKQYTW	Cw*0714	Cw*0727	Cw*0801	Cw*0802	Cw*0803	Cw*0804	Cw*0805
		Cw*0807	Cw*0808	Cw*0809	Cw*0810	Cw*0811	Cw*0812	Cw*1202
		Cw*1203	Cw*1204	Cw*1205	Cw*1206	Cw*1207	Cw*1209	Cw*1212
		Cw*1213	Cw*1214	Cw*1215	Cw*1405	Cw*1511	Cw*1601	Cw*1602
		Cw*1604	Cw*1607					
		Cw*0401	Cw*0403	Cw*0404	Cw*0405	Cw*0406	Cw*0407	Cw*0409
9	YEKQFTW	Cw*0410	Cw*0411	Cw*0412	Cw*0413	Cw*0414	Cw*0416	Cw*0708
		Cw*1210	Cw*1402	Cw*1403	Cw*1404	Cw*1406	Cw*1801	Cw*1802
10	YEKQFLW	Cw*0408						
11	YEKQYMW	Cw*0502						
		Cw*0701	Cw*0706	Cw*0707	Cw*0709	Cw*0716	Cw*0718	Cw*0720
		Cw*0721	Cw*0722	Cw*0724	Cw*0726	Cw*0728	Cw*1208	Cw*1502
12	YENQYTW	Cw*1503	Cw*1504	Cw*1505	Cw*1506	Cw*1507	Cw*1508	Cw*1509
		Cw*1510	Cw*1512	Cw*1513	Cw*1514			
13	YEKQSLW	Cw*0703						
14	YEKQSEW	Cw*0715						
15	YENQSTW	Cw*0719						
16	YEKQYAW	Cw*0806						

b.

Group	Pocket B	HLA Allele						
1	YFSREKQVC	Cw*0102	Cw*0103	Cw*0104	Cw*0106	Cw*0107	Cw*0108	Cw*0109
		Cw*0110	Cw*0111					
2	YFSREKQVS	Cw*0105						
		Cw*0202	Cw*0203	Cw*0205	Cw*0206	Cw*0207	Cw*0208	Cw*0209
		Cw*0210	Cw*0211	Cw*0212	Cw*0302	Cw*0303	Cw*0304	Cw*0305
3	YYAREKQVY	Cw*0306	Cw*0307	Cw*0309	Cw*0310	Cw*0312	Cw*0313	Cw*0314
		Cw*0315	Cw*0316	Cw*0317	Cw*0318	Cw*0319	Cw*1202	Cw*1203

		Cw*1204	Cw*1205	Cw*1206	Cw*1207	Cw*1209	Cw*1213	Cw*1214
		Cw*1215	Cw*1511	Cw*1601	Cw*1602	Cw*1604	Cw*1607	Cw*1701
		Cw*1702	Cw*1703					
4	CYAREKQVY	Cw*0204						
5	YYARENQVY	Cw*0308	Cw*1208	Cw*1502	Cw*1503	Cw*1504	Cw*1505	Cw*1506
		Cw*1507	Cw*1508	Cw*1509	Cw*1510	Cw*1512	Cw*1513	Cw*1514
6	YYSREKQVY	Cw*0311	Cw*0603					
		Cw*0401	Cw*0404	Cw*0405	Cw*0407	Cw*0408	Cw*0409	Cw*0410
7	YSAREKQVF	Cw*0411	Cw*0412	Cw*0413	Cw*0414	Cw*1402	Cw*1403	Cw*1404
		Cw*1406						
8	YYAREKQVF	Cw*0403	Cw*0406	Cw*0416	Cw*1210			
9	YSAREKQVY	Cw*0415	Cw*0417	Cw*1405				
		Cw*0501	Cw*0502	Cw*0503	Cw*0504	Cw*0505	Cw*0506	Cw*0508
10	YYAQEKQVY	Cw*0509	Cw*0510	Cw*0511	Cw*0801	Cw*0802	Cw*0803	Cw*0804
		Cw*0805	Cw*0806	Cw*0807	Cw*0808	Cw*0809	Cw*0810	Cw*0811
		Cw*0812						
		Cw*0602	Cw*0604	Cw*0606	Cw*0607	Cw*0608	Cw*0609	Cw*0611
11	YDSREKQVY	Cw*0612	Cw*0613	Cw*0704	Cw*0705	Cw*0711	Cw*0712	Cw*0714
		Cw*0727						
12	YDSQEKQVY	Cw*0605						
13	YDPREKQVY	Cw*0610						
14	YDSREKQVY	Cw*0701	Cw*0706	Cw*0707	Cw*0709	Cw*0716	Cw*0718	Cw*0721
		Cw*0722	Cw*0724	Cw*0728				
15	YDSREKQVS	Cw*0702	Cw*0703	Cw*0710	Cw*0713	Cw*0715	Cw*0717	Cw*0723
		Cw*0725	Cw*0729					
16	YDSREKQVF	Cw*0708	Cw*1801	Cw*1802				
17	YDSREKQVS	Cw*0719						
18	YDARENQVY	Cw*0720						
19	YYSREKQVY	Cw*0726						
20	YYAREKQVS	Cw*1211	Cw*1606					
21	YYAREKQVY	Cw*1212						

C.

Group	Pocket	C	HLA Allele					
1	FQTDW	Cw*0102	Cw*0103	Cw*0104	Cw*0106	Cw*0107	Cw*0108	Cw*0109
		Cw*0110	Cw*0111					
2	FQTDR	Cw*0105						
		Cw*0202	Cw*0203	Cw*0204	Cw*0206	Cw*0207	Cw*0208	Cw*0209
		Cw*0210	Cw*0211	Cw*0302	Cw*0303	Cw*0304	Cw*0306	Cw*0307
		Cw*0308	Cw*0309	Cw*0310	Cw*0311	Cw*0312	Cw*0313	Cw*0314
		Cw*0316	Cw*0318	Cw*0319	Cw*0501	Cw*0502	Cw*0503	Cw*0504
3	YQTDR	Cw*0505	Cw*0506	Cw*0508	Cw*0509	Cw*0510	Cw*0511	Cw*0801
		Cw*0802	Cw*0803	Cw*0804	Cw*0806	Cw*0807	Cw*0809	Cw*0810
		Cw*0811	Cw*1502	Cw*1504	Cw*1505	Cw*1506	Cw*1507	Cw*1508
		Cw*1509	Cw*1510	Cw*1511	Cw*1512	Cw*1513	Cw*1514	Cw*1606
		Cw*1607						
4	YQTDW	Cw*0205	Cw*0317	Cw*0812	Cw*1205	Cw*1209	Cw*1601	Cw*1602
		Cw*1604						
5	YQADR	Cw*0212	Cw*0315	Cw*0403	Cw*0406	Cw*0416	Cw*0726	Cw*0805
		Cw*1202	Cw*1208	Cw*1210	Cw*1214	Cw*1503	Cw*1701	Cw*1702

		Cw*1703							
6	YQTDS	Cw*0305	Cw*0808						
7	SQADR	Cw*0401	Cw*0404	Cw*0405	Cw*0407	Cw*0408	Cw*0409	Cw*0412	
		Cw*0413	Cw*0414	Cw*0415	Cw*0417				
8	SQTDR	Cw*0410	Cw*0411						
9	DQADW	Cw*0602	Cw*0604	Cw*0606	Cw*0607	Cw*0608	Cw*0609	Cw*0610	
		Cw*0611	Cw*0612	Cw*0613					
10	YQADW	Cw*0603	Cw*1203	Cw*1204	Cw*1206	Cw*1207	Cw*1211	Cw*1212	
		Cw*1213	Cw*1215						
11	DQTDW	Cw*0605							
		Cw*0701	Cw*0702	Cw*0703	Cw*0704	Cw*0706	Cw*0707	Cw*0708	
		Cw*0709	Cw*0710	Cw*0711	Cw*0712	Cw*0713	Cw*0714	Cw*0715	
12	DQADR	Cw*0716	Cw*0717	Cw*0718	Cw*0719	Cw*0720	Cw*0721	Cw*0722	
		Cw*0723	Cw*0724	Cw*0725	Cw*0727	Cw*0728	Cw*0729	Cw*1801	
		Cw*1802							
13	DQADN	Cw*0705							
14	SQTDW	Cw*1402	Cw*1403	Cw*1405	Cw*1406				
15	SQADW	Cw*1404							

d.

Group Pocket D		HLA Allele							
1	FDR	Cw*0102	Cw*0105	Cw*0106	Cw*0107	Cw*0108	Cw*0110	Cw*0111	
2	FNR	Cw*0103							
3	FDW	Cw*0104	Cw*0109						
		Cw*0202	Cw*0204	Cw*0205	Cw*0206	Cw*0208	Cw*0209	Cw*0210	
4	YDW	Cw*0211	Cw*0212	Cw*0603	Cw*1202	Cw*1203	Cw*1204	Cw*1205	
		Cw*1206	Cw*1207	Cw*1208	Cw*1210	Cw*1211	Cw*1212	Cw*1213	
		Cw*1215	Cw*1604						
		Cw*0203	Cw*0302	Cw*0303	Cw*0304	Cw*0305	Cw*0307	Cw*0308	
		Cw*0309	Cw*0310	Cw*0311	Cw*0312	Cw*0313	Cw*0314	Cw*0315	
5	YDL	Cw*0316	Cw*0317	Cw*0318	Cw*0319	Cw*0726	Cw*0809	Cw*0811	
		Cw*1214	Cw*1502	Cw*1503	Cw*1504	Cw*1505	Cw*1506	Cw*1507	
		Cw*1509	Cw*1510	Cw*1511	Cw*1512	Cw*1513	Cw*1514		
6	YHW	Cw*0207							
7	YVL	Cw*0306							
		Cw*0401	Cw*0405	Cw*0407	Cw*0408	Cw*0409	Cw*0410	Cw*0411	
8	SNR	Cw*0412	Cw*0415	Cw*0417					
		Cw*0403	Cw*0416	Cw*0501	Cw*0502	Cw*0503	Cw*0505	Cw*0506	
9	YNR	Cw*0509	Cw*0510	Cw*0802	Cw*0805	Cw*0807	Cw*0812		
10	SNL	Cw*0404	Cw*0413						
		Cw*0406	Cw*0511	Cw*0801	Cw*0803	Cw*0804	Cw*0806	Cw*0808	
11	YNL	Cw*0810	Cw*1701	Cw*1702	Cw*1703				
12	SNQ	Cw*0414							
13	YDR	Cw*0504	Cw*1209	Cw*1508					
14	YNW	Cw*0508							
		Cw*0602	Cw*0607	Cw*0608	Cw*0610	Cw*0611	Cw*0612	Cw*0613	
15	DDW	Cw*0722							

16	DDL	Cw*0604	Cw*0701	Cw*0702	Cw*0703	Cw*0705	Cw*0706	Cw*0707
		Cw*0708	Cw*0709	Cw*0710	Cw*0713	Cw*0714	Cw*0715	Cw*0716
		Cw*0717	Cw*0718	Cw*0719	Cw*0720	Cw*0721	Cw*0723	Cw*0724
		Cw*0725	Cw*0727	Cw*0729				
17	DDR	Cw*0605	Cw*0606					
18	DNW	Cw*0609						
19	DDD	Cw*0704	Cw*0711	Cw*0712				
20	DNL	Cw*0728						
21	SDR	Cw*1402	Cw*1403	Cw*1404	Cw*1405			
22	SDL	Cw*1406						
23	YDQ	Cw*1601	Cw*1602	Cw*1606	Cw*1607			
24	DNR	Cw*1801	Cw*1802					

e.

Group	Pocket	E	HLA Allele						
1	WDER		Cw*0102	Cw*0107	Cw*0108	Cw*0110	Cw*0111	Cw*0605	Cw*0606
			Cw*1209	Cw*1402	Cw*1403	Cw*1404	Cw*1405		
2	WNER		Cw*0103	Cw*0812					
			Cw*0104	Cw*0109	Cw*0205	Cw*0602	Cw*0603	Cw*0607	Cw*0608
3	WDEW		Cw*0610	Cw*0611	Cw*0612	Cw*0613	Cw*1203	Cw*1204	Cw*1205
			Cw*1206	Cw*1207	Cw*1211	Cw*1212	Cw*1213	Cw*1215	
4	RDER		Cw*0105	Cw*0504	Cw*1508				
5	WDVR		Cw*0106						
6	RDEW		Cw*0202	Cw*0204	Cw*0206	Cw*0208	Cw*0209	Cw*0210	Cw*0211
			Cw*0212	Cw*1202	Cw*1208	Cw*1210			
7	RDVL		Cw*0203						
8	RHEW		Cw*0207						
			Cw*0302	Cw*0303	Cw*0304	Cw*0307	Cw*0308	Cw*0309	Cw*0310
9	RDEL		Cw*0311	Cw*0312	Cw*0313	Cw*0315	Cw*0316	Cw*0318	Cw*0319
			Cw*1502	Cw*1503	Cw*1504	Cw*1505	Cw*1506	Cw*1507	Cw*1509
			Cw*1510	Cw*1511	Cw*1512	Cw*1513	Cw*1514		
10	SDEL		Cw*0305						
11	RVEL		Cw*0306						
			Cw*0314	Cw*0701	Cw*0702	Cw*0703	Cw*0706	Cw*0707	Cw*0708
12	RDAL		Cw*0709	Cw*0710	Cw*0713	Cw*0714	Cw*0715	Cw*0716	Cw*0717
			Cw*0718	Cw*0719	Cw*0720	Cw*0721	Cw*0723	Cw*0724	Cw*0725
			Cw*0726	Cw*0727	Cw*0729	Cw*1214			
13	WDEL		Cw*0317	Cw*0604	Cw*1406				
			Cw*0401	Cw*0403	Cw*0405	Cw*0407	Cw*0408	Cw*0409	Cw*0410
14	RNER		Cw*0411	Cw*0412	Cw*0415	Cw*0416	Cw*0417	Cw*0501	Cw*0502
			Cw*0503	Cw*0505	Cw*0506	Cw*0509	Cw*0510	Cw*0802	Cw*0805
			Cw*0807	Cw*1801	Cw*1802				
15	RNEL		Cw*0404	Cw*0406	Cw*0511	Cw*0804	Cw*1701	Cw*1702	Cw*1703
16	RNAL		Cw*0413	Cw*0728					
17	RNEQ		Cw*0414						

18	RNEW	Cw*0508
19	WNEW	Cw*0609
20	RDAD	Cw*0704 Cw*0711 Cw*0712
21	NDAL	Cw*0705
22	RNAW	Cw*0722
23	RNTL	Cw*0801 Cw*0803 Cw*0806 Cw*0810
24	SNTL	Cw*0808
25	RDTL	Cw*0809 Cw*0811
26	WDAQ	Cw*1601 Cw*1602
27	WDAW	Cw*1604
28	RDAQ	Cw*1606 Cw*1607

f.

Group Pocket F		HLA Allele						
1	SNLY	Cw*0102	Cw*0105	Cw*0106	Cw*0107	Cw*0108	Cw*0109	Cw*0110
		Cw*0111	Cw*0303	Cw*0304	Cw*0305	Cw*0306	Cw*0308	Cw*0309
		Cw*0311	Cw*0313	Cw*0317	Cw*0318	Cw*0729		
2	SNLF	Cw*0103	Cw*0411	Cw*0704	Cw*0711	Cw*0712	Cw*0713	Cw*0728
		Cw*0801	Cw*0802	Cw*0803	Cw*0804	Cw*0805	Cw*0806	Cw*0807
		Cw*0808	Cw*0812					
3	SNLS	Cw*0104	Cw*0302	Cw*0314	Cw*0316	Cw*0611	Cw*0701	Cw*0702
		Cw*0703	Cw*0705	Cw*0706	Cw*0708	Cw*0710	Cw*0714	Cw*0715
		Cw*0716	Cw*0717	Cw*0718	Cw*0719	Cw*0720	Cw*0721	Cw*0722
		Cw*0723	Cw*0724	Cw*0725	Cw*0726	Cw*0727	Cw*0809	Cw*0811
		Cw*1202	Cw*1203	Cw*1206	Cw*1208	Cw*1210	Cw*1211	Cw*1212
		Cw*1213	Cw*1214	Cw*1402	Cw*1403	Cw*1405	Cw*1406	Cw*1601
		Cw*1604	Cw*1606	Cw*1607				
4	NKLS	Cw*0202	Cw*0203	Cw*0204	Cw*0205	Cw*0208	Cw*0209	Cw*0210
		Cw*0211	Cw*0315	Cw*0504	Cw*0602	Cw*0603	Cw*0604	Cw*0605
		Cw*0606	Cw*0607	Cw*0608	Cw*0610	Cw*0612	Cw*0613	Cw*0707
		Cw*0709	Cw*1204	Cw*1205	Cw*1209	Cw*1504	Cw*1509	Cw*1602
5	NKLL	Cw*0206	Cw*1502	Cw*1503	Cw*1508	Cw*1510	Cw*1511	Cw*1512
		Cw*1513	Cw*1514					
6	NKLD	Cw*0207						
7	SKLS	Cw*0212						
8	NKLY	Cw*0307	Cw*1506					
9	SKLY	Cw*0310						
10	SNLL	Cw*0312	Cw*0319	Cw*1215	Cw*1507			
		Cw*0401	Cw*0403	Cw*0404	Cw*0405	Cw*0406	Cw*0407	Cw*0408
11	NKLF	Cw*0409	Cw*0410	Cw*0412	Cw*0413	Cw*0414	Cw*0415	Cw*0416
		Cw*0417	Cw*0501	Cw*0502	Cw*0503	Cw*0505	Cw*0506	Cw*0508
		Cw*0509	Cw*0510	Cw*0511	Cw*0609	Cw*0810	Cw*1505	Cw*1701
		Cw*1702	Cw*1703	Cw*1801	Cw*1802			
12	GNLS	Cw*1207						
13	NNLS	Cw*1404						

The classification of class I HLA is based on IMGT/HLA Sequence Database Release 2.10 (released in July 2005).

Group: The pocket group with same positional residue constitution.

CPFR: critical polymorphic functional residue, the polymorphic amino acid residues that directly contact with binding peptides in the pockets, placed according to their original positions on the HLA chains, designated in single capital letter.

HLA Alleles: The class I HLA alleles that belong to the group, with sequence based nomenclature begin with “Cw*” and followed by a four digit number.

APPENDIX 4

PUBLICATIONS

PUBLICATIONS

1. Zhao B, Png AE, Ren EC, Kolatkar PR, Mathura VS, Sakharkar MK & Kanguane P (2003). Compression of functional space in HLA-A sequence diversity. *Human Immunology* 64(7): 718-728.
2. Zhao B, Mathura VS, Rajaseger G, Moochhala S, Sakharkar MK & Kanguane P (2003). A novel MHCp binding prediction model. *Human Immunology* 64(12): 1123-1143.
3. Kanguane P, Sakharkar MK, Rajaseger G, Bolisetty S, Sivasekari B, Zhao B, Ravichandran M, Shapshak P & Subbiah S (2005). A framework to sub-type HLA supertypes. *Frontiers in Bioscience* 10: 879-886.

Compression of Functional Space in HLA-A Sequence Diversity

Bing Zhao, Adrian Eak H. Png, Ee Chee Ren,
Prasanna R. Kolatkar,
Venkatarajan Subramaniam Mathura,
Meena Kishore Sakharkar, and
Pandjassaram Kanguane

ABSTRACT: The major histocompatibility complex (MHC) is highly polymorphic and more than 1500 human MHC alleles are known to date. These alleles do not bind to a given peptide with identical affinity. Although MHC alleles are functionally related, it is difficult to quantify the functional variation between them. Three-dimensional structures of known MHC-peptide (MHCp) complexes suggest that specific peptide residues bind selectively to functional pockets in the binding groove. From a set of known MHCp structures we identified 21

critical polymorphic functional residue positions (CPFRP) that significantly reduced functional pocket variability to just 189 among 212 HLA-A alleles. Interestingly 101 HLA-A alleles clustered into 29 clusters such that the six functional pockets formed by the CPFRPs are identical within the cluster. *Human Immunology* 64, 718–728 (2003). © American Society for Histocompatibility and Immunogenetics, 2003. Published by Elsevier Inc.

KEYWORDS: MHC; peptide binding; functional pocket

ABBREVIATIONS

ASA accessible surface area
CPFRP critical polymorphic functional residue position
CTL cytotoxic T lymphocyte
ERP essential residue position

HLA human leukocyte antigen
MHC major histocompatibility complex
MHCp MHC peptide
TCR T-cell receptor

INTRODUCTION

The formation of class I MHCp complex is a rate-limiting step in cytotoxic T-lymphocyte (CTL) mediated immune response [1]. Peptides (8–11 residues) degraded from cytosolic proteins are bound to class I molecules in the endoplasmic reticulum and the MHCp complexes are recognized by T-cell receptors (TCR) [2, 3]. Peptide sequence pooling and MHCp structures indicate that the

sequence-specific binding of MHC and peptide is largely due to polymorphism in the binding grooves [3–8]. The governing principles of MHCp interactions are roughly understood from known crystal structures [4, 5, 9–13]. Peptide side chains at anchor positions are accommodated in the pockets of the groove with different geometry and chemical properties [13–16].

The relationship between peptide anchor residues and MHC pockets has been extensively studied [9, 15–19]. Certain types of residues at specific peptide anchor positions generally comply with the cumulative physical and chemical properties of the residues in the corresponding binding pockets. It has been found that residues at some anchor positions binding to selective pockets in specific MHC alleles are primarily characterized by closely related residues [3, 20–23]. The side chains at the second and third positions of the peptides are oriented in

From the School of MPE & NCSV (B.Z., M.K.S., P.K.), Nanyang Technological University, the Department of Microbiology (A.E.H.P., E.C.R.), National University of Singapore, and the Genome Institute of Singapore (E.C.R., P.R.K.), the Capricorn, Singapore; and the Department of Human Genetics (V.S.M.), University of Texas Medical Branch, Galveston, TX, USA.

Address reprint requests to: Dr. Pandjassaram Kanguane, School of Mechanical and Production Engineering, 50 Nanyang Avenue, Nanyang Technological University, Singapore 639 798; Tel: +65 (6) 790-5836; Fax: +65 (6) 774-4340; ; E-mail: mpandjassaram@ntu.edu.sg.

Received January 8, 2003; revised March 26, 2003; accepted March 27, 2003.

Human Immunology 64, 718–728 (2003)

© American Society for Histocompatibility and Immunogenetics, 2003
Published by Elsevier Inc.

0198-8859/03/\$—see front matter
doi:10.1016/S0198-8859(03)00078-8

TABLE 1 Major histocompatibility complex functional property matrix

Pockets	A							B								
CPFRP	7	63	66	70	99	163	167	7	9	24	35	63	66	70	76	99
ALA (A)	0	0	0	0	0	0	0	0	0	211	0	0	0	0	32	0
GLY (G)	0	0	0	0	0	0	43	0	0	0	0	0	0	0	0	0
LEU (L)	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
VAL (V)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	125	0
TRP (W)	0	0	0	0	0	1	169	0	0	0	0	0	0	0	0	0
CYC (C)	2	0	0	0	2	0	0	2	0	0	0	0	0	0	0	2
ILE (I)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
MET (M)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
PHE (F)	0	0	0	0	37	0	0	0	72	0	0	0	0	0	0	37
PRO (P)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ASP (D)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
GLU (E)	0	150	0	0	0	4	0	0	0	0	0	150	0	0	55	0
ARG (R)	0	0	0	0	0	42	0	0	0	0	211	0	0	0	0	0
HIS (H)	0	0	1	154	1	0	0	0	0	0	0	0	1	154	0	1
LYS (K)	0	0	88	0	0	0	0	0	0	0	0	0	88	0	0	0
ASN (N)	0	53	122	0	0	0	0	0	0	0	0	53	122	0	0	0
GLN (Q)	0	9	1	58	0	0	0	0	0	0	1	9	1	58	0	0
SER (S)	0	0	0	0	1	0	0	0	49	1	0	0	0	0	0	1
THR (T)	0	0	0	0	0	164	0	0	18	0	0	0	0	0	0	0
TYR (Y)	210	0	0	0	171	0	0	210	73	0	0	0	0	0	0	171
Total	212	212	212	212	212	212	212	212	212	212	212	212	212	212	212	212
Variability	2	3	4	2	5	5	2	2	4	2	2	3	4	2	3	5
Hydrophilic	1	3	4	2	3	3	0	1	3	1	2	3	4	2	1	3
Hydrophobic	1	0	0	0	2	2	2	1	1	1	0	0	0	0	2	3

(Continued)

specific directions and held by the pockets B and D, respectively [17]. The C-terminal side chain of the peptide is enclosed in pocket F, which is close to one end of the groove [9, 15, 18, 19]. The fourth residue is always solvent exposed in 8 to 9 residues long bound peptides [24]. It is also known that the side-chains of the fifth or sixth residues interact with pockets in the middle of the groove [9, 10, 20, 21].

Although our knowledge on the nature of peptide anchor and binding pockets has improved over the last few years, a complete understanding is still lacking. Zhang and coworkers [25] completed the first classification of 408 pockets from 68 HLA molecules into 272 families based on the physical/chemical properties and peptide side-chain selectivity. The results provided an approach of classifying the class I pockets based on structural principles that have more functional generality than the DNA/protein sequence themselves. However, the method has a relatively low efficiency in reducing pocket variability. In this article, we describe a new approach of classification that produced a high efficiency for compressing functional pocket variability.

MATERIALS AND METHODS

Our recent study on MHCp crystal complexes suggested that residues at 60 positions in the α_1 , α_2 domains are predominantly involved in peptide binding [24]. This is true for a set of MHCp complexes characterized by varying MHC alleles and peptide sequences [24]. These residue positions are called essential residue positions (ERP). The residues at ERPs demonstrated a change in solvent accessibility (ΔASA) of $> 0 \text{ \AA}^2$ upon complex formation in a set of MHCp structures [24]. Among the ERPs, 21 positions revealed at least one amino acid polymorphism among 212 HLA-A alleles (IMGT/HLA release 1.14). These 21 residue positions are called critical polymorphic functional residue positions (CPFRP). The CPFRPs are then classified into six pockets (A–F) based on the MHCp structure published by Bjorkman *et al.* [5]. A set of discontinuous residue stretches corresponding to each pocket in the groove is generated. The discontinuous residue stretches in each pocket are compared among themselves to group the 212 HLA-A alleles into different clusters. The pockets generated based on res-

TABLE 1 Major histocompatibility complex functional property matrix (*continued*)

C					D			E				F			
9	70	73	74	97	99	114	156	97	114	152	156	77	80	81	116
0	0	1	0	0	0	0	0	0	0	22	0	0	0	45	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	96	0	0	0	96	0	0	167	0
0	0	0	0	0	0	0	0	0	0	138	0	0	0	0	1
0	0	0	0	0	0	0	54	0	0	3	54	0	0	0	0
0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0
0	0	15	0	47	0	0	0	47	0	0	0	0	45	0	0
0	0	0	0	86	0	0	0	86	0	0	0	0	0	0	0
72	0	0	0	0	37	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	154	0	0	0	0	0	0	0	0	125	0	0	109
0	0	0	0	0	0	13	0	0	13	43	0	0	0	0	0
0	0	0	0	79	0	62	11	79	62	6	11	0	0	0	0
0	154	0	57	0	1	86	0	0	86	0	0	0	0	0	14
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	1	0	0	0	0	0	0	0	0	76	0	0	1
0	58	0	0	0	0	51	51	0	51	0	51	0	0	0	0
49	0	1	0	0	1	0	0	0	0	0	0	11	0	0	0
18	0	195	0	0	0	0	0	0	0	0	0	0	167	0	0
73	0	0	0	0	171	0	0	0	0	0	0	0	0	0	87
212	212	212	212	212	212	212	212	212	212	212	212	212	212	212	212
4	2	4	3	3	5	4	4	3	4	5	4	3	2	2	5
3	2	2	3	1	3	4	2	1	4	2	2	3	1	0	4
1	0	2	0	2	2	0	2	2	0	3	2	0	1	2	1

Variability: the number of types of residues that occur at CPFRPs.

Hydrophilic: the number of types of hydrophilic residues that occur at CPFRPs.

Hydrophobic: the number of types of hydrophobic residues that occur at CPFRPs.

Abbreviation: CPFRP = critical polymorphic functional residue position.

idue compositions at CPFRPs are thereafter called functional pockets.

RESULTS

The binding groove between α_1 and α_2 domains of class I HLA molecules accommodates different types of short peptides. We analyzed the binding grooves of 212 HLA-A alleles for identical functional pockets among them. Using MHCp crystal complexes, we identified 60 ERPs and 21 of these positions are CPFRPs among 212 HLA-A alleles. The frequency of a particular residue at the CPFRPs for the 212 alleles is illustrated in Table 1. The 21 CPFRPs are unevenly distributed in the six functional pockets (named A to F) in the groove. For example, pocket B contains 9 CPFRPs and pocket D contains only 3 CPFRPs. A maximum variability of 5 is observed at 2 CPFRPs in pocket A, and at 1 CPFRP in each of the pockets B, D, E, and F. The uneven distribution of CPFRPs among functional pockets contributes to specific peptide selection in a combinatorial manner. Chemical and physical properties of binding pockets have been previously utilized in peptide selection stud-

ies. In this analysis we find that the residues at the CPFRPs are prevalently hydrophilic among the 212 HLA-A alleles and, therefore, the functional pockets are predominantly hydrophilic. Hydrophilic to hydrophilic substitution is found to be more common than hydrophobic to hydrophobic or hydrophilic to hydrophobic substitutions among HLA-A alleles (Table 1). Some CPFRPs are shared between pockets because residues at these positions constitute the pocket rims. Understanding the chemical properties of individual functional pockets and clustering HLA-A alleles by common functional pockets will provide insights to define metrics for individual pocket function.

We grouped the 212 HLA-A alleles based on the residue composition at the CPFRPs in each pocket (Table 2). The grouping produced 36 clusters for pocket A; 49 for pocket B; 34 for pocket C; 20 for pocket D; 38 for pocket E; and 12 for pocket F. Thus a total of 189 different clusters are generated. Pocket B has the maximum number of clusters, whereas pocket F has the minimum number of clusters. The result suggests that the 212 HLA-A alleles vary only by 49 different types at the CPFRPs in pocket B. Interestingly, 64 alleles are

TABLE 2 Classification of pockets based on CPFRs

Group	Pocket A	HLA alleles	Count
1	YENHYRG	A*0101, 0102, 0103, 0106, 0107, 0109	6
2	YENHYWG	A*0108	1
3	YEKHYTW	A*0201, 0202, 0203, 0204, 0205, 0206, 0209, 0211, 0212, 0213, 0214, 0221, 0222, 0224, 0225, 0226, 0227, 0228, 0230, 0231, 0240, 0241, 0242, 0244, 0245, 0246, 0247, 0248, 0249, 0250, 0251, 0252, 2416, 3007, 3102, 3205, 7404	37
4	YEKHCTW	A*0207, 0218	2
5	YENHYTW	A*0208, 0220, 0308, 3002, 3003, 3004, 3006, 3009, 3012, 3101, 3103, 3104, 3106, 3201, 3202, 3203, 3204, 3206, 3207, 3601, 3602, 3603, 7401, 7402, 7403, 7405, 7406, 7407, 7408	29
6	YEKHFTW	A*0210, 0217, 0239, 2304, 2403, 2418, 2422, 2423, 2433	9
7	YEKHYEW	A*0216	1
8	YEKHYTG	A*0219, 0236, 0237, 0254, 2414, 2415	6
9	YEQHYTW	A*0229	1
10	YEKHSTW	A*0233	1
11	YEKQYTW	A*0234, 0235	2
12	YEKHyrw	A*0238, 2607	2
13	YNNHYTW	A*0255, 2609, 3301, 3303, 3304, 3305, 3306, 6803, 6804, 6805, 6815, 6820	12
14	YENQYTW	A*0301, 0302, 0304, 0305, 0306, 0307, 0309, 1104, 3001, 3008, 3011, 6810, 6813, 6814	14
15	YENQYRW	A*1101, 1102, 1103, 1105, 1106, 1107, 1108, 1109, 1112, 1113	10
16	YNNQYRW	A*1110, 2502, 2613, 6601, 6604	5
17	YQNQYRW	A*1111	1
18	YEKHFTG	A*2301, 2302, 2303, 2306, 2402, 2404, 2405, 2406, 2408, 2413, 2417, 2420, 2421, 2426, 2427, 2428, 2429, 2430, 2432, 2434, 2435	21
19	CEKHFTG	A*2305, 2425	2
20	YEKQFTG	A*2407, 2419	2
21	YEKHFRW	A*2410	1
22	YNNQFTG	A*2424	1
23	YQKHFTG	A*2431	1
24	YNNHYRW	A*2501, 2503, 2504, 2601, 2602, 2603, 2605, 2606, 2608, 2610, 2612, 2614, 2615, 2616, 2617, 2618	16
25	YNNHYLW	A*2604	1
26	YQNQYTW	A*2901, 2902, 2905, 2906	4
27	YQNQYTG	A*2903	1
28	YQHQYTW	A*2904	1
29	YENHHTW	A*3010	1
30	YENHYTG	A*3105	1
31	YNKQYTW	A*3401	1
32	YNNQYTW	A*3402, 3403, 3404, 6801, 6802, 6806, 6807, 6808, 6809, 6812, 6816, 6817, 6819, 6821, 6901	15
33	YQNHYRW	A*4301	1
34	YNNQYEW	A*6602	1
35	YNNHYEW	A*6603	1
36	YENHYEG	A*8001	1

Group	Pocket B	HLA alleles	Count
1	YFARENHAY	A*0101, 0103, 0106, 0108, 0109, 3601, 3602, 3603	8
2	YSARENHAY	A*0102	1
3	YFARENHEY	A*0107, 3201, 3202, 3203, 3204, 3206	6
4	YFAREKHVY	A*0201, 0202, 0203, 0204, 0209, 0211, 0212, 0213, 0216, 0219, 0222, 0224, 0225, 0226, 0227, 0230, 0231, 0236, 0237, 0238, 0240, 0245, 0246, 0247, 0248, 0249, 0250, 0252, 7404	29
5	YYAREKHVY	A*0205, 0206, 0214, 0221, 0228, 0241, 0244, 0251, 0254	9
6	YFAREKHVC	A*0207, 0218	2
7	YYARENHVY	A*0208	1
8	YYAREKHVF	A*0210	1
9	YFAREKHVF	A*0217, 0239	2
10	YFARENHVY	A*0220, 0308, 7401, 7402, 7403, 7405, 7406, 7407, 7408	9
11	YFAREQHVY	A*0229	1
12	YFAREKHVS	A*0233	1
13	YFAREKQVY	A*0234, 0235	2
14	YFSREKHVY	A*0242	1
15	YFARNNHVY	A*0255	1
16	YFARENQVY	A*0301, 0302, 0304, 0305, 0306, 0307, 0309	7

(Continued)

TABLE 2 Classification of pockets based on CPMRs (*continued*)

Group	Pocket B	HLA alleles	Count
17	YYARENQVY A*1101, 1102, 1103, 1104, 1105, 1106, 1107, 1108, 1109, 1112, 1113, 3008, 6810, 6813, 6814		15
18	YYARNNQVY A*1110, 3402, 3403, 3404, 6601, 6602, 6604, 6801, 6802, 6806, 6807, 6808, 6809, 6812, 6816, 6817, 6819, 6821, 6901		19
19	YYARQNQVY A*1111		1
20	YSAREKHEF A*2301, 2302, 2303, 2304, 2306, 2402, 2403, 2405, 2406, 2408, 2410, 2413, 2417, 2418, 2420, 2421, 2422, 2423, 2426, 2427, 2429, 2430, 2432, 2433, 2434, 2435		26
21	CSAREKHEF A*2305, 2425		2
22	YSAREKHAF A*2404		1
23	YSAREKQEF A*2407		1
24	YSAREKHEY A*2414, 2415, 3007		3
25	YTAREKHEY A*2416		1
26	YSAREKQVF A*2419		1
27	YSARNNQEF A*2424		1
28	YSAREKHVF A*2428		1
29	YSARQKHEF A*2431		1
30	YYARNNHEY A*2501, 2504, 2605		3
31	YYARNNQEY A*2502		1
32	YFARNNHEY A*2503		1
33	YYARNNHAYA*2601, 2602, 2604, 2608, 2609, 2610, 2612, 2614, 2615, 2617, 2618		11
34	YYARNNHVYA*2603, 2606, 6603, 6803, 6804, 6805, 6815, 6820		8
35	YYAREKHAY A*2607		1
36	YYARNNQAYA*2613		1
37	YSARNNHAY A*2616		1
38	YTARQNQAY A*2901, 2902, 2903, 2905, 2906		5
39	YTARQHQAY A*2904		1
40	YSARENQVY A*3001, 3011		2
41	YSARENHEY A*3002, 3003, 3004, 3006, 3009, 3012, 3207		7
42	YSARENHEH A*3010		1
43	YTARENHVY A*3101, 3103, 3104, 3105, 3106		5
44	YTAREKHVY A*3102		1
45	YFAREKHEY A*3205		1
46	YTARNNHVYA*3301, 3303, 3304, 3305, 3306		5
47	YYARNKQVY A*3401		1
48	YYARQNHAYA*4301		1
49	YFAQENHAY A*8001		1
Group	Pocket C	HLA alleles	Count
1	FHTDI A*0101, 0106, 0107, 0108, 0109, 0308, 3204, 3601, 3602		9
2	SHTDI A*0102, 3002, 3003, 3004, 3006, 3007, 3009, 3010, 3012		9
3	FHTDM A*0103, 3201, 3202, 3203, 3205, 3206, 3603, 7401, 7402, 7403, 7404, 7408		12
4	FHTHR A*0201, 0202, 0203, 0207, 0209, 0212, 0213, 0216, 0218, 0219, 0220, 0222, 0224, 0225, 0226, 0227, 0229, 0230, 0231, 0233, 0236, 0237, 0238, 0239, 0240, 0242, 0245, 0246, 0249, 0250, 0252, 0255		32
5	FHTHM A*0204, 0217, 7406		3
6	YHTHR A*0205, 0206, 0208, 0210, 0214, 0221, 0228, 0241, 0244, 0251, 0254, 2603, 2606, 6815		14
7	FHIDR A*0211		1
8	FQTHR A*0234		1
9	FQTDR A*0235		1
10	FHSHR A*0247		1
11	FHTDR A*0248, 2503		2
12	FQTDI A*0301, 0302, 0304, 0305, 0306, 0307		6
13	FQTHI A*0309		1
14	YQTDI A*1101, 1102, 1103, 1104, 1105, 1107, 1108, 1109, 1110, 1111, 1112, 1113, 3008, 3402, 3403, 3404		16
15	YQTHI A*1106		1
16	SHTDM A*2301, 2302, 2303, 2304, 2305, 2306, 2402, 2403, 2404, 2405, 2406, 2408, 2410, 2413, 2415, 2417, 2418, 2420, 2421, 2422, 2423, 2425, 2426, 2427, 2429, 2431, 2432, 2433, 2435, 3207		30
17	SQTDI A*2407, 2419, 2424		3
18	SHTDR A*2414, 2616		2

(Continued)

TABLE 2 Classifications of pockets based on CPFRs (*continued*)

Group	Pocket C	HLA alleles	Count
19	THTDM	A*2416	1
20	SHTHM	A*2428, 2430	2
21	SHIDM	A*2434	1
22	YHTDR	A*2501, 2504, 2601, 2602, 2604, 2605, 2607, 2608, 2609, 2610, 2612, 2614, 2615, 2617, 2618, 4301, 6603	17
23	YQTDR	A*2502, 2613, 3401, 6601, 6602, 6604, 6802, 6901	8
24	TQTDM	A*2901, 2902, 2903, 2904, 2905, 2906	6
25	SQTDI	A*3001, 3011	2
26	THIDM	A*3101, 3102, 3105, 3106, 3301, 3303, 3304, 3305, 3306	9
27	THIDI	A*3103, 3104	2
28	YQTDM	A*6801, 6806, 6807, 6808, 6809, 6810, 6812, 6813, 6814, 6816, 6817, 6819, 6821	13
29	YHTDM	A*6803	1
30	YHIDM	A*6804	1
31	YHTHM	A*6805, 6820	2
32	FHADM	A*7405	1
33	FHIDM	A*7407	1
34	FHTNI	A*8001	1

Group	Pocket D	HLA alleles	Count
1	YRR	A*0101, 0102, 0103, 0107, 0108, 0109, 1108, 3601, 3602, 3603	10
2	YRL	A*0106, 0301, 0304, 0305, 0306, 0307, 0308, 0309, 2901, 2902, 2903, 2904, 2906, 3103, 3104, 3106, 3204, 3402, 3403, 3404, 6808, 8001	22
3	YHL	A*0201, 0204, 0206, 0209, 0211, 0214, 0216, 0220, 0221, 0224, 0225, 0226, 0228, 0229, 0230, 0231, 0234, 0235, 0236, 0240, 0242, 0245, 0246, 0248, 0251, 0255, 6901	27
4	YHW	A*0202, 0203, 0205, 0208, 0222, 0247, 0250, 6802, 6815	9
5	CHL	A*0207, 0218	2
6	FHL	A*0210, 0217, 0239, 2301, 2303, 2304, 2305, 2306, 2413, 2418, 2424	11
7	YHQ	A*0212, 0213, 0219, 0227, 0237, 0238, 0244, 0254, 2414, 2415	10
8	SHL	A*0233	1
9	YQL	A*0241, 2416, 3101, 3102, 3105, 3201, 3203, 3205, 3206, 3207, 3301, 3303, 3304, 3305, 3306, 7401, 7402, 7403, 7404, 7405, 7406, 7407, 7408	23
10	YHR	A*0249	1
11	YEL	A*0252, 3001, 3002, 3003, 3007, 3008, 3009, 3011, 3012	9
12	YRQ	A*0302, 1101, 1102, 1103, 1104, 1105, 1106, 1107, 1109, 1110, 1111, 1112, 1113, 2905, 6809	15
13	FHW	A*2302, 2406, 2422	3
14	FHQ	A*2402, 2403, 2404, 2405, 2407, 2408, 2410, 2419, 2420, 2421, 2423, 2425, 2426, 2427, 2428, 2429, 2430, 2431, 2432, 2433, 2434, 2435	22
15	FRQ	A*2417	1
16	YQW	A*2501, 2502, 2503, 2601, 2602, 2603, 2604, 2605, 2606, 2607, 2609, 2610, 2612, 2613, 2614, 2615, 2616, 2617, 2618, 3401, 4301, 6601, 6602, 6603, 6604	25
17	YQQ	A*2504, 2608, 3202	3
18	YEW	A*3004, 3006, 6806	3
19	HEL	A*3010	1
20	YRW	A*6801, 6803, 6804, 6805, 6807, 6810, 6812, 6813, 6814, 6816, 6817, 6819, 6820, 6821	14

Group	Pocket E	HLA alleles	Count
1	IRAR	A*0101, 0102, 0107, 0108, 0109, 3601, 3602	7
2	MRAR	A*0103, 3603	2
3	IRAL	A*0106	1
4	RHVL	A*0201, 0206, 0207, 0209, 0210, 0211, 0214, 0216, 0218, 0220, 0221, 0224, 0225, 0228, 0229, 0230, 0231, 0233, 0234, 0235, 0236, 0239, 0240, 0242, 0245, 0246, 0248, 0251, 0255, 6901	30
5	RHVV	A*0202, 0205, 0208, 0222, 0247, 0250, 6802, 6815	8
6	RHEW	A*0203	1
7	MHVL	A*0204, 0217, 2301, 2303, 2304, 2305, 2306, 2413, 2424	9
8	RHVQ	A*0212, 0219, 0237, 0244, 0254, 2414	6
9	RHEQ	A*0213, 0238	2
10	RHEL	A*0226	1
11	RHAQ	A*0227	1

(Continued)

TABLE 2 Classification of pockets based on CPFs (*continued*)

Group	Pocket E	HLA alleles	Count
12	RQVL	A*0241	1
13	RHVR	A*0249	1
14	REVL	A*0252	1
15	IREL	A*0301, 0304, 0305, 0306, 0308, 0309, 3204, 3402, 3403, 3404	10
16	IRVQ	A*0302	1
17	IRVL	A*0307, 3103, 3104	3
18	IRAQ	A*1101, 1102, 1104, 1105, 1106, 1107, 1109, 1110, 1111, 1112, 1113	11
19	IREQ	A*1103	1
20	IRER	A*1108	1
21	MHVW	A*2302, 2406, 2422	3
22	MHVQ	A*2402, 2403, 2404, 2405, 2407, 2408, 2410, 2415, 2419, 2420, 2421, 2423, 2425, 2426, 2427, 2428, 2429, 2430, 2431, 2432, 2433, 2434, 2435	23
23	MQVL	A*2416, 3101, 3102, 3105, 3201, 3203, 3205, 3206, 3207, 3301, 3303, 3304, 3305, 3306, 7401, 7402, 7403, 7404, 7405, 7406, 7407, 7408	22
24	MRVQ	A*2417, 6809	2
25	MHEL	A*2418	1
26	RQEW	A*2501, 2502, 2503, 2601, 2602, 2603, 2604, 2605, 2606, 2607, 2609, 2610, 2613, 2614, 2615, 2616, 2617, 3401, 4301, 6601, 6602, 6603, 6604	23
27	RQEQ	A*2504, 2608	2
28	RQVW	A*2612, 2618	2
29	MRVL	A*2901, 2902, 2903, 2904, 2906, 3106, 6808	7
30	MREQ	A*2905	1
31	IEWL	A*3001, 3008, 3011	3
32	IERL	A*3002, 3003, 3007, 3010, 3012	5
33	IEVW	A*3004, 3006	2
34	IEVL	A*3009	1
35	MQVQ	A*3202	1
36	MRVW	A*6801, 6803, 6804, 6805, 6807, 6810, 6812, 6813, 6814, 6816, 6817, 6819, 6820, 6821	14
37	MEVW	A*6806	1
38	IRRL	A*8001	1

Group	Pocket F	HLA alleles	Count
1	NTLD	A*0101, 0102, 0103, 0106, 0107, 0108, 0109, 2601, 2604, 2605, 2607, 2608, 2609, 2610, 2612, 2613, 2614, 2615, 2616, 2617, 2618, 2901, 2902, 2903, 2904, 2905, 2906, 3601, 3602, 3603, 4301, 8001	32
2	DTLY	A*0201, 0202, 0203, 0204, 0205, 0206, 0207, 0208, 0209, 0210, 0211, 0212, 0213, 0214, 0216, 0217, 0218, 0219, 0220, 0221, 0222, 0224, 0225, 0226, 0227, 0228, 0229, 0230, 0231, 0233, 0234, 0235, 0236, 0237, 0238, 0239, 0240, 0241, 0242, 0244, 0245, 0246, 0247, 0248, 0249, 0250, 0251, 0254, 0255, 2419, 2428, 6802, 6815, 6901	54
3	DTLH	A*0252, 3001, 3008, 3011, 6806, 6807	6
4	DTLD	A*0301, 0302, 0304, 0305, 0306, 0307, 0308, 0309, 1101, 1102, 1103, 1104, 1105, 1106, 1107, 1108, 1109, 1110, 1111, 1112, 1113, 2603, 2606, 3101, 3102, 3103, 3104, 3105, 3106, 3301, 3303, 3304, 3305, 3306, 3401, 3402, 3403, 3404, 6601, 6602, 6603, 6604, 6801, 6803, 6804, 6805, 6808, 6809, 6810, 6812, 6813, 6814, 6816, 6819, 6820, 6821, 7401, 7402, 7403, 7404, 7405, 7406, 7407, 7408	64
5	NIAY	A*2301, 2302, 2303, 2304, 2305, 2306, 2402, 2403, 2405, 2406, 2407, 2408, 2410, 2413, 2414, 2415, 2418, 2420, 2421, 2422, 2423, 2424, 2425, 2426, 2427, 2429, 2430, 2431, 2433, 2434, 2435	31
6	NTLY	A*2404	1
7	NIAD	A*2416, 2417, 3203	3
8	SIAY	A*2432	1
9	SIAD	A*2501, 2502, 2503, 2504, 3201, 3202, 3204, 3205, 3206, 3207	10
10	NTLN	A*2602	1
11	NTLH	A*3002, 3003, 3004, 3006, 3007, 3009, 3010, 3012	8
12	DTLV	A*6817	1

Cluster: Clusters of pockets with same positional residue composition at the CPFs. HLA alleles: the class I HLA alleles that belong to the group, with sequence-based nomenclature, begin with "A*" and are followed by a four digit number.

Abbreviation: CPF = critical polymorphic functional residue position; HLA = human leukocyte antigen.

found to share a particular type of residue composition (D-T-L-D) at the CPFs in pocket F (Table 2). Of the

189 unique pockets, 114 pockets have no more than 3 alleles in each cluster. About 8 clusters have more than

30 alleles and 20 clusters have more than 20 alleles. It should be noted that the binding of peptides to MHC molecules is neither because of any one independent functional pocket nor because of any one independent peptide residue. However, their binding is defined by the cumulative physical, chemical properties of all the residues in the peptide and all the pockets in the MHC molecule. Knowledge on the similarity and dissimilarity of individual functional pockets will enable the development of metrics for cumulative pocket effects towards specific peptide binding.

When pockets A–F are rearranged using residues at CPFRPs for the 212 HLA-A alleles 140 clusters were produced. In other words 140 different patterns of discontinuous sequence stretches at the CPFRPs are identified among the 212 HLA-A alleles. From the above grouping procedure we identified 101 HLA alleles clustered into 29 clusters such that alleles in the clusters have identical A–F functional pockets (Table 3). These clusters have at least two alleles. The remaining 111 alleles produce 111 clusters with each cluster containing 1 allele. This finding provides the key to define HLA subtypes with similar peptide binding functions as observed by serologic investigations.

DISCUSSION

HLA molecules are highly polymorphic and this is evident from the growth in HLA sequence data at IMGT/HLA [26]. HLA-A and -B are most polymorphic with currently more than 200 and 400 alleles defined by DNA sequence analysis, respectively. HLA molecules are interesting pharmaceutical targets for vaccination and immuno-therapeutics through the design of peptide ligands. The characteristic interaction between peptide and HLA provide the molecular basis for allele-specific recognition of antigenic peptides. The peptide residues that are held by a certain pocket are determined by the physical and chemical properties of the pocket. Nearly 6.1×10^{10} combinations of 9 residue long peptide stretches can be formed from the naturally occurring 20 amino acids. Only a very small fraction of the peptide pool binds to each of the known alleles and the interactions between peptide and alleles are highly combinatorial. Therefore, combinatorial reduction approaches will provide roadmaps to quantify functional variations among HLA alleles.

The application of molecular modeling in MHCp binding prediction utilized hydrophobic and hydrophilic interaction functions for scoring. Estimation of hydrophilic interaction between peptide and HLA still remains a great challenge. From our analysis, we find that the residues at the CPFRPs in the 212 HLA-A alleles are mostly hydrophilic and a preference for hydrophilic to

hydrophilic residue substitution is observed. At positions 99, 163, 152, and 116 a very high variability of 5 is observed, and therefore pockets containing these CPFRPs contribute more towards the functional difference between alleles. Although, the alleles show considerable residue identity at the CPFRPs, the subtle dissimilarity at the CPFRPs among alleles is the cause for considerable differences in peptide binding function. Allele specific knowledge on the chemical properties and residue variability of the CPFRPs constituting the functional pockets are crucial in MHCp binding quantification. Thus, a systematic approach to quantify the critical difference between alleles will help to extrapolate the binding affinities of known MHCp interactions to new HLA alleles with unknown binding information.

In our approach, we used residues at the 21 CPFRPs to differentiate HLA-A alleles, and thus classified the 212 HLA-A alleles into different clusters based on the residue compositions in each of the functional pockets (Table 2). Each molecule has 6 pockets and therefore, 212 HLA-A alleles constitute 1272 pockets. On the contrary, the identification and utilization of 21 CPFRPs generated only 189 unique pockets from 212 HLA-A alleles. Our methodology produced a sevenfold reduction for functional pocket variability among HLA-A alleles. This implies that the HLA-A alleles are functionally and quantitatively not identical but similar. The results also suggest that functional pocket variations between alleles are mainly in specific pockets. The limited number of unique pockets among HLA-A alleles presents a promising solution for reducing the large possibility of peptides binding to HLA-A molecules. When the 212 A alleles are classified based on the compositions at the CPFRPs in each pocket, we found 36 different types of pocket A; 49 types of pocket B; 34 types of pocket C; 20 types of pocket D; 38 types of pocket E and 12 types of pocket F. The grouping of HLA-A molecules based on the individual pocket can provide a strong tool for the understanding of pocket features. For example, pocket F is the most conserved with just 12 different clusters and pocket B is the least conserved with 49 clusters. A cluster for pocket F with residue composition “D-T-L-D” occurs in 64 different alleles. This implies that many alleles share functionally relevant residues at different pockets among alleles. The classification presented in Table 2 will enable one to compare and contrast the functional similarities and differences between functional pockets among HLA-A alleles in a combinatorial manner. In other words one can find clusters of alleles that share all the 6 pockets, only 5 pockets, only 4 pockets, only 3 pockets, only 2 pockets, and only 1 pocket. This exercise will result in large combinations of alleles with different functional pocket category. Based on the functional

TABLE 3 List of HLA alleles sharing common CPFRs in all the six pockets

Group	CPFR	HLA alleles
1	YFARENHTDANTLIYRDARRG	A*0101, A*0109
2	YFAREKHTHVDTLRYHYVLTW	A*0201, A*0209, A*0224, A*0225, A*0230, A*0231, A*0240, A*0245, A*0246
3	YFAREKHTHVDTLRYHYVWTW	A*0202, A*0222, A*0250
4	YYAREKHTHVDTLRYHYVLTW	A*0206, A*0214, A*0221, A*0228, A*0251
5	YFAREKHTHVDTLRCHYVLTW	A*0207, A*0218
6	YFAREKHTHVDTLRYHYVQTG	A*0219, A*0237
7	YFARNQTDVDTLIYRDELTW	A*0301, A*0304, A*0305, A*0306
8	YYARENQTDVDTLIYRDAQRW	A*1101, A*1102, A*1105, A*1107, A*1109, A*1112, A*1113
9	YSAREKHTDENIAMFHYVLTG	A*2301, A*2303, A*2306, A*2413
10	YSAREKHTDENIAMFHYVWTG	A*2302, A*2406
11	YSAREKHTDENIAMFHYVQTG	A*2402, A*2405, A*2408, A*2420, A*2421, A*2426, A*2427, A*2429, A*2435
12	YSAREKHTDENIAMFHYVQTW	A*2403, A*2423, A*2433
13	YYARNNHTDANTLRYQDEWRW	A*2601, A*2610, A*2614, A*2615, A*2617
14	YYARNNHTHVDTLRYQDEWRW	A*2603, A*2606
15	YYARNNHTDANTLRYQDVWRW	A*2612, A*2618
16	YTARQNQTDANTLMYRDVLTW	A*2901, A*2902, A*2906
17	YSARENQTDVDTLIYEHVLTW	A*3001, A*3011
18	YSARENHTDENTLIYEHRLTW	A*3002, A*3003, A*3012
19	YSARENHTDENTLIYEHVWTW	A*3004, A*3006
20	YTARENHIDVDTLIYRDVLTW	A*3103, A*3104
21	YFARENHTDESIAMYQDVLTW	A*3201, A*3206
22	YTARNNHIDVDTLMYQDVLTW	A*3301, A*3303, A*3304, A*3305, A*3306
23	YYARNNQTDVDTLIYRDELTW	A*3402, A*3403, A*3404
24	YFARENHTDANTLIYRDARTW	A*3601, A*3602
25	YYARNNQTDVDTLRYQDEWRW	A*6601, A*6604
26	YYARNNQTDVDTLMYRDVWTW	A*6801, A*6812, A*6816, A*6819, A*6821
27	YYARNNHVDVDTLMYRDVWTW	A*6805, A*6820
28	YYARENQTDVDTLMYRDVWTW	A*6810, A*6813, A*6814
29	YFARENHTDVDTLMYQDVLTW	A*7401, A*7402, A*7403, A*7408

Abbreviations: CPFR = critical polymorphic functional residue; HLA = human leukocyte antigen.

pocket similarity strategies can be designed to extrapolate properties from known cases to unknown cases.

We find that 101 HLA-A alleles are grouped into 29 clusters by performing the simplest exercise of finding clusters of alleles with common A–F functional pockets. This implies that a given peptide will bind to alleles in the cluster with similar binding affinities. In other words, alleles in the same group in the 29 clusters will have similar binding patterns with a defined peptide. If the binding data of a given peptide is known for one allele in each cluster then the predicted binding of that peptide to other alleles can be easily and efficiently extrapolated. However this inference needs experimental testing and validation. With the help of effective cross-validation, predicted binding data of a defined peptide for nearly 50% of HLA-A alleles can be generated using binding data for that peptide to at least one allele in each of the 29 clusters.

By performing a random validation experiment, we looked for peptides with known binding information for more than one HLA alleles in MHCBN [27] and JEN-PEP [28]. A limited number of such peptides are documented in these databases and we identified two peptides

from MHCBN with known binding information for two alleles in the clusters presented in Table 3. The peptides FLWGPRALV and AAGIGILTV are documented to have “high/moderate” binding affinity to A*0201 and A*0209. Two other peptides GILGFVFTL and YLEPG-PVTA are documented to have similar T-cell activity in JENPEP [28]. It is also known that A*0201 and A*0209 are in one cluster (Table 3). This roughly suggests that the functional repertoires of HLA molecules are as a result of changes in the CPFRs and it is possible to maintain functional similarity by the conservation of CPFRs. It is also worth mentioning that we could hardly find a single peptide with contrasting binding affinities for alleles grouped in the same clusters for proving the proposed hypothesis otherwise (Table 3).

The results suggest that it is possible to quantify the functional difference between HLA-A alleles and develop better methodologies for HLA-peptide binding predictions. The work provides a framework to classify class I pockets based on peptide binding principles, which have more functional generality than the alleles themselves. The current approach demonstrates a sevenfold reduction in pocket variability among HLA-A alleles and the re-

duction approach will help in improving our capability to predict peptide binding to a range of HLA alleles. The challenge lies in generating CPFRPs for each allele, clustering into different pockets, storing in the form of lookup tables and establishing the common relation between functional pockets for the accurate scanning of binding functions between them. The method thus demonstrates the useful application of structural information in understanding the functional consequences of sequence polymorphism in HLA molecules.

CONCLUSION

The current challenge in synthetic peptide vaccine design is the identification of a minimum number of short antigenic peptides capable of binding to multiple HLA alleles for efficient and broad range immune response. This can be achieved by testing compatibility between each possible peptide against multiple alleles. The binding of a given peptide to a HLA molecule ideally depends on the conformation of the peptide and the cumulative chemical properties of the peptide sequence. Solution thermodynamics and molecular mechanical calculations are traditionally used to estimate such compatibility with limited success. Our findings help to demonstrate the functional differences and similarities among HLA-A alleles. Knowledge on the similarity and dissimilarity of individual functional pockets will enable the development of metrics for cumulative pocket effects towards specific peptide binding.

We find that residues at the 21 CPFRPs play a more crucial role in determining functional diversity among HLA molecules for a given peptide than those at other positions in the grooves. The identification of 189 unique pockets among 212 HLA-A alleles suggests that many HLA alleles share considerable number of functional pockets among them. In other words, the functional diversity of individual pockets is sevenfold less than that of the primary sequences themselves. The 212 HLA-A alleles are also grouped into 140 clusters based on residues at CPFRPs with 29 clusters containing at least two alleles. Thus, nearly a twofold functional reduction is observed among 212 HLA-A alleles. We propose that alleles in the same cluster should bind with similar affinity (if not identical) for a given peptides. If the binding strength of a given peptide to any of the allele is known in the cluster, then its binding strength to other alleles in the cluster can be easily inferred. However, this hypothesis has to be validated by experimental data for at least one given peptide to all the alleles in the clusters. The results from the current investigation clearly provide the functional pocket differences and similarities among HLA-A alleles. The next challenging task is to develop metrics to quantify functional differences among alleles

and work is in progress to develop metrics for cumulative peptide residues functional pockets compatibility estimation.

ACKNOWLEDGMENTS

The authors wish to thank the anonymous reviewers for their critical comments and intriguing questions.

REFERENCES

1. Gaur A, Fathman CG: Immunotherapeutic strategies directed at the trimolecular complex. *Adv Immunol* 56:219, 1994.
2. Townsend A, Bodmer H: Antigen recognition by class-I restricted T lymphocyte. *Annu Rev Immunol* 7:601, 1989.
3. Falk K, Rötzschke O, Stevanovic S, Jung G, Rammensee HG: Allele-specific motifs revealed by sequencing of self-peptides eluted from MHC molecules. *Nature* 351:290, 1991.
4. Bjorkman PJ, Saper MA, Samraoui B, Bennett WS, Strominger JL, Wiley DC: Structure of the human class I histocompatibility antigen HLA-A2. *Nature* 329:506, 1987.
5. Bjorkman PJ, Saper MA, Samraoui B, Bennett WS, Strominger JL, Wiley DC: The foreign antigen binding site and T cell recognition regions of class I histocompatibility antigens. *Nature* 329:512, 1987.
6. Rammensee HG, Falk K, Rötzschke O: MHC molecules as peptide receptors. *Curr Opin Immunol* 5:35, 1993.
7. Rammensee HG: Chemistry of peptides associated with MHC class I and class II molecules. *Curr Opin Immunol* 7:85, 1995.
8. van Bleek GM, Nathenson SG: Isolation of an endogenously processed immunodominant viral peptide from the class I H-2Kb molecule. *Nature* 348:213, 1990.
9. Fremont DH, Matsumura M, Stura EA, Peterson PA, Wilson IA: Crystal structures of two viral peptides in complex with murine MHC class I H-2Kb. *Science* 257:919, 1992.
10. Zhang W, Young ACM, Imarai M, Nathenson SG, Sacchettini JC: Crystal structure of the MHC class I H-2Kb molecule containing a single viral peptide: Implications for peptide binding and T cell receptor recognition. *Proc Natl Acad Sci USA* 89:8403, 1992.
11. Silver ML, Guo HC, Strominger JL, Wiley DC: Atomic structure of a human MHC molecule presenting an influenza virus peptide. *Nature* 360:367, 1992.
12. 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* 75:693, 1993.
13. Saper MA, Bjorkman PJ, Wiley DC: Refined structure of the human histocompatibility antigen HLA-A2 at 2.6 Å resolution. *J Mol Biol* 219:277, 1991.

14. Garrett TPJ, Saper MA, Bjorkman PJ, Strominger JL, Wiley DC: Specificity pockets for the side chains of peptide antigens in HLA-Aw68. *Nature* 342:692, 1989.
15. Guo HC, Madden DR, Silver ML, Jardetzky TS, Gorga JC, Strominger JL, Wiley DC: Comparison of the P2 specificity pocket in three human histocompatibility antigens: HLA-A*6801, HLA-A*0201, and HLA-B8*2705. *Proc Natl Acad Sci USA* 90:8053, 1993.
16. Madden DR, Gorga JC, Strominger JL, Wiley DC: The structure of HLA-B27 reveals nonamer self-peptides bound in an extended conformation. *Nature* 353:321, 1991.
17. Matsumura M, Fremont DH, Peterson PA, Wilson IA: Emerging principles for the recognition of peptide antigens by MHC class I molecules. *Science* 257:927, 1992.
18. Bouvier M, Wiley DC: Importance of peptide amino and carboxyl termini to the stability of MHC class I molecules. *Science* 265:398, 1994.
19. DiBrino M, Parker KC, Margulies DH, Shiloach J, Turner RV, Biddison WE, Coligan JE: The HLA-B14 peptide binding site can accommodate peptides with different combinations of anchor residues. *J Biol Chem* 269:32426, 1994.
20. Young ACM, Zhang W, Sacchettini JC, Nathenson SG: The three-dimensional structure of H-2Db at 2.4 Å resolution: Implications for antigen-determinant selection. *Cell* 76:39, 1994.
21. Reid SW, McAdam S, Smith KJ, Klenerman P, O'Callaghan CA, Harlos K, Jakobsen BK, McMichael AJ, Bell JI, Stuart DI, Jones EY: Antagonist HIV-1 gag peptides induce structural changes in HLA B8. *J Exp Med* 184:2279, 1996.
22. Jardetzky TS, Lane WS, Robinson RA, Madden DR, Wiley DC: Identification of self peptides bound to purified HLA-B27. *Nature* 353:326, 1991.
23. Hunt DF, Henderson RA, Shabanowitz J, Sakaguchi K, Michel H, Sevilir N, Cox AL, Appella E, Engelhard VH: Characterization of peptides bound to the class I MHC molecule HLA-A2.1 by mass spectrometry. *Science* 255:1261, 1992.
24. Kangueane P, Sakharkar MK, Kolatkar PR, Ren EC: Towards the MHC-peptide combinatorics. *Hum Immunol* 62:539, 2001.
25. Zhang C, Anderson A, DeLisi C: Structural principles that govern the peptide binding motifs of class I MHC molecules. *J Mol Biol* 281:929, 1998.
26. Robinson J, Waller MJ, Parham P, Bodmer JG, Marsh SGE: IMGT/HLA database – a sequence database for the human major histocompatibility complex. *Nucleic Acids Res* 29:210, 2001.
27. <http://www.imtech.res.in/raghava/mhcbn/>.
28. Blythe MJ, Doytchinova IA, Flower DR: JenPep: a database of quantitative functional peptide data for immunology. *Bioinformatics* 18:434, 2002.

A Novel MHCp Binding Prediction Model

Bing Zhao, Venkatarajan Subramanian Mathura,
Ganapathy Rajaseger, Shabbir Moochhala,
Meena Kishore Sakharkar, and
Pandjassaram Kanguane

ABSTRACT: Many statistical and molecular mechanics models have been developed and tested for major histocompatibility complex peptide (MHCp) binding predictions during the last decade. The statistical model prediction using pooled peptide sequence data and three-dimensional modeling prediction by molecular mechanics calculations have been assessed for efficiency and human leukocyte antigen diversity coverage. We describe a novel predictive model using information gleaned from 29 human MHCp crystal structures. The validation for the new model is performed using four different sets of data: (1) MHCp crystal structures, (2) peptides with known IC₅₀ binding values, (3) peptides tested positive by tetramer staining, (4) peptides with known binding information at the MHCBN database. The model produces high predic-

tion efficiencies (average 60 %) with good sensitivity (~50%–73%) and specificity (52%–58%) values. The average positive predictive value of the model is 89%, while the average negative predictive value is only 18%. The efficiency is very high in predicting binders and very low in predicting nonbinders. This model is superior to many existing methods because of its potential application to any given MHC allele whose sequence is clearly defined. *Human Immunology* 64, 1123–1143 (2003). © American Society for Histocompatibility and Immunogenetics, 2003. Published by Elsevier Inc.

KEYWORDS: MHC; peptide binding; amino acids property matrix; prediction

ABBREVIATIONS

ASA	accessible surface area
CS	circumsporozoite
CTL	cytotoxic T lymphocyte
DS1	dataset 1
DS2	dataset 2
DS3	dataset 3
DS4	dataset 4
EBV	Epstein-Barr virus
ERP	essential residue position
HBV	hepatitis B virus
HCMV	human cytomegalovirus
HCV	hepatitis C virus
HIV	human immunodeficiency virus
HLA	human leukocyte antigen
IE-1	protein in the immediate early phase of HCMV replication cycle

MHC	major histocompatibility complex
MHCp	MHC-peptide
MOMP	major outer membrane protein
NS3	nonstructural protein 3
NS4B	nonstructural protein 4B
NS5B	nonstructural protein 5B
PBMC	peripheral blood mononuclear cell
PDB	protein databank
PDM	property distance matrix
pp65	HCMV matrix protein
TCR	T-cell receptor
RT	reverse transcriptase
SEHPR	solvent exposed hydrophobic peptide residues
vdWC	van der Waals clashes

INTRODUCTION

Data from protein x-ray crystallography [1–18], competitive binding assay [19–23], and biochemical binding

experiment [24–27] indicates the binding of major histocompatibility complex (MHC) molecules to short an-

From the School of Mechanical and Production Engineering (B.Z., M.K.S., P.K.), Nanyang Centre for Supercomputing and Visualization, Nanyang Technological University, Singapore; the Department of Human Genetics (V.S.M.), University of Texas Medical Branch, Galveston, TX, USA; and the Defense Medical Research Institute (G.R., S.M), Singapore.

Address reprint requests to: Dr. Pandjassaram Kanguane, School of

Mechanical and Production Engineering, 50 Nanyang Avenue, Nanyang Technological University, Singapore 639 798; Tel: +65 (67) 90 5836; Fax: +65 (67) 74 4340; E-mail: mpandjassaram@ntu.edu.sg.

Received June 17, 2003; revised August 4, 2003; accepted August 12, 2003.

tigenic peptides. Information from MHC-peptide (MHCp) T-cell receptor (TCR) x-ray crystallography data [28–35], tetramer staining experiment [36–62], and ^{51}Cr release assay [63, 64] suggests the specific binding of MHCp complexes to TCR molecules. Therefore, it is well established that the specific binding of peptides to MHC molecules is one among the many crucial steps in cell-mediated immune response. It is also known that MHCp binding does not fully guarantee immunogenicity [65]. It should also be noted that tetramer negative cells may still retain peptide-specific cytotoxicity and interferon- γ (INF γ) production [66]. Despite these appealing observations and cautious inferences, the MHC biology still remains unclear because of high polymorphism among MHC molecules.

The current International ImMunoGeneTics information system (IMGT)/human leukocyte antigen (HLA) database (January 2003) contains over 1620 HLA alleles [67, 68]. Several billion 9 residues long peptide stretches can be generated from 20 naturally occurring amino acids. Only a small fraction of this large peptide pool binds specifically to each of the known alleles and the MHCp binding is highly combinatorial. For MHCp binding predictions, a number of mathematical models have been proposed and tested using computer-aided calculations in the last decade [69–90]. Many common models utilize features extracted from pooled peptide sequences and peptide residue anchor position assignment procedures [69–78]. The efficiency of these models has been significantly improved by the careful utilization of the extracted features in nonlinear statistical models such as artificial neural network (ANN) [79, 80], Hidden Markov Model (HMM) [81], and Support Vector machine (SVM) [82]. A bottleneck in anchor position assignment is the need for a large number of peptide sequences with predetermined binding affinity to each of the known MHC alleles. The efficiency of all data driven predictive models is traditionally judged by the quality and quantity of binding data generated for each MHC allele. Generation of such binding data for each allele is experimentally laborious and economically expensive.

Concurrent progress in computational structural biology during the last decade has motivated many of us to assess the usefulness of molecular mechanics and modeling calculations for MHCp binding predictions [83–90]. All known MHC molecules have at least 70% sequence identity to one of the known MHC structures [91]. Homologous proteins share a common fold, even when the overall sequence identity is less than 10% [92]. Homology modeling and side chain packing calculations are commonly performed to predict the structures of all known MHC alleles [83–90]. The drawback in this approach is the current limitation in backbone predictions as peptide backbones in the binding groove play a

very significant role [83, 93]. This is further complicated due to the inadequate knowledge on the physicochemical characteristics of MHCp molecular recognitions [94, 95]. Also, a large number of interaction functions and scoring schemes have been validated over the last few years but their efficiency is far from satisfactory.

We have been studying the nature of MHCp molecular interactions using crystal structures during the past years [91, 93, 96, 97]. The results from our previous investigations have significantly improved our understanding of MHCp interactions in relation to the polymorphism of MHC molecules. We exhaustively used information from human MHCp structures and developed a novel model for MHCp binding predictions. The use of structural properties in the development of the model and its predictive efficiency for different datasets are discussed in this article.

MATERIALS AND METHODS

Description of the Model

We describe here the development and validation of a MHCp binding prediction model. The model is developed in the following six steps.

Step 1: Creation of human MHCp x-ray structure dataset.

The human MHCp three-dimensional (3D) structures solved by x-ray crystallography are obtained from the protein databank (PDB). We selected 29 (class I) MHCp structures from PDB (Table 1). The selected structures represent 22 unique MHCp sequences. In other words, the combination of MHC and peptide sequences in them is unique. It should be noted that peptides in selected structures are of length 9 residues long. MHCp structures with bound peptides of length less or greater than 9 residues are removed from the dataset. The length of the peptide throughout the investigation is fixed at 9 residues to avoid possible bias due to dissimilar MHCp binding patterns among peptides of different length [91]. Our previous study demonstrates that 8, 9, and 10 residues long peptides have different binding patterns (in physical terms), especially at the C terminal of the peptides. Moreover, MHCp structures with bound peptides of length 9 residues are best represented among solved structures. It should also be stated that MHCp structures with peptides of length 8 and 10 residues are not as represented as 9 residues long peptides. Therefore, we fixed the length of the peptide to 9 residues. This will not affect the application of the model in peptide vaccine design, because the length of the peptide in the design of a synthetic vaccine can be fixed. Issues related to class I MHC bound peptides of length greater than 9 residues have to be addressed in future as we have access to more structural data of its kind. In this investigation we

TABLE 1 Class I MHCp complexes in PDB

Number	PDB code	MHC allele	CIM	Peptide set		CIP	PL	Source	R [Å]	Release year	Reference
				Redundant	Nonredundant						
1	1JHT	A*0201	{a}	ALGIGILTV	ALGIGILTV	{c}	9	Mart-1	2.15	2001	[1]
2	1B0G	A*0201	{a}	ALWGFPPVL	ALWGFPPVL	{c}	9	Human-peptide P0149	2.60	1998	[2]
3	1I7U	A*0201	{a}	ALWGFVPVL	ALWGFVPVL	{c}	9	Synthetic construct	1.80	2001	[3]
4	1I7T	A*0201	{a}	ALWGVFPVL	ALWGVFPVL	{c}	9	Synthetic construct	2.80	2001	[3]
5	1I7R	A*0201	{a}	FAPGFFPYL	FAPGFFPYL	{c}	9	Synthetic construct	2.20	2001	[3]
6	1I1F	A*0201	{a}	FLKEPVHGV	FLKEPVHGV	{c}	9	HIV RT	2.80	2000	[4]
7	1HHI	A*0201	{a}	GILGFVFTL	GILGFVFTL	{c}	9	Synthetic	2.50	1993	[5]
8	1AKJ	A*0201	{a}	ILKEPVHGV	ILKEPVHGV	{c}	9	HIV-1 RT	2.65	1997	[6]
9	1HHJ	A*0201	{a}	ILKEPVHGV	—	{c}	9	Synthetic	2.50	1993	[5]
10	1QRN	A*0201	{a}	LLFGYAVYV	LLFGYAVYV	{c}	9	Tax peptide P6A	2.80	1999	[7]
11	1QSE	A*0201	{a}	LLFGYPRYV	LLFGYPRYV	{c}	9	Tax peptide V7R	2.80	1999	[7]
12	1QSF	A*0201	{a}	LLFGYPVAV	LLFGYPVAV	{c}	9	Tax peptide Y8A	2.80	1999	[7]
13	1AO7	A*0201	{a}	LLFGYPVYV	LLFGYPVYV	{c}	9	HTLV-1 Tax	2.60	1997	[8]
14	1BD2	A*0201	{a}	LLFGYPVYV	—	{c}	9	HTLV-1 Tax	2.50	1998	[9]
15	1DUZ	A*0201	{a}	LLFGYPVYV	—	{c}	9	HTLV-1 Tax	1.80	2000	[10]
16	1HHK	A*0201	{a}	LLFGYPVYV	—	{c}	9	Synthetic	2.50	1993	[5]
17	1IM3	A*0201	{a}	LLFGYPVYV	—	{c}	9	HTLV-1 Tax	2.20	2001	[11]
18	1HHG	A*0201	{a}	TLTSCNTSV	TLTSCNTSV	{c}	9	HIV-1 gp120	2.60	1993	[5]
19	1I1Y	A*0201	{a}	YLKEPVHGV	YLKEPVHGV	{c}	9	HIV-1 RT	2.20	2000	[4]
20	1HSA	B*2705	{a}	ARAAAAAIAK	ARAAAAAIAK	{c}	9	N/A	2.10	1992	[12]
21	1JGE	B*2705	{a}	GRFAAAIAK	GRFAAAIAK	{c}	9	Synthetic (M9)	2.10	2002	[13]
22	1K5N	B*2709	{a}	GRFAAAIAK	—	{c}	9	Synthetic (M9)	1.09	2002	[13]
23	1A9B	B*3501	{a}	LPPLDITPY	LPPLDITPY	{c}	9	EBNA-3C	3.20	1998	[14]
24	1A9E	B*3501	{a}	LPPLDITPY	—	{c}	9	EBV-Ebna3c	2.50	1998	[14]
25	1E27	B*5101	{a}	LPPVVAKEI	LPPVVAKEI	{c}	9	HIV-1 Kml	2.20	2000	[15]
26	1A1M	B*5301	{a}	TPYDINQML	TPYDINQML	{c}	9	HIV-2 gag	2.30	1998	[16]
27	1A1O	B*5301	{a}	KPIVQYDNF	KPIVQYDNF	{c}	9	HIV-1 Nef	2.30	1998	[16]
28	1EFX	Cw*0304	{a}	GAVDPLLAL	GAVDPLLAL	{c}	9	Importin α -2	3.00	2000	[17]
29	1IM9	Cw*0401	{a}	QYDDAVYKL	QYDDAVYKL	{c}	9	Synthetic	2.80	2001	[18]

Duplicate peptides were removed in the nonredundant peptide set.

Abbreviations: CIM = chain identifier for MHC; CIP = chain identifier for peptide; MHC = major histocompatibility complex; PDB = protein databank; PL = peptide length; R = resolution; Release year = the year in which the entry was released by PDB.

strictly restricted our analysis to class I specific peptides and we plan to extend the work to class II specific peptides in the future.

Step 2: Identification of highly essential residue positions. We used 29 human MHCp crystal structures from PDB (Table 1) and identified 39 residue positions in the α_1 , α_2 domains that are predominantly involved in peptide binding (Figure 1A; Table 2). These residue positions are called highly essential residue positions (HERP). The residues at the HERPs (except at position 24) revealed a mean change in solvent accessibility (Δ ASA) of $\geq 1 \text{ \AA}^2$ upon MHCp complex formation (Figure 1B; Table 2). This is true for a set of MHCp complexes characterized by varying MHC alleles and peptide sequences (Table 1). We included position 24 into the set of HERPs because residues at this position indicate three inter-residue interactions with the peptides in the dataset. The HERPs defined in this investigation is a subset of the essential

residue positions (ERPs) defined in our previous investigation [98]. The difference is in the selection of cutoff values used for selecting HERPs and ERPs. The ERPs reveal a Δ ASA of $> 0 \text{ \AA}^2$, whereas the HERPs exhibit a Δ ASA of $\geq 1 \text{ \AA}^2$ in the dataset. In this investigation we neglected those positions that characterized a very low Δ ASA of $< 1 \text{ \AA}^2$ and $> 0 \text{ \AA}^2$ because these positions have negligible involvement in MHCp binding. We excluded these residues using an arbitrary Δ ASA range of 0 to $< 1 \text{ \AA}^2$ because we believed their inclusion will add noise to the model. It should also be noted that the definition HERPs stated in article are different from CPERP's defined in our previous investigation; CPERP refers to polymorphic positions among 212 HLA-A alleles [98].

Step 3: Definition of virtual binding pockets. The MHC residue positions interacting with each of the peptide residue positions are identified by counting the number

FIGURE 1 Essential residues in the MHC binding groove are illustrated. (a) MHC residue positions demonstrating change in accessibility upon complex formation, these residue positions are identified using a set of 29 MH-C-peptide crystal structures; (b) mean change in accessibility in the dataset; (c) standard deviation about the mean change in accessibility in the dataset. Abbreviations: ASA = accessible surface area; HLA = human leukocyte antigen; MHC = major histocompatibility complex.

of inter-residue pairwise contacts between the MHC and the peptide (Table 2). By definition, MHCp residue pairs are considered to be interacting if the distance between any atom in the peptide residue and any atom in the

MHC residue is less than or equal to an arbitrary cutoff distance of 4 Å. This procedure established the relationship between peptide residue positions and MHC residue positions (Table 3). The residue pairs at the correspond-

TABLE 2 List of HERPs

HERP	Δ ASA		SUM	Contacts	
	Mean	SD		Mean	SD
5	3.59	1.39	19	0.62	0.49
7	18.33	2.81	61	2.00	0.27
9	6.26	3.88	29	1.00	0.60
24	0.71	1.76	3	0.14	0.35
45	4.94	3.03	20	0.66	0.48
59	4.91	3.78	13	0.45	0.51
62	8.67	16.53	11	0.38	0.73
63	21.12	4.12	58	1.93	0.26
65	5.23	8.76	4	0.21	0.49
66	67.99	11.29	91	2.97	0.98
67	5.47	3.84	8	0.28	0.45
69	16.67	12.05	7	0.24	0.44
70	34.31	14.51	49	1.55	0.99
72	1.96	4.71	3	0.10	0.31
73	50.37	10.67	62	2.14	0.95
74	2.70	5.45	7	0.24	0.58
76	16.51	8.69	18	0.62	0.49
77	35.45	7.29	61	2.10	0.31
80	15.27	4.94	30	1.03	0.42
81	7.75	2.92	17	0.59	0.50
84	3.63	1.90	29	1.00	0.00
95	3.87	5.80	7	0.24	0.51
97	22.18	10.65	28	0.97	0.82
99	25.79	4.73	34	1.17	0.38
114	2.68	3.71	6	0.21	0.41
116	11.93	7.79	26	0.83	0.66
123	7.38	5.17	5	0.17	0.38
143	13.40	1.63	8	0.28	0.45
146	21.39	7.58	8	0.28	0.53
147	44.31	5.67	55	1.76	0.44
150	7.77	7.66	4	0.14	0.35
152	26.85	7.69	19	0.62	0.56
155	46.49	18.96	40	1.38	0.86
156	21.70	9.90	27	0.86	0.69
159	46.86	5.12	41	1.34	0.77
163	11.55	6.30	3	0.10	0.31
167	24.96	8.47	11	0.38	0.49
171	4.34	1.33	1	0.03	0.19

Mean change in accessible surface area (Δ ASA) and standard deviation (SD) about the mean is given for each highly essential residue position (HERP). The inter-residue major histocompatibility complex peptide contacts are estimated at a cutoff distance of 4 Å for each HERP.

SUM = total number of inter residue contacts between a particular HERP and peptide residues in a dataset of MHC-peptide structures.

ing MHCp positions have at least 1 inter-residue interaction (Table 4). The set of MHC residue positions interacting with each peptide position forms the virtual binding pockets. As mentioned in step 2, these MHC residues positions (except position 24) demonstrate a Δ ASA of $\geq 1 \text{ Å}^2$ in the dataset. This suggests the residue positions forming the virtual pockets are none other than the HERPs defined in step 2 with the inclusion of position 24. Using the mapping that is established between the peptide residue positions and MHC virtual

binding pocket forming residue positions, the compatibility of a particular residue type in a specific peptide position to the residues forming the virtual binding pockets in the groove can be determined using a scoring scheme. A similar strategy is adopted to define binding pockets (derived from four unique human structures and two murine class I MHCp structures) by Chelvanayagam [99]. Chelvanayagam [100] extended the strategy to HLA-DR molecules (derived from two unique human DR1 structures and one murine I-Ek structure). It should be noted that the virtual pockets defined in this model are not strictly identical to the A–F pockets defined by Zhao *et al.* [98], Zhang *et al.* [101], and Guan *et al.* [76] (A–F pockets for A3 supertypes). The A–F pockets defined in these reports are certainly related (not strictly identical) to the virtual pockets defined in our method, but a one-to-one correspondence has to be established in future for further improvement. Current attempts to develop this one-to-one correspondence is pampered because of A–F pocket overlapping for relating virtual pockets defined in this article.

Step 4: Scoring scheme. The scoring scheme is one of the important steps in MHCp binding calculations. For a given MHC allele, the virtual binding pockets (1–9) can be defined using step 3. The binding of a given peptide to the MHC allele has to be evaluated using a scoring metric system. The overall binding affinity (predicted) is the sum of compatibility measures for peptide residues at each peptide residue position and its corresponding virtual binding pockets. We evaluated the compatibility of each peptide positional residues with the virtual pockets made of HERPs using the Q matrix developed elsewhere [102]. The Q matrix provides a measure to cluster the 20 amino acids based on 237 physicochemical properties. The overall ability of a particular peptide to bind to a particular MHC is a weighted sum of the amino acid preferences of the individual binding pockets. The binding affinity between the peptide and the MHC allele is estimated as the cumulative sum of their amino acid property compatibility values in the Q matrix.

In this calculation, the effect of peptide residue at position 4 is neglected because the residue at this position is always solvent exposed in all the MHCp crystal structures (Figures 2A and 2B). The standard deviation caused by amino acids changes at this position is also minimal (Figure 2C). This suggests that property of the amino acid at this position does not affect MHCp binding.

Step 5: Assessment of model performance. The performance of the model is assessed by calculating sensitivity (SE), specificity (SP), positive predictive value (PPV), negative predictive value (NPV), and accuracy (AC) using true-

TABLE 3 Relationship between HERP and peptide residue position

HERP	Peptide residue position									Total
	1	2	3	4	5	6	7	8	9	
5	19	0	0	0	0	0	0	0	0	19
7	30	30	1	0	0	0	0	0	0	61
9	0	26	1	0	1	0	1	0	0	29
24	0	3	0	0	0	0	0	0	0	3
45	0	20	0	0	0	0	0	0	0	20
59	13	0	0	0	0	0	0	0	0	13
62	5	2	0	4	0	0	0	0	0	11
63	29	29	0	0	0	0	0	0	0	58
65	0	0	0	3	0	1	0	0	0	4
66	15	25	23	24	1	3	0	0	0	91
67	0	8	0	0	0	0	0	0	0	8
69	0	0	0	1	1	5	0	0	0	7
70	0	3	22	2	7	13	2	0	0	49
72	0	0	0	0	0	0	0	3	0	3
73	0	0	0	0	3	22	23	19	0	67
74	0	0	0	0	3	0	2	0	2	7
76	0	0	0	0	0	0	0	18	0	18
77	0	0	0	0	0	0	3	30	30	63
80	0	0	0	0	0	0	0	3	28	31
81	0	0	0	0	0	0	0	0	18	18
84	0	0	0	0	0	0	0	0	30	30
95	0	0	0	0	0	0	1	0	6	7
97	0	0	8	0	3	5	12	0	2	30
99	0	3	30	0	1	0	0	0	0	34
114	0	0	3	0	0	0	3	0	0	6
116	0	0	0	0	0	0	6	0	20	26
123	0	0	0	0	0	0	0	0	5	5
143	0	0	0	0	0	0	0	0	8	8
146	0	0	0	0	0	0	1	7	0	8
147	0	0	0	0	0	0	26	29	0	55
150	0	0	0	0	0	0	4	0	0	4
152	0	0	0	0	2	1	16	0	0	19
155	0	0	14	2	23	1	4	0	0	44
156	0	0	21	0	4	2	0	0	0	27
159	24	0	16	0	1	0	0	0	0	41
163	3	0	0	0	0	0	0	0	0	3
167	11	0	0	0	0	0	0	0	0	11
171	1	0	0	0	0	0	0	0	0	1
Total	150	149	139	36	50	53	104	109	149	939

Abbreviation: HERP = highly essential residue positions.

The value in each cell indicates the number of inter-residue contacts between a residue at a HERP and a residue at a specific peptide residue position. The inter-residue contacts are calculated using a dataset of 29 MHCp structures at a cutoff distance of 4 Å.

positive (TP), true-negative (TN), false-positive (FP), and false-negative (FN).

$$SE = \frac{TP}{(TP + FN)} * 100,$$

$$SP = \frac{TN}{TN + FP} * 100,$$

$$PPV = \frac{TP}{(TP + FP)} * 100,$$

$$NPV = \frac{TN}{(TN + FN)} * 100, \text{ and}$$

$$\%AC = \frac{TP + TN}{TP + FP + TN + FN} * 100.$$

Step 6: Crossvalidation. The predictive efficiency of this procedure is assessed using four datasets.

Dataset 1 (DS1) is presented in Table 1. The characteristic features of this dataset are described in step 1.

TABLE 4 Map between PRP and virtual pockets in the binding groove

	PRP								
	P1	P2	P3	P4	P5	P6	P7	P8	P9
HERP	5	7	7	62	9	65	9	72	74
	7	9	9	65	66	66	70	73	77
	59	24	66	66	69	69	73	76	80
	62	45	70	69	70	70	74	77	81
	63	62	97	70	73	73	77	80	84
	66	63	99	155	74	97	95	146	95
	159	66	114		97	152	97	147	97
	163	67	155		99	155	114		116
	167	70	156		152	156	116		123
	171	99	159		155		146		143
				156		147			
				159		150			
						152			
						155			

Abbreviations: HERP = highly essential residue positions; PRP = peptide residue positions; P1 to P9 = residue positions in a 9 mer peptide.

This dataset contains sequence information represented by 29 MHCp structures with bound nanomer peptides. There are 8 unique MHC alleles and 22 unique nanomer peptides in the dataset. The model, developed in steps 1–5, is used to predict the binding affinity of these nanomer peptides with the corresponding MHC alleles. The predicted values are presented in Table 5. All the peptides tested in this dataset are binders as their structure is solved with their corresponding alleles. Using this information, the predicted peptides are identified as TP and FN. FP and TN are not available for this dataset because this dataset does not contain nonbinders. Internal crossvalidation indicates that TP is 20 and FN is 9 for this dataset. Therefore, the SE of the model is 69% for this dataset.

Dataset 2 (DS2) is obtained from literature and presented in Table 6. Peptides in this dataset are a subset of peptides used in our previous investigation [86]. This dataset contains only nanomer peptides and does not include octamer and decamer peptides that are used in our previous study [86]. In this dataset, 102 binding data are present, of which 24 are tested for A*0201; 11 each tested for A*0202, A*0203, A*0206, A*6802; 6 each tested for A*0301, A*1101, A*3101, A*3301, A*6801; and 1 each tested for B*3501, B5101, B*5301, and B5401. The definition of binders and nonbinders (on the basis of IC₅₀ cutoff value) stated in our previous study is followed here [86]. The model, described in steps 1–4, is validated using this dataset and the results are presented (Table 7) and summarized in Table 6. Validation reveals that TP is 39, TN is 18, FP is 9, and FN is 36 for this dataset. Therefore, the SE is 67%, SP is 52%, AC is 55%, PPV is 81%, and NPV is 33% for this dataset. For the same dataset we applied the sidechain

packing, and the van der Waals clashes (vdWC)/solvent exposed hydrophobic peptide residue (SEHP) scoring procedures as described elsewhere [86]. The SE is 70%, SP is 53%, AC is 66%, PPV is 80%, and NPV is 40% (recalculated for 9 mer peptides) for predictions based the sidechain packing and vdWC/SEHPR scoring procedures (Table 7).

Dataset 3 (DS3) is obtained from literature and presented in Table 8. This process is laborious and time consuming. Peptides in this dataset are tested positive by tetramer staining in peripheral blood mononuclear cells (PBMC) obtained from infected patients. In this dataset 46 tetramer positive peptide data are present, of which 15 are tested for A*0201; 1 tested for A*6802; 14 tested for A2; 2 tested each for B*3501, B14, B27, and B53; and 1 tested each for A3, B*0702, B*0801, B*5801, B18, B35, B7, and Cw4. These peptides are present in the MHCBN database [22]. Qualitative information on binding (BI), relative binding (RB), CTL activity, and relative CTL activity (RCTL) are available for most peptides. Unknown information are indicated by a “?” mark (this notation applies throughout this text). The model described in steps (1–5) is validated using this dataset and the results are presented in Table 9 and summarized in Table 8. Validation indicates that TP is 34 and FN is 10 for this model. Therefore, the SE is 78% and AC is 78% for this dataset. This dataset does not contain tetramer negative peptides in PBMC, and TN and FP are not available for this dataset.

Dataset 4 (DS4) is presented in Tables 10 and 11. Peptides in this dataset are strictly 9 residues long and are retrieved from MHCBN [22]. Qualitative binding information is available for these peptides in MHCBN. The dataset contains 1624 HLA-A specific peptides

FIGURE 2 Involvement of peptide residue positions in binding are illustrated (9 residues long). (a) Position specific involvement in the dataset are shown; (b) mean involvement of each peptide residue position; (c) standard deviation about the mean involvement of each peptide residue position.

(1574 binders and 50 nonbinders) and 861 HLA-B specific peptides (859 binders and 0 nonbinders). In this dataset, the HLA-A peptides are spread over 33 alleles

and the HLA-B peptides are spread over 52 alleles. The model described in steps (1–5) is validated using this dataset and the results are summarized (Tables 10 and

TABLE 5 Performance of model in dataset 1 (MHC_p structural dataset)

Serial number	PDB code	HLA allele	Peptide	Predicted value	Inference from prediction
1	1JHT	A*0201	ALGIGILTV	−1439.95	FN
2	1BOG	A*0201	ALWGFFPVL	−46.64	FN
3	1I7U	A*0201	ALWGFVPVL	156.61	TP
4	1I7T	A*0201	ALWGVFPVL	−140.40	FN
5	1I7R	A*0201	FAPGFFPYL	−224.39	FN
6	1I1F	A*0201	FLKEPVHGV	754.14	TP
7	1HHI	A*0201	GILGFVFTL	685.15	TP
8	1AKJ	A*0201	ILKEPVHGV	578.75	TP
9	1HHJ	A*0201	ILKEPVHGV	578.75	TP
10	1QRN	A*0201	LLFGYAVYV	1220.34	TP
11	1QSE	A*0201	LLFGYPRYV	1012.92	TP
12	1QSF	A*0201	LLFGYPVAV	775.34	TP
13	1AO7	A*0201	LLFGYPVYV	831.96	TP
14	1BD2	A*0201	LLFGYPVYV	831.96	TP
15	1DUZ	A*0201	LLFGYPVYV	831.96	TP
16	1HHK	A*0201	LLFGYPVYV	831.96	TP
17	1IM3	A*0201	LLFGYPVYV	831.96	TP
18	1HHG	A*0201	TLTSCNTSV	66.71	TP
19	1I1Y	A*0201	YLKEPVHGV	1092.96	TP
20	1HSA	B*2705	ARAAAAAIAK	−361.18	FN
21	1JGE	B*2705	GRFAAAIAK	−753.20	FN
22	1K5N	B*2709	GRFAAAIAK	−683.80	FN
23	1A9B	B*3501	LPPLDITPY	753.25	TP
24	1A9E	B*3501	LPPLDITPY	753.25	TP
25	1E27	B*5101	LPPVVAKEI	1990.63	TP
26	1A1M	B*5301	TPYDINQML	1150.85	TP
27	1A1O	B*5301	KPIVQYDNF	225.88	TP
28	1EFX	C*0304	GAVDPLLAL	−2708.77	FN
29	1IM9	C*0401	QYDDAVYKI	−410.90	FN

The predicted values are estimated using the model described in this article.

Abbreviations: FN = false-negative value; HLA = human leukocyte antigen; MHC_p = major histocompatibility complex–peptide; PDB = protein databank; TP = true-positive value.

TABLE 6 Summary of model/performance in dataset 2 (dataset with known binding values [IC₅₀])

HLA allele	Number of peptides	TP	TN	FP	FN	% SE	% SP	% AC
A*0201	24	15	0	0	9	—	63	63
A*0202	11	5	1	2	3	33	63	55
A*0203	11	6	0	0	5	—	55	55
A*0206	11	3	2	2	4	50	43	45
A*0301	6	2	0	0	4	—	33	33
A*1101	6	4	0	0	2	—	67	67
A*3101	6	0	2	2	2	50	—	33
A*3301	6	0	6	0	0	100	—	100
A*6801	6	2	1	0	3	100	40	50
A*6802	11	1	5	3	2	63	33	55
B*3501	1	0	0	0	1	—	—	—
B*5101	1	1	0	0	0	—	100	—
B*5301	1	0	0	0	1	—	—	—
B*5401	1	0	1	0	0	100	—	100
Total	102	39	18	9	36	67	52	55

Abbreviations: AC = accuracy; FN = false-negative value; FP = false-positive value; HLA = human leukocyte antigen; SE = sensitivity; SP = specificity; TN = true-negative value; TP = true-positive value.

TABLE 7 Performance of model in dataset 2 (with Known binding values [IC₅₀])

HLA	Peptide	Predicted value	IC ₅₀	Inference	Modeling results [86]				References
					vdWC	SEHPR	Remark	Inference	
A*0201	ALSTGLIHL	−1090.68	312.5	FN	0	0	G	TP	[19]
A*0201	ALCRWGGLL	−457.59	165	FN	0	1	G	TP	[20]
A*0201	ILAGYGAGV	−167.45	116.3	FN	0	0	G	TP	[19]
A*0201	DLMGYIPLV	−95.25	80	FN	0	1	G	TP	[19]
A*0201	LLTFWNPPT	−83.77	178.6	FN	1	1	G	TP	[20]
A*0201	VHRDDLLEA	−70.81	365	FN	4	0	P/NB	FN	[20]
A*0201	VLHDDLLEA	−67.64	30	FN	0	0	G	TP	[20]
A*0201	KIFGSLAFL	−64.74	33.3	FN	0	1	G	TP	[20]
A*0201	CLTSTVQLV	−17.97	147.1	FN	0	1	G	TP	[20]
A*0201	LLTFWNPV	62.46	16	TP	5	2	P/NB	FN	[20]
A*0201	KVAELVHFL	66.11	68.4	TP	0	1	G	TP	[20]
A*0201	TLTSCNTSV	66.71	+	TP	0	0	G	TP	[8]
A*0201	LMTFWNPV	94.78	4.5	TP	0	1	G	TP	[20]
A*0201	ILHNGAYSL	392.04	74.6	TP	0	0	G	TP	[20]
A*0201	YLSGANLNL	490.7	27.8	TP	0	0	G	TP	[20]
A*0201	LLFGYPVYP	546.52	+	TP	0	0	G	TP	[8]
A*0201	ILKEPVHGV	578.75	+	TP	0	0	G	TP	[6]
A*0201	GILGFVFTL	685.15	+	TP	0	1	G	TP	[5]
A*0201	FLLADARV	785.92	18.2	TP	1	1	G	TP	[19]
A*0201	HMWNFISGI	802.42	15.2	TP	2	1	P/NB	FN	[19]
A*0201	LLFGYPVYV	831.96	+	TP	0	0	G	TP	[5]
A*0201	IMIGVLGV	1022.36	68.5	TP	0	0	G	TP	[20]
A*0201	FLWGPRALV	1152.63	31.3	TP	0	1	G	TP	[20]
A*0201	YLVAYQATV	1554.11	20.4	TP	0	0	G	TP	[19]
A*0202	KVAELVHFL	−266.54	29	FN	2	1	P/NB	FN	[20]
A*0202	CLTSTVQLV	−256.05	457	FN	3	1	P/NB	FN	[20]
A*0202	KIFGSLAFL	−87.74	9	FN	2	1	P/NB	FN	[20]
A*0202	LLTFWNPPT	124.29	1720	FP	0	1	G	FP	[20]
A*0202	LMTFWNPV	270.97	779	FP	1	1	G	FP	[20]
A*0202	ALCRWGGLL	−84.74	>10000	TN	0	1	G	FP	[20]
A*0202	LLTFWNPV	238.65	297	TP	0	1	G	TP	[20]
A*0202	ILHNGAYSL	259.27	358	TP	7	0	P/NB	FN	[20]
A*0202	YLSGANLNL	467.9	165	TP	1	0	G	TP	[20]
A*0202	IMIGVLGV	578.93	62	TP	1	0	G	TP	[20]
A*0202	FLWGPRALV	1443.05	43	TP	1	1	G	TP	[20]
A*0203	KVAELVHFL	−882.17	14	FN	2	1	P/NB	FN	[20]
A*0203	ALCRWGGLL	−655.4	278	FN	0	0	G	TP	[20]
A*0203	CLTSTVQLV	−385.35	6.7	FN	1	0	G	TP	[20]
A*0203	KIFGSLAFL	−347.79	9	FN	11	2	P/NB	FN	[20]
A*0203	IMIGVLGV	−245.8	13	FN	1	0	G	TP	[20]
A*0203	ILHNGAYSL	110.75	100	TP	5	0	P/NB	FN	[20]
A*0203	LLTFWNPPT	273.6	67	TP	5	1	P/NB	FN	[20]
A*0203	YLSGANLNL	295.31	2.4	TP	3	0	P/NB	FN	[20]
A*0203	LLTFWNPV	419.83	26	TP	1	1	G	TP	[20]
A*0203	LMTFWNPV	452.15	7.6	TP	3	0	P/NB	FN	[20]
A*0203	FLWGPRALV	1784.82	14	TP	15	1	P/NB	FN	[20]
A*0206	KIFGSLAFL	−386.24	23	FN	0	1	G	TP	[20]
A*0206	KVAELVHFL	−280.73	168	FN	0	1	G	TP	[20]
A*0206	CLTSTVQLV	−37.7	308	FN	1	1	G	TP	[20]
A*0206	LLTFWNPV	−1.04	56	FN	0	1	G	TP	[20]
A*0206	YLSGANLNL	160.02	804	FP	1	0	G	FP	[20]
A*0206	ILHNGAYSL	373.76	567	FP	0	0	G	FP	[20]
A*0206	ALCRWGGLL	−843.22	8863	TN	2	1	P/NB	TN	[20]
A*0206	LLTFWNPPT	−147.27	755	TN	1	1	G	FP	[20]
A*0206	LMTFWNPV	53.23	33	TP	0	1	G	TP	[20]
A*0206	IMIGVLGV	565.73	106	TP	0	0	G	TP	[20]
A*0206	FLWGPRALV	955.52	336	TP	15	1	P/NB	FN	[20]
A*0301	VAGALVAFK	−1180.4	45.8	FN	0	1	G	TP	[21]
A*0301	LGFGAYMSK	−570.56	135.8	FN	3	0	P/NB	FN	[21]
A*0301	RLGVRATRK	−270.22	11.7	FN	1	0	G	TP	[21]

TABLE 7 Performance of model in dataset 2 (with Known binding values [IC₅₀]) (*Continued*)

HLA	Peptide	Predicted value	IC ₅₀	Inference	Modeling results [86]				References
					vdWC	SEHPR	Remark	Inference	
A*0301	KTTERSQPR	−208.66	68.8	FN	0	0	G	TP	[21]
A*0301	QLFTFSRR	715.52	14.7	TP	0	1	G	TP	[21]
A*0301	LIFCHSKKK	1293.56	20.4	TP	0	0	G	TP	[21]
A*1101	VAGALVAFK	−1678.12	6.7	FN	0	1	G	TP	[21]
A*1101	LGFGAYMSK	−1192.45	20.7	FN	0	0	G	TP	[21]
A*1101	QLFTFSRR	146.67	182	TP	0	0	G	TP	[21]
A*1101	RLGVRATRK	199.22	207	TP	0	0	G	TP	[21]
A*1101	KTTERSQPR	456.85	93.8	TP	0	0	G	TP	[21]
A*1101	LIFCHSKKK	747.33	1.6	TP	3	0	P/NB	FN	[21]
A*3101	KTTERSQPR	−798.98	66.7	FN	0	0	G	TP	[21]
A*3101	RLGVRATRK	−533.25	428.6	FN	0	1	G	TP	[21]
A*3101	LIFCHSKKK	14.02	2535.6	FP	2	0	P/NB	TN	[21]
A*3101	VAGALVAFK	195.19	3750	FP	1	0	G	FP	[21]
A*3101	LGFGAYMSK	−266.01	2950.8	TN	10	0	P/NB	TN	[21]
A*3101	QLFTFSRR	−61.93	620.7	TN	0	1	G	FP	[21]
A*3301	KTTERSQPR	−659.65	1812.5	TN	0	0	G	FP	[21]
A*3301	RLGVRATRK	−418.56	NB	TN	1	0	G	FP	[21]
A*3301	LGFGAYMSK	−315.45	NB	TN	3	0	P/NB	TN	[21]
A*3301	LIFCHSKKK	−161.54	NB	TN	14	0	P/NB	TN	[21]
A*3301	QLFTFSRR	−140.78	3766.2	TN	0	0	G	FP	[21]
A*3301	VAGALVAFK	−69.22	NB	TN	5	1	P/NB	TN	[21]
A*6801	VAGALVAFK	−1537.89	258.1	FN	2	1	P/NB	FN	[21]
A*6801	LGFGAYMSK	−576.33	222.2	FN	4	1	P/NB	FN	[21]
A*6801	KTTERSQPR	−250.71	145.5	FN	1	0	G	TP	[21]
A*6801	RLGVRATRK	−506.61	NB	TN	4	0	P/NB	TN	[21]
A*6801	QLFTFSRR	277.71	2.6	TP	1	0	G	TP	[21]
A*6801	LIFCHSKKK	723.53	333.3	TP	0	0	G	TP	[21]
A*6802	KVAELVHFL	−1379.03	17	FN	0	1	G	TP	[20]
A*6802	IMIGVLGV	−500.95	89	FN	2	1	P/NB	TN	[20]
A*6802	LMTFWNPPV	1.52	3448	FP	3	0	P/NB	TN	[20]
A*6802	YLSGANLNL	64.77	>10000	FP	2	2	P/NB	TN	[20]
A*6802	ILHNGAYSL	98.77	>10000	FP	0	1	G	FP	[20]
A*6802	ALCRWGILL	−1334.68	>10000	TN	2	1	P/NB	TN	[20]
A*6802	KIFGSLAFL	−654.82	3333	TN	3	2	P/NB	TN	[20]
A*6802	CLTSTVQLV	−205.81	8000	TN	2	0	P/NB	TN	[20]
A*6802	LLTFWNPPPT	−139.57	>10000	TN	1	0	G	FP	[20]
A*6802	LLTFWNPPV	−131.91	9442	TN	3	0	P/NB	TN	[20]
A*6802	FLWGPRAIV	1026.62	40	TP	18	1	P/NB	FN	[20]
B*3501	LPGCSFSIF	−643.83	90.4	FN	2	1	P/NB	FN	[21]
B*5101	LPGCSFSIF	172.98	100	TP	2	1	P/NB	FN	[21]
B*5301	LPGCSFSIF	−109.41	114	FN	0	1	G	TP	[21]
B*5401	LPGCSFSIF	−577.51	6666	TN	4	1	P/NB	TN	[21]

Abbreviations: FN = false-negative value; FP = false-positive value; G = good binder; P/NB = poor/nonbinder; SEHPR = solvent exposed hydrophobic peptide residue; vdWC = van der Waals clash.

11). Validation reveals that TP is 770 | 539, TN is 29 | ?, FP is 21 | ? and FN is 804 | 320 for HLA-A | HLA-B peptides. The SE is 49% | 63, SP is 58% | ?, AC is 49% | 63, PPV is 97% | ?, and NPV is 3% | ? for HLA-A | HLA-B peptides from MHCBN. It should be noted that this dataset contains a predominant number of binders and a very small number of nonbinders (only 50 nonbinders).

RESULTS AND DISCUSSION

The critical issue in the design of a T-cell epitope vaccine is the identification of peptide cocktails with proven ability to bind all or most known class I or class II HLA molecules. Experimental identification of such a peptide mixture is arduous, time consuming, and economically not feasible. For example, a 200 residues long surface antigen can ideally produce 192 overlapping 9 residues

TABLE 8 Summary of model performance in dataset 3 (peptides that tested positive by tetramer staining)

Serial number	HLA allele	Number of peptides	TP	TN	FP	FN	SE (%)	AC (%)
1	A*0201	15	13	—	—	2	86	86
2	A2	14	12	—	—	2	86	86
3	A3	1	1	—	—	0	100	100
4	A*6802	1	0	—	—	1	—	—
5	B*3501	2	0	—	—	2	—	—
6	B7	1	0	—	—	1	—	—
7	B*0702	1	1	—	—	0	100	100
8	B*0801	1	1	—	—	0	100	100
9	B14	2	1	—	—	1	50	50
10	B18	1	1	—	—	0	100	100
11	B27	2	2	—	—	0	100	100
12	B35	1	1	—	—	0	100	100
13	B53	2	1	—	—	1	50	50
14	B*5801	1	1	—	—	0	100	100
15	Cw4	1	1	—	—	0	100	100
Total		46	36	—	—	10	78	78

Abbreviations: AC = accuracy; FN = false-negative value; FP = false-positive value; HLA = human leukocyte antigen; SE = sensitivity; SP = specificity; TN = true-negative value; TP = true-positive value.

long short peptides. Testing the binding ability of these short peptides to more than 1500 HLA molecules (current status) requires cloning, expression, and purification of these molecules followed by binding assay. A reasonably validated mathematical model for the prediction of these binding peptides to multiple HLA molecules is handy and many such mathematical models are currently available. Each of the currently available methods has its own merits and demerits [69–90]. In this article we propose a novel predictive model that estimates the binding affinity of peptides to MHC molecules as a sum of compatibility measures between each residue at every peptide position to the corresponding residues forming the virtual binding pocket in the receptor molecule. The measure of compatibility between the peptide residues and the virtual binding pockets is determined using a high dimensional amino acids property matrix called the Q matrix [98]. Other salient features of the model discussed below include: (1) identification of HERPs from 29 MHCp structures, (2) mapping of individual peptide residue positions to virtual binding pockets, (3) exclusion of fourth residue in the peptide for MHCp binding calculations, (4) generic nature of the model and its application to any allele whose sequence is clearly defined.

The HERPs are identified using 29 MHCp structures (Table 1). The dataset contains at least one complex from A*0201, B*2705, B*2709, B*3501, B*5101, B*5301, Cw*0304, and Cw*0401 (Table 1). The dataset also contains 22 unique 9 residue long peptides (Table 1). Although this dataset is not exhaustive, it is reasonably representative for the current analysis. The definition of HERPs can be refined as and when unique MHCp struc-

tures are available in the PDB. It is generally agreed that homologous proteins share a common fold, even when the overall sequence identity is less than 10% [92]. HLA alleles share at least 70% sequence identity within a class [83]. Therefore, all the HLA molecules will have identical structural folds and binding grooves. MHC alleles share a very high homology as reflected from the sequence identity measure (greater than 70%). Studies show that proteins with such high percentage sequence similarity have negligible difference in their protein structures (3D backbone trace). The HERP defined in our study is based on MHCp structures in the PDB and the HERP's determinants will not be significantly different for homologous structures (other HLA alleles). However, this could be tested and validated as and when a large number of unique MHCp structures are available in the PDB.

The functional difference between HLA alleles is caused by the polymorphism among them in the binding groove. This alters the side chain orientation of residues in the binding groove affecting the pocket surfaces and their chemical properties. Although the side chain orientations of these residues are expected to differ between HLA molecules, the backbone trace of such residues in space is largely similar [83]. Our previous study also indicates that MHC side-chain atoms have a significant role in MHCp binding than MHC backbone atoms [93]. Despite the conservation of backbone trace among HLA alleles at the binding groove, different HLA alleles bind the same peptide with different affinity because of the variation in the physicochemical contour (properties of side-chain atoms) caused by the subtle polymorphism in the groove. The HERPs identified in this study provide

TABLE 9 Performance of model in dataset 3 (peptides tested positive by tetramer staining)

Serial number	HLA allele	Peptide	PB	IN	MHCBN				Pathogen	Peptide source	References
					BI	RB	CTL	RCTL			
1	A*0201	VLGPISGHV	−741.6	FN	Y	H	Y	?	HCMV	pp65 (14–22)	[37]
2	A*0201	MLNIPSINV	141.19	TP	Y	H	Y	?	HCMV	pp65 (120–128)	[37]
3	A*0201	ILAGYGAGV	−167.45	FN	Y	M	N	?	HCV	NS4 (1851–1859)	[38]
4	A*0201	YLNKIQNSL	1123.41	TP	Y	H	?	?	<i>Plasmodium falciparum</i>	CS (327–335)	[39]
5	A*0201	YLKKIKNSL	1801.47	TP	Y	M	?	?	<i>Plasmodium falciparum</i>	CS (334–342) 7G8 isolate	[39]
6	A*0201	YLKKIKNSI	1978.02	TP	?	?	?	?	<i>Plasmodium falciparum</i>	CS 7G8 isolate	[39]
7	A*0201	YLKKIQNSL	1526.57	TP	Y	H	?	?	<i>Plasmodium falciparum</i>	CS (327–335) WEL, It2G1, T9-101, T4R	[39]
8	A*0201	YLKTIQNSL	1526.57	TP	Y	?	?	?	<i>Plasmodium falciparum</i>	CS (327–335) LE5 isolate	[39]
9	A*0201	YLKTIKNSL	1801.47	TP	?	?	?	?	<i>Plasmodium falciparum</i>	CS (327–335) T9-98, 366a, 399,406,419	[39]
10	A*0201	YLQKIQNSL	1379.21	TP	?	?	?	?	<i>Plasmodium falciparum</i>	CS (327–335) 366b isolate	[39]
11	A*0201	YLQKIKNSL	1654.11	TP	?	?	?	?	<i>Plasmodium falciparum</i>	CS (327–335) 406, 419 isolates	[39]
12	A*0201	YLNTIQNSL	1123.41	TP	?	?	?	?	<i>Plasmodium falciparum</i>	CS (327–335) 427 isolate	[39]
13	A*0201	YLQKIRNSL	1705.06	TP	?	?	?	?	<i>Plasmodium falciparum</i>	CS (327–335) 366c isolate	[39]
14	A*0201(A2)	CINGVCWTV	654.95	TP	Y	?	Y	?	HCV	NS3 (1073–1081)	[40,41]
15	A*0201(A2)	ALYDVVTKL	749.08	TP	Y	?	Y	?	HCV	NS5B (2594–2602)	[40,41]
16	A*6802	ETAYFILKL	−1566.39	FN	Y	?	Y	H	HIV	pol (744–752)	[42]
17	A2	FMVFLQTHI	1190.78	TP	Y	H	?	?	EBV	BZLF-1	[43]
18	A2	VLQWASLAV	308.12	TP	Y	?	Y	?	EBV	gp350	[43]
19	A2	SLVIVTTFV	178.68	TP	Y	?	Y	?	EBV	gp85SV	[43]
20	A2	LMIIPLINV	401.59	TP	Y	?	Y	?	EBV	gp85LV	[43]
21	A2	TLFIGSHVV	295.08	TP	Y	?	Y	?	EBV	gp85TV	[43]
22	A2	YVLDHLIVV	1343.23	TP	?	?	?	?	EBV	BRLF-1 (109–117)	[44]
23	A2	GLCTLVAML	−275.63	FN	Y	H	Y	?	EBV	BMLF-1 (190–197)	[44,45]
24	A2	WLSLLVPFV	167.27	TP	Y	?	Y	?	HBV	envelope protein (335–343)	[46,47]
25	A2(01)	RLNMFTPYI	220.71	TP	Y	?	Y	?	<i>Cblamydia trachomatis</i>	MOMP (258–266)	[48]
26	A2(01)	NMFTPYIGV	186.48	TP	Y	?	Y	?	<i>Cblamydia trachomatis</i>	MOMP (260–268)	[48]
27	A2(01)	NLVPMVATV	397.35	TP	Y	H	Y	?	HCMV	pp65 (495–503)	[44, 49, 50]
28	A2(0201)	SLYNTVATL	415.61	TP	Y	H	Y	?	HIV	gag p17 (77–85)	[51, 52, 53, 54]
29	A2(0202)	ILKEPVHGV	578.75	TP	Y	M	?	?	HIV	RT (9–17)	[51, 55, 56, 57]
30	A2(0202)	YTAFTIPSI	−276.38	FN	Y	H	?	?	HIV	pol (476–484)	[52, 53]
31	A3	RVRAYTYSK	244.48	TP	?	?	?	?	EBV	BRLF-1	[43]
32	B*0702	TPRVTGGGA	335.27	TP	?	?	?	?	HCMV	pp65 (417–425)	[50]
33	B*0801	DANDIYRIF	−1261.3	FN	?	?	?	?	HCMV	pp65	[50]
34	B*3501	HPNIEEVAL	−1015.99	FN	Y	H	Y	H	HCV	NS3 (1359–1367)	[58]
35	B*3501	DPNPQEVVL	−1735.42	FN	Y	?	Y	H	HIV	env (77–85)	[59]
36	B*5801	IVLPEKDSW	288.89	TP	?	?	?	?	HIV	pol (244–252)	[42, 55]
37	B14	ERYLRDQQL	1574.94	TP	Y	?	Y	?	HIV	gp41(589–597)	[60]
38	B14	DLNMMLNIV	1101.96	FN	Y	?	Y	H	HIV	p24(51–59)	[60]
39	B18	YPLTFGWCF	496.22	TP	?	?	?	?	HIV	nef (139–147)	[60]
40	B27(05)	YRLLLTRVL	1376.78	TP	Y	?	Y	?	<i>Cblamydia trachomatis</i>	ygeD	[61]
41	B27(05)	NRELIQQEL	960.49	TP	Y	?	Y	?	<i>Cblamydia trachomatis</i>	CT (847–855)	[61]
42	B35	IPSINVHHY	906.54	TP	Y	?	Y	?	HCMV	pp65 (123–131)	[50]
43	B53	TPQDLNMML	−458.2	FN	Y	?	Y	?	HIV	gag(180–188)	[62]

TABLE 9 Performance of model in dataset 3 (peptides tested positive by tetramer staining) (*Continued*)

Serial number	HLA allele	Peptide	PB	IN	MHCBN				Pathogen	Peptide source	References
					BI	RB	CTL	RCTL			
44	B53(5301)	TPYDINQML	1150.85	TP	Y	?	Y	?	HIV	HIV-2	[62]
45	B7	LPCVLW/PVL	-1272.42	FN	?	?	?	?	EBV	BZLF-1	[43]
46	Cw4	KYRLKHLVW	2039.97	TP	Y	?	Y	?	HIV	gag p17 (28–36)	[42]

Abbreviations: BI = binding; CTL = cytotoxic T lymphocyte; EBV = Epstein-Barr virus; FN = false-negative value; HBV = hepatitis B virus; HCMV = human cytomegalovirus; HCV = hepatitis C virus; HIV = human immunodeficiency virus; HLA = human leukocyte antigen; IN = inference; MOMP = major outer membrane protein; PB = predicted binding; RB = relative binding; RCTL = relative CTL activity; TP = true-positive value.

a methodology to define virtual pockets and localize functions in these pockets.

The 38 HERPs reveal a mean change in accessibility of at least $1 \text{ \AA}^2$ during MHCp complex formation and/or a minimum of 1 inter-residue contact between MHC and peptide at a cutoff $4 \text{ \AA}$ (Table 2). A caveat in this

approach is that all residue positions that exhibit a mean change in accessibility of greater than $0 \text{ \AA}^2$ and less than $1 \text{ \AA}^2$ are neglected. We believe that these residue positions may not significantly affect the MHCp binding calculations due to their negligible contact with the peptide (Figure 1B). It should also be noted that the

TABLE 10 Performance of model in dataset 4 (HLA-A binding peptides derived from MHCBN)

Serial number	HLA allele	Number of peptides	TP	TN	FP	FN	SP (%)	SE (%)	AC (%)
1	A*0101	9	4	0	0	5	—	44	44
2	A*0201	794	424	18	11	341	62	55	56
3	A*0202	78	35	1	2	40	33	47	46
4	A*0303	65	21	0	0	44	—	32	32
5	A*0204	38	20	0	0	18	—	53	53
6	A*0205	23	5	0	0	18	—	22	22
7	A*0206	83	50	1	0	32	100	61	61
8	A*0207	19	8	0	0	11	—	42	42
9	A*0209	5	2	0	0	3	—	40	40
10	A*0210	3	1	0	0	2	—	33	33
11	A*0211	4	2	0	0	2	—	50	50
12	A*0214	8	8	0	0	0	—	100	100
13	A*0217	1	1	0	0	0	—	100	100
14	A*0301	66	34	0	0	32	—	52	52
15	A*0302	2	1	0	0	1	—	50	50
16	A*1101	77	13	0	0	64	—	17	17
17	A*2402	118	83	0	1	34	0	71	70
18	A*2601	4	3	0	0	1	—	75	75
19	A*2602	2	1	0	0	1	—	50	50
20	A*2603	2	0	0	0	2	—	—	—
21	A*2902	20	5	0	0	15	—	25	25
22	A*3001	1	1	0	0	0	—	100	100
23	A*3002	10	7	0	0	3	—	70	70
24	A*3003	1	1	0	0	0	—	100	100
25	A*3101	19	3	0	0	16	—	16	16
26	A*3201	1	1	0	0	0	—	100	100
27	A*3301	33	9	0	0	24	—	27	27
28	A*3402	1	0	0	0	1	—	0	—
29	A*6601	10	4	0	0	6	—	40	40
30	A*6801	63	13	0	0	50	—	21	21
31	A*6802	60	10	9	7	34	56	23	32
32	A*6901	3	0	0	0	3	—	—	—
33	A*7401	1	0	0	0	1	—	—	—
Total		1624	770	29	21	804	58	0.49	49

Abbreviations: AC = accuracy; FN = false-negative value; FP = false-positive value; HLA = human leukocyte antigen; MHCBN = major histocompatibility complex binding and nonbinding peptides database; SE = sensitivity; SP = specificity; TN = true-negative value; TP = true-positive value.

TABLE 11 Performance of model in data set 4 (HLA-B binding peptides derived from MHCBN)

Serial Number	HLA allele	Number of peptides	TP	TN	FP	FN	SE (%)	AC (%)
1	B*0702	72	32	0	0	40	44	44
2	B*0801	15	12	0	0	3	80	80
3	B*0802	4	2	0	0	2	50	50
4	B*1501	11	7	0	0	4	64	64
5	B*1502	4	2	0	0	2	50	50
6	B*1503	3	2	0	0	1	67	67
7	B*1508	2	1	0	0	1	50	50
8	B*1509	9	3	0	0	6	33	33
9	B*1510	9	3	0	0	6	33	33
10	B*1512	1	0	0	0	1	0	—
11	B*1513	6	4	0	0	2	67	67
12	B*1516	7	4	0	0	3	57	57
13	B*1517	4	2	0	0	2	50	50
14	B*1518	4	1	0	0	3	25	25
15	B*2701	6	2	0	0	4	33	33
16	B*2702	13	5	0	0	8	38	38
17	B*2703	32	19	0	0	13	59	59
18	B*2704	17	12	0	0	5	71	71
20	B*2705	108	77	0	0	31	71	71
21	B*2706	23	16	0	0	7	70	70
22	B*2707	4	2	0	0	2	50	50
23	B*2709	4	2	0	0	2	50	50
24	B*3501	155	91	0	0	64	59	59
25	B*3503	1	1	0	0	0	100	100
26	B*3701	4	2	0	0	2	50	50
27	B*3801	6	4	0	0	2	67	67
28	B*3901	8	8	0	0	0	100	100
29	B*3909	3	1	0	0	2	33	33
30	B*4001	8	6	0	0	2	75	75
31	B*4002	16	6	0	0	10	38	38
32	B*4006	4	3	0	0	1	75	75
33	B*4201	1	1	0	0	0	100	100
34	B*4402	8	3	0	0	5	38	38
35	B*4403	7	1	0	0	6	14	14
36	B*4501	1	0	0	0	1	0	—
37	B*4601	5	3	0	0	2	60	60
38	B*5101	48	38	0	0	10	79	79
39	B*5102	32	26	0	0	6	81	81
40	B*5103	30	25	0	0	5	83	83
41	B*5201	6	5	0	0	1	83	83
42	B*5301	59	42	0	0	17	71	71
43	B*5401	57	41	0	0	16	72	72
44	B*5501	2	0	0	0	2	0	—
45	B*5502	5	1	0	0	4	20	20
46	B*5601	3	0	0	0	3	0	—
47	B*5701	10	6	0	0	4	60	60
48	B*5702	4	3	0	0	1	75	75
49	B*5801	11	7	0	0	4	64	64
50	B*5802	3	3	0	0	0	100	100
51	B*7301	2	1	0	0	1	50	50
52	B*8101	2	1	0	0	1	50	50
Total		859	539	0	0	320	63	63

Abbreviations: AC = accuracy; FN = false-negative values; FP = false-positive values; HLA = human leukocyte antigen; MHCBN = major histocompatibility complex binding and nonbinding peptides database; SE = sensitivity; SP = specificity; TN = true-negative value; TP = true-positive value.

standard deviation about the mean change in accessibility is large at the HERPs which is caused by the polymorphism of the MHC molecules (Figure 1C). The

fourth residue position in the peptide is neglected in the MHCp binding calculations because of its minimum involvement in binding irrespective of the amino acid

type at that position (Figures 2A and 2B). The standard deviation about the mean change at each peptide position is illustrated in Figure 2C. Interestingly, the deviation at peptide residue positions 5, 6, and 7 are large.

The next challenging step is to map the HERPs to the corresponding peptide residue positions so as to define peptide residue position specific virtual pockets in the MHC. Inter-residue contacts are used to map the peptide residue positions with the virtual binding pockets of the MHC molecules (Table 3 and Table 4). For example, the first peptide residue position has at least one inter-residue contact with positions 5, 7, 59, 62, 63, 66, 159, 163, 167, and 171 of the MHC in the dataset (Table 4). These HERPs form the virtual pockets for binding to the first peptide residue. It should be noted that some of these HERPs are also shared with other neighboring virtual pockets. Thus, nine virtual pockets corresponding to each peptide residue positions are defined for class I HLA molecules. Due to high sequence homology among HLA alleles, the HERPs are extrapolated to other alleles and virtual binding pockets are defined for each known allele.

The peptide residue position specific virtual binding pockets are defined to each known allele. Once this mapping is established, it is equally important to estimate the binding affinity of a 9 mer peptide with these virtual binding pockets. In this method, we describe the binding affinity as a sum of compatibility measure of each peptide residue with its corresponding virtual pocket. The compatibility is calculated using a property matrix called the Q matrix [102]. The Q matrix is a high dimensional 20×20 property matrix developed using 237 physicochemical properties of amino acids. When 9 residue long peptides bind to MHC molecules, each residue in the peptide interacts with a virtual pocket. The overall binding between peptide and MHC is a cumulative measure of the physical and chemical compatibility between each residue in the peptide and the residues forming the virtual pockets. The predicted binding score is the sum of compatibility scores of residues at each peptide residue positions and the corresponding virtual pockets. In this procedure, we consider those peptides with negative scores as nonbinders and those with positive scores as binders.

The efficiency of our model is evaluated using four different binding datasets, namely DS1, DS2, DS3, and DS4. The characteristic features of the datasets are described in the materials and methods section. A large number of manually collected (laborious process) peptides with known binding/binding-related information are included in DS2 and DS4. DS4 is one of the largest repositories of MHCp binding data (developed from several primary repositories and disparate sources) available in the public domain. It should be noted that the data in

DS4 is less homogenous and the information noise (data quality, consistency, experimental procedure to get data, annotation) is quantifiably high. However, these datasets (DS1–DS4) provide the basis to validate the proposed model in a fair and transparent manner. DS1 contains 29 MHCp structures and the model performs well with sensitivity (SE) of $\sim 69\%$ in this dataset (Table 5). The 9 FNs observed in this validation is probably introduced by the noise in the Q matrix used for scoring as Q matrix is a similarity matrix. Improvements in the Q matrix by fine tuning the interacting pairs will remove false negatives and improve sensitivity significantly. Work is in progress to reevaluate and accordingly modify the Q matrix for further developments. In our previous study, we evaluated the application of 3D side-chain packing procedures for MHCp binding predictions. Encouraging results are obtained by the study and the main drawback is the cumbersome modeling procedure involving energy calculations. The same dataset is used evaluate the performance of the new predictive model and its characteristics features are described in DS2. DS2 contains 102 peptides (Table 6) with known MHCp binding information (mostly from competitive binding assay). The model produces an overall sensitivity (SE) of 67% and specificity (SP) of 52% with an average accuracy (AC) of 56% in DS2 (Table 8). The PPV is 81% and NPV is 33%. These results are comparable to those (results recalculated only for 9 mer peptides) obtained from the side-chain packing procedures described elsewhere [86]. The PPV in the new model is 81% and the NPV is only 33%. The low NPV is because of high FN (36) and low TN (18) produced by the model. The high FN is introduced by the noise in the scoring Q matrix. The low TN is because of the nature of DS2 which contains only 27 nonbinders (only 26%). NPV can be improved by the re-evaluation of the scoring matrix and incorporation of more nonbinders in the validation set. It should be noted that the amount of known nonbinders for a given allele is significantly low as knowledge on non-binders is of less importance in inferring immunologically relevant phenomena. However, nonbinding information is critical in the development and assessment of mathematical models. Despite the limitations stated, the current model is superior to the previous procedure because of easy implementation.

It is known that MHCp binding does not fully guarantee immunogenicity [65]. In order to establish a better understanding of MHCp binding and immune functions, the model is assessed using 46 peptides (9 residues long) that are documented as positive to tetramer staining in PBMC. These peptides are positive to tetramer staining in PBMC obtained from peptide specific infected patients (DS3). It should be noted that this does not contain peptides negative to tetramer staining. The

model performs with a sensitivity and accuracy of 78% in DS3 (Table 8). The true measure of PPV cannot be calculated for DS3 as this dataset lacks non-binders or peptides negative to tetramer staining. FN in this dataset is 10 and, as discussed in the previous section, the FN is because of noise introduced in the scoring matrix. We would like to state that this does not provide any additional information other than they are tetramer positive and our model performance in this dataset is good (SE = 78%). We also evaluated the model using a relatively large number of peptides (known binding information) described in DS4. DS4 contains 1624 HLA-A binding peptides (Table 10) and 859 HLA-B binding peptides (Table 11). The SE of the model is 49% and 63% for HLA-A and HLA-B binding peptides, respectively. A specificity of 58% is obtained for HLA-A binding peptides. The PPV is 97% for HLA-A binding peptides. The FN is 804 in this dataset and this large number of FN is because of noise in the scoring matrix. As in other datasets, the number of nonbinders in DS4 is very less (only 50 HLA-A specific nonbinders). The number of nonbinders is negligible compared to the binders (1574 HLA-A peptides and 859 HLA-B peptides) in DS4. Therefore, the NPV calculated for DS4 (3%) is not a real measure of the model's negative predictive value. The average specificity for HLA-A peptides in DS4 is 58%.

The model produced an average accuracy of 60% in all the four datasets used for validation. The average PPV is 89% for the model with good sensitivity (~50%–73%) and specificity (52%–58%) values. Although, this model produced only moderate predictive accuracy, the approach is novel, simple, and superior to other models because of its potential application to any known MHC molecule whose sequence is clearly defined. The predictive power of the described model can be improved in future by the careful reexamination of HERPs for the definition of virtual pockets.

CONCLUSION

The current challenge in synthetic vaccine design is the development of a methodology to identify and test short antigen peptides as potential T-cell epitopes. Many allele specific data driven statistical models and 3D structure based procedures are available for MHCp binding prediction and calculations. Data driven statistical methods are generally available for A*0201, A*1101, A*0301, B*0801, DR1, and DR4. The 3D structure based procedures are generally adopted for any given allele because the prediction is based on molecular energy functions developed using first principles. Structural procedures utilizing structure prediction by energy minimization followed by protein-ligand interaction calculations are computationally intensive. Moreover, robust protein-li-

gand interaction functions are not currently available. Our method circumvents these problems because this model is simple, fast, and robust with easy implementation.

The method described in this article is based on the definition of virtual binding pockets, mapping of virtual pockets to position specific peptide residue anchors, and estimation of peptide residue–virtual binding pocket compatibility. Virtual pockets are defined using information gleaned from eight unique HLA alleles and the mapping of virtual pocket to position specific residue anchors is done using the 29 MHCp structures analyzed in this study. The virtual pockets defined by this procedure is extrapolated to other class I alleles because HLA molecules are highly homologous, and homologous proteins have conserved backbone trace. The peptide residue–virtual binding pocket compatibility is estimated using the Q matrix. The average PPV of the model is 89%, whereas the average NPV is only 18%. The low NPV can be improved using (1) redefinition of virtual pockets as and when more MHCp structures are available; (2) a modified Q matrix; (3) validation of the model with a dataset containing more nonbinders; (4) performance of blindfold prediction and validation. We plan to improve the NPV of the model by addressing the above issues. Despite the limitations set by the model, the method described in this article is novel, generic, and simple. Although the model is validated for class I alleles in this report, extension of the model for class II alleles is trivial.

REFERENCES

1. Sliz P, Michielin O, Cerottini JC, Luescher I, Romero P, Karplus M, Wiley DC: Crystal structures of two closely related but antigenically distinct HLA-A2/melanocyte-melanoma tumor-antigen peptide complexes. *J Immunol* 167:3276, 2001.
2. Zhao R, Loftus DJ, Appella E, Collins EJ: Structural evidence of T cell xeno-reactivity in the absence of molecular mimicry. *J Exp Med* 189:359, 1999.
3. Buslepp J, Zhao R, Donnini D, Loftus D, Saad M, Appella E, Collins EJ: T cell activity correlates with oligomeric peptide–major histocompatibility complex binding on T cell surface. *J Biol Chem* 276:47320, 2001.
4. Kirksey TJ, Pogue-Caley RR, Frelinger JA, Collins EJ: The structural basis for the increased immunogenicity of two HIV-reverse transcriptase peptide variant/class-I major histocompatibility complexes. *J Biol Chem* 274: 37259, 1999.
5. 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* 75:693, 1993.
6. Gao GF, Tormo J, Gerth UC, Wyer JR, McMichael AJ, Stuart DI, Bell JI, Jones EY, Jakobsen BK: Crystal structure of the complex between human CD8 α (alpha) and HLA-A2. *Nature* 387:630, 1997.
 7. Ding YH, Baker BM, Garboczi DN, Biddison WE, Wiley DC: Four A6-TCR/peptide/HLA-A2 structures that generate very different T cell signals are nearly identical. *Immunity* 11:45, 1999.
 8. Garboczi DN, Ghosh P, Utz U, Fan QR, Biddison WE, Wiley DC: Structure of the complex between human T-cell receptor, viral peptide and HLA-A2. *Nature* 384:134, 1996.
 9. Ding YH, Smith KJ, Garboczi DN, Utz U, Biddison WE, Wiley DC: Two human T cell receptors bind in a similar diagonal mode to the HLA-A2/tax peptide complex using different TCR amino acids. *Immunity* 8:403, 1998.
 10. 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* 164:6398, 2000.
 11. Gewurz BE, Gaudet R, Tortorella D, Wang EW, Ploegh HL, Wiley DC: Antigen presentation subverted: structure of the human cytomegalovirus protein Us2 bound to the class I molecule HLA-A2. *Proc Nat Acad Sci USA* 98:6794, 2001.
 12. Madden DR, Gorga JC, Strominger JL, Wiley DC: The three-dimensional structure of HLA-B27 at 2.1 Å resolution suggests a general mechanism for tight peptide binding to MHC. *Cell* 70:1035, 1992.
 13. Hulsmeier M, Hillig RC, Volz A, Ruhl M, Schroder W, Saenger W, Ziegler A, Uchanska-Ziegler B: HLA-B27 subtypes differentially associated with disease exhibit subtle structural alterations. *J Biol Chem* 277:47844, 2002.
 14. Menssen R, Orth P, Ziegler A, Saenger W: Decamer-like conformation of a nona-peptide bound to HLA-B*3501 due to non-standard positioning of the C terminus. *J Mol Biol* 285:645, 1999.
 15. Maenaka K, Maenaka T, Tomiyama H, Takiguchi M, Stuart DI, Jones EY: Non-standard peptide binding revealed by crystal structures of HLA-B*5101 complexed with HIV immunodominant epitopes. *J Immunol* 165:3260, 2000.
 16. Smith KJ, Reid SW, Harlos K, McMichael AJ, Stuart DI, Bell JI, Jones EY: Bound water structure and polymorphic amino acids act together to allow the binding of different peptides to MHC class I HLA-B53. *Immunity* 4:215, 1996.
 17. Boyington JC, Motyka SA, Schuck P, Brooks AG, Sun PD: Structure of an NK cell immunoglobulin-like receptor in complex with its class-I MHC ligand. *Nature* 405:537, 2000.
 18. Fan QR, Long EO, Wiley DC: Crystal structure of the human natural killer cell inhibitory receptor KIR2DL1-HLA-Cw4 complex. *Nat Immunol* 2:452, 2001.
 19. Alexander J, Del Guercio MF, Fikes JD, Chesnut RW, Chisari FV, Chang KM, Appella E, Sette A: Recognition of a novel naturally processed, A2 restricted, HCV-NS4 epitope triggers IFN-gamma release in absence of detectable cytopathicity. *Hum Immunol* 59:776, 1998.
 20. Kawashima I, Hudson SJ, Tsai V, Southwood S, Takesako K, Appella E, Sette A, Celis E: Multi-epitope approach for immunotherapy for cancer: identification of several CTL epitopes from various tumor-associated antigens expressed on solid epithelial tumors. *Hum Immunol* 59:1, 1998.
 21. Chang KM, Gruener NH, Southwood S, Sidney J, Pape GR, Chisari FV, Sette A: Identification of HLA-A3 and HLA-B7-restricted CTL response to hepatitis C virus in patients with acute and chronic hepatitis C. *J Immunol* 162:1156, 1999.
 22. Bashin M, Singh H, Raghava PS: MHCBN: a comprehensive database of MHC binding and non-binding peptides. *Bioinformatics* 19:665, 2003.
 23. Blythe MJ, Doytchinova IA, Flower DR: JenPep: a database of quantitative functional peptide data for immunology. *Bioinformatics* 18:434, 2002.
 24. Joshi RV, Zarutskie JA, Stern LJ: A three-step kinetic mechanism for peptide binding to MHC class II proteins. *Biochemistry* 39:3751, 2000.
 25. Kasson PM, Rabinowitz JD, Schmitt L, Davis MM, McConnell HM: Kinetics of peptide binding to the class II MHC protein I-Ek. *Biochemistry* 39:1048, 2000.
 26. Natarajan SK, Assadi M, Sadegh-Nasseri S: Stable peptide binding to MHC class II molecule is rapid and is determined by a receptive conformation shaped by prior association with low affinity peptides. *J Immunol* 162:4030, 1999.
 27. Springer S, Doring K, Skipper JC, Townsend AR, Cerundolo V: Fast association rates suggest a conformational change in the MHC class I molecule H-2Db upon peptide binding. *Biochemistry* 37:3001, 1998.
 28. Reiser JB, Gregoire C, Darnault C, Mosser T, Guimezanes A, Schmitt-Verhulst AM, Fontecilla-Camps JC, Mazza G, Malissen B, Housset D: A T-cell receptor CDR3 beta loop undergoes conformational changes of unprecedented magnitude upon binding to a peptide/MHC class I complex. *Immunity* 16:345, 2002.
 29. Reiser JB, Darnault C, Guimezanes A, Gregoire C, Mosser T, Schmitt-Verhulst AM, Fontecilla-Camps JC, Malissen B, Housset D, Mazza G: Crystal structure of a T cell receptor bound to an allogeneic MHC molecule. *Nat Immunol* 1:291, 2000.
 30. Degano M, Garcia KC, Apostolopoulos V, Rudolph MG, Teyton L, Wilson IA: A functional hot spot for antigen

- recognition in a super agonist TCR/MHC complex. *Immunity* 12:251, 2000.
31. Luz JG, Huang M, Garcia KC, Rudolph MG, Apostolopoulos V, Teyton L, Wilson IA: Structural comparison of allogeneic and syngeneic T cell receptor-peptide major histocompatibility complex complexes: a buried alloreactive mutation subtly alters peptide presentation substantially increasing V(beta) Interactions. *J Exp Med* 195: 1175, 2002.
 32. Garcia KC, Degano M, Pease LR, Huang M, Peterson PA, Teyton L, Wilson IA: Structural basis of plasticity in T cell receptor recognition of a self peptide-MHC antigen. *Science* 279:1166, 1998.
 33. Hennecke J, Carfi A, Wiley DC: Structure of a covalently stabilized complex of a human alpha beta T-cell receptor, influenza HA peptide and MHC class II molecule, HLA-DR1. *EMBO J* 19:5611, 2000.
 34. Hennecke J, Wiley DC: Structure of a complex of the human alpha/beta T cell receptor (TCR) Ha17, Influenza hemagglutinin peptide, and major histocompatibility complex class II molecule: HLA-DR4 (DRA*0101 and DRB*10401). Insight into TCR cross-restriction and alloreactivity. *J Exp Med* 195:571, 2002.
 35. Reinherz EL, Tan K, Tang L, Kern P, Liu J, Xiong Y, Hussey RE, Smolyar A, Hare B, Zhang R, Joachimiak A, Chang HC, Wagner G, Wang J: The crystal structure of a T cell receptor in complex with peptide and MHC class II. *Science* 286:1913, 1999.
 36. Altman JD, Moss PAH, Goulder PJR, Barouch DH, McHeyzer-Williams MG, Bell JI, McMichael AJ, Davis MM: Phenotypic analysis of antigen specific T lymphocytes. *Science* 274:94, 1996.
 37. Aubert G, Hassan-Walker AF, Madrigal JA, Emery VC, Morre C, Grace S, Koh MB, Potter M, Prentice HG, Dodi IA, Travers PJ: Cytomegalovirus specific cellular immune responses and viremia in recipients of allogeneic stem cell transplants. *J Infect Dis* 184:955, 2001.
 38. Prezzi C, Casciaro MA, Francavilla V, Schiaffella E, Finocchi L, Chircu LV, Bruno G, Sette A, Abrignani S, Barnaba V: Virus-specific CD8(+) T cells with type 1 or type 2 cytokine profile are related to different disease activity in chronic hepatitis C virus infection. *Eur J Immunol* 31:894, 2001.
 39. Bonelo A, Valmori D, Triponez F, Tiercy JM, Mentha G, Oberholzer J, Champagne P, Romero JF, Esposito F, Nebie I, Barbey C, Romero P, Herrera S, Corradin G, Lopez JA: Generation and characterization of malaria specific human CD8(+) lymphocyte clones: effect of natural polymorphism on T cell recognition and endogenous cognate antigen presentation by liver cells. *Eur J Immunol* 30:3079, 2000.
 40. Grabowska AM, Lechner F, Klenerman P, Tighe PJ, Ryder S, Ball JK, Thomson BJ, Irving WL, Robin RA: Direct ex vivo comparison of the breadth and specificity of the T cells in the liver and peripheral blood of patients with chronic HCV infection. *Eur J Immunol* 31:2388, 2001.
 41. Lechner F, Wong DK, Dunbar PR, Chapman R, Chung RT, Dohrenwend P, Robbins G, Phillips R, Klenerman P, Walker BD: Antibodies to the junctional adhesion molecule cause disruption of endothelial cells and do not prevent leukocyte influx into the meninges after viral or bacterial infection. *J Exp Med* 191:1499, 2000.
 42. Appay V, Nixon DF, Donahoe SM, Gillespie GM, Dong T, King A, Ogg GS, Spiegel HM, Conlon C, Spina CA, Havlir DV, Richman DD, Waters A, Easterbrook P, McMichael AJ, Rowland-Jones SL: HIV specific CD8(+) T cells produce antiviral cytokines but are impaired in cytolytic function. *J Exp Med* 192:63, 2000.
 43. Catalina MD, Sullivan JL, Bak KR, Luzuriaga K: Differential evolution and stability of epitope specific CD8(+) T cell responses in EBV infection. *J Immunol* 167: 4450, 2001.
 44. Faint JM, Annels NE, Curnow SJ, Shields P, Pilling D, Hislop AD, Wu L, Akbar AN, Buckley CD, Moss PA, Adams DH, Rickinson AB, Salmon M: Memory T cells constitute a subset of the human CD8+ CD45 RA+ pool with distinct phenotypic and migratory characteristics. *J Immunol* 167:212, 2001.
 45. Hislop AD, Annels NE, Gudgeon NH, Leese AM, Rickinson AB: Epitope-specific evolution of human CD8(+) T cell responses from primary to persistent phases of Epstein Barr virus infection. *J Exp Med* 195:893, 2002.
 46. Boni C, Penna A, Ogg GS, Bertoletti A, Pilli M, Cavallo C, Cavalli A, Urbani S, Boehme R, Panebianco R, Fiaccadori F, Ferrari C: Lamivudine treatment can overcome cytotoxic T cell hypo-responsiveness in chronic hepatitis B: new perspectives for immune therapy. *Hepatology* 33:963, 2001.
 47. Reignat S, Webster GJ, Brown D, Ogg GS, King A, Seneviratne SL, Dusheiko G, Williams R, Maini MK, Bertoletti A: Escaping high viral load exhaustion: CD8 cells with altered tetramer binding in chronic hepatitis B virus infection. *J Exp Med* 195:1089, 2002.
 48. Kim SK, Devine L, Angevine M, DeMars R, Kavathas PB: Direct detection and magnetic isolation of *Chlamydia trachomatis* major outer membrane protein-specific CD8+ CTLs with HLA class I tetramers. *J Immunol* 165:7285, 2000.
 49. Vaz Santiago J, Lule J, Rohrlach P, Jacquier C, Gibert N, Le Roy E, Betbeder D, Davignon JL, Davrinche C: Ex vivo stimulation and expansion of both CD4(+) and CD8(+) T cells from peripheral blood mononuclear cells of human cytomegalovirus sero-positive blood donors by using a soluble recombinant chimeric protein, IE1-pp65. *J Virol* 75:7840, 2001.
 50. Hassan-Walker AF, Vargas Cuero AL, Mattes FM, Klenerman P, Lechner F, Burroughs AK, Griffiths PD, Phillips RE, Emery VC: CD8+ cytotoxic lymphocyte responses against cytomegalovirus after liver transplan-

- tation: correlation with time from transplant to receipt of tacrolimus. *J Infect Dis* 183:835, 2001.
51. Ogg GS, Kostense S, Klein MR, Jurriaans S, Hamann D, McMichael AJ, Miedema F: Longitudinal phenotypic analysis of human immunodeficiency virus type-1 specific cytotoxic T lymphocytes: correlation with disease progression. *J Virol* 73:9153, 1999.
 52. Shankar P, Russo M, Harnisch B, Patterson M, Skolnik P, Lieberman J: Longitudinal phenotypic analysis of human immunodeficiency virus type-1 specific cytotoxic T lymphocytes: correlation with disease progression. *Blood* 96:3094, 2000.
 53. Trimble LA, Shankar P, Patterson M, Daily JP, Lieberman J: Human immunodeficiency virus specific circulating CD8(+) T lymphocytes have down modulated CD3 zeta and CD28, key signaling molecules for T-cell activation. *J Virol* 74:7320, 2000.
 54. Migueles SA, Sabbaghian MS, Shupert WL, Bettinotti MP, Marincola FM, Martino L, Hallahan CW, Selig SM, Schwartz D, Sullivan J, Connors M: HLA-B*5701 is highly associated with restriction of virus replication in a subgroup of HIV infected long term non-progressors. *Proc Natl Acad Sci USA* 97:2709, 2000.
 55. Kostense S, Ogg GS, Manting EH, Gillespie G, Joling J, Vandenberghe K, Veenhof EZ, van Baarle D, Jurriaans S, Klein MR, Miedema F: High viral burden in the presence of major HIV specific CD8(+) T cell expansions: evidence for impaired CTL effector function. *Eur J Immunol* 31:677, 2001.
 56. Seth A, Markee J, Hoering A, Sevin A, Sabath DE, Schmitz JE, Kuroda MJ, Lifton MA, Hirsch MS, Collier AC, Letvin NL, McElrath MJ: Alterations in T cell phenotype and human immunodeficiency virus type-1 specific cytotoxicity after potent anti-retroviral therapy. *J Infect Dis* 183:722, 2001.
 57. Rinaldo CR Jr, Huang XL, Fan Z, Margolick JB, Borowski L, Hoji A, Kalinyak C, McMahon DK, Riddler SA, Hildebrand WH, Day RB, Mellors JW: Anti-human immunodeficiency virus type 1 (HIV-1) CD8(+) T-lymphocyte reactivity during combination anti-retroviral therapy in HIV-1 infected patients with advanced immunodeficiency. *J Virol* 74:4127, 2000.
 58. Sobao Y, Tomiyama H, Nakamura S, Sekihara H, Tanaka K, Takiguchi M: Visual demonstration of hepatitis C virus specific memory CD8(+) T-cell expansion in patients with acute hepatitis C. *Hepatology* 33:287, 2001.
 59. Ogg GS, Jin X, Bonhoeffer S, Dunbar PR, Nowak MA, Monard S, Segal JP, Cao Y, Rowland-Jones SL, Cerundolo V, Hurley A, Markowitz M, Ho DD, Nixon DF, McMichael AJ: Quantitation of HIV-1 specific cytotoxic T lymphocytes and plasma load of viral RNA. *Science* 279:2103, 1998.
 60. Kaul R, Rowland-Jones SL, Gillespie G, Kimani J, Dong T, Kiama P, Simonsen JN, Bwayo JJ, McMichael AJ, Plummer FA: Gonococcal cervicitis is associated with reduced systemic CD8+ T cell responses in human immunodeficiency virus type-1 infected and exposed: uninfected sex workers. *J Infect Dis* 185:1525, 2002.
 61. Kuon W, Holzhtutter HG, Appel H, Grolms M, Kollnberger S, Traeder A, Henklein P, Weiss E, Thiel A, Lauster R, Bowness P, Radbruch A, Kloetzel PM, Sieper J: Identification of HLA-B27 restricted peptides from the *Chlamydia trachomatis* proteome with possible relevance to HLA-B27 associated diseases. *J Immunol* 167: 4738, 2001.
 62. Dorrell L, Willcox BE, Jones EY, Gillespie G, Njai H, Sabally S, Jaye A, DeGleria K, Rostron T, Lepin E, McMichael A, Whittle H, Rowland-Jones S: Cytotoxic T lymphocytes recognize structurally diverse: clade-specific and cross-reactive peptides in human immunodeficiency virus type-1 gag through HLA-B53. *Eur J Immunol* 31:1747, 2001.
 63. Brunner KT, Mauel J, Cerottini JC, Chapuis B: Quantitative assay of the lytic action of immune lymphoid cells on ⁵¹Cr labelled allogeneic target cells in vitro: inhibition by isoantibody and by drugs. *Immunology* 14:181, 1968.
 64. Cerottini JC, Brunner KT: Cell-mediated cytotoxicity, allograft rejection, and tumor immunity. *Adv Immunol* 18:67, 1974.
 65. 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* 31:1391, 1994.
 66. Lawson TM, Man S, Wang EC, Williams S, Amos N, Gillespie GM, Moss PA, Borysiewicz LK: Functional differences between influenza A specific cytotoxic T lymphocyte clones expressing dominant and subdominant TCR. *Int Immunol* 13:1383, 2001.
 67. <http://www.ebi.ac.uk/imgt/hla>
 68. Robinson J, Waller MJ, Parham P, de Groot N, Bontrop R, Kennedy LJ, Stoehr P, Marsh SGE: IMGT/HLA and IMGT/MHC: sequence databases for the study of the major histocompatibility complex. *Nucleic Acids Res* 31:311, 2003.
 69. Hammer J, Takacs B, Sinigaglia F: Identification of a motif for HLA-DR1 binding peptides using M13 display libraries. *J Exp Med* 176:1007, 1992.
 70. Hammer J, Bono E, Gallazzi F, Belunis C, Nagy Z, Sinigaglia F: Precise prediction of major histocompatibility complex class II peptide interaction based on peptide side chain scanning. *J Exp Med* 180:2353, 1994.
 71. Sturniolo T, Bono E, Ding J, Raddrizzani L, Tuereci O, Sahin U: Generation of tissue-specific and promiscuous HLA ligand databases using DNA microarrays and virtual HLA class II matrices. *Nat Biotechnol* 17:555, 1999.
 72. Singh H, Raghava GPs: ProPred: prediction of HLA-DR binding sites. *Bioinformatics* 17:1236, 2001.
 73. Singh H, Raghava GPs: ProPred1: prediction of promiscuous

- cuous MHC class-I binding sites. *Bioinformatics* 19:1009, 2003.
74. 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* 152:163, 1994.
 75. Rammensee HG, Bachmann J, Emmerich NP, Bachor OA, Stevanovic S: SYFPEITHI: atabase for MHC ligands and peptide motifs. *Immunogenetics* 50:213, 1999.
 76. Guan PP, Doytchinova IA, Flower DR: HLA-A3 supermotif defined by quantitative structure-activity relationship analysis. *Protein Eng* 16:11, 2003.
 77. Doytchinova IA, Flower DR: Toward the quantitative prediction of T-cell epitopes: coMFA and coMSIA studies of peptides with affinity for the class-I MHC molecule HLA-A*0201. *J Med Chem* 44:3572, 2001.
 78. Doytchinova IA, Blythe MJ, Flower DR: Additive method for the prediction of protein-peptide binding affinity. Application to the MHC class I molecule HLA-A*0201. *J Proteome Res* 1:263, 2002.
 79. Brusic V, Rudy G, Honeyman MC, Hammer J, Harrison LC: Prediction of MHC class-II binding peptides using an evolutionary algorithm and artificial neural network. *Bioinformatics* 14:121, 1998.
 80. Honeyman MC, Brusic V, Stone N, Harrison LC: Neural network based prediction of candidate T-cell epitope. *Nat Biotechnol* 16:966, 1998.
 81. Mamitsuka H: Predicting peptides that bind to MHC molecules using supervised learning of hidden Markov models. *Proteins* 33:460, 1998.
 82. Dönnes P, Elofsson A: Prediction of MHC class I binding peptides, using SVMHC. *BMC Bioinformatics* 3:25, 2002.
 83. Schueler-Furman O, Elber R, Margalit H: Knowledge-based structure prediction of MHC class I bound peptides: a study of 23 complexes. *Fold Des* 3:549, 1998.
 84. 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 Sci* 9:1838, 2000.
 85. Rognan D, Lauemoller SL, Holm A, Buus S, Tschinke V: Predicting binding affinities of protein ligands from three-dimensional models: application to peptide binding to class I major histocompatibility proteins. *J Med Chem* 42:4650, 1999.
 86. Kangueane P, Sakharkar MK, Lim KS, Hao H, Lin K, Chee RE, Kolatkar PR: Knowledge-based grouping of modeled HLA peptide complexes. *Hum Immunol* 61:460, 2000.
 87. Lee C, McConnell HM: A general model of invariant chain association with class II major histocompatibility complex proteins. *Proc Natl Acad Sci USA* 92:8269, 1995.
 88. Altuvia Y, Schueler O, Margalit H: Ranking potential binding peptides to MHC molecules by a computational threading approach. *J Mol Biol* 249:244, 1995.
 89. Altuvia Y, Sette A, Sidney J, Southwood S, Margalit H: A structure-based algorithm to predict potential binding peptides to MHC molecules with hydrophobic binding pockets. *Hum Immunol* 58:1, 1997.
 90. Rognan D, Scapozza L, Folkers G, Daser A: Molecular dynamics simulation of MHC-peptide complexes as a tool for predicting potential T cell epitopes. *Biochemistry* 33:11476, 1994.
 91. Kangueane P, Sakharkar MK, Kolatkar PR, Ren EC: Towards the MHC-peptide combinatorics. *Hum Immunol* 62:539, 2001.
 92. Abagyan RA, Batalov S: Do aligned sequences share the same fold? *J Mol Biol* 273:355, 1997.
 93. Adrian P, Rajaseger G, Mathura V, Sakharkar M, Kangueane P: Types of inter-atomic interactions at the MHC-peptide interface: identifying commonality from accumulated data. *BMC Struct Biol* 2:2, 2002.
 94. Miyazawa S, Jernigan RL: Estimation of effective inter-residue contact energies from protein crystal structure, quasi-chemical approximation. *Macromolecules* 18:534, 1985.
 95. 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* 256:623, 1996.
 96. Govindarajan KR, Kangueane P, Tan TW, Ranganathan S: MPID: MHC-peptide interaction database for sequence structure function information on peptides binding to MHC molecules. *Bioinformatics* 19:309, 2003.
 97. Ren EC, Kangueane P, Kolatkar P, Lin MT, Tseng LH, Hansen JA: Molecular modeling of the minor histocompatibility antigen HA-1 peptides binding to HLA-A alleles. *Tissue Antigens* 55:24, 2000.
 98. Zhao B, Png AEH, Ren EC, Kolatkar PR, Mathura VS, Sakharkar MK, Kangueane P: Compression of functional space in HLA-A sequence diversity. *Hum Immunol* 64:718, 2003.
 99. Chelvanayagam G: A roadmap for HLA-A, HLA-B and HLA-C peptide binding specificities. *Immunogenetics* 45:15, 1996.
 100. Chelvanayagam G: A roadmap for HLA-DR peptide binding specificities. *Hum Immunol* 58:61, 1997.
 101. Zhang C, Anderson A, DeLisi C: Structural principles that govern the peptide binding motifs of class I MHC molecules. *J Mol Biol* 281:929, 1998.
 102. Venkatarajan MS, Braun W: New quantitative descriptor of amino acids based on multi-dimensional scaling of a large number of physical-chemical properties. *J Mol Model* 7:445, 2001.

[Frontiers in Bioscience 10, 879-886, January 1, 2005]

A FRAMEWORK TO SUB-TYPE HLA SUPERTYPES

Pandjassaram Kanguane ¹, Meena Kishore Sakharkar ¹, Ganapathy Rajaseger ², Subhashini Bolisetty ³, Balasubramanian Sivasekari ³, Bing Zhao ¹, Manickam Ravichandran ⁴, Paul Shapshak ⁵, Subramanian Subbiah ⁶

¹ School of Mechanical and Production Engineering, Nanyang Technological University, Nanyang Centre for Supercomputing and Visualization, Singapore 639798, ²DMERI, Defense Science Organization, Singapore, ³ Centre for Biotechnology, Anna University, India 600 025, ⁴Department of Microbiology and Parasitology, School of Medical Sciences, University Science Malaysia, Malaysia, ⁵Department of Psychiatry and Behavior Science, University of Miami Medical School, USA, ⁶ Department of Applied Physics, Stanford University, USA

TABLE OF CONTENTS

1. Abstract
2. Introduction
3. Materials and methods
 - 3.1 Dataset
 - 3.2 Structural basis for HLA supertypes
 - 3.3 Definition of HLA sub-supertypes
 - 3.4 Validation of predictive sub-supertypes
4. Results and discussion
 - 4.1 HLA supertypes
 - 4.2 Perplexing issues with HLA supertypes
 - 4.3 Reduction of HLA sequence diversity
 - 4.4 HLA sub-supertypes
 - 4.5 Validation of predictive sub-supertypes
5. Conclusions
6. Abbreviations
7. Acknowledgements
8. References

1. ABSTRACT

The human leukocyte antigen (HLA) alleles are extremely polymorphic among ethnic population and the peptide binding specificity varies for different alleles in a combinatorial manner. However, it has been suggested that majority of alleles can be covered within few HLA supertypes, where different members of a supertype bind similar peptides, yet exhibiting distinct repertoires. Since the overlap between different members of a supertype appears to be extensive, it is crucial to develop a framework for grouping alleles into supertypes just from sequence information.

In this report, we define sub supertypes, where members show functional overlap with identical repertoire, and describe a strategy to group HLA-A, B and C alleles into different categories of sub supertypes. The strategy grouped 47% of 295 A alleles, 44% of 540 B alleles and 35% of 156 C alleles to just 36, 71 and 18 groups, respectively. The grouping is moderately validated using available binding data. However, the validation is limited due to lack of binding data. Hence, the data presented in this article serve as a framework to test specific functional overlap between alleles.

The grouping of HLA alleles into different categories of sub supertypes has profound use in the

understanding of antigenic peptide selection, degeneration and discrimination during T-cell mediated immune response. A complete knowledge of this phenomenon finds utility in epitope design for the development of HLA based vaccines and immuno-therapeutics.

2. INTRODUCTION

The human leukocyte antigen (HLA) alleles are highly polymorphic among ethnic population. Today, more than 1,800 HLA alleles are known and about a 1,000 of them refer to the class I loci (1). Class I alleles bind peptides of length 8-10 residues during T-cell mediated immune response (2). Therefore, the huge combination of specific HLA-peptide binding is clearly beyond our realization. However, it has been suggested that a majority of alleles can be grouped within few "HLA supertypes", where the members of a supertype bind similar peptides, yet distinct binding repertoires (3). The functional overlap between different alleles within defined supertypes will significantly reduce peptide binding diversity. A catalogue of functional overlap is critical for grouping alleles into supertypes from sequence information. In recent years, a number of supertypes have been defined by comparing peptide binding data. Thus, HLA- A1 (4), A2 (3, 5), A3 (5), A24 (4), B7 (5), B27 (4), B44 (6), B58 (4), and B62 (4) supertypes have been defined. Classification of alleles into

Subtyping HLA supertypes

supertypes using binding data is seldom comprehensive and conclusive. Moreover, a complete grouping of all known alleles using binding data is laborious and expensive. It is also practically impossible to cluster all known alleles using binding data at multiple levels of functional overlaps. Therefore, it is critical to develop theoretical procedures for grouping alleles into supertypes. However, such grouping procedures require rigorous validation prior to routine application. Chelvanayagam *et al.* (7), Zhang *et al.* (8), Zhao *et al.* (9) and Doytchinova *et al.* (10) grouped HLA alleles into functionally overlapping clusters from sequence data. Chelvanayagam *et al.* (7) identified interaction pockets from MHC-peptide (MHCp) crystal structures, Zhang *et al.* (8) defined A-F structural binding pockets, Zhao *et al.* (9) defined functional pockets made of critical polymorphic functional residue positions (CPFRP), Doytchinova *et al.* (10) used molecular interaction fields (MIF), hierarchical clustering (HC) and principal component analysis (PCA) and Lund *et al.* (11) used clustering procedures for grouping HLA alleles into putative supertypes. In this report, we utilize the procedure described by Zhao *et al.* (9) to define HLA 'sub-supertypes' with overlapping function of identical binding repertoire. Using this approach, we grouped 991 alleles (class I) into several groups of 'sub-supertypes'. The grouping is further validated using binding data extracted from MHCBN (12). The importance of HLA sub-supertypes in establishing a conclusive framework for functional overlaps between alleles is discussed.

3. MATERIALS AND METHOD

3.1. Dataset

The protein sequences of HLA-A (295 alleles), HLA-B (540 alleles) and HLA-C (156 alleles) were obtained from IMGT/HLA (release 2.5) for this analysis (1).

3.2. Structural basis for HLA supertypes

HLA allele sequences show high degree of homology among themselves. Therefore, their structures have similar fold and peptide binding groove, where the 3-dimensional spatial orientations of residue backbone atoms are similar for various alleles at different residue positions. However, HLA alleles exhibit extreme polymorphism among themselves and their peptide binding specificity varies between them. Nonetheless, the peptide binding residues show similarity among certain alleles and these alleles binding identical peptides through the concept of HLA supertypes (Table 1).

3.3. Definition of HLA sub-supertypes

In this approach, we define critical polymorphic functional residue positions (CPFRP) for each allele using a methodology described elsewhere (9). Here, CPFRPs are described as those positions that are predominantly involved in peptide binding in known MHCp crystal structures and show polymorphism at least once among known A, B and C allele sequences. Thus, 21 CPFRPs were defined for each allele. The physical and chemical properties of residues at the CPFRPs characterize the binding difference between alleles. It has been shown that it

is possible to sub-group members of a supertype using just 21 residues polymorphic in pockets A-F of the peptide binding groove (9). Hence, the 21 CPFRP residues were extracted for HLA A, B and C alleles. The extracted discontinuous residue segment patterns formed by the CPFRPs were compared and alleles having identical CPFR segments were grouped together. Thus, alleles clustered within a group share identical CPFR segments and are proposed to form functional pockets capable of binding similar peptides with identical repertoire (Tables 2 - 4). Therefore, members of a group bind similar peptides and the groups represent "sub-supertypes". It is proposed that the "sub-supertypes" would have the predictive power previously promised for the supertype itself.

3.4. Validation of predictive sub-supertypes

The grouping is reasonably validated using binding data (Tables 5 - 6) extracted from MHCBN (12). Here, we show several peptides, found in the literature, that are bound by two members of the same sub-supertype. In this article, a total of five pairs of alleles in five sub-supertypes (three of which are HLA-A, and two of which are HLA-B) are found to have similar peptide binding. It should be noted that the utility of this approach awaits more rigorous experimental validation.

4. RESULTS AND DISCUSSION

4.1. HLA Supertypes

More than 1,800 HLA alleles have been defined (1). Therefore, the number of theoretically possible combinations of HLA-peptide complexes is extremely large. However, the immune system maintains a homogenous balance by specific selection, degeneration and discrimination (self/non-self) of short peptides using HLA molecules. Although, HLA molecules are polymorphic in ethnic population, they exhibit a substantial amount of functional overlap through the phenomenon of 'HLA supertypes', where members bind similar peptides and yet display distinct repertoires. A number of 'HLA supertypes' have already been defined using binding data (Table 1). Table 1 shows six peptides binding to all members of the A2 supertypes (A*0201, A*0202, A*0203, A*0206 and A*6802). The functional overlap between different members of the supertype is intriguing. However, it also shows several peptides binding to some members but not all members of the A2 supertypes (Table 1).

4.2. Perplexing issues with HLA supertypes

The concept of HLA supertypes is that alleles belonging to supertypes bind a highly shared set of peptides; in principle it should be possible to predict peptide binding of other members of a supertype using experimental results based on just one member of the type. However, as illustrated in Table I, this promise does not hold in the major supertypes A and B. Hence, the binding of peptides to different members of the A2 supertype is combinatorial in selection and degeneration. Moreover, this grouping is inconclusive given the known number of HLA alleles. Therefore, we devised a theoretical procedure to group HLA alleles into clusters of overlapping function from sequence information (9).

Subtyping HLA supertypes

Table 1. HLA supertype definition

Peptide	Supertype	A*0201	A*0202	A*0203	A*0206	A*6802	Reference
LLFNILGGWV	A2	b	b	b	b	b	13
YLVAIQATV	A2	b	b	b	b	b	13
KVAELVHFL	A2	b	b	b	b	b	14
FLWGPRLV	A2	b	b	b	b	b	14
FLLLADARV	A2	b	b	b	b	b	13
IMIGVLVGV	A2	b	b	b	b	b	14
KIFGSLAFL	A2	b	b	b	b	nb	14
CLTSTVQLV	A2	b	b	b	b	nb	14
RLIVFPDLGV	A2	b	b	b	b	nb	13
YLQLVFGIEV	A2	b	b	b	b	nb	14
LLTFWNPPV	A2	b	b	b	b	nb	14
VLVGGVLA	A2	b	b	b	b	nb	13
WMNRLIAFA	A2	b	b	b	nb	b	13
DLMGYIPLV	A2	b	nb	b	b	b	13
ILHNGAYSL	A2	b	b	b	nb	nb	14
YLSGANLNL	A2	b	b	b	nb	nb	14
VMAGVGSPYV	A2	b	b	b	nb	nb	14
ILAGYGAGV	A2	b	b	b	nb	nb	13
LMTFWNPPV	A2	b	nb	b	b	nb	14
YLVTRHADV	A2	b	nb	b	b	nb	13
HMWNFISGI	A2	b	nb	b	b	nb	14
YLLPRRGPR	A2	b	nb	b	b	nb	13
LLFLLADA	A2	b	b	nb	nb	nb	13
LLTFWNPPPT	A2	b	nb	b	nb	nb	14
ALCRWGLLL	A2	b	nb	b	nb	nb	13
		A*0301	A*1101	A*3101	A*3301	A*6801	
KTSEKQPR	A3	b	b	b	nb	b	13
RMVGGVEHR	A3	b	b	b	nb	nb	13
QLFTFSPPR	A3	b	b	nb	nb	b	13
LGFGAYMSK	A3	b	b	nb	nb	b	13
LIFCHSKKK	A3	b	b	nb	nb	b	13
GVAGALVAFK	A3	b	b	nb	nb	b	13
VAGALVAFK	A3	b	b	nb	nb	b	13
RLGVRATRK	A3	b	b	b	?	?	15
		B*0702	B*3501	B*5101	B*5301	B*5401	
LPGCSFSIF	B7	b	b	b	b	nb	15

b = binder; nb = non-binder; ? = undetermined

4.3. Reduction of HLA sequence diversity

Our previous analysis grouped 101 HLA-A alleles into 29 clusters such that each group contains at least two alleles with identical functional binding CPFR segments (9). In this report, we extended the grouping to HLA-A, -B and -C alleles. Here, we used 991 alleles (class I) of which, 295 are 'A alleles', 540 are 'B alleles' and 156 are 'C alleles'. The discontinuous CPFR residue segment patterns were extracted for each allele (9). Comparison of these segment patterns among HLA-A, B and C alleles grouped A alleles into 192 unique clusters, B alleles into 373 unique clusters and C alleles into 119 unique clusters. Thus, 991 alleles were grouped into 684 clusters. The total diversity among the 991 alleles is reduced to 69% with a 'reduction in variability' of 31%. The diversity in A, B and C is reduced to 65%, 69% and 76%, respectively with a 'reduction in variability' of 35%, 31% and 24%, respectively. An understanding of such reduction among the functionally important binding residues is critical to decipher the molecular basis for HLA-supertypes.

4.4. HLA sub-supertypes

In this study, 139 A alleles (47% of all known A alleles) were clustered into 36 groups such that each group contains at least two alleles (Table 2). Members in each of these groups have identical CPFR segment patterns. Thus, alleles within a group have identical functional pockets. The proposed hypothesis is that alleles in a group bind to similar peptides and members in the group form 'sub-

upertypes' with identical binding repertoire. If binding of a peptide is known for one representative allele in a group, its binding to other members in the group can be deduced. However, this clustering has to be validated using experimentally determined binding data for a representative set of alleles. This requires large scale generation of binding data and it is difficult to generate binding data at such scale due to limitations in the synthesis and purification of HLA alleles for binding assay. If this hypothesis is reasonably validated using binding data, it is possible to infer the binding property of a peptide to 47% of A alleles using binding values for 36 representative A alleles. The 36 groups clustered A alleles into A1, A2, A3, A11, A23, A24, A26, A29, A30, A31, A32, A33, A34, A36, A66, A68 and A74 'predictive sub-supertypes' where members of a group is proposed to bind similar peptides (Table 2). We extended our analysis to B alleles and showed that 238 B alleles (44% of all known B alleles) were clustered into 71 groups such that each group contains at least two alleles (Table 3). We also show that 55 C alleles (35% of all known C alleles) were grouped into 18 groups such that each group contains at least two alleles (Table 4). Thus, we demonstrate that the strategy (9) used to group alleles into sub-supertypes is extremely powerful in reducing functional diversity.

4.5. Validation of predictive sub-supertypes

In order to validate the proposed hypothesis, we extracted peptides from MHCBN (12). This experiment

Subtyping HLA supertypes

Table 2. HLA-A alleles (139) grouped into putative supertypes

Group	CPFR	HLA-A alleles
1	YFARENHTDANTLIYRDARRG	A*0101 A*0104 A*0109
2	YFAREKHTHVDTLRYHYVLTW	A*0201 A*0209 A*0224 A*0225 A*0231 A*0240 A*0243 A*0245 A*0246 A*0258 A*0259 A*0266 A*0267 A*0268
3	YFAREKHTHVDTLRYHYVWTW	A*0202 A*0222 A*0250 A*0263
4	YYAREKHTHVDTLRYHYVLTW	A*0206 A*0214 A*0221 A*0228 A*0251 A*0261
5	YFAREKHTHVDTLRCHYVLTW	A*0207 A*0215 A*0218
6	YFAREKHTHVDTLRYHYVQTG	A*0219 A*0237
7	YFARENQTHVDTLRYHYVLTW	A*0256 A*0262
8	YFARENQTDVDTLIYRDELTW	A*0301 A*0304 A*0305 A*0306 A*0313 A*0314
9	YFARENQTDVDTLIYRDVQTW	A*0302 A*0310
10	YYARENQTDVDTLIYRDAQRW	A*1101 A*1102 A*1105 A*1107 A*1109 A*1112 A*1113 A*1115 A*1116
11	YSAREKHTDENIAMFHYVLTG	A*2301 A*2303 A*2306 A*2307 A*2413
12	YSAREKHTDENIAMFHYVWTG	A*2302 A*2406
13	YSAREKHTDENIAMFHYVQTG	A*2402 A*2405 A*2409 A*2411 A*2421 A*2426 A*2427 A*2429 A*2435 A*2437 A*2439 A*2440
14	YSAREKHTDENIAMFHYVQTW	A*2403 A*2423 A*2433
15	YSAREKHTDENIAMFRDVQTG	A*2417 A*2441
16	YSAREKHTDENIAMFHYVQTG	A*2430 A*2442
17	YYARNNHTDANTLRYQDEWRW	A*2601 A*2610 A*2614 A*2615 A*2617
18	YYARNNHTHVDTLRYQDEWRW	A*2603 A*2606
19	YYARNNHTDANTLRYQDVWRW	A*2612 A*2618
20	YTARQNQTDANTLMYRDVLTW	A*2901 A*2902 A*2906 A*2909 A*2910 A*2911
21	YSARENQTDVDTLIYEHVLTW	A*3001 A*3011
22	YSARENHTDENTLIYEHRLTW	A*3002 A*3003 A*3012
23	YSARENHTDENTLIYEHVWTW	A*3004 A*3006
24	YTARENHIDVDTLMYQDVLTW	A*3101 A*3109
25	YTARENHIDVDTLIYRDVLTW	A*3103 A*3104
26	YTAREKHTDENIAMYQDVLTW	A*3107 A*3108
27	YFARENHTDESIAMYQDVLTW	A*3201 A*3206
28	YTARNNHIDVDTLMYQDVLTW	A*3301 A*3303 A*3304 A*3305 A*3306 A*3307
29	YYARNKQTDVDTLRYQDEWTW	A*3401 A*3405
30	YYARNNQTDVDTLIYRDELTW	A*3402 A*3403 A*3404
31	YFARENHTDANTLIYRDARTW	A*3601 A*3602
32	YYARNNQTDVDTLRYQDEWRW	A*6601 A*6604
33	YYARNNQTDVDTLMYRDVWTW	A*6801 A*6816 A*6819 A*6821 A*6822 A*6824 A*6825
34	YYARNNHTHVDTLMYRDVWTW	A*6805 A*6820
35	YYARENQTDVDTLMYRDVWTW	A*6810 A*6813 A*6814
36	YFARENHTDVTDLMYQDVLTW	A*7401 A*7402 A*7403 A*7408 A*7409

CPFR = critical polymorphic functional residue

identified five peptides: (1) FLWGPRALV, (2) AAGIGILTV, (3) GILGFVFTL, (4) VLYRYGSFSV and (5) YLEPGPVTA binding A*0201 and A*0209 (Table 5). In Table 2, A*0201 and A*0209 are categorized together (Group #2). Comparison of Table 5 with Table 2 suggest that peptides binding to A*0201, also binds to A*0209. This observation is very interesting. However, it is important to establish the binding of these peptides to other members of the group such as A*0224, A*0225, A*0231, A*0240, A*0243, A*0245, A*0246, A*0258, A*0259, A*0266, A*0267 and A*026. It should be noted that it is labor intensive to clone and express all of these alleles to determine binding to the above peptides. Nonetheless, if experimentally determined binding data show similar binding to all these alleles, then extrapolation of the proposed hypothesis to other predictive sub-supertype is trivial.

We found yet another peptide GILGFVFTL which binds to A*0206 and A*0214 (Table 5). These two alleles are categorized together in 'Group #4' (Table 2). This observation is also interesting. However, if similar binding of this peptide to other members of 'Group 4' such as A*0221, A*0228, A*0251 and A*0261 is established, extrapolation of this strategy to other predictive sub-supertypes will be trivial (Table 2). Interestingly, GILGFVFTL binds A*0201, A*0209, A*0206 and A*0214 (Table 5). Although, GILGFVFTL binds all the above four alleles, our strategy grouped these alleles into two distinct groups (Group #2 (A*0201 & A*0209) and Group #4 (A*0206 & A*0214)). We further probed into

their functional overlap but examining their CPFR segments. Examination of CPFR segments in Group #2 and Group #4 (Table 2) shows a single residue mutation (9F to 9Y) between these two groups. F (phenylalanine) and Y (tyrosine) are aromatic and the mutation is synonymous. This explains binding overlap between members of Group #2 and Group #4. By definition, our strategy groups members with similar peptide binding and members across groups (for example, Group #2 and Group #4) do not always bind similar peptides. We also found that a number of peptides (LLFNILGGWV, YLVAYQATV, KVAELVHFL, FLWGPRALV, FLLADARV, IMIGVLGV) bind A*0201 (Group #2) and A*0206 (Group #4). A*0201 and A*0206 are grouped as members of A2 supertype using binding data (Table 1). The data implies that some peptides binding to A*0201 are also found to bind A*0206. However, this is not always true. Peptides (WMNRLIAFA, ILHNGAYSL, YLSGANLNL, VMAGVGSPYV, ILAGYGAGV) that bind to A*0201 do not always bind to A*0206 (Table 1). This suggests that A*0201 and A*0206 are not strict members of the A2 'sub-supertype'. On the other hand, we propose that peptides that bind to A*0201 should strictly bind to A*0209 as they share identical CPFR segments (Table 2). Similarly, A*0206 and A*0214 should always bind similar peptides as they share identical CPFR segments (Table 2).

The CPFR segments in A*0201 (Group 2) and A*0202 (Group 3) shows single residue mutation (156L to

Subtyping HLA supertypes

Table 3. HLA-B alleles (238) grouped into putative supertypes

Group	CPFR	HLA-B alleles
1	YYSRNIQTDESNSLYDEREW	B*0702 B*0710 B*0721 B*0722 B*0723 B*0730 B*0733 B*0735
2	YYSRNINTDESNSLYDEREW	B*0703 B*0708
3	YYSRNIQTDESNSLYDEREW	B*0705 B*0706
4	YYSRNIQTDESNLRYDEREW	B*0707 B*0712 B*0714 B*0718
5	YYSRNIQTDESNSLYDSEREW	B*0709 B*0717
6	YYSRNINTYESNSLYDEREW	B*0716 B*0737
7	YDSRNINTDESNSLYNVDTW	B*0801 B*0804 B*0818 B*0819
8	YYTREINTYENTARYNLVLEW	B*1301 B*1311
9	YYTREINTYENTATYNLVLEW	B*1302 B*1308
10	YYSRNINTDESNLWYNFELTW	B*1401 B*1402
11	YYAREINTYESNLRYDSEWLW	B*1501 B*1528 B*1533 B*1534 B*1538 B*1546 B*1556 B*1560 B*1566 B*1578 B*1581 B*1582
12	YYARNINTYESNLRYDSELLW	B*1502 B*1521 B*3511 B*3521
13	YYSREINTYESNLRYDSELLW	B*1503 B*1562 B*1574
14	YYAREINTYESNLTYDSEWLW	B*1504 B*1535
15	YYAREINTYESNLRYDSVLLW	B*1505 B*1520 B*3528
16	YYARNINTYESNLRYDSEWLW	B*1508 B*1511 B*1515 B*3514 B*3543
17	YYSRNINTYESNLRYDYELLW	B*1510 B*1537
18	YYAREINTYESNLRYDSEWLW	B*1512 B*1519
19	YYARNINTYENIARYDSELLW	B*1513 B*5306 B*5308
20	YYSRNINTYESNLRYDSELLW	B*1518 B*1529 B*1564 B*1572 B*1580
21	YYAREINTYESNLRYDSELLW	B*1525 B*1539
22	YYSREINTYESNLRYDSEREW	B*1547 B*1549
23	YYARNINTYESNSLYDSVLLW	B*1555 B*3505 B*3517 B*3530
24	YYARNIQTDESNLRYDSEWLW	B*1576 B*5603
25	YHSRNINTYESNLRYDSVLTW	B*1801 B*1805 B*1807 B*1811
26	YYARNINTYESNLRYDSVLTW	B*1804 B*3535
27	YHTREIKTDEDTLNHYDVLW	B*2703 B*2705 B*2713 B*2717
28	YYARNINTYESNLRYDSVLLW	B*3501 B*3507 B*3519 B*3520 B*3524 B*3526 B*3532 B*3540 B*3541 B*3542 B*3546 B*3547 B*3549
29	YYARNINTYESNLRYNYVLLW	B*3502 B*3504 B*3509 B*3512
30	YYARNINTYESNLRYDFVLLW	B*3503 B*3536
31	YYARNINTYESNLRYDYVLLW	B*3534 B*3539
32	YHSREINTYEDTLRSNFVDTW	B*3701 B*3704
33	YYSRNINTYENIARYNFVLTW	B*3801 B*3805 B*3806 B*3807 B*3809
34	YYSRNINTDESNLRYNFVLTW	B*3901 B*3904 B*3910 B*3916 B*3919 B*3926
35	YYSREINTDESNLRYNFVLTW	B*3902 B*3922 B*3923
36	YYSRNINTDESNSLYNFVLTW	B*3903 B*3924
37	YYSRNINTDESNLTYNFVLTW	B*3906 B*3928
38	YHTREINTYESNLRYNYVLEW	B*4001 B*4004 B*4007 B*4011 B*4014 B*4046
39	YHTREINTYESNSLYNYVLEW	B*4002 B*4029 B*4035 B*4045
40	YHTREINTYESNSLYDYVLEW	B*4009 B*4018 B*4024 B*4031
41	YYAREINTYESNLRYNYVLEW	B*4010 B*4021
42	YYSREINTYESNLRYNYVLEW	B*4012 B*4803
43	YHTREINTYENIASYNYVLEW	B*4013 B*4019
44	YHTREINTYESNSLYNYEREW	B*4015 B*4016
45	YHTREINTYESNLRYNYELLW	B*4026 B*4028
46	YHTREINTYESNLVYNYVLEW	B*4030 B*4034
47	YHTREINTYESNLRYDYVLEW	B*4033 B*4042
48	YHTREINTYESNLRYNYVDTW	B*4101 B*4103 B*4106
49	YYSRNIQTDESNSLYNVDTW	B*4201 B*4205
50	YYTREINTYENTARYDDVDLS	B*4402 B*4408 B*4422 B*4424 B*4427 B*4433
51	YYTREINTYENTARYDDVLLS	B*4403 B*4407 B*4413 B*4426 B*4429 B*4430 B*4432 B*4436 B*4438
52	YHTREINTYESNLRYNLVDLS	B*4501 B*4503 B*4505
53	YYTREINTYESNLRFHVDLEW	B*4702 B*4703
54	YYSREINTYESNSLYNYVLEW	B*4801 B*4804
55	YHTREINTYESNLRYNLELLW	B*5001 B*5004
56	YYARNINTYENIATYNYELLW	B*5101 B*5102 B*5107 B*5111 B*5112 B*5117 B*5118 B*5122 B*5124 B*5126 B*5128 B*5130 B*5132
57	YYARNINTYENIARYNYELLW	B*5104 B*5106
58	YYARNINTYENIATYNYVLLW	B*5109 B*5119
59	YYARNINTYENIATYNYELEW	B*5116 B*5134
60	YYAREINTYENIATYNYELLW	B*5201 B*5202 B*5204 B*5205
61	YYARNINTYENIARYDSVLLW	B*5301 B*5302
62	YYARNIQTDESNLTYNLVLTW	B*5401 B*5502 B*5507
63	YHARNIQTDESNLTYNLVLTW	B*5402 B*5516
64	YYARNIQTDESNLTYNLELTW	B*5501 B*5505 B*5515
65	YYARNIQTDESNLRYNLVLLW	B*5602 B*5604
66	YYARNIQTDESNLTYNYELLW	B*5605 B*5606
67	YYARENSTYENIAYDSVLLW	B*5701 B*5706 B*5708
68	YYARENSTYENIARYDSVLLW	B*5801 B*5804 B*5809
69	YYARNINTDESNLTYNYELLW	B*7801 B*7803
70	YYSRNIQTDESNSLYNYVLEW	B*8101 B*8102
71	YYSRNIQTDESNLRFNLVDLS	B*8201 B*8202

CPFR = critical polymorphic functional residue

Subtyping HLA supertypes

Table 4. HLA-C alleles (55) grouped into putative supertypes

Group	CPFR	HLA-Cw alleles
1	YYAREKQTDVNKLRYDSEWEW	Cw*0202 Cw*0208 Cw*0209
2	YYAREKQTDVSNLRYDYELLW	Cw*0303 Cw*0304 Cw*0309 Cw*0313
3	YSAREKQADVKNLRFNFERTW	Cw*0401 Cw*0405 Cw*0407 Cw*0409 Cw*0412
4	YYAQEKQTDVNKLRYNFERTW	Cw*0501 Cw*0503 Cw*0505 Cw*0506 Cw*0509
5	YDSREKQADVKNLWYDSEWTW	Cw*0602 Cw*0607
6	YDSRENQADVSNLRYDSALTW	Cw*0701 Cw*0706 Cw*0716 Cw*0718 Cw*0721 Cw*0724
7	YDSREKQADVSNLRSDSALTW	Cw*0702 Cw*0710 Cw*0717 Cw*0723 Cw*0725
8	YDSREKQADVSNLRYDFADTW	Cw*0704 Cw*0711 Cw*0712
9	YDSRENQADVKNLRYDSALTW	Cw*0707 Cw*0709
10	YYAQEKQTDVSNLRYNFRTLTW	Cw*0801 Cw*0803
11	YYAQEKQTDVSNLRYNFERTW	Cw*0802 Cw*0807
12	YYAQEKQTDVSNLRYDSTLTW	Cw*0809 Cw*0811
13	YYAREKQADVSNLWYDSEWTW	Cw*1203 Cw*1206
14	YSAREKQTDVSNLWFDSEWTW	Cw*1402 Cw*1403
15	YYARENQTDVNKLRYDLELTW	Cw*1502 Cw*1510 Cw*1512
16	YYARENQTDVNKLRYDSELTW	Cw*1504 Cw*1509
17	YYAREKQADVKNLRYNFELEW	Cw*1701 Cw*1702 Cw*1703
18	YDSREKQADVKNLRFNFERTW	Cw*1801 Cw*1802

CPFR = critical polymorphic functional residue

Table 5. Validation of HLA-A supertypes

Group	Peptide	Allele pair	Binding affinity	T-cell activity
Category #1				
2	FLWGPRLV	A*0201 A*0209	YES (HIGH/?)	YES
	AAGIGILTV	A*0201 A*0209	YES (MOD/HIGH)	YES
	GILGFVFTL	A*0201 A*0209	YES (HIGH)	YES
	VLYRYGSFSV	A*0201 A*0209	YES (HIGH)	YES
	YLEPGFVTA	A*0201 A*0209	YES	YES
Category #2				
4	GILGFVFTL	A*0206 A*0214	YES (?/LOW)	MOD
Category #3				
22	RISGVDRYY	A*3002 A*3003	YES	?

MOD = moderate; ? = undetermined, The group numbers indicated in this Table refers to the group numbers designated in Table 2.

156W). The residues leucine (L) and tryptophan (W) are hydrophobic and the mutation is synonymous. Therefore, some peptides that bind A*0201 are also found to bind A*0202 (Table 1). This trend is not always true (Table 1). Therefore, A*0201 and A*0202 are not strict members of the A2 'sub-supertype'. In yet another case, the peptide RASGVDRYY binds A*3002 and A*3003 (Table 5). Table 2 shows that A*3002 and A*3003 are grouped together with another allele A*3012 (Group #22) and members of this group are proposed to bind similar peptides. Therefore, the binding of this peptide to A*3012 will be of great significance to validate this hypothesis. We further validated the grouping for B alleles using binding data for 2 pairs of alleles (B*2703/B*2705 and B*5101/B*5102) and the data is given in Table 6. The pairs B*2703/B*2705 (Group #27) and B*5101/B*5102 (Group #56) strictly bind similar peptides (Tables 3 and 6). More than 20 peptides were shown to bind each of the allele pairs clustered in Group #27 and Group #56. However, the binding of other members of the group is required for further extrapolation.

Here, we demonstrate using binding data for 3 pairs of HLA-A alleles (A*0201/A*0209 (Group #2), A*0206/A*0214 (Group #4) and A*3002/A*3003 (Group #22) that alleles within a group (Table 2) bind similar peptides (Table 5). The validation is extended to two pairs of B alleles B*2703/B*2705 and B*5101/B*5102 (Table 3 and Table 6). Thus, our methodology groups alleles into strict supertypes, where members always bind similar peptides (Tables 2 and 5). We also note that there may be

some degree of functional overlap (Table 6) across members of different groups (A*0201 (Group 2) / A*0206 (Group #4)) due to synonymous residue substitutions at the CPFRP (Table 2). Several peptide binders of A*0201 are non-binders to A*0206 (Table 1). Hence, the functional overlap across groups (Table 1 and Table 2) is not always true. This warrants that grouping of HLA alleles into supertypes based on binding data is seldom conclusive and comprehensive (Table 1). The validation of the grouping strategy is limited in the current report. Further validation is required to apply this methodology to group alleles into 'sub-supertypes' from sequence information on a large scale. It is our hope that the clustering provided here will serve as a theoretical framework for investigating the phenomenon of HLA supertypes using binding data.

5. CONCLUSIONS

Knowledge on all possible combinations of MHCp binding is useful in the design of peptide vaccine candidates, immuno-therapeutic targets and diagnostics agents. The theoretically possible combinations are overwhelmingly large. However, the functional overlap between alleles and the grouping of alleles into 'sub-supertypes' is extremely powerful in understanding peptide selection and degeneration. Grouping of alleles into supertypes using binding data is seldom conclusive and comprehensive. The strategy described in this report, grouped 47% of known A alleles (295), 44% of known B alleles (540) and 35% of known C alleles (156) to just 36, 71 and 18 groups,

Subtyping HLA supertypes

Table 6. Validation of HLA-B supertypes

Group	Peptide	Allele pairs		Binding affinity	T-cell activity
Category #1					
27	ARRHGFLPRH	B*2703	B*2705	YES (MOD)	?
	ARTAHYGSL	B*2703	B*2705	YES	?
	ARYQKSTEL	B*2703	B*2705	YES	?
	FQYNGLIHR	B*2703	B*2705	YES	?
	FRYNGLIHR	B*2703	B*2705	YES	?
	GRAFVTIGA	B*2703	B*2705	YES	?
	GRAFVTIGK	B*2703	B*2705	YES	?
	GRERFEMER	B*2703	B*2705	YES (MOD)	?
	GRFFGGDRG	B*2703	B*2705	YES	?
	GRGLSLSRF	B*2703	B*2705	YES	?
	GRIDKPILA	B*2703	B*2705	YES	?
	GRIDKPILK	B*2703	B*2705	YES	?
	SRAHSSHLK	B*2703	B*2705	YES	?
	SRFSWGAEG	B*2703	B*2705	YES (LOW)	?
	SRHKKLMFK	B*2703	B*2705	YES (HIGH)	?
	SRSGSPMAR	B*2703	B*2705	YES (MOD)	?
	SRYWAITR	B*2703	B*2705	YES	?/YES
	VRRCPHHER	B*2703	B*2705	YES (LOW/MOD)	?
	VRVCACPGR	B*2703	B*2705	YES (MOD)	?
	RRYQKSTEL	B*2703	B*2705	YES	?
	QRHGSKYLA	B*2703	B*2705	YES (LOW/MOD)	?
	RRIKEIVKK	B*2703	B*2705	YES (MOD)	?
	RRTEENLR	B*2703	B*2705	YES (MOD)	?
Category #2					
56	FPISPIETV	B*5101	B*5102	YES (HIGH/?)	?
	FPVRPQVPL	B*5101	B*5102	YES	?
	FPVRPQVPL	B*5101	B*5102	YES	?
	CPKVSFEPI	B*5101	B*5102	YES	?
	CPSGHAVGI	B*5101	B*5102	YES	?
	DARAYDTEV	B*5101	B*5102	YES (HIGH/?)	NO/?
	EPLDLPQIIB	B*5101	B*5102	YES	?
	YPFKPPKVB	B*5101	B*5102	YES	?
	IPLGDAKLV	B*5101	B*5102	YES	?
	IPTSGDVVI	B*5101	B*5102	YES	?
	LPALSTGLI	B*5101	B*5102	YES	?
	LPCRIKQIIB	B*5101	B*5102	YES (LOW/?)	?
	LPEKDSWTV	B*5101	B*5102	YES	?
	LPPLERLTL	B*5101	B*5102	YES	?
	LPPTTGPPIB	B*5101	B*5102	YES	?
	LPPVVAKEI	B*5101	B*5102	YES (HIGH)	?
	NALFRNLDV	B*5101	B*5102	YES	?
	NANPDCKTI	B*5101	B*5102	YES	?
	NPPIPVGEIB	B*5101	B*5102	YES	?
	QGWKGSPAI	B*5101	B*5102	YES	?
	TAVQMAVFI	B*5101	B*5102	YES	?
	TGYLNTVTV	B*5101	B*5102	YES	?
	VAQRAYRAI	B*5101	B*5102	YES	?
	VGCLVGLRI	B*5101	B*5102	YES	?
	VPVKLKPGM	B*5101	B*5102	YES	?
	YAPPIGGQI	B*5101	B*5102	YES	?
	YPCTVNFTI	B*5101	B*5102	YES	?
	YPLASLKSL	B*5101	B*5102	YES	?
	YPLTSLRSL	B*5101	B*5102	YES	?
	APTLWARMI	B*5101	B*5102	YES	?

MOD = moderate; ? = undetermined, The group numbers indicated in this Table refers to the group numbers designated in Table 3.

respectively. This grouping procedure is useful because the binding of a peptide to ~50% of all known alleles can be inferred using a handful of binding data representing all predictive sub-supertypes. However, a comprehensive validation is required for large scale extrapolation. Some members across groups show overlapping function. However, the overlap across group is not always true. It should be noted that the methodology described here (9) is different from the procedure described by Doytchinova *et al.* (10) and Lund *et al.* (11) and they differ among themselves. We hope to establish consensus among these procedures in future investigations.

7. ACKNOWLEDGEMENTS

This research is supported by Nanyang Technological University, Singapore. P.K and M.K.S

acknowledge the support from A*STAR-BMRC, Grant # 03/1/22/19/242. PS is supported by NIH grants DA014533 and GM056529.

8. REFERENCES

- Robinson J., M.J. Waller, P. Parham, N. de Groot, R. Bontrop, L.J. Kennedy, P. Stoeckl & S.G.E. Marsh IMGT/HLA and IMGT/MHC - sequence databases for the study of the major histocompatibility complex. *Nucleic Acids Res.* 31(1) 311-314 (2003)
- Yewdell J.W., E. Reits & J. Neefjes: Making sense of mass destruction – quantitating MHC class I antigen presentation. *Nat. Rev. Immunol.* 3(12) 952-961 (2003)

Subtyping HLA supertypes

3. Del Guercio M.F., J. Sidney, G. Hermanson, C. Perez, H.M. Grey, R.T. Kubo & A. Sette: Binding of a peptide antigen to multiple HLA alleles allows definition of an A2-like supertype. *J. Immunol.* 154(2) 685-693 (1995)
4. Sette A. & J. Sidney: Nine Major HLA class I supertypes account for the vast preponderance of HLA-A and -B polymorphism. *Immunogenetics* 50(3-4) 201-212 (1999)
5. Sette A. & J. Sidney: HLA supertypes and supermotifs - a functional perspective on HLA polymorphism. *Curr. Opin. Immunol.* 10(4) 478-482 (1998)
6. Sidney J., S. Southwood, V. Pasquetto & A. Sette: Simultaneous prediction of binding capacity for multiple molecules of the HLA B44 supertype. *J. Immunol.* 171(11) 5964- 5974 (2003)
7. Chelvanayagam G: A roadmap for HLA-A, HLA-B, and HLA-C peptide binding specificities. *Immunogenetics* 45(1) 15-26 (1996)
8. Zhang C, A. Anderson & C. DeLisi: Structural principles that govern the peptide-binding motifs of class I MHC molecules. *J. Mol. Biol.* 281(5) 929-947 (1998)
9. Zhao B, A.E.H. Png, E.C Ren, P.R. Kolatkar, V.S Mathura, M.K Sakharakar & P. Kanguane: Compression of functional space in HLA-A sequence diversity. *Hum. Immunol.* 64(7) 718-728 (2003)
10. Doytchinova IA, P. Guan & D.R. Flower: Identifying human MHC supertypes using bioinformatics methods. *J. Immunol.* 172(7), 4314-4323 (2004)
11. Lund O., M. Nielsen, C. Kesmir, A.G. Petersen, C. Lundegaard, P. Worning, C. Sylvester-Hvid, K. Lamberth, G. Roder, S. Justesen, S. Buus & S. Brunak: Definition of supertypes for HLA molecules using clustering of specificity matrices. *Immunogenetics* 55(12), 797-810 (2004)
12. Bhasin M, H. Singh & G.P. Raghava: MHCBN - a comprehensive database of MHC binding and non-binding peptides. *Bioinformatics* 19(5), 665-666 (2003)
13. Scognamiglio P., D. Accapezzato, M.A. Casciaro, A. Cacciani, M. Artini, G. Bruno, ML Chircu, J Sidney, S Southwood, S Abrignani, A Sette & V Barnaba. Presence of effector CD8+ T cells in hepatitis C virus-exposed healthy seronegative donors. *J. Immunol.* 162(11), 6681-6689 (1999)
14. Kawashima I, S.J. Hudson, V. Tsai, S. Southwood, K. Takesako, E. Appella, A. Sette & E. Celis: The multi-epitope approach for immunotherapy for cancer: identification of several CTL epitopes from various tumor-associated antigens expressed on solid epithelial tumors. *Hum. Immunol.* 59(1), 1-14 (1998)
15. Chang K, M., N.H. Gruener, S. Southwood, J. Sidney, G.R. Pape, F.V. Chisari & A. Sette: Identification of HLA-

A3 and -B7-restricted CTL response to hepatitis C virus in patients with acute and chronic hepatitis C. *J. Immunol.* 162(2), 1156-1164 (1999)

Abbreviations: HLA: human leukocyte antigen, IC₅₀: inhibitory concentration 50, TCR: T-cell receptor, CPFR: critical polymorphic functional residue, CPFRP: critical polymorphic functional residue position, IMGT: ImMunoGeneTics, MHC: major histocompatibility complex, MHCBN: major histocompatibility complex binding and non-binding peptides, MHCp: MHC-peptide complex, MIF: molecular interaction fields, PCA: principal component analysis, HC: hierarchical clustering

Key Words: HLA, Supertypes, Sub-Supertypes, Functional Overlap, Vaccine Design, Review

Send correspondence to: Dr. Pandjassaram Kanguane, School of Mechanical and Production Engineering, 50 Nanyang Avenue, Nanyang Technological University, Singapore 639 798; Tel: +65 (6) 790-5836; Fax: +65 (6) 774-4340, E-mail: mpandjassaram@ntu.edu.sg

<http://www.bioscience.org/current/vol10.htm>