UNDERSTANDING PEPTIDE SPECIFICITY THROUGH STRUCTURAL IMMUNOINFORMATICS

TONG JOO CHUAN

(*B. Sc. (Hons.), NUS*)

A THESIS SUBMITTED

FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

DEPARTMENT OF BIOCHEMISTRY

NATIONAL UNIVERSITY OF SINGAPORE

2007

Acknowledgement

Many people have helped me in the duration of this research and the write up of this dissertation. I would like to take the opportunity to thank the following people for their invaluable help and guidance.

- Professor Shoba Ranganathan, Head, Biotechnology Research Institute, Macquarie University, Australia and Department of Biochemistry, for her invaluable advice, guidance and support throughout the course of research.
- Associate Professor Tan Tin Wee, Department of Biochemistry, National University of Singapore, for extending all possible help and support throughout my candidature. In addition, I also thank him for providing priceless advice and encouragement.
- Professor Vladimir Brusic, School of Land and Food Sciences and the Institute for Molecular Biosciences, The University of Queensland, Australia, for valuable advice.
- Associate Professor Kunchithapadam Swaminathan, Department of Biological Sciences, for giving the valuable opportunity to carry out my candidature.
- Dr. Gan Yunn Hwen, Department of Biochemistry, for the chance to pursue my candidature.
- Professor Animesh Sinha, Department of Dermatology, Weill Medical College of Cornell University, for providing experimental data set.

- Dr. Ng See Kiong, Manager, Knowledge Discovery Department, Institute for Infocomm Research for his kind understanding and support.
- My fellow post-graduate friends Justin Choo Khar Heng, Gopalan Vivek, Kong Lesheng, Bernett Lee Teck Kwong and Paul Tan Thiam Joo, Department of Biochemistry, National University of Singapore for many discussions on the subject of this thesis.
- Mark de Silva and Lim Kuan Siong for systems support and the people in the Department of Biochemistry and the Dean's office of the Faculty of Medicine for their assistance in administrative matters.
- My colleagues Judice Koh Lie Yong, Zhang Guanglan, Zhang Zong Hong, Anita Veeramani, Menaka Rajapakse, Ang Wee Tiong and Kavitha Gopalakrishnan for their comradeship.
- My fiancée Lai Phui Ching for her love and encouragement during difficult times.
- My family for their love, support and encouragement.

Table of Contents

Summary

Major histocompatibility complex (MHC) molecules bind peptides of diverse sequences in order to generate maximal immunological protection by covering the spectrum of peptides that may be seen by a host over the course of its lifetime. However, in many circumstances the immune system malfunctions and incorrectly recognizes a self-peptide. This results in disease characterized by recognition and attack of self. Pemphigus vulgaris (PV) is an example for such autoimmune disorders. In such a situation, identifying disease-implicated alleles and their respective T cell epitope repertoire is valuable in the definition of qualities such as antigenicity and immunodominance, and is an essential preliminary step towards effective immunotherapeutical treatments. However, experimental determination of binding peptides for every disease-implicated allele is prohibitively expensive in terms of labour, time and cost; and is not feasible to studies involving large numbers of protein sequences.

This thesis describes original findings from the application of bioinformatic tools to the study of peptide/MHC interactions and subsequent application to PV. Several novel aspects are presented in this thesis. This is, to the author's knowledge, the first study of its kind, where structural interaction parameters have been used for the analysis of MHC supertypes or superfamilies. Conserved interaction characteristics among different MHC supertypes have been discovered.

The first study on the use of structural principles to discriminate between peptide binders and non-binders, for a number of disease-implicated and non-disease-implicated alleles, is presented. By focusing on known

peptide binders, this bioinformatic approach can discriminate between alleles implicated in the disorder and those that are not. Insights into structural features that underlie the immune response provided by protective alleles for PV have also been obtained.

A new docking protocol and a complementary scoring function, developed as part of this work, has been applied successfully for modeling the bound conformation of peptide ligands to both class I and class II molecules. High prediction accuracy of MHC-binding peptides was validated by existing experimental biochemical and functional data. This approach successfully identified peptide binders which lack conserved binding motifs. The first reported evidence on the possibility of multiple binding registers within a candidate class II binding peptide provides new insights to the binding specificities of class II alleles. The ability of a candidate peptide to bind a specific class II allele is affected by both the binding registers and flanking peptide residues.

In the context of PV, the results of analysis reveal the possibility that the disease-implicated alleles DRB1*0402 and DQB1*0503 share similar specificities by binding peptides at different core recognition regions. The target antigen of PV, desmoglein-3 (Dsg3), is a 130-kDa transmembrane glycoprotein within the desmosomes of the spinous layer of the skin. Little is known about the function of Dsg3 in the normal structure and function of hair. Although it had been postulated that the protein plays a key role in providing cell adhesion between keratinocytes, few in vivo models exist that confirm their actual function. The discovery of multiple initial shared immunodominant

epitopes and intracellular specificities within the desmoglein-3 (Dsg3) self-antigen shed new light into the pathology of PV.

The study of peptide specificity through immunoinformatics facilitates the discovery of T cell epitopes and bears the potential to expedite the vaccine discovery process. Because MHC alleles are diverse in nature, with a spectrum of binding specificities to a restricted range of peptides, the methodology presented in this thesis is suitable for the analysis of alleles where experimental binding data is lacking. The immunoinformatics lessons will be useful for the study of sequence-structure-function relationships involved in the selection of specific antigenic peptides by the different MHC alleles and their implications for disease pathogenesis.

List of Figures

List of Tables

Abbreviations

3D	Three-dimensional
APC	Antigen presenting cell
APS-II	Autoimmune polyendocrine syndrome type II
ANN	Artificial neural network
ASA	Accessible surface area
β_2m	β_2-microglobulin
CD	Celiac disease
Dsg	Desmoglein
ECD	Extracellular domain
ECEPP/3	Empirical conformational energy program for peptide 3
ER	Endoplasmic reticulum
ERAAP	ER aminopeptidase associated with antigen processing
FN	False negative
FP	False positive
HLA	Human leukocyte antigen
HMM	Hidden Markov model
ICM	Internal Coordinate Mechanics
IDDM	Insulin-dependent diabetes mellitus
IFN	Interferon
Ig	Immunoglobulin
Ii	Invariant chain
LMP	Low molecular weight protein
MHC	Major histocompatibility complex
MIIC	MHC class II compartment

PDB	Protein Databank
PV	Pemphigus vulgaris
QSAR	Quantitative structure-affinity relationship
RMSD	Root mean square deviation
SE	Sensitivity
SP	Specificity
SVM	Support vector machine
TAP	Transporter associated with antigen processing
TCR	T cell receptor
TN	True negative
TP	True positive

Chapter 1: Introduction

Major histocompatibility complex (MHC) molecules play critical roles in adaptive immune responses. They bind a variety of small and medium-sized molecules including short peptide fragments and present them on the surface of antigen-presenting cells for recognition by T cell receptors. Two classes of MHC molecules are responsible for antigen presentation: class I and class II. MHC class I molecules are present in all nucleated cells except neurons, while MHC class II molecules are present in dendritic cells, endothelial cells, monocytes and B-cells for MHC. The presentation of MHC-bound peptides to T cell receptors (Lefranc and Lefranc, 2001) on the surface of T cells is responsible for T cell activation and stimulation of adaptive immune response. Hence, MHC-peptide binding studies are invaluable for designing vaccines and immunotherapeutic strategies for controlling allergic or autoimmune responses.

Pemphigus vulgaris (PV) is a potentially life-threatening form of autoimmune blistering skin disorder due to loss of integrity of normal intercellular attachments within the epidermis and mucosal epithelium. The target antigen of PV, desmoglein (Dsg) 3, is a 130-kDa transmembrane glycoprotein that belongs to the cadherin superfamily of cell adhesion molecules (Amagai *et al*., 1991). In early disease (mucosal PV), patients demonstrate autoimmunity only to Dsg3 and develop mucosal blisters; while at the later stage (mucocutaneous PV), patients exhibit non-cross-reactive immunity to both Dsg3 and Dsg1 (Salato *et al*., 2005). Strong association of PV to the MHC class II alleles have been reported in the literature (Ahmed *et* *al*., 1990, 1991; Scharf *et al*., 1989; Sinha *et al*., 1988). However, much remains unknown with regards to the exact nature of the interactions between Dsg3 derived peptides and PV-implicated alleles. Improved understanding of peptide binding to the disease-implicated alleles is important for elucidating its role in disease progression.

Bioinformatic tools are now a standard methodology in facilitating T cell epitope discovery (Schirle *et al*., 2001; Yu *et al*., 2002, Srinivasan *et al*., 2004). Computational methods for predicting MHC-binding peptides include procedures based on sequence motifs (Wucherpfennig *et al*., 1995), quantitative matrices (Parker *et al*., 1994; Davenport *et al*., 1995; Gulukota *et al*., 1997), decision trees (Savoie *et al*., 1999; Segal *et al*., 2001), artificial neural networks (ANNs) (Brusic *et al*., 1994, 1998), hidden Markov models (HMMs) (Mamitsuka, 1998) and support vector machines (SVMs) (Dönnes and Elofsson, 2002; Bhasin and Raghava, 2004; Bozic *et al*., 2005). Despite recent advances in sequence-based predictive techniques, effective computational models for MHC class II molecules are still lacking. This deficiency is attributed to the lack of training data as well as the presence of register shifts and polymorphisms in the binding registers. Up to now, few prediction techniques for MHC class II molecules have been developed using three-dimensional models due to complexities in development as the dual issues of model quality and discriminative technique must be addressed.

1.1 Research issues investigated in this thesis

There is a need to characterize disease-implicated antigens efficiently and speedily to gain information for a global perspective in experimental design.

As experimental determination of binding peptides for every disease-implicated allele is prohibitively expensive in terms of labour, time and cost, bioinformatic analysis supports experimental studies by assisting in planning of critical experiments. The specific objectives of this thesis were to focus on alleles that have not been extensively studied. These include the following sub-projects:

1. Build a database of (TCR/) peptide/MHC crystallographic structures.
2. Identify common structural characteristics of peptide/MHC complexes using existing crystallographic data.
3. Develop a fast and efficient protocol for docking peptide ligands to MHC receptors.
4. Develop methodologies for effective discrimination of binding peptides from the background using three-dimensional models of peptide/MHC complexes.
5. Application of research to analyze MHC molecules implicated in the autoimmune disorder pemphigus vulgaris (PV).

1.2 Contributions of this thesis

The author's original contributions in this thesis include:

1) extraction of crystallographic structures of (TCR/) peptide/MHC complexes from the Protein Data Bank (PDB) (Berman *et al*., 2000). Interaction parameters between bound peptides and their corresponding receptors were computed to facilitate the characterization of peptide/MHC interface. The structures and

computed parameters are deposited into a new database termed MHC-Peptide Interaction Database version T (MPID-T).

2) identification of the existence of different MHC supertype structural interaction characteristics using existing crystallographic structures.
3) development of a new docking protocol to model the bound conformation of peptide ligands to both MHC class I and class II molecules. The predictive performance of the protocol was validated by existing experimental data.
4) development of a complementary scoring scheme for functional prediction of MHC-binding peptides. This approach has been successfully applied to identify peptide binders which lack conserved binding motifs.
5) identification of multiple binding registers within a candidate class II binding peptide, supporting existing evidence that the ability of a candidate peptide to bind a specific class II allele is affected by both the core and the flanking peptide residues.
6) discrimination of PV-implicated alleles from non-implicated alleles using structural principles. By focusing on known peptide binders, this bioinformatic approach successfully discriminated alleles implicated in PV from those that are not.
7) recognition of multiple initial epitopes to be responsible for disease initiation and progression in PV. At the present time, much remains unknown with regards to disease progression in PV.
8) identification of T cell epitope repertoire of Dsg3 glycoprotein for both DRB1*0402 and DQB1*0503. The predictive performance of

the protocol was validated by existing experimental Dsg3 binding data. At present, few Dsg3 epitopes for both alleles have been identified.

9) discovery of similar specificities for binding peptides in DRB1*0402 and DQB1*0503, but with different core recognition regions.

1.3 A summary of this thesis

This thesis is divided into eight chapters. Chapter 1 provides an introduction to the problems in identifying peptide epitopes in autoimmunity-implicated alleles with specific reference to PV, and the research issues investigated in this thesis. This is followed by a literature survey (Chapter 2) on MHC biology and diversity; the complexities involved in identifying T cell epitopes, and existing bioinformatic resources and applications that are available for the study of MHC molecules and prediction of T cell epitopes.

Crystallographic structures of (TCR/) peptide/MHC complexes were extracted from the Protein Data Bank (PDB) (Berman *et al*., 2000) and deposited into a new database termed MHC-Peptide Interaction Database version T (MPID-T; Chapter 3). The collected crystallographic structures were systematically clustered into superfamilies and analyzed for conserved structural interaction characteristics.

A new generic protocol for docking peptide ligands to MHC class I and class II receptors is described in Chapter 4. This procedure forms the basis for the prediction of peptides that will bind to specific MHC alleles and hence facilitate the design of peptide vaccines. The accuracy of the docking protocol was assessed against a large dataset of non-redundant peptide/MHC complexes in which three-dimensional information is available.

Chapter 5 describes the structural analysis of ten PV associated, non-associated and protective MHC class II receptors (DR4: DRB1*0401, *0402, *0404, *0406, DR6 (also classified now as DR14): DRB1*1401, *1404, *1405, DQ2: DQB1*0201, *0202 and DQ5: DQB1*0503) in an attempt to understand the functional correlation between MHC class II alleles and PV. Nine previously identified epitopes capable of stimulating patient derived T cells, were docked into the binding groove of each model to analyze the structural aspects of allele-specific binding. The results of this study indicate that the perfect fitting of a binding register within the binding groove of MHC class II alleles may not guarantee perfect fitting of the entire peptide. In addition, the results also indicate that flanking residues outside the binding groove appear to play a critical role in peptide selection. The PV-implicated alleles DRB1*0402 and DQB1*0503 share similar binding specificities and no single epitope may be responsible for both disease initiation and propagation in PV. In addition, it is discovered that the protective alleles DQB1*0201, *0202 may be capable of binding to most peptides with greater affinity than PV susceptible alleles, allowing for efficient deletion of autoreactive T cells.

Chapter 6 details the development of a complementary scoring function for functional prediction of MHC class II binding peptides. High prediction accuracy of MHC class II binding peptides was validated by experimental biochemical and functional data. This approach successfully identified peptide binders which lack conserved binding motifs. Further analysis of the binding characteristics of class II binding peptides revealed the possible existence of multiple binding registers within a candidate class II binding peptide, suggesting recognition via flexible fitting may play a critical role in binding to class II alleles.

The developed docking protocol and its complementary scoring functions served as the basis for further analysis of DRB1*0402- and DQB1*0503-specific T cell epitope repertoire of the Dsg3 autoantigen (Chapter 7). Interestingly, the T cell epitope repertoire of DRB1*0402 and DQB1*0503 exhibit extensive overlap, indicating the existence of multiple initial immunodominant epitopes responsible for both disease initiation and propagation in PV. Further analysis on the high level of cross-reactivities between DRB1*0402 and DQB1*0503 revealed that both alleles share similar specificities by binding peptides at different binding registers. The inability of some Dsg3 peptides to be recognized by autoreactive cells may be at the level of T cell recognition rather than the level of epitope selection by MHC molecules as a result of clonal deletion or anergic response.

Chapter 8 draws conclusions from the bioinformatic-based approach to peptide/MHC analysis and also discusses future directions. The work presented in this thesis has been published in a series of journal articles and book chapters including a review on state-of-the-art techniques for predicting immunogenic epitopes (Chapter 2); the development of an interaction database for (TCR/) peptide/MHC crystallographic structures (Chapter 3); Chapter 4 where the peptide/MHC docking protocol was developed; Chapter 5 where the difference in structural organization of the binding grooves of ten PV associated, non-associated and protective alleles is discussed; Chapter 6 where a scoring function for MHC class II allele was developed; Chapter 7 where large-scale screening of Dsg3 epitopes for PV-implicated alleles was performed. In addition, several other research works have been published during the course of study. These include – a review on the application of HMM on computational biology, and the development of an automated comparative modeling server for small disulphide-bonded proteins.

Chapter 2: Literature Survey

2.1 Introduction

MHC molecules are cell surface glycoproteins that play a vital role in the adaptive immune responses (Rammensee *et al*., 1993). Two classes of MHC molecules (class I and II) are responsible for antigen presentation. MHC class I molecules are synthesized in the endoplasmic reticulum (ER) and are present on the surface of virtually all nucleated cells, except neurons, in human. Similarly, MHC class II molecules are also synthesized in the ER but are present only in specific antigen presenting cells such as dendritic cells, endothelial cells, monocytes and B-cells. In order to help stimulate immune responses against a large repertoire of possible pathogens, MHC receptors can bind to a wide variety of peptides. The interaction of peptide/MHC complexes with T cell receptors (Lefranc and Lefranc, 2001) on the surface of T cells is responsible for T cell activation and stimulation of adaptive immune response. Hence, knowledge of the structure and biosynthesis of MHC molecules is fundamental to understanding how T cells recognize foreign antigens. This chapter introduces the basics of MHC biology and diversity; its relevance in clinical medicine and the complexities involved in the identification of potential peptides that bind to specific MHC molecules.

2.2 Discovery of the MHC

The discovery of the MHC dates back to 1936 by Peter Gorer where he identified a blood group locus in mice and showed that blood type segregated with susceptibility and resistance to a transplantable tumor (Gorer, 1937). This was the first case of individual identification of a histocompatibility locus. Subsequently, George Snell introduced the term histocompatibility (H) antigen to describe antigens provoking graft rejection and demonstrated that, of all the potential H antigens, differences at the H-2 locus provoked the strongest graft rejection seen among various mouse strains. The designation MHC was not introduced until the early 1970s, when it became known that systems genetically homologous to H-2 existed in many other vertebrates.

2.3 Genetic organization of the MHC

The MHC genes in human, termed human leukocyte antigen (HLA) are found on chromosome 6. Today, HLA is organized into three major genetic regions or loci designated class I, II and III. Class III genes primarily encode components of the serum complement system. Class I and class II loci, on the other hand, encode a number of highly polymorphic cell-surface proteins responsible for antigen presentation. The HLA class I locus is subdivided into HLA-A, -B, and -C sub-regions, each encoding class I α chain genes. The class II HLA locus, HLA-D, is sub-divided into at least six sub-regions, namely HLA-DR, -DQ, -DP, -DX, -DO, and -DZ. The class I and class II genes are highly polymorphic genes in the human genome; for some of these genes over 200 allelic variants have been identified. HLA specificities are identified by an identifier for locus and a number (e.g. A1, DR4, and DQ5) and the haplotypes are identified by individual specificities. Specificities which are defined by genomic analysis are names beginning with an identifier for the locus followed by a four digit code (e.g. A*0101, Cw*0401, and DRB1*0503). Despite considerable MHC polymorphism, a single individual expresses a finite number of MHC alleles and is heterozygous for each MHC gene in humans. The work presented here focus on MHC molecules that are responsible for antigen presentation. Therefore, the use of MHC for the rest of the text is restricted to only the class I and class II genes.

2.4 Structure of the MHC

MHC class I and class II molecules assist in immune surveillance by presenting peptide fragments of potential antigens to circulating T cells. In general, all MHC molecules share certain structural characteristics that are critical for their role in peptide display and recognition by T cells. T cell recognition of antigen is said to be MHC restricted, as T cell receptors (TCRs) will only bind to fragments of antigen that are associated with products of the MHC. Each MHC molecule contains an extracellular peptide-binding cleft which is composed of paired α-helices resting on a floor consisting of an eight-stranded anti-parallel β-sheet. This portion of the MHC molecule binds antigenic peptides for display to T cells, and the TCRs interact with the displayed peptide and with the helices of the MHC molecules. The amino acid residues located in and around this cleft are highly polymorphic and they are responsible for the different peptide binding specificities among different MHC alleles. A non-polymorphic determinant on the MHC molecules acts as the binding site for the T cell co-receptor molecules CD4 and CD8. CD4 and CD8 are expressed on distinct subpopulations of mature T cells and together with the antigen receptors, participate in the recognition of antigen. CD8 binds selectively to class I MHC molecules, and CD4 binds to class II MHC molecules. Hence, CD8$^+$ T cells recognize only peptides displayed by class I molecules, and CD4$^+$ T cells recognize only peptides presented by class II molecules. Most CD8$^+$ T cells function as cytotoxic T cells and CD4$^+$ cells are helper cells.

2.4.1 MHC class I molecules

MHC class I molecules are ternary complexes composed of a heavy glycosylated transmembrane protein non-covalently linked to a smaller polypeptide β_2-microglobulin (β_2m). The complete molecule has four globular domains; three formed by the heavy chain (α_1, α_2, α_3) and one by β_2m as shown in Figure 1. Both the α_1 and α_2 domains adopt similar structure: starting from the N-terminus each region of the chain forms four anti-parallel β-strands followed by a helical region across the β-strands on one side of the β-sheet. The two domains associate in such a way that their β-sheets are hydrogen-bonded to each other forming a platform of a continuous eight-stranded anti-parallel β-sheet. The *β*-sheet is relatively flat with a small propeller twist. The sides of this cleft are formed by two α-helices, one from α_1 and one from α_2. It is within this cleft that antigen fragments are held and presented to T cells. The α_3 domain consists of a transmembrane segment and a short cytoplasmic tail that anchors the molecule in the membrane.

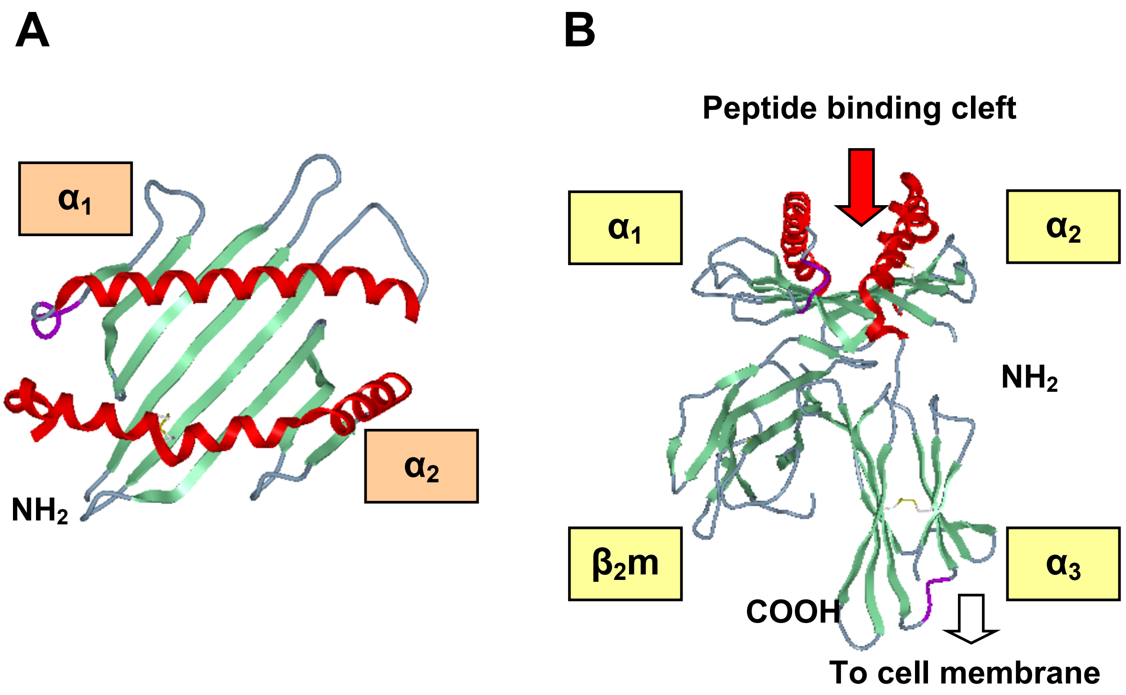


Figure 1 Schematic of MHC class I structure. (A) Top view of class I molecule (HLA-A*0201) based on X-ray crystallographic structure. (B) Side view of the same molecule clearly showing the anatomy of the peptide binding cleft formed by α-helices (red) sitting on a platform of a β-sheet (green).

2.4.2 MHC class II molecules

MHC class II molecules are also transmembrane glycoproteins, consisting of two polypeptide chains (α, β) held together by non-covalent interactions. Similar to class I MHC, the complete class II MHC molecule has four globular domains, two on each chain (α_1, $\alpha_{2,}$ β_1, β_2). The α_1 and β_1 domains mimic the class I α_1 and α_2 domains in forming a peptide binding groove bounded by two α-helices and a β-sheet floor (Figure 2).

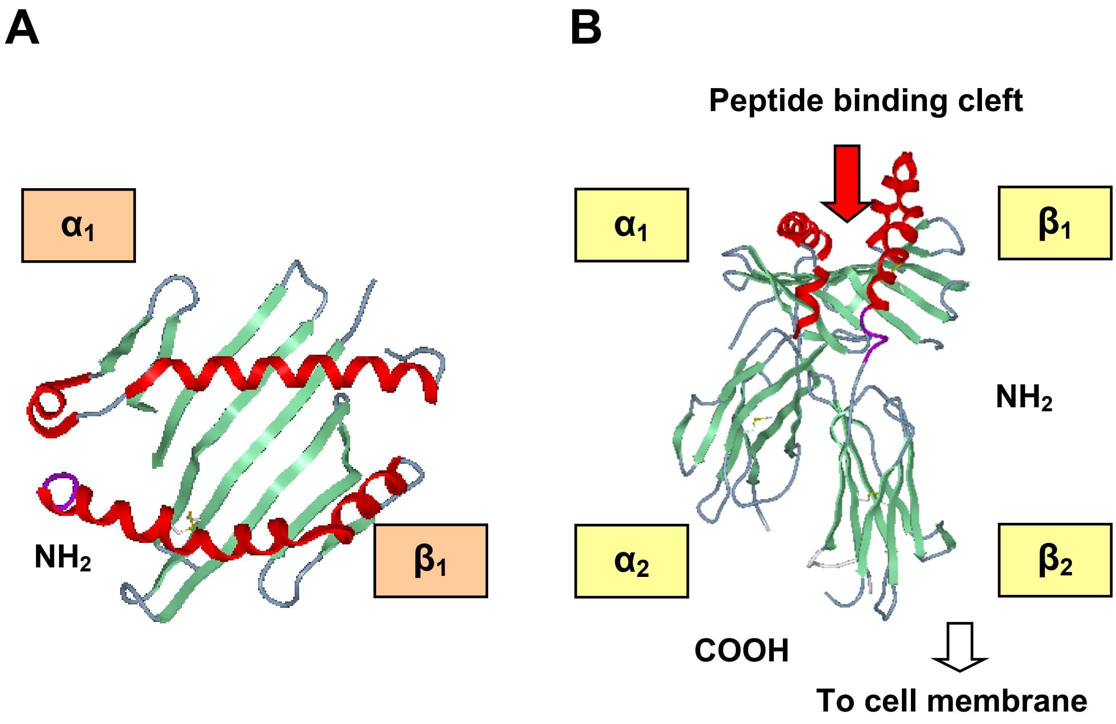


Figure 2 Schematic of MHC class II structure. (A) Top view of MHC class II molecule (HLA-DQB1*0302) based on X-ray crystallographic structure. (B) Side view of the same molecule clearly showing the anatomy of the peptide binding cleft formed by α-helices (red) sitting on a platform of a β-sheet (green).

2.5 Function of the MHC

2.5.1 MHC class I molecules

The class I MHC-restricted antigen processing and presentation pathway provides a sophisticated surveillance mechanism aimed at detecting viral infections in cells. MHC class I molecules are synthesized in the endoplasmic reticulum (ER) and are present on the surface of virtually all nucleated cells, except neurons, in human. Their function is to bind peptides derived from endogenous antigens within the cell, transport them to the cell surface, and present the bound peptide ligands to cytotoxic T cells through the TCR and

CD8. Most class I peptide ligands are derived from proteins that are degraded by proteasomes. The proteasome has broad substate specificity and can generate a wide variety of peptides from cytosolic proteins. Exposure of cells to interferon (IFN)-γ induces the synthesis of three proteolytic proteasome subunits − low molecular weight proteins (LMP)-2, LMP-7, and multicatalytic endopeptidase complex (MECL)-1 − which are incorporated into an alternative form of proteasome, called immunoproteasome, displacing the constitutive subunits β1, β2, and β5, respectively (Palmowski *et al.*, 2006). At the present time, much remains unknown regarding how the products of such endopeptidase activity are related to the final MHC class I ligands. One possibility is that the proteasomes directly produce peptides of appropriate size. Alternatively, the proteasomes may generate longer peptides that require further processing. It is also possible that two short non-continuous peptide fragment can be fused together to create the final class I ligand via post-translational protein splicing (Hanada *et al.*, 2004). In any case, majority of these peptides are transported from the cytosol into the ER by the transporter associated with antigen processing (TAP) (Yewdell *et al.*, 2003). TAP consists of two structurally related subunits, which interact to form a functional peptide-transporting complex. Before peptide translocation by TAP, peptides bind to the membrane-proximal, cytosolic surface of TAP1/TAP2 complexes (Androlewicz *et al.*, 1994). Hydrolysis of ATP results in peptide translocation into the ER lumen (Momburg and Hämmerling, 1998). Within the ER lumen, precursor peptides may be further trimmed by an ER-resident amino peptidase ERAAP (ER aminopeptidase associated with antigen processing) before loading onto MHC class I molecules (Hammer *et al.*, 2005). The class I peptide/MHC complexes eventually exit the ER by association with B cell-associated protein Bap31 (Spiliotis *et al*., 2000).

2.5.2 MHC class II molecules

Similar to MHC class I molecules, MHC class II molecules are also synthesized in the ER. In addition to the polypeptide α and β chains, an invariant chain (Ii) is also produced within the ER, which associates with MHC class II molecules before they reach the cell surface (Cresswell, 1994). Unlike MHC class I expression, which encompasses most cells, class II MHC expression is limited to specific antigen presenting cells (APCs) such as dendritic cells, endothelial cells, monocytes and B-cells. They present exogenous peptide antigens to helper T cells through the TCR and CD4. Exogenous foreign antigen is processed through the MHC class II pathway. Antigen is internalized and degraded enzymatically in endosomes and lysosomes into peptide fragments. MHC class II molecules remain competent for peptide loading by binding fragments of Ii in the ER. These fragments remain bound while Ii targets the MHC class II molecule to a lysosomal-like compartment termed MHC class II compartment (MIIC) (Peters *et al*., 1995; Rudensky *et al*., 1994). Within the MIIC, the Ii is removed from MHC class II molecules by the combined action of proteolytic enzymes and HLA-DM molecule, and the peptides are able to bind to the available peptide binding clefts of the class II molecules. Newly loaded class II molecules are subsequently translocated to the surface of APCs where their interactions with helper T cells stimulate effector response by the production of cytokines.

2.6 Binding of peptides to MHC

2.6.1 MHC class I molecules

The binding of peptides to MHC class I molecules is a non-covalent interaction mediated by residues both in the peptides and in the clefts of the MHC molecules. Peptides of seven to fourteen residues bind to class I MHC molecules in an extended conformation. The amino-acid residues of a peptide may contain side-chains that fit into polymorphic cavities (or ‘pockets’) and bind to complementary amino acids in the MHC molecule. These residues are called anchor residues because they ‘anchor’ the peptide firmly in the MHC binding cleft and contribute most of the favorable interactions of the binding. There are typically two anchor residues at the second (or fifth) and final peptide position. The termini of the peptide are buried deep in the cleft and are bound by a set of conserved hydrogen bonds (Madden *et al*., 1992). Interestingly, this arrangement does not limit the length of the peptide. Longer peptides may zigzag (Madden *et al*., 1993) or bulge (Collins *et al*., 1995; Guo *et al*., 1992) to allow peptides of greater length to maintain the relative position of the termini, and peptides without the presence of canonical anchors have also been discovered to bind with high avidity to their respective MHC molecules (Chen *et al*., 1994; Jameson and Bevan, 1992; Ruppert *et al*., 1993; Doytchinova *et al*., 2004a). The peptide-binding cleft can be subdivided into various pockets (A to F) (Garrett *et al*., 1989). The polymorphic residues that line the peptide-binding cleft determine the individual specificity of peptide/MHC interaction.

2.6.2 MHC class II molecules

Unlike class I, where the allele-independent hydrogen bonding to the peptide is focused at the N- and C- termini, the class II MHC forms hydrogen bond along the entire length of the peptide with links to the atoms forming the main chain. The open nature of the class II binding cleft places no constraint on the length of the peptide, which can extend out of the binding cleft unlike class I ligand site. Thus, each class II molecule can accommodate peptides with a spectrum of lengths ranging from nine to thirty amino acid residues. Similar to class I molecules, the peptide-binding cleft of class II molecules can be subdivided into a series of pockets (1 to 9) (Stern and Wiley, 1994; Murthy and Stern, 1997).

2.7 Relevance for clinical medicine

Many common and severe diseases depend on the function of the cellular immune system and consequently on its mechanisms for specific recognition. Although this naturally applies to infectious diseases, this is also true for a number of autoimmune disorders and chronic inflammatory conditions such as rheumatic diseases, diabetes and multiple sclerosis. It is estimated that between 1–5% of all peptides can bind a particular MHC molecule (Brusic and Zeleznikow, 1999). Where infectious diseases are concerned, clear knowledge of allele specific disease-inducing peptides as relevant T cell epitopes provides a better platform for the construction of new vaccines; one can ascertain exactly which parts of a microorganism are recognized by the cellular immune system and specifically focus the production of vaccine on those regions (Buus, 1999; Corradin and Demotz, 1997; Uebel and Tampe, 1999; Ferrari *et al*., 2000).

In many autoimmune disorders, better explanations have been provided for the associations between disease susceptibility and the histocompatibility antigen type carried by an individual (Singh, 2000). Detailed understanding of the binding specificities of relevant alleles implicated in disease facilitates the development of immunotherapeutic strategies for selectively diminishing or altering immune reactions that may a central role in autoimmune disorders. As strides have been made in identifying auto-antigens capable of provoking disease, the use of immunogenic epitopes of these antigens to prevent or ameliorate disease has also been reported (Evavold *et al*., 1993; Bielekova and Martin, 2001).

2.8 Complexities in identifying T cell epitopes

The identification of T cell epitopes is beset with a number of inherent difficulties. Complexities to be addressed include high polymorphism of HLA alleles, as well as allele specificity of candidate peptides (Williams 2001). Within the human population there is a great diversity of HLA genes with more than 2745 known variants identified as of February 2007 (http://www.anthonynolan.org.uk/HIG/). Binding studies show that each HLA allele has a unique spectrum of binding specificities to a restricted set of peptides and efficient peptide/MHC binding is required for immunogenicity (Sette *et al*., 1994). However, there is also evidence indicating efficient peptide/MHC binding does not guarantee immunogenicity (Feltkamp *et al*., 1994). Thus, binding of antigenic peptides to specific MHC alleles is an important rate-limiting step in T cell activation. Wet-lab verification of binding peptides for every allele is a time-consuming and costly process; and not applicable to studies involving large numbers of protein sequences (Doytchinova and Flower, 2005). There is presently a limited set of experimental data on HLA-binding peptides. For the majority of HLA variants experimental data do not exist at all. Furthermore, imprecision, errors, and biases are prevalent in existing experimental data. Computing approaches with tolerance for imprecision, uncertainty and partial truth are in great demand to accelerate the T cell epitope discovery process (Yu *et al*., 2002; Schirle *et al*., 2001).

2.9 Bioinformatic resources for peptide/MHC interactions

In recent years, bioinformatic tools modeling the immune system network have played an instrumental role in advancing peptide vaccine discovery, with reported successes in melanoma (Roberts *et al*., 2006), multiple sclerosis (Bourdette *et al*., 2005), malaria (Lopez *et al*., 2001) and anti-tumor vaccines (Knutson *et al*., 2001). The availability of general and specialized boutique databases have enabled the development of bioinformatic tools for the analysis and prediction of peptide/MHC interactions. The most important databases are discussed next with some of the implications they have for the study of peptide/MHC interactions.

2.9.1 General databases

2.9.1.1 Swiss-Prot

Swiss-Prot (Bairoch *et al*., 1998, 2004) is a protein sequence database that endeavors to provide high quality annotation through manual curation with minimum redundancy. As of February 2007, Swiss-Prot contains 257,964 records totally 93,947,433 amino acids. Swiss-Prot records are deposited by biologist and further validated by domain experts. As a result of manual curation, the coverage of Swiss-Prot is not as wide as one would hope for. Therefore to counter this limitation, TrEMBL (Translated EMBL) was created as a computer-annotated supplement to Swiss-Prot (Bairoch and Apweiler, 1998). This supplement consists of all translation of EMBL nucleotide sequences that are not available in Swiss-Prot. As of February 2007, TrEMBL consists of 3,745,801 records encompassing 1,218,084,224 amino acids.

2.9.1.2 Protein Data Bank (PDB)

PDB (http://www.rcsb.org/pdb/) (Berman *et al*., 2000) is the single worldwide archive of structural data of biological macromolecules. It contains structures of proteins, nucleic acids, and a few carbohydrates. The PDB assigns a four-character identifier to each structure deposited. The first character is a number from 1–9. In many cases several entries correspond to one protein, either solved in different states of ligation, or in different crystal forms, or re-solved using better crystals or more accurate data collection techniques. As of February 2007, PDB contains a total of 41,527 structures.

2.9.2 Specialized databases

2.9.2.1 IMmunoGeneTics HLA (IMGT/HLA) Sequence Database

The IMGT/HLA Sequence Database (http://www.ebi.ac.uk/imgt/hla/) is a specialist database for HLA sequences and includes the official sequences for the WHO HLA Nomenclature Committee for Factors of the HLA System. As of February 2007, the database contains 2,745 allele sequences as well as detailed information concerning the material from which the sequence was derived and data on the validation of the sequences. Additionally IMGT/HLA also publishes monthly HLA nomenclature updates both in journals and online.

2.9.2.2 NCBI dbMHC

NCBI dbMHC database (http://www.ncbi.nih.gov/mhc/MHC.cgi?cmd=init) was designed to provide a platform where the HLA community can submit, edit, view, and exchange MHC data. It currently consists of an interactive Alignment Viewer for HLA and related genes, an MHC microsatellite database, a sequence interpretation site for Sequencing Based Typing (SBT), and a Primer/Probe database. The MHC database is fully integrated with other NCBI resources, and provides links to the IMGT/HLA database.

2.9.2.3 MHCPEP

MHCPEP (http://wehih.wehi.edu.au/mhcpep/) is a manually curated database that contains more than 13,000 experimentally validated MHC-binding peptide sequences (Brusic *et al*., 1998a). Entries are compiled from published reports and direct submissions of experimental data. Each record contains the peptide sequence, its MHC specificity and where available, experimental method, observed activity, binding affinity, source protein and anchor positions, as well as publication references.

2.9.2.4 MHCBN

MHCBN (http://bioinformatics.uams.edu/mirror/mhcbn/index.html) is a database of MHC binding and non-binding peptides compiled from published literature and existing databases (Bhasin *et al.*, 2003). As of February 2007, the database contains 20,717 MHC binders and 4,022 MHC non-binders for over 400 MHC molecules. The database also contains other information such as TAP binding and non-binding peptides, as well as sequence and structure data of source proteins of peptides and MHC molecules. MHCBN also provides hypertext links to major databases including Swiss-Prot, PDB, IMGT/HLA, and PubMed, among others.

2.9.2.5 AntiJen

AntiJen (http://www.jenner.ac.uk/antijen) is a database containing experimentally determined quantitative binding data for MHC-binding ligands, T cell epitopes, and TAP-binding peptides, among others (Toseland *et al.*, 2005). The database also archives continuous quantitative data on a variety of immunological molecular interactions including thermodynamic and kinetic measures of peptide binding to TAP and MHC, peptide/MHC complexes binding to T cell receptors, antibodies binding to protein antigens and general immunological protein-protein interactions. As of February 2007, the database contains over 24,000 entries.

2.9.2.6 SYFPEITHI

SYFPEITHI (http://www.syfpeithi.de) is a database for MHC class I and class II ligands and peptide motifs of humans and other species, such as apes, cattle, chicken, and mouse obtained from published data (Rammensee *et al*., 1999). All motifs currently available are accessible as individual entries. Searches for MHC alleles, MHC motifs, natural ligands, T cell epitopes, source proteins/organisms and references are possible. The database includes hyperlinks to EMBL and PubMed as well as ligand predictions for a number of MHC alleles.

2.10 Computational methods for predicting T cell epitopes

Two main categories of specialized bioinformatic tools are available for prediction of MHC-binding peptides – methods based on identifying patterns in sequences of binding peptides, and those that employ three-dimensional (3D) structures to model peptide/MHC interactions. The first group includes procedures based on binding motifs, matrices, decision trees, artificial neural networks, hidden Markov models and support vector machines. In contrast, the second category corresponds to techniques with distinct theoretical lineage and includes the use of homology modeling, docking, 3D-QSAR and 3D threading techniques. This section provides an overview of these methods, their strengths and their weaknesses.

2.10.1 Sequence-based approach

2.10.1.1 Discovery of anchor residues and sequence motifs

The earliest attempt to predict MHC-binding peptides started with the discovery that peptides binding to specific MHC alleles are functionally related and share residues with similar properties at various positions of their primary sequences. Class I and class II binding peptides contain residues with side-chains that fit into polymorphic cavities (or 'pockets') and bind to complementary residues of specific MHC alleles. These residues are called anchor residues because they 'anchor' the peptides firmly at various positions in the MHC binding cleft (Falk *et al*., 1991a,b; Jardetzky *et al*., 1991; Hunt *et al*., 1992) and contribute to most of the binding interactions. This led to the definition of "peptide motif" (Falk *et al*., 1991b; Roetzschke *et al*., 1991) for an array of class I and class II alleles. Numerous research groups, including Zhang *et al*. (1993), Lipford *et al*. (1993), Sette *et al*. (1993), Sidney *et al*. (1994), Parker *et al*. (1994), Hammer *et al*. (1994), Rammensee *et al*. (1995), Meister *et al*. (1995), D'Amaro *et al*. (1995) and Rajapakse *et al*. (2006) developed computational tools that scan peptides that fit these motifs.

It was later discovered that residues along other positions of a peptide also play a vital role to binding and sequence motifs alone are inadequate to account for the comprehensive binding ability of a candidate peptide (Chen *et al*., 1994; Jameson and Bevan, 1992; Ruppert *et al*., 1993; Doytchinova *et al*., 2004a). Immunodominant peptides without the required binding motifs were identified (Scott *et al*., 1998) and not all motif-conforming peptides do bind to the respective MHC alleles (Martin *et al*., 2003). In an attempt to investigate the role of motifs in binding, Ruppert *et al*. (1993) performed binding assays on peptides which are motif-positive for HLA-A*0201 and found that only about 30% of motif-conforming peptides were actual binders. In practice, binding motif models have proven to be both non-sensitive and non-specific (Martin *et al*., 2003). This approach fails to detect binders not conforming to existing motifs and includes non-binding sequences that fit the required patterns (Meister *et al*., 1995). However, despite these limitations, this approach is still a useful alternative to random guessing or use of a complete overlapping set of peptides for selection of candidate binders (Yu *et al*., 2002).

2.10.1.2 Binding matrices

Binding matrices represent an enhancement of simple motif models by correlating peptide residue positions to binding. This approach employs the use of tables containing *l*×20 coefficients where *l* corresponds to the length of the binding motif and 20 for each amino acid symbol (Davenport *et al*., 1995; Gulukota *et al*., 1997). Consensus scores are obtained by summing, multiplying or averaging the matrix coefficients and compared against a predetermined threshold. In general, matrices are constructed using amino acid frequencies at different position of known binders or quantitative MHC-binding data. The former indicates the binding likelihood of a peptide sequence to the MHC molecule, while the later provides means of quantifying the peptide binding affinity. Examples of matrices derived from simple counting of amino acid frequencies at different position of known peptide binders include EpiMatrix (Schafer *et al*., 1998) and SYFPEITHI (Rammensee *et al*., 1999), while BIMAS (Parker *et al*., 1994) was developed by fitting of MHC-binding data.

More complex forms of matrix-based models have been developed to detect weak binding patterns and to account for noisy and collinear data. Reche *et al.* (2002) employed the use of position specific scoring matrices from a set of aligned binding peptides to predict binders to an array of MHC class I and II molecules. Peters *et al*. (2003) introduced the use of Stabilized Matrix Method (SMM) as predictor for HLA-A2 binding peptides. Nielsen *et al*. (2004) applied a Gibbs sampler to detect weak sequence motifs in class I and class II binding peptides. Rajapakse *et al.* (2005) utilized a multi-objective evolutional algorithm to identify a consensus motif for I-A^{g7}. Guan *et al*. (2003) and Doytchinova *et al*. (2002) employed the use of multivariate statistics to improve the predictive performance of their matrices. An additive equation was formulated to account for individual amino acid contributions at each position and interactions with neighboring amino acids. The matrix was subsequently solved through the use of partial least square regression.

2.10.1.3 Decision trees

Decision trees are rule-based models that classify patterns using a sequence of well-defined rules (Duda *et al*., 2001). Position-specific binding motifs are converted into rules and embedded within the nodes of a decision tree. The resulting tree structure indicates amino acid properties that are strongly correlated with physicochemical properties of binding peptides. Peptide sequences are threaded through a series of nodes and the result of all node-to-node transitions are used to determine the outcome of prediction. Because of its capability to elucidate both linear and non-linear problems, this approach has been adopted by several groups to identify higher-level rules for binding. Savoie *et al*. (1999) constructed a decision tree using the BONSAI program to investigate T cell preference and adverse motifs for HLA-A*0201 binding peptides. Segal *et al*. (2001) adopted a similar tree-structured technique to predict peptides binding to H2-K^b. An example of a decision tree network is shown in Figure 3.

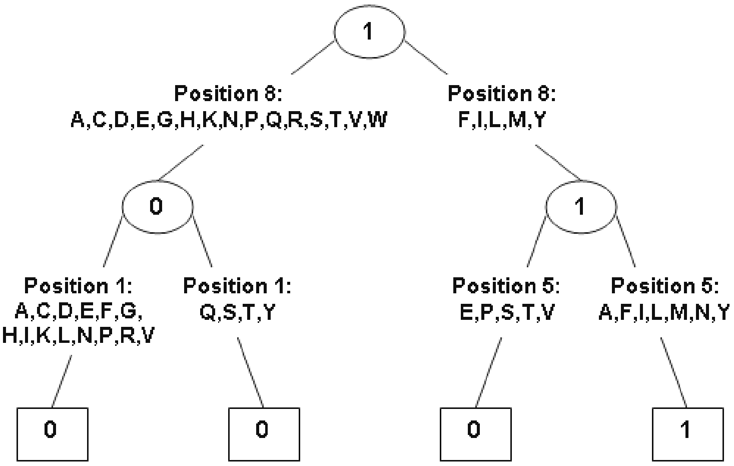


Figure 3 Subset of decision tree network employed by Segal *et al*. (2001). Each node represents grouping of preferential/non-preferential amino acid residues at various positions of H2-K^b binding peptides. Predicted class at each node (ellipses – internal; rectangles – terminal) is given by the 0 (non-binding) or 1 (binding) within each node.

2.10.1.4 Artificial neural networks

Artificial Neural Networks (ANNs) are connectionist models particularly well suited to perform classification and complex pattern recognition tasks (Zurada, 1999). ANNs can encode non-linear data and have been used extensively for prediction of peptide binding to both class I and class II alleles

(Brusic *et al*., 1994; Honeyman *et al*., 1998; Gulukota *et al*., 1997; Adams and Koziol, 1995; Milik *et al*., 1998; Buus *et al*., 2003; Nielsen *et al*., 2003). Peptide features are represented by amino acid descriptors such as composition, hydrophobicity, volume and charge. The descriptors are used to train an ANN for classifying peptides into binders and non-binders. An example of ANN architecture is illustrated in Figure 4. An investigation on the predictive performance of ANNs revealed that this approach gradually outperforms motifs, matrices and hidden Markov models (HMMs) with increasing peptide data (Yu *et al*., 2002). A major drawback of ANN is the requirement of a fixed input length. As such, a given ANN model can only predict binding peptides that are of the same length as those in the training dataset. This constraint restricts the ability of ANN to predict epitopes with length that differ from those used in the trained network.

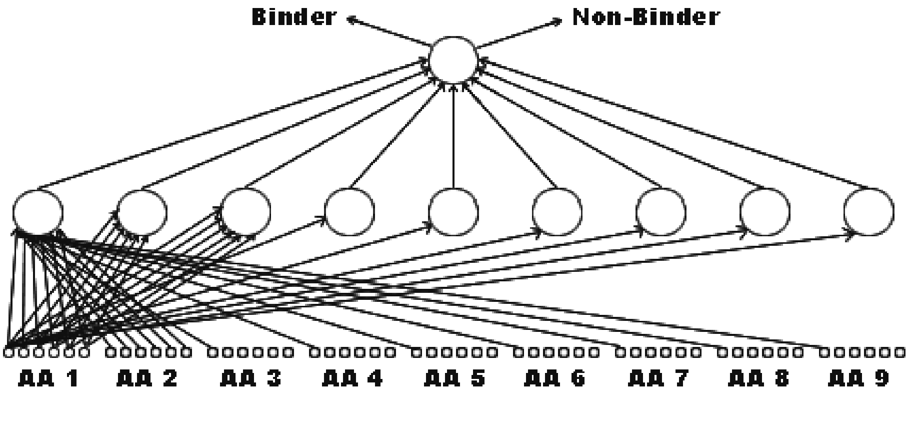


Figure 4 A three-layer ANN for predicting class I binding peptides by Brusic *et al*. (1994). The first layer represents input nodes with the number of nodes corresponding to the length of input peptide; the number of second (hidden) layer nodes equals to the ideal length of binding peptides; and a single output node predicts binding versus non-binding.

Various groups have developed hybrid versions of ANN for peptide/MHC prediction. Nielsen *et al*. (2003) described a combination of a series of neural networks using several sequence coding strategies including a hidden Markov model encoding to improve the predictive power of the system. Brusic *et al*. (1998b) integrates the strength of matrix models and evolutionary algorithm (EA) for processing ANN training set. New alignment matrices were selected by EA based on evolutionary principles. Each parent (matrix) produces two children consisting of an exact copy of itself and a mutant copy, and passes the child with the higher fitness value to the next generation. The highest scoring alignments from the final generation matrices were subsequently fed into ANN for training.

2.10.1.5 Hidden Markov models

Hidden Markov model (HMM) belongs to a type of probabilistic graphical models that have been successfully applied to a wide range of applications in statistical pattern recognition and classification (Rabiner, 1989). In order to overcome the potential limitations of ANNs, HMMs have been applied to predict peptides binding to MHC (Mamitsuka, 1998; Brusic *et al*., 2002). Similar to decision trees and ANNs, HMMs have the ability to cope with non-linear data and are suitable for representing sequences having flexible lengths. Associated with each HMM is a series of discrete-state, time-homologous, first-order Markov chain (MC) with suitable transition probabilities between states and an initial distribution. Each state consists of a discrete or continuous distribution over possible emissions or outputs. These outputs are generated when the particular state is visited or during transition from state to state. Transitions between states follow a set of transition and emission probabilities. The transition probability is the probability of moving from one state to another via a connected edge, and the emission probability is the probability of emitting a particular symbol at a state. The sequences of states underlying the MC are hidden and cannot be observed, hence the name hidden Markov model. The probability of any sequence, given the model, is computed by multiplying the emission and transition probabilities along the path.

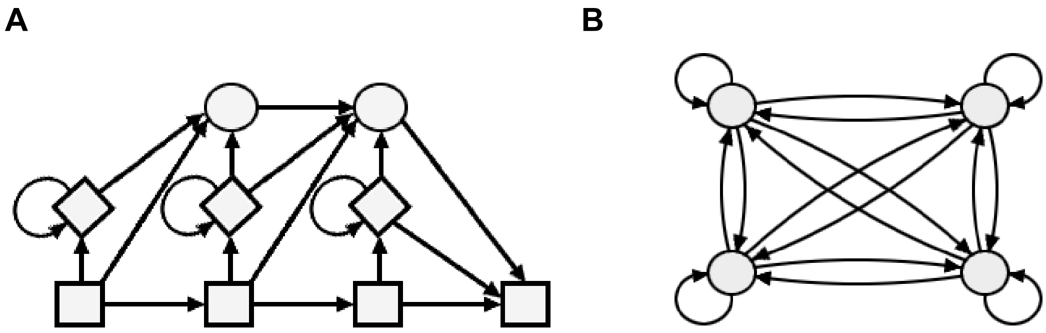


Figure 5 HMM topologies adopted for peptide/MHC prediction by Mamitsuka, (1998). (A) A profile HMM, (B) A fully connected HMM.

The use of HMM for peptide/MHC prediction was first reported in the literature (Mamitsuka, 1998) using two different HMM topologies: profile HMM and fully connected HMM. Profile HMMs (Figure 5a) are linear left-right models where the underlying directed graph is acyclic, with the exception of loops, hence supporting a partial order of the states. The profile HMM architecture (Durbin *et al*., 1998) consists of three classes of states: the match state, the insert state and the delete state; and two sets of parameters: transition probabilities, and emission probabilities. The match and insert states always emit a symbol, whereas the delete states are silent states without emission probabilities. A fully connected HMM (Figure 5b) consists of states that are pairwise connected such that the underlying digraph is complete. There are no distinguished starting and terminating states and the transition matrix does not contain any zero entries with the exception of diagonal entries, which correspond to loops or self-transitions. Because there is no constraint on the structure of a fully connected HMM, this model permits the representation of more than one sequence pattern concealed in the training data.

2.10.1.6 Support vector machines

Support vector machines (SVM) are statistical learning methods based on the structural risk minimization principle (Han *et al*., 2004). Similar to decision tree, ANN and HMM, it has the ability to handle both linear and non-linear data. Every peptide sequence is represented by specific feature vector assembled from encoded representations of residue properties such as amino acid composition, hydrophobicity, polarity, charge, bulkiness and solvent accessibility. Parameters are trained by mapping input vectors into a high dimensional feature space and maximizing the margin between the binders and non-binders with an optimal separating hyperplane. SVM outperforms ANN and decision tree in the absence of large training dataset (Zhao *et al*., 2003) and has been embraced by several groups including Dönnes and Elofsson (2002), Bhasin and Raghava (2004a) and Bozic *et al*. (2005) for predicting class I and class II binding peptides. Hybrid models based on ANN and SVM have also been developed by Bhasin and Raghava (2004b) for consensus and combined prediction of T cell epitopes.

2.10.2 Structure-based approach

2.10.2.1 Protein threading

Protein threading (Akutsu and Sim, 1999) or side-chain conformational search (Sezerman *et al*., 1996) involves computing an alignment between a target amino acid sequence and the spatial positions of a 3D structure. In the context of peptide/MHC modeling, this involves substituting the backbone coordinates of a source peptide (P_1, P_2 … P_n) that is bound to a MHC molecule of interest with the target peptide sequence (S_1, S_2 … S_n) by replacing P_i with S_i. A search for the best side-chain conformations is usually performed, and a scoring scheme is subsequently applied to discriminate the binders from non-binders.

Altuvia *et al*. (1995) demonstrated the use of protein threading to detect binding peptides not conforming to HLA-A*0201 binding motifs using the statistical pairwise potential table of Miyazawa and Jernigan (1985, 1996). This was subsequently extended to the analysis of peptides binding to an array of class I alleles (Altuvia *et al*., 1997; Schueler-Furman *et al*., 1998). This approach successfully identified peptides binding to MHC molecules with hydrophobic binding pockets but not to MHC molecules with hydrophilic, charged pockets. In order to circumvent the problem, Kangueane *et al*. (2000) introduced the use of knowledge-based rules to discriminate binders from non-binders based on the number of observed atomic clashes between the MHC and its bound peptide, and the number of solvent exposed hydrophobic residues on the modeled peptide. The problem was later solved by Schueler-Furman *et al*. (2000) through the use of a different pairwise potential table (Betancourt and Thirumalai, 1999) that described hydrophilic interactions more appropriately.

2.10.2.2 Homology modeling

Homology modeling (Swindells and Thornton, 1991; Sali and Blundell, 1993) employs the use of available homologous protein structure(s) to predict the unknown structure of a related amino acid sequence. In the context of peptide/MHC prediction, the aim is to model the bound conformation of a peptide sequence with an unknown structure given the 3D structure of other bound peptides to homologous MHC molecules. Hammer *et al*. (1995) constructed a series of synthetic peptide/HLA-DRB1*0402 models from HA peptide/HLA-DRB1*0101 crystallographic structure to identify specific patterns of peptide binding. Michielin *et al*. (2000) successfully developed a model of T1 TCR/PbCS/H2-K^d complex based on its homology with the 2C TCR, the A6 TCR/Tax/HLA-A2 complex, the 1934.4 TCR Vα chain, the 14.3.d TCR Vβ chain, and the H2-K^b ovalbumine peptide. Buoyed by the excellent results, Michielin *et al*. (2002) applied the methodology to identify critical residues of the A6 TCR that interacts with peptide/HLA-A2 complex. Rognan *et al*. (1999) and Logean *et al*. (2001) applied a similar two-step approach to construct the bound conformation of peptides to an array of class I alleles. Their modeling procedure begins by selecting peptide termini residues based on homology to the most similar MHC-bound peptide with available crystallographic structure.

The remaining residues were subsequently constructed by satisfaction of spatial restraints using a knowledge-based loop search procedure.

2.10.2.3 Docking

Computer-simulated ligand binding or docking is a powerful technique for investigating intermolecular interactions. In general, the purpose of docking simulation is two-fold – (i) to find the most probable translational, rotational, and conformational juxtaposition of a given ligand-receptor pair, and (ii) to evaluate the relative goodness-of-fit or how well a ligand can bind to the receptor. Several docking techniques have been developed to address the peptide/MHC combinatorial problem. Caflisch *et al*. (1992) developed a combinatorial buildup algorithm to dock the influenza matrix peptide 58-68 to HLA-A*0201. Rosenfeld *et al*. (1993, 1995) utilized a multiple copy algorithm to identify probable termini peptide conformations and constructed the intervening sequence using a loop closure algorithm. Molecular dynamics (MD) simulations have also been applied by various groups (Antes *et al*., 2006; Lim *et al*., 1996; Wan *et al*., 2004; Tzakos *et al*., 2004; Zacharias and Springer, 2004) to simulate the interactions of MHC and its corresponding bound peptides. This approach simulates the motion of a molecule by computing the changes of the atomic coordinates as a function of time. The simulation begins with a static structure, usually corresponding to a low energy conformation of the molecule. At each successful step, the position and the velocities of the atoms are computed using previous coordinates, velocities and accelerations, until predetermined criteria such as energy minima are achieved.

2.10.2.4 3D-QSAR

The three-dimensional quantitative structure-affinity relationship (3D-QSAR) approach employs the use of peptide structures (in the absence of MHC) to predict the affinity of peptides to MHC molecules. Zhihua *et al.* (2004) constructed a A*0201 3D-QSAR model to study the relationship between 3D structural parameters of the HLA-A*0201 binding peptide and the HLA-A*0201/peptide binding affinities. Fickel and del Carpio (2000) applied the methodology to study the structural deviations of HLA-A24 complexes. An alternative discrimination scheme was also introduced by Doytchinova and coworkers that employed similarity indices to study peptides binding to an array of MHC molecules (Doytchinova and Flower, 2001; Guan *et al.*, 2003; Doytchinova *et al.*, 2004a; Doytchinova *et al.*, 2005; Hattotuwagama *et al.*, 2005).

2.11 Computational methods for predicting MHC supertypes

The classification of MHC alleles into supertypes or superfamilies is important for the development of epitope-based vaccine (Sette *et al.* 2001, 2002). By clustering MHC alleles on the basis of their structural features and/or peptide binding specificities, promiscuous T cell epitopes that bind multiple MHC alleles can be identified. Such peptides are key targets for the design of vaccines and immunotherapies because they are applicable to higher proportions of human population. However, experimental determination of binding specificities for even a single MHC allele is an expensive, laborious and time consuming process; and not practical for the study of MHC supertypes which involve large numbers of alleles (Doytchinova and Flower, 2005).

In silico, bioinformatics is emerging as an alternative and viable approach for MHC supertype classification. Two groups of clustering techniques can be recognized in the literature reviewed – methods based on peptide specificities, and those that classify MHC alleles using 3D structural features. This section provides an overview of existing strategies for MHC supertype classification.

2.11.1 Clustering using peptide specificities

A strategy for the development of epitope-based vaccines with wide population coverage is to identify HLA alleles that are present in most individuals from all major ethnic groups and ensuring that these alleles bind to at least one of the peptides in the vaccine. Accordingly, promiscuous peptides that bind more than one HLA allele are ideal for such purpose. By clustering MHC alleles on the basis of their peptide binding specificities, promiscuous T cell epitopes that are representative of large proportion of human population can be identified. Sturniolo *et al*. (1999) demonstrated the use of multiple quantitative matrices for predicting promiscuous peptides binding to HLA-DR alleles. Brusic *et al*. (2002) combined peptide and MHC interaction sequences with a HMM to predict peptide binding to the HLA-A2 supertype. Guan *et al*. (2003) employed the use of 2D-QSAR to investigate peptide specificities to four HLA-A3 alleles and formulated a refined HLA-A3 supertype motif. Lund *et al*. (2004) constructed weight matrices representing the specificities of several HLA-DR alleles as well as all HLA-A and -B alleles in the SYFPEITHI database using a Gibbs sampling procedure. Distance matrices were clustered using the neighbour-joining method of Saitou and Nei (1987). This approach characterized HLA-A, -B and -DR alleles into five, seven and nine clusters respectively according to their peptide binding specificities.

2.11.2 Clustering using MHC structural features

An alternative approach for HLA supertype definition is to identify alleles with similar binding specificities from a structural view point. HLA alleles with similar binding specificities share common structural features within the peptide binding cleft. The binding clefts contain cavities (or anchor “pockets”) that correspond to primary and secondary anchor positions on the binding peptide. Doytchinova *et al*. (2004b, 2005) demonstrated that only one to three amino acids within these binding pockets are sufficient to classify an allele to a particular class I or class II supertype. HLA-A, -B, -C, -DR, -DQ, -DP alleles were subsequently grouped into three, three, two, five DRs, three DQs and four DPs clusters respectively.

2.12 Modeling Issues

The accuracy of a prediction model is highly dependent on the quantity and quality of available experimental data. This section discusses the issues related to peptide data which have implications for the selection and performance of prediction model.

2.12.1 Data Quantity

The availability of known peptide binders to specific alleles has a direct impact on the choice and quality of prediction model. When little or no data are available, structure-based predictive techniques are preferred. However, the development of computational tools under this category is severely impeded by inherent complexities in terms of model building, data fitting and computational speed. As the number of known peptide binders increases, sequence-based predictive techniques become more useful predictors. SVM outperforms ANNs and decision trees using small training dataset of 36 binders and 167 non-binders (Zhao *et al*., 2003). An investigation on the predictive performance of ANNs revealed that this approach gradually outperforms motifs, matrices and HMMs with increasing peptide data (Yu *et al*., 2002). ANN and HMM are the predictive methods of choice for MHC alleles with more than 100 known binders (Yu *et al*., 2002).

2.12.2 Data Quality

Noise and errors in the datasets have an adverse effect on the construction of useful predictive models. Brusic *et al*. (1997) investigated the impact of noise in datasets for constructing matrix-based models. They demonstrated that 5% of errors in a dataset will double the number of data points, relative to a 'clean' dataset, required to build a matrix-based model of a pre-set accuracy. On the contrary, the same magnitude of error does not significantly affect the performance of ANNs due to their ability to handle imperfect or incomplete data (Hammerstrom, 1993).

2.12.3 Data Bias

Overfitting occurs when a predictive model adapts too well to the training data and includes random disturbances in the training set as being significant. As these disturbances do not reflect the underlying distribution, the performance of the machine learning techniques on the given dataset is affected. This overfitting problem is typically avoided by using a regularizer (Karplus, 1995) that replaces the observed amino acid distribution by its estimator.

2.13 Summary

- The functional responses of T cells are initiated by the recognition of peptide/MHC complexes on the surfaces of APCs. Two classes of MHC molecules, class I and class II, are responsible for antigen presentation to TCRs.
- The human form of MHC genes, HLA are highly polymorphic with more than 2000 known variants identified at the present time.
- Each HLA allele has a unique spectrum of binding specificities to a restricted set of peptides and an efficient peptide/MHC binding is required for immunogenicity. However, there is also evidence that indicates that efficient peptide/MHC binding does not guarantee immunogenicity (Feltkamp *et al.*, 1994).
- Detailed understanding of the binding specificities of relevant alleles implicated in disease facilitates the development of epitope-based vaccines and immunotherapeutic strategies.
- The experimental determination of binding peptides for every allele is prohibitively expensive in terms of labour, time and cost.

- At present, there is a limited set of experimental data on HLA-binding peptides. For majority of HLA variants experimental data do not exist.
- The availability of general and specialized boutique databases such as Swiss-Prot, PDB, MHCPEP, MHCBN, IMGT/HLA, NCBI dbMHC, and AntiJen, have facilitated the development of bioinformatic tools for the analysis and prediction of peptide/MHC interactions.
- T cell epitope prediction tools help researchers identify allele-specific binding peptides and reduce the number of peptides to be synthesized and assayed.
- MHC supertype classification tools facilitate the identification of alleles with similar structural features and/or peptide specificities. Such tools are important for the identification of promiscuous epitopes that can bind to multiple MHC alleles.
- Bioinformatic tools for scanning for candidate T cell epitopes from protein antigens help researchers to identify regions with high concentrations of T cell epitopes or immunological ‘hot spots’ and focus upon relevant experiments.
- Data availability has a direct impact on the choice and quality of prediction model. When experimental data is limited or absent, structure-based techniques are preferred. As the dataset increases, sequence-based predictive techniques become more useful predictors.
- The accuracy of a predictive model is highly dependent on the availability of good quality data. Noise and errors in the datasets have an adverse effect on the construction of useful predictive models.
- Overfitting may result in inaccurate modeling of a prediction tool.

Chapter 3: MHC-Peptide Interaction Database version T (MPID-T)

3.1 Introduction

The experimentally determined three-dimensional structures of TCR/peptide/MHC and peptide/MHC complexes are available in the PDB (Berman *et al*., 2000), with some interaction parameters reported as significant for peptide/MHC interactions (Kangueane et al., 2001). A preliminary peptide/MHC interaction database termed MPID (MHC-Peptide Interaction Database) was developed by Govindarajan *et al*. (2003) and consisted of 86 entries of classical peptide/MHC complexes with standard residues derived mainly from human and rodents. Thereafter, new structures have become available in the PDB. In this study, new crystallographic structures are collected and together with existing MPID records, the crystallographic structures and computed interaction parameters are stored in a new database termed MPID-T (MHC-Peptide Interaction Database version T).

MPID-T is a curated structure-derived database containing interaction information on 187 peptide/MHC complexes (represented by 40 human, murine and rat alleles), and 16 TCR/peptide/MHC complexes (13 class I and 3 class II alleles). The database is available at http://surya.bic.nus.edu.sg. Information for each MPID-T entry is classified into four main groups: MHC (allele, source, class), bound peptide (length, source, redundancy), computed interaction parameters (intermolecular hydrogen bonds, gap volume, gap index, interface area), and links to related external databases, particularly to IMGT/3Dstructure-DB (Kaas *et al*., 2004) that provides detailed annotations and expertise on TCR and MHC sequences involved in the 3D structures, and IMGT peptide/MHC contact analysis IMGT Colliers de Perles for TCR/peptide/MHC and TCR/peptide/MHC entries (http://imgt.cines.fr) (Robinson *et al*., 2001). The ultimate purpose of MPID-T is to enhance the understanding of the binding mechanism underlying TCR/peptide/MHC and peptide/MHC interactions by mapping the TCR footprint on the MHC and its bound peptide, as this eventually determines T cell recognition and binding.

3.2 Resource Description

3.2.1 Capabilities

MPID-T is a curated MySQL (http://www.mysql.com) database hosted on a UNIX server (IRIX 6.5, Apache 1.3.12). Currently, MPID-T contains only experimentally determined structures available in the PDB. For PDB entries with multiple molecular assemblies, the first TCR/peptide/MHC or peptide/MHC complex is stored as a single entity, for rapid visualization, characterization and comparison. Each structure is manually verified, classified and analyzed for intermolecular interactions (i) between the MHC and its corresponding bound peptide and (ii) between a TCR and its bound peptide/MHC complex where TCR structural information is available. Included in MPID-T are non-classical structures and complexes with non-standard residues, which have implications for vaccine design. The non-redundant set of peptides bound to a particular allele is selected using the most accurate and complete structures.

3.2.2 Definition of interaction parameters

Specific interaction parameters have been identified as being significant for the characterization of peptide/MHC interface (Kangueane *et al*., 2001; Govindarajan *et al*., 2003) and may also be applied for describing the relatively large and variable interface between peptide/MHC and TCR. These descriptors can be computed from the three-dimensional coordinates of a peptide/MHC complex. These include (i) the number of intermolecular hydrogen bonds, (ii) the interface area between associating molecules, (iii) the gap volume and (iv) the gap index. Although the gap volume is computed as described by Kangueane *et al*. (2001), the accessible surface area (ASA) required for calculating the other three parameters, is now computed using Naccess program (http://wolf.bms.umist.ac.uk/naccess/). A brief outline of the MPID-T interaction parameters follows.

3.2.2.1 Intermolecular hydrogen bonds

The number of intermolecular hydrogen bonds between the bound peptide and MHC molecule was calculated using HBPLUS (McDonald and Thornton, 1994) in which hydrogen bonds are defined in accordance to standard geometric criteria of maximum distances (D–A = 3.9 Ǻ, H–A = 2.5 Ǻ and S–S = 3.0 Ǻ) with minimum angles (D–H–A = 90°, H–A–AA = 90° and D–H–AA = 90°), where participating atoms are represented as D for donor, A for acceptor, H for hydrogen, AA for acceptor antecedent and S for sulphur.

3.2.2.2 Gap volume

The volume enclosed by the MHC molecule and its corresponding bound peptide is calculated using the SURFNET program (Laskowski, 1991). The algorithm places a series of spheres (maximum radius 5.00 Å) midway between the surfaces of each pair of subunit atoms, such that its surface is in contact with the surfaces of the atoms in the pair. The size of each sphere is reduced accordingly whenever it is intercepted by other atoms and subsequently discarded if it falls below a minimum allowed radius (1.00 Å). The gap volume between the two subunits is computed based on the volume enclosed by all the allowable gap-spheres.

3.2.2.3 Gap index

One essential feature in receptor-ligand binding is the electrostatic and geometric complementarity observed between associating molecules. In this study, we adopted the use of gap index (reviewed in Jones and Thornton, 1996) as a means to evaluate complementarity of interacting interfaces between the bound peptide and HLA molecule expressed by equation 1.

$$\text{Gap index (Å)} = \frac{\text{Gap volume between peptide/MHC (Å}^3\text{)}}{\text{Interface ASA (Å}^2\text{) (per complex)}}$$

(*Equation 1*)

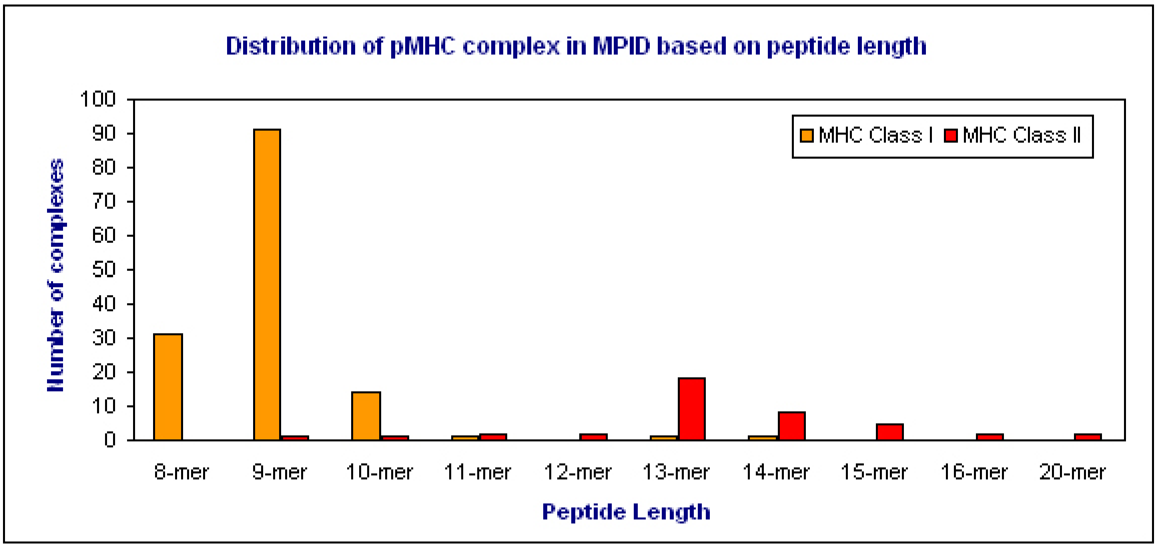


Figure 6 Distribution of peptide/MHC complex in MPID-T based on peptide length shows predominance of 9-mer peptide/MHC complexes.

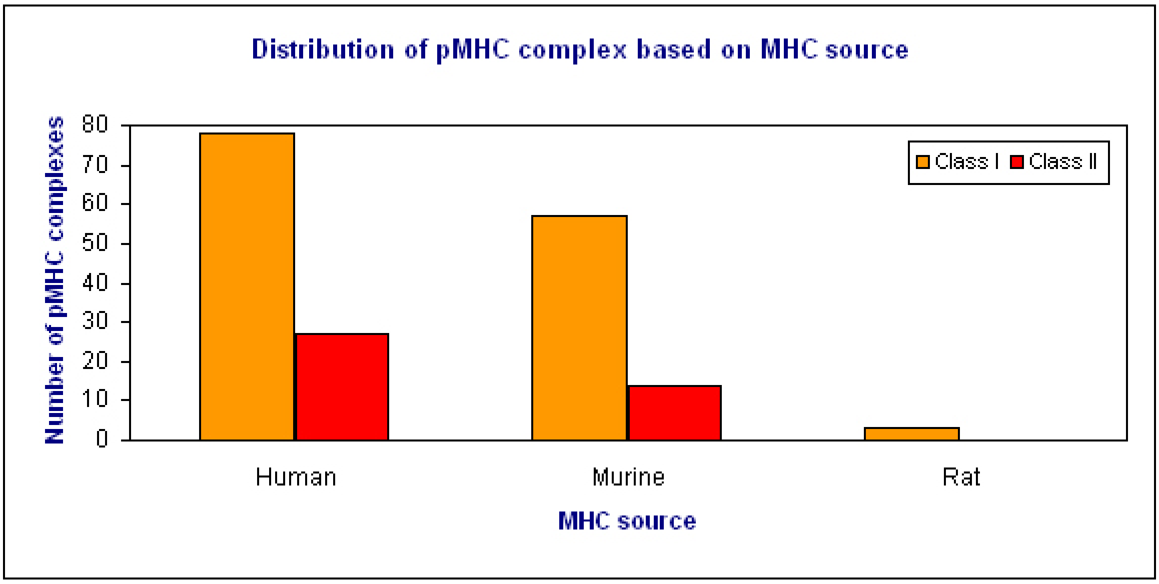


Figure 7 Distribution of peptide/MHC complex in MPID-T based on MHC source contains the maximum of peptide/MHC complex structures from human source.

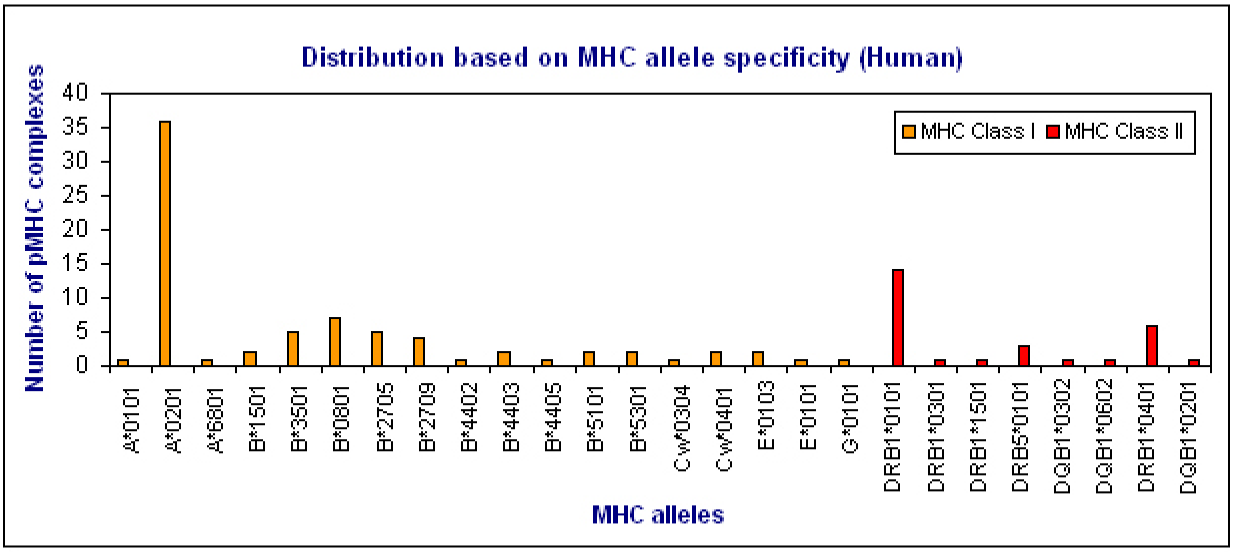


Figure 8 Distribution of human MHC allele in MPID-T shows A*0201 (class I) and DRB1*0101 (class II) have the maximum peptide/MHC complex structures.

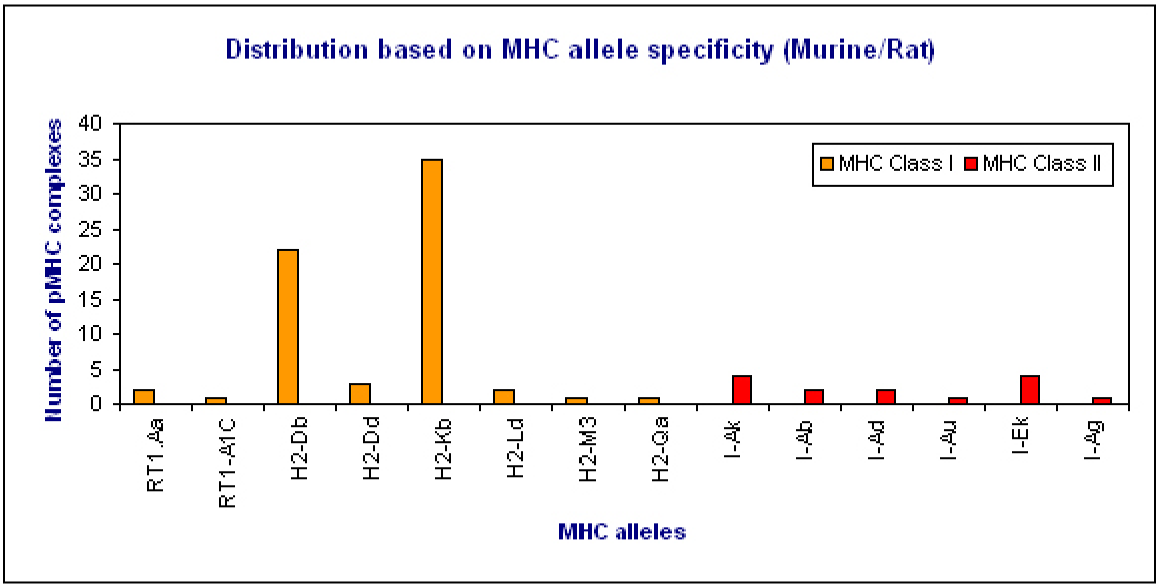


Figure 9 Distribution of rodent (murine and rat) MHC allele in MPID-T shows H2-Db (class I) and H2-Kb (class I) have the maximum peptide/MHC complex structures.

Table 1 MHC class I dataset

MHC Source	PDB ID	Allele	Peptide Length	Peptide Source	Peptide Sequence	MCI	PCI	Res. (Ǻ)	Release Year
Human	1DUY	A*0201	8	Tax peptide	LFGYPVYV	A	C	2.15	2000
Human	1W72	A*0101	9	Melanoma-Associated Antigen 1	EADPTGHSY	A	C	2.15	2004
Human	1AKJ	A*0201	9	HIV-1 RT	ILKEPVHGV	A	C	2.65	1997
Human	1AO7	A*0201	9	HTLV-1 Tax	LLFGYPVYV	A	C	2.60	1997
Human	1B0G	A*0201	9	Human-peptide P0149	ALWGFFPVL	A	C	2.60	1998
Human	1B0R	A*0201	9	Influenza matrix	GILGFVFTL	A	C	2.90	1998
Human	1BD2	A*0201	9	HTLV-1 Tax	LLFGYPVYV	A	C	2.50	1998
Human	1DUZ	A*0201	9	HTLV-1 Tax	LLFGYPVYV	A	C	1.80	2000
Human	1EEY	A*0201	9	Gp2 Peptide Variant	ILSALVGIV	A	C	2.25	2003
Human	1EEZ	A*0201	9	Gp2 Peptide Variant	ILSALVGIL	A	C	2.30	2003
Human	1HHG	A*0201	9	HIV-1 gp 120	TLTSCNTSV	A	C	2.60	1993
Human	1HHI	A*0201	9	Synthetic	GILGFVFTL	A	C	2.50	1993
Human	1HHJ	A*0201	9	Synthetic	ILKEPVHGV	A	C	2.50	1993
Human	1HHK	A*0201	9	Synthetic	LLFGYPVYV	A	C	2.50	1993
Human	1I1F	A*0201	9	HIV-RT	FLKEPVHGV	A	C	2.80	2000
Human	1I1Y	A*0201	9	HIV-1RT	YLKEPVHGV	A	C	2.20	2000
Human	1I7R	A*0201	9	Synthetic	FAPGFFPYL	A	C	2.20	2001
Human	1I7T	A*0201	9	Synthetic	ALWGVFPVL	A	C	2.80	2001
Human	1I7U	A*0201	9	Synthetic	ALWGFVPVL	A	C	1.80	2001
Human	1IM3	A*0201	9	HTLV-1 Tax	LLFGYPVYV	A	C	2.20	2001
Human	1LP9	A*0201	9	Self peptide	ALWGFFPVL	A	C	2.00	2003
Human	1OGA	A*0201	9	Synthetic	GILGFVFTL	A	C	1.40	2003
Human	1QR1	A*0201	9	Gp2 Peptide	IISAVVGIL	A	C	2.40	2000
Human	1QRN	A*0201	9	Tax peptide P6A	LLFGYAVYV	A	C	2.80	1999
Human	1QSE	A*0201	9	Tax peptide	LLFGYPRYV	A	C	2.80	1999
Human	1QSF	A*0201	9	Tax peptide	LLFGYPVAV	A	C	2.80	1999
Human	1S9W	A*0201	9	Ny-Eso-1 Peptide	SLLMWITQC	A	C	2.20	2004
Human	1S9X	A*0201	9	Ny-Eso-1 Peptide Analogue S9A	SLLMWITQA	A	C	2.50	2004
Human	1S9Y	A*0201	9	Ny-Eso-1 Peptide Analogue S9S	SLLMWITQS	A	C	2.30	2004

Human	1TVB	A*0201	9	Melanocyte Protein Pmel 17 epitope	ITDQVPFSV	A	C	1.80	2005
Human	1TVH	A*0201	9	Melanocyte Protein Pmel 17 epitope	IMDQVPFSV	A	C	1.80	2005
Human	1JHT	A*0201	9	Mart-1	ALGIGILTV	A	C	2.15	2001
Human	1P7Q	A*0201	9	Pol Polyprotein	ILKEPVHGV	A	C	3.40	2003
Human	1JF1	A*0201	10	Mart-1	ELAGIGILTV	A	C	1.85	2001
Human	1I4F	A*0201	10	MAGE-4 Antigen	GVYDGREHTV	A	C	1.40	2001
Human	1HHH	A*0201	10	HBV nucleocapsid	FLPSDFFPSV	A	C	3.00	1993
Human	2CLR	A*0201	10	Synthetic	MLLSVPLLIG	A	C	2.00	1998
Human	1TMC	A*6801	10	Synthetic	EVAPPEYHRK	A	C	2.30	1995
Human	1AGB	B*0801	8	HIV-1 gag	GGRKKYKL	A	C	2.20	1997
Human	1AGC	B*0801	8	HIV-1 gag	GGKKKYQL	A	C	2.10	1997
Human	1AGD	B*0801	8	HIV-1 gag	GGKKKYKL	A	C	2.05	1997
Human	1AGE	B*0801	8	HIV-1 gag	GGRKKYKL	A	C	2.30	1997
Human	1AGF	B*0801	8	HIV-1 gag	GGKKRYKL	A	C	2.20	1997
Human	1M05	B*0801	9	Ebna-3 Nuclear Protein	FLRGRAYGL	A	E	1.90	2003
Human	1MI5	B*0801	9	Epstein Barr Virus Peptide	FLRGRAYGL	A	C	2.50	2003
Human	1XR8	B*1501	9	Ebna-3 Nuclear Protein	LEKARGSTY	A	C	2.30	2005
Human	1XR9	B*1501	9	Ubiquitin-Conjugating Enzyme E2 E1	ILGPPGSVY	A	C	1.79	2005
Human	1HSA	B*2705	9	N.A.	ARAAAAAAA	A	C	2.10	1992
Human	1JGE	B*2705	9	Peptide M9	GRFAAAIAK	A	C	2.10	2002
Human	1OF2	B*2705	9	Vasoactive Intestinal Polypeptide Receptor	RRKWRRWHL	A	C	2.20	2004
Human	1OGT	B*2705	9	Vasoactive Intestinal Polypeptide Receptor	RRKWRRWHL	A	C	1.47	2004
Human	1W0V	B*2705	9	Butyrate Response Factor 2	RRLPIFSRL	A	C	2.27	2005
Human	1K5N	B*2709	9	Nonameric Model Peptide M9	GRFAAAIAK	A	C	1.09	2002
Human	1UXW	B*2709	9	Membrane Protein Lmp-2A/Lmp-2B	RRRWRRLTV	A	C	1.71	2004
Human	1W0W	B*2709	9	Butyrate Response Factor 2	RRLPIFSRL	A	C	2.10	2005
Human	1JGD	B*2709	10	Peptide S10R	RRLLRGHNQY	A	C	1.90	2003
Human	1A1N	B*3501	8	HIV-1 Nef	VPLRPMTY	A	C	2.00	1998
Human	1A9E	B*3501	9	EBV-Ebna3c	LPPLDITPY	A	C	2.50	1998
Human	1QEW	B*3501	9	Melanoma-Associated Antigen 3	FLWGPRALV	A	C	2.20	2003
Human	1A9B	B*3501	9	EBNA-3C	LPPLDITPY	A	C	3.20	1998
Human	1XH3	B*3501	14	Aa 4-17 of M-Csf	LPAVVGLSPGEQEY	A	C	1.48	2004
Human	1M6O	B*4402	9	Hla Dpa*0201 Peptide	EEFGRAFSF	A	C	1.60	2003

Human	1SYS	B*4403	9	Sorting Nexin 5	EEPTVIKKY	A	C	2.40	2004
Human	1N2R	B*4403	9	Hla Dpa*0201 Peptide	EEFGRAFSF	A	C	1.70	2004
Human	1SYV	B*4405	9	Self Ligand	EEFGRAFSF	A	C	1.70	2004
Human	1E28	B*5101	8	HIV-1 Km2	TAFTIPSI	A	C	3.00	2000
Human	1E27	B*5101	9	HIV-1 Kml	LPPVVAKEI	A	C	2.20	2000
Human	1A1O	B*5301	9	HIV-1 Nef	KPIVQYDNF	A	C	2.30	1998
Human	1A1M	B*5301	9	HIV-2 gag	TPYDINQML	A	C	2.30	1998
Human	1EFX	Cw*0304	9	Importin alpha 2	GALVDPLLAL	A	C	3.00	2000
Human	1QQD	Cw*0401	9	Synthetic	QYDDAVYKL	A	C	2.70	1999
Human	1IM9	Cw*0401	9	Synthetic	QYDDAVYKL	A	C	2.80	2001
Human	1MHE	E*0101	9	Synthetic	VMAPRTVLL	A	P	2.85	1999
Human	1KPR	E*0103	9	Synthetic	VMAPRTVLL	A	P	2.80	2003
Human	1KTL	E*0103	9	Peptide B27	VTAPRTLLL	A	P	3.10	2003
Human	1YDP	G*0101	9	Histone 2A	RIIPRHLQL	A	P	1.90	2005
Murine	1JPG	H2-Db	9	LCMV peptide	FQPQNGQFI	A	C	2.20	2001
Murine	1FFP	H2-Db	9	Gp33 Peptide	SAVYNFATM	A	C	2.60	2002
Murine	1FG2	H2-Db	9	Gp33 Peptide	KAVYNFATC	A	C	2.75	2000
Murine	1BZ9	H2-Db	9	Peptide P1027	FAPGVFPYM	A	P	2.80	1998
Murine	1CE6	H2-Db	9	SV nucleoprotein	FAPGNYPAL	A	C	2.90	1999
Murine	1FFN	H2-Db	9	Gp33 Peptide	KAVYNFATM	A	C	2.70	2002
Murine	1FFO	H2-Db	9	Gp33 Peptide	AAVYNFATM	A	C	2.65	2002
Murine	1HOC	H-2Db	9	Influenza virus nucleoprotein	ASNENMETM	A	C	2.40	1994
Murine	1INQ	H2-Db	9	H13A	SSVVGVWYL	A	C	2.20	2002
Murine	1JUF	H2-Db	9	H13B	SSVIGVWYL	A	C	2.00	2002
Murine	1S7U	H2-Db	9	Lcmv- Derived Gp33 Index Peptide	KAVYNFATM	A	C	2.20	2004
Murine	1S7V	H2-Db	9	Lcmv- Derived Gp33 Index Peptide	KAVYNLATM	A	C	2.20	2004
Murine	1S7W	H2-Db	9	Lcmv- Derived Gp33 Index Peptide	KALYNFATM	A	C	2.40	2004
Murine	1S7X	H2-Db	9	Lcmv- Derived Gp33 Index Peptide	KAVFNFATM	A	C	2.41	2004
Murine	1N5A	H2-Db	9	Gp33 Derived From Lcmv	KAVYNFATM	A	C	2.85	2003
Murine	1QLF	H2-Db	9	SV-nucleoprotein	FAPSNYPAL	A	C	2.65	1999
Murine	1WBX	H2-Db	10	Influenza A Peptide	SQLKNNAKEI	A	C	1.90	2005
Murine	1WBY	H2-Db	10	Influenza A Peptide	SSLENFRAYV	A	C	2.30	2005
Murine	1DDH	H2-Db	10	HIV-1 gp120	RGPGRAFVTI	A	P	3.10	1999
Murine	1N3N	H2-Db	10	Mycobacterial Hsp60 Decameric Epitope	SNLQNAASIA	A	I	3.00	2003

Murine	1JPF	H2-Db	11	LCMV peptide	SGVENPGGYCL	A	C	2.18	2001
Murine	1BII	H2-Dd	10	HIV-1 P18-100	RGPGRAFVTI	A	P	2.40	1998
Murine	1QO3	H2-Dd	10	HIV Envelope Glycoprotein 120 Peptide	RGPGRAFVTI	A	P	2.30	2000
Murine	1KJ3	H2-Kb	8	Naturally processed	KVITFIDL	H	P	2.30	2002
Murine	1FO0	H2-Kb	8	Natural peptide	INFDFNTI	H	P	2.50	2000
Murine	1FZJ	H2-Kb	8	VSV nucleoprotein	RGYVYQGL	A	P	1.90	2001
Murine	1FZK	H2-Kb	9	SV nucleoprotein	FAPGNYPAL	A	P	1.70	2001
Murine	1FZM	H2-Kb	8	VSV nucleoprotein	RGYVYQGL	A	P	1.80	2001
Murine	1BQH	H2-Kb	8	Vesicular stomatitis virus	RGYVYQGL	A	C	2.80	1998
Murine	1G6R	H2-Kb	8	Syir protein	SIYRYYGL	H	P	2.80	2000
Murine	1KJ2	H2-Kb	8	Naturally processed	KVITFIDL	H	P	2.71	2002
Murine	1G7Q	H2-Kb	8	mucin1,transmembrane	SAPDTRPA	A	P	1.60	2002
Murine	1KBG	H2-Kb	8	VSV nucleoprotein	RGYVYXGL	H	P	2.20	1999
Murine	1KPU	H2-Kb	8	VSV8, Nucleocapsid Fragment	RGYVYQGL	A	P	1.50	2003
Murine	1LEG	H2-Kb	8	Dev 8	EQYKFYSV	A	P	1.75	2002
Murine	1LEK	H2-Kb	8	Dev 8	EQYKFYSV	A	P	2.15	2002
Murine	1LK2	H2-Kb	8	Synthetic	GNYSFYAL	A	P	1.35	2003
Murine	1MWA	H2-Kb	8	Dev 8	EQYKFYSV	H	P	2.40	2002
Murine	1N59	H2-Kb	8	Gp33 Derived From Lcmv	AVYNFATM	A	P	2.95	2003
Murine	1T0M	H2-Kb	8	Glycoprotein B	SSIEFARL	A	P	2.00	2004
Murine	1T0N	H2-Kb	8	Glycoprotein B	SSIEFARL	A	P	1.80	2004
Murine	1VAC	H2-Kb	8	Ovalbumin	SIINFEKL	A	P	2.50	1996
Murine	1NAM	H2-Kb	8	VSV Nucleoprotein Fragment	RGYVYQGL	H	P	2.70	2003
Murine	1NAN	H2-Kb	8	Riken Cdna 2410004N11	INFDFNTI	H	M	2.30	2003
Murine	1OSZ	H2-Kb	8	VSV nucleoprotein	RGYLYQGL	A	C	2.10	1999
Murine	2CKB	H2-Kb	8	Dev 8	EQYKFYSV	H	P	3.20	1998
Murine	2MHA	H2-Kb	8	Vesicular stomatitis virus	RGYVYQGL	A	E	2.80	1993
Murine	1KPV	H2-Kb	9	SEV9, Nucleoprotein Fragment	FAPGNYPAL	A	P	1.71	2003
Murine	1FZO	H2-Kb	9	SV nucleoprotein	FAPGNYPAL	A	P	1.80	2001
Murine	1G7P	H2-Kb	9	alpha-glucosidase p1	SRDHSRTPM	A	P	1.50	2002
Murine	1VAD	H2-Kb	9	Yeast alpha glucosid	SRDHSRTPM	A	P	2.50	1996
Murine	1WBZ	H2-Kb	9	Influenza A Peptide	SSYRRPVGI	A	P	2.00	2005
Murine	1S7R	H2-Kb	9	Lcmv- Derived Gp33 Index Peptide	KAVYNLATM	A	C	2.95	2004

Murine	1S7S	H2-Kb	9	Lcmv- Derived Gp33 Index Peptide	KALYNFATM	A	C	1.99	2004
Murine	1S7T	H2-Kb	9	Lcmv- Derived Gp33 Index Peptide	KAVFNFATM	A	C	2.30	2004
Murine	1S7Q	H2-Kb	9	Lcmv- Derived Gp33 Index Peptide	KAVYNFATM	A	C	1.99	2004
Murine	2VAA	H2-Kb	9	VSV nucleoprotein	FAPGNYPAL	A	P	2.30	1996
Murine	2VAB	H2-Kb	9	SV nucleoprotein	FAPGNYPAL	A	P	2.50	1996
Murine	1LD9	H2-Ld	9	Synthetic	YPNVNIHNF	A	C	2.40	1998
Murine	1LDP	H2-Ld	9	Natural peptide	APAAAAAAM	H	P	3.10	1998
Murine	1MHC	H2-M3	10	Rat Nadh Dehydrogenase	MYFINILTL	A	C	2.10	1996
Murine	1K8D	Qa-2	9	60S Ribosomal Protein	ILMEHIHKL	A	P	2.30	2001
Rat	1KJM	RT1.Aa	9	B6 Peptide	AQFSASASR	A	P	2.35	2002
Rat	1ED3	RT1.Aa	13	Rat atapase	ILPSSERLISNR	A	C	2.55	2000
Rat	1KJV	RT1-A1C	9	Peptide Npr	NPRAMQALL	A	P	1.48	2002

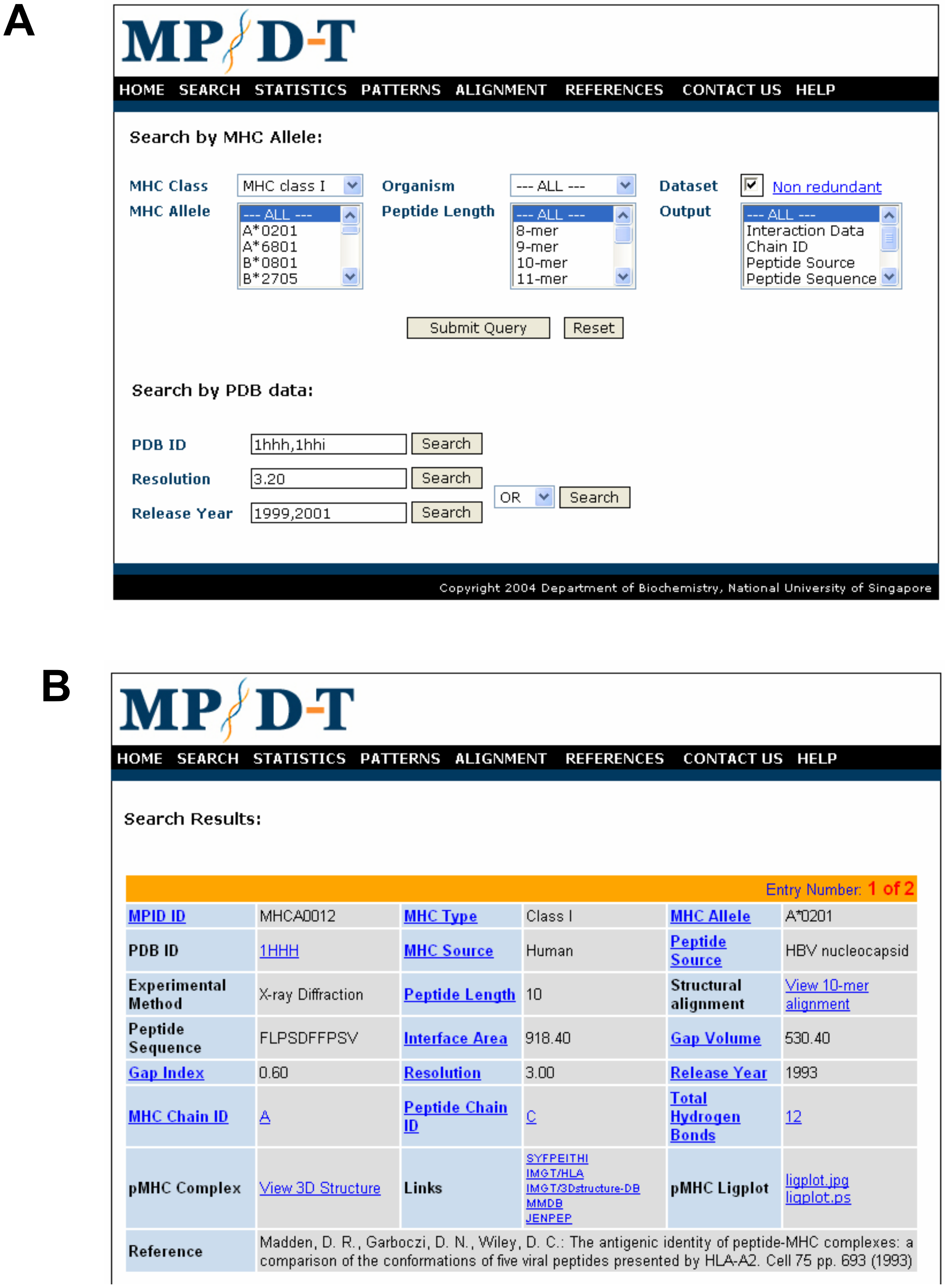
MCI = MHC chain identifier, PCI = peptide chain identifier, Res = resolution, Release year = the year in which the entry was released by PDB.

Table 2 MHC class II dataset

MHC Source	PDB ID	Allele	Peptide Length	Peptide Source	Peptide Sequence	MCI	PCI	Res. (Å)	Release Year
Human	1S9V	DQB1*0201	11	Alpha-I Gliadin	LQPFPQPELPY	A, B	C	2.22	2004
Human	1JK8	DQB1*0302	14	Insulin B Peptide	LVEALYLVCGERGG	A, B	C	2.4	2001
Human	1UVQ	DQB1*0602	20	Hypocretin Peptide	MNLPSTKVSWAAVGGGGSLV	A, B	C	1.8	2004
Human	1AQD	DRB1*0101	14	Endogeneous Peptide	GSDWRFLRGYHQYA	A, B	C	2.45	1998
Human	1DLH	DRB1*0101	13	Influenza Virus Peptide	PKYVKQNTLKLAT	A, B	C	2.8	1994
Human	1FYT	DRB1*0101	13	Hemagglutinin Ha1 Peptide Chain	PKYVKQNTLKLAT	A, B	C	2.6	2000
Human	1HXY	DRB1*0101	13	Hemagglutinin	PKYVKQNTLKLAT	A, B	C	2.6	2001
Human	1JWM	DRB1*0101	13	Ha Peptide	PKYVKQNTLKLAT	A, B	C	2.7	2003
Human	1JWS	DRB1*0101	13	Ha Peptide	PKYVKQNTLKLAT	A, B	C	2.6	2003
Human	1JWU	DRB1*0101	13	Ha Peptide	PKYVKQNTLKLAT	A, B	C	2.3	2003
Human	1KG0	DRB1*0101	13	Hemagglutinin Ha Peptide	PKYVKQNTLKLAT	A, B	D	2.65	2002
Human	1KLG	DRB1*0101	15	Triosephosphate Isomerase Peptide	GELIGILNAAKVPAD	A, B	C	2.4	2002
Human	1KLU	DRB1*0101	15	Triosephosphate Isomerase Peptide	GELIGTLNAAKVPAD	A, B	C	1.93	2002
Human	1LO5	DRB1*0101	13	Hemagglutinin Peptide	PKYVKQNTLKLAT	A, B	C	3.2	2002
Human	1PYW	DRB1*0101	9	Influenza Virus Hemagglutinin Related Peptide	FVKQNAXAL	A, B	C	2.1	2003
Human	1R5I	DRB1*0101	13	Hemagglutinin Peptide	PKYVKQNTLKLAT	A, B	C	2.6	2004
Human	1SEB	DRB1*0101	13	Endogeneous Peptide	AAAAAAAAAAAAA	A, B	C	2.7	1996
Human	1SJE	DRB1*0101	15	Gag Polyprotein	PEVIPMFSALSEGAT	A, B	C	2.45	2004
Human	1SJH	DRB1*0101	13	Gag Polyprotein	PEVIPMFSALSEG	A, B	C	2.25	2004
Human	1T5W	DRB1*0101	13	Fragment Of Regulatory Protein Mig1	AAYSDQATPLLLS	A, B	C	2.4	2004
Human	1T5X	DRB1*0101	13	Fragment Of Regulatory Protein Mig1	AAYSDQATPLLLS	A, B	C	2.5	2004
Human	1A6A	DRB1*0301	15	Clip	PVSKMRMATPLLMQA	A, B	C	2.75	1998
Human	1D5M	DRB1*0401	7	Peptide Inhibitor	RAMXSX	A, B	D	2	2000
Human	1D5X	DRB1*0401	5	Dipeptide Mimetic Inhibitor	RXXX	A, B	D	2.45	2000
Human	1D5Z	DRB1*0401	6	Peptidomimetic Inhibitor	RAXSX	A, B	D	2	2000
Human	1D6E	DRB1*0401	7	Peptidomimetic Inhibitor	RXMASX	A, B	D	2.45	2000

Human	1J8H	DRB1*0401	13	Hemagglutinin Ha1 Peptide Chain	PKYVKQNTLKLAT	A, B	C	2.4	2002
Human	2SEB	DRB1*0401	12	Collagen II Peptide	AYMRADAAAGGA	A, B	E	2.5	1998
Human	1BX2	DRB1*1501	14	Human Myelin Basic Protein	ENPVVHFFKNIVTP	A, B	C	2.6	1998
Human	1FV1	DRB5*0101	20	Myelin Basic Protein	NPVVHFFKNIVTPRTPPPSQ	A, B	C	1.9	2000
Human	1H15	DRB5*0101	14	Epstein Barr Vius (Ebv) DNA Polymerase	GGVYHFVKKHVHES	A, B	C	3.1	2002
Human	1HQR	DRB5*0101	10	Myelin Basic Protein	VHFFKNIVTP	A, B	C	3.2	2001
Murine	1D9K	I- Ak	16	Conalbumin Peptide	GNSHRGAIEWEGIESG	C, D	P	3.2	1999
Murine	1ES0	I-A(G7)	14	Glutamic Acid Decarboxylase Peptide	YEIAPVFVLLEYVT	A, B	B	2.6	2000
Murine	1LNU	I-Ab	13	Ealpha3K Peptide	FEAQKAKANKAVD	A, B	B	2.5	2002
Murine	1MUJ	I-Ab	15	Clip Peptide	PVSKMRMATPLLMQA	A, B	C	2.15	2003
Murine	1IAO	I-Ad	13	Ovalbumin Peptide	RGISQAVHAAHAE	A, B	B	2.6	1998
Murine	2IAD	I-Ad	14	Influenza Hemagglutinin Peptide	GHATQGVTAASSHE	A, B	B	2.4	1998
Murine	1F3J	I-Ak	14	Gallus Gallus	AMKRHGLDNYRGYS	A, B	P	3.1	2000
Murine	1IAK	I-Ak	13	Hen Eggwhite Lysozyme Peptide	STDYGILQINSRW	A, B	P	1.9	1998
Murine	1JL4	I-Ak	16	Ovotransferrin	GNSHRGAIEWEGIESG	A, B	C	4.3	2001
Murine	1K2D	I-Au	11	Myelin Basic Protein Peptide	SRGGASQYRPS	A, B	P	2.2	2003
Murine	1KT2	I-Ek	12	Moth Cytochrome C Peptide	ADLIAYLKQATK	A, B	B	2.8	2002
Murine	1KTD	I-Ek	14	Pigeon Cytochrome C Peptide	AADLIAYLKQASAK	A, B	B	2.4	2002
Murine	1R5V	I-Ek	13	Artificial Peptide	ADLIAYPKAATKF	A, B	E	2.5	2004
Murine	1R5W	I-Ek	13	Artificial Peptide	ADLIAYFKAATKF	A, B	E	2.9	2004

MCI = MHC chain identifier, PCI = peptide chain identifier, Res = resolution, Release year = the year in which the entry was released by PDB.


A
MPID-T
HOME SEARCH STATISTICS PATTERNS ALIGNMENT REFERENCES CONTACT US HELP
Submit Query
Reset
Search
B
MPID-T
HOME SEARCH STATISTICS PATTERNS ALIGNMENT REFERENCES CONTACT US HELP

Search by MHC Allele:

MHC Class: MHC class I

Organism: --- ALL ---

Dataset: ☑ Non redundant

MHC Allele: --- ALL ---, A*0201, A*6801, B*0801, B*2705

Peptide Length: --- ALL ---, 8-mer, 9-mer, 10-mer, 11-mer

Output: --- ALL ---, Interaction Data, Chain ID, Peptide Source, Peptide Sequence

Search by PDB data:

PDB ID: 1hhh,1hhi

Resolution: 3.20

Release Year: 1999,2001

OR

Copyright 2004 Department of Biochemistry, National University of Singapore

Search Results:

Entry Number: 1 of 2

MPID ID	MHCA0012	MHC Type	Class I	MHC Allele	A*0201
PDB ID	1HHH	MHC Source	Human	Peptide Source	HBV nucleocapsid
Experimental Method	X-ray Diffraction	Peptide Length	10	Structural alignment	View 10-mer alignment
Peptide Sequence	FLPSDFFPSV	Interface Area	918.40	Gap Volume	530.40
Gap Index	0.60	Resolution	3.00	Release Year	1993
MHC Chain ID	A	Peptide Chain ID	C	Total Hydrogen Bonds	12
pMHC Complex	View 3D Structure	Links	SYFPEITHI IMGT/HLA IMGT/3Dstructure-DB MMDB JENPEP	pMHC Ligplot	ligplot.jpg ligplot.ps
Reference	Madden, D. R., Garboczi, D. N., Wiley, D. C.: The antigenic identity of peptide-MHC complexes: a comparison of the conformations of five viral peptides presented by HLA-A2. Cell 75 pp. 693 (1993)				

Figure 10 Sample (a) input and (b) output web interface from MPID-T with user defined input parameters (MHC Class, Organism, Data redundancy, MHC allele, peptide length and output format).

3.2.2.4 Interface area

The interface area refers to the accessible surface area (ASA) between the bound peptide and MHC molecule. It is defined as the change in solvent-accessible surface area (∆ASA) on complexation from an unbound MHC molecule to a bound peptide/MHC complex state and is represented by the equation:

$$(\Delta\text{ASA}) = \frac{\text{ASA of MHC + ASA of peptide - ASA of peptide/MHC molecule}}{2}$$

(*Equation 2*)

3.2.3 Implementation

The web interface permits searching the molecular complexes stored in the database based on MHC allele or PDB information, as shown in Figure 10. Structural visualization of the T cell receptor/peptide/MHC complex, peptide/MHC complex, MHC or the bound peptide can be performed using freely available graphics applications such as RasMol (http://www.openrasmol.org) or Chime (http://www.mdlchime.com), whereas 3D alignment of structures (based on MHC class and peptide length) (May and Johnson, 1995) can be viewed using the Jmol molecular viewer (http://www.jmol.org) or a Chime-compatible web browser client. Each MPID-T entry bears a unique identifier, with sequence data hyperlinked to external databases that include IMGT/HLA (for the human MHC sequences) (Robinson *et al*., 2001), IMGT/3Dstructure-DB (for peptide/MHC and T cell receptor/peptide/MHC sequences and structures) (Kaas *et al*., 2004), SYFPEITHI (for MHC ligands and peptide motifs) (Rammensee *et al*., 1999) and AntiJen (for experimental binding affinity) (Toseland *et al*., 2005). Related sequences and structures for the relevant protein chains can be accessed via the National Center for Biotechnology Information (NCBI) Structure database (http://www.ncbi.nlm.nih.gov/Structure) and bibliographic references from PubMed. Pre-computed schematic diagrams based on the plotting program LIGPLOT (Wallace *et al*., 1995) are provided to illustrate explicit peptide/MHC interactions. Consensus patterns among peptides of the same length or allele are also available in MPID-T generated using the program WebLogo (Crooks *et al*., 2004). Other useful sources of information for researchers in vaccine design and immunology as referenced in Rammensee *et al*. (1999) are also provided under MHC resources on the MPID-T help page.

3.3 Data Analysis

For all supertypes investigated in this study, the gap index, which measures geometric and electrostatic complementation between the bound peptide and HLA molecule, inversely correlates with increasing interface area (Figures 11B, 12B, 13B, 14B, and 15B). The implication is that complexes with larger interface area have better geometric and electrostatic complementarity (i.e. smaller gap index), resulting in the formation of more intermolecular hydrogen bonds (Figures 11A, 12A, 13A, 14A, and 15A), which in turn contribute to the stability of the bound complexes. The mean Cα root mean square deviation (RMSD) of HLA class I peptides is 1.41 Ấ (Table 3). The peptide N- and C-terminal residues are highly conserved with mean Cα RMSD of 0.08 Ǻ and 0.09 Ǻ respectively. A similar highly conserved backbone conformation is observed at the ends of the core peptide fragments in the binding cleft of DR1 molecules with mean Cα RMSD of 0.08 Å and 0.09 Å for the two peptide termini, respectively.

3.3.1 HLA-A2

The mean interface area for A2 is 846.3 ± 48.9 Ǻ^2. On average, the number of intermolecular hydrogen bonds and gap index for A2 is 11.1 ± 1.9 and 0.9 ± 0.2 respectively. Extensive hydrogen bonding is found in binding pockets A, B and F. No clear difference is observed in the number of intermolecular hydrogen bonds for 9-mer (11.0 ± 1.8) and 10-mer (11.8 ± 2.2) complexes. The gap index for 9-mer and 10-mer complexes are 1.0 ± 0.2 and 0.8 ± 0.3 respectively. The results indicate that the interacting surfaces of 10-mer complexes are generally more complementary than 9-mer complexes. On average, the Cα RMSD of A2 peptides is 1.72 Ǻ. The peptide N- and C-terminal residues are highly conserved with mean Cα RMSD of 0.05 Ǻ and 0.09 Ǻ respectively.

3.3.2 HLA-B7

The average number of intermolecular hydrogen bonds, gap index, and interface area for B7 are 14.3 ± 2.3, 1.0 ± 0.1, and 876.7 ± 72.4 Ǻ^2 respectively. In general, 9-mer complexes (gap index = 0.9 ± 0.1) are more complementary than 8-mer complexes (gap index = 1.0 ± 0.1) and intermolecular hydrogen bonds are well distributed throughout the entire complex. The corresponding numbers of hydrogen bonds for 8-mer and 9-mer complexes are 15.6 ± 2.3 and 16.5 ± 2.7. Correlations between the number of intermolecular hydrogen bonds with gap volume (Figure 12C; r=-0.03) and gap index (Figure 12D; r=-0.07) are insignificant, indicating that the binding mechanism underlying peptide/B7 interactions may be different from peptide/A2 interactions. In general, the Cα RMSD of B7 peptides is 1.04 Ǻ. As in A2, the peptide N- and C-terminal residues are highly conserved with mean Cα RMSD of 0.16 Ǻ and 0.08 Ǻ respectively.

3.3.3 HLA-B27

B27 has the largest average interface area (934.0 ± 136.0 $Ǻ^2$) among all class I supertypes investigated in this study. In contrast to A2 and B44, the number of intermolecular hydrogen bonds inversely correlates with gap volume (Figure 13C; r=-0.22) and gap index (Figure 13D; r=-0.38). High concentrations of hydrogen bonds are observed in pockets A, B and F. More hydrogen bonds are formed at smaller gap index (higher geometric and electrostatic complementation) compared to the complexes of A2 and B44. The average Cα RMSD for B27 peptides is 1.00 Ǻ. Again, the peptide N- and C-terminal residues are highly conserved with mean Cα RMSD of 0.07 Ǻ and 0.09 Ǻ respectively.

3.3.4 HLA-B44

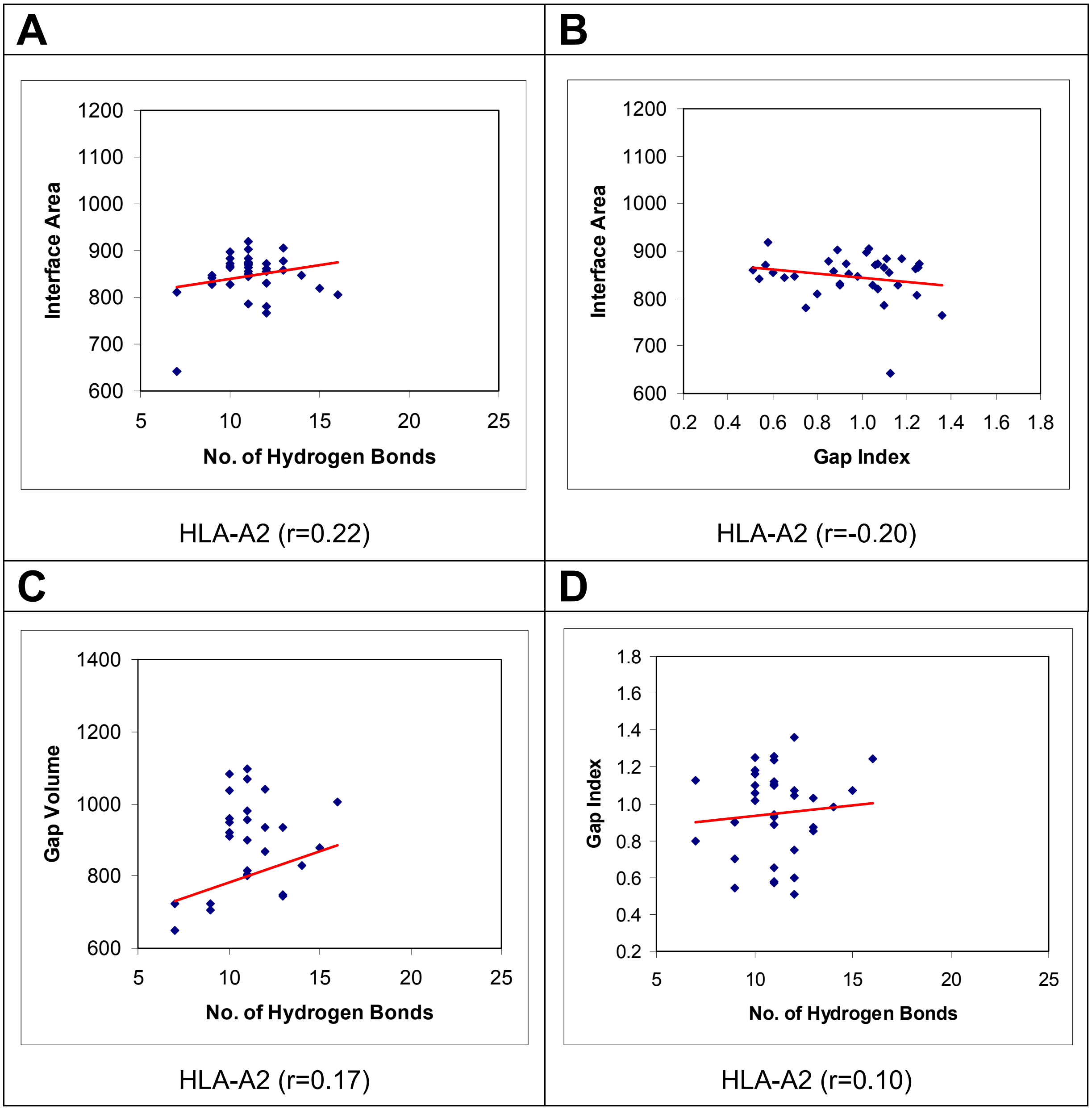
The interaction characteristics of B44 are similar to A2 (Figure 14). Intermolecular hydrogen bonds are primarily concentrated in pockets A, B and F. On average, the number of intermolecular hydrogen bonds, interface area, gap volume, and gap index is 12.3 ± 3.5, 892.6 ± 57.3, 891.0 ± 114.7, and 1.0 ± 0.1 respectively. Correlations between the number of intermolecular hydrogen bonds with gap volume (Figure 14C; r=0.79) and gap index (Figure 14D; r=0.60) are strong, indicating that more hydrogen bonds are formed in B44 complexes as geometric and electrostatic complementarity decreases (i.e. gap index increases). The mean Cα RMSD of B44 peptides is 1.02 Ǻ. High conservation of the peptide N- and C-terminal residues is detected with mean Cα RMSD of 0.07 Ǻ and 0.10 Ǻ respectively.

3.3.5 HLA-DR1

The mean number of intermolecular hydrogen bonds, interface area, gap volume, and gap index is 15.6 ± 2.0, 1079.9 ± 86.4, 1092.8 ± 129.3, and 1.0 ± 0.1 respectively. In general, intermolecular hydrogen bonds are well distributed throughout the entire bound binding register. The number of intermolecular hydrogen bonds directly correlates with gap volume (Figure 15C; r=0.43) but correlation between the number of hydrogen bonds with gap index is insignificant (Figure 15D; r=-0.02). The mean Cα RMSD of DR1 binding registers is 0.60 Ǻ. Similar to class I peptides, the N- and C-terminal residues of the binding registers are highly conserved with mean Cα RMSD of 0.08 Ǻ and 0.09 Ǻ respectively.

3.4 Discussion

MPID-T is a manually curated specialist database for sequence-structure-function information on peptide/MHC and TCR/peptide/MHC interactions. The aim of developing MPID-T is to define structural descriptors for in-depth characterization of TCR/peptide/MHC and peptide/MHC interactions. Such descriptors should better reflect TCR/peptide/MHC and peptide/MHC interactions than just sequence alone. Together with other databases containing MHC- or antigen- related data such as AntiJen (which contains experimental binding affinities) (Toseland *et al*., 2005), MHCBN (which contains MHC binding and non-binding peptide sequences) (Bhasin *et al*., 2003), FIMM (which contains fully referenced data on protein antigens, MHC, peptide/MHC and relevant disease associations) (Schönbach *et al*., 2000), MPID-T aim to facilitate the extraction of high-level relationships hidden within (TCR/) peptide/MHC interaction data by mapping the TCR footprint on the MHC and its bound peptide as this eventually determines T cell recognition and binding. Identification of such descriptors will enhance the understanding of the binding mechanism underlying TCR/peptide/MHC and peptide/MHC interactions and facilitate the development of algorithms (Kangueane *et al*., 2000) for predicting whether a peptide sequence can bind to a specific MHC allele and subsequently whether the bound peptide/MHC complex can evoke a response to TCR. Future developments will include computed data on additional structural parameters characterizing the TCR/peptide/MHC and peptide/MHC interaction region.



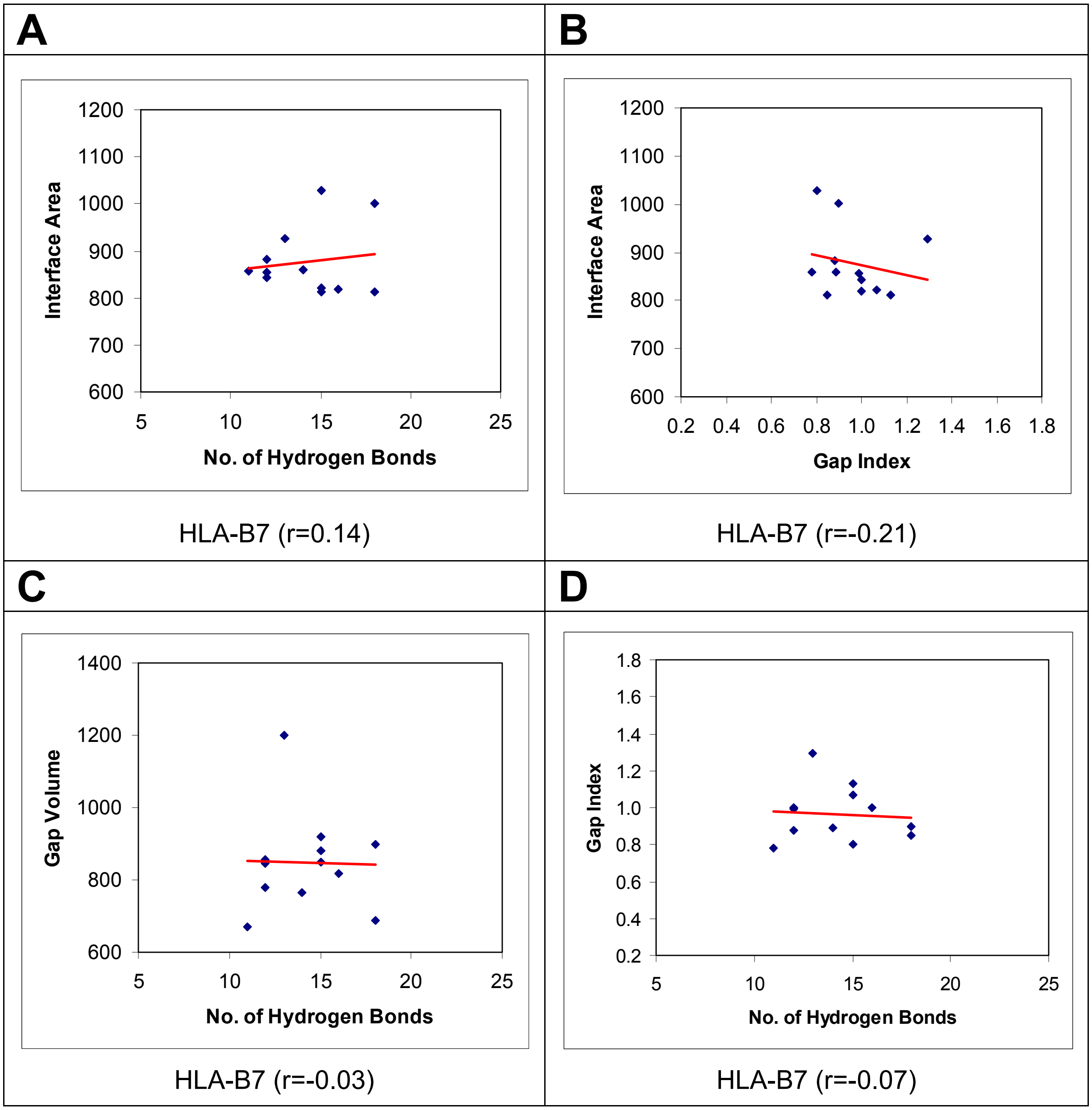
HLA-A2 (r=0.22)

HLA-A2 (r=-0.20)

HLA-A2 (r=0.17)

HLA-A2 (r=0.10)

Figure 11 Graphs of different structural interaction parameters for HLA-A2 supertype and their respective correlation coefficients r investigated in this study: (A) Interface area vs. number of intermolecular hydrogen bonds; (B) Interface area vs. gap index; (C) Gap volume vs. number of intermolecular hydrogen bonds; and (D) Gap index vs. number of intermolecular hydrogen bonds.



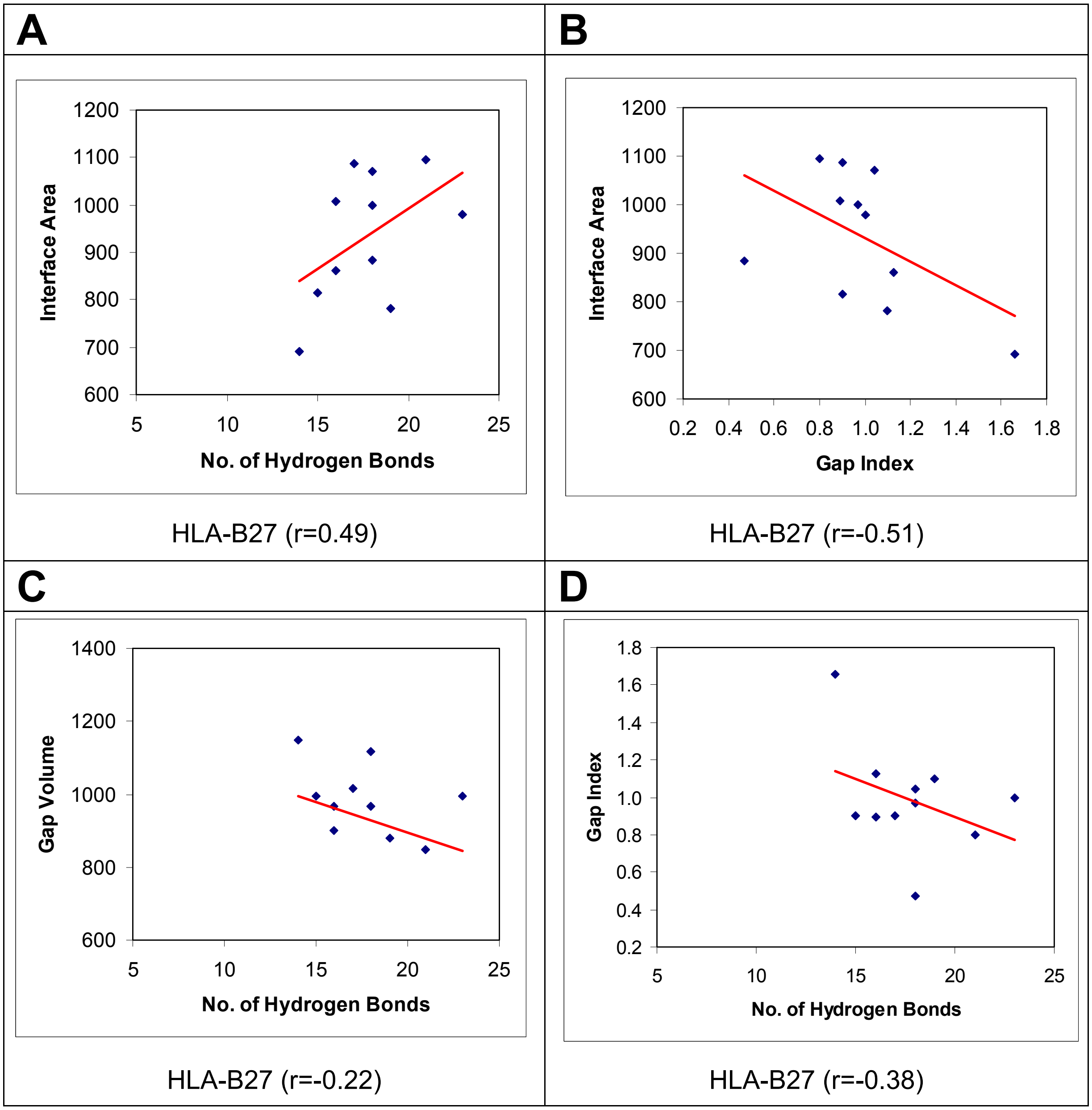
HLA-B7 (r=0.14)

HLA-B7 (r=-0.21)

HLA-B7 (r=-0.03)

HLA-B7 (r=-0.07)

Figure 12 Graphs of different structural interaction parameters for HLA-B7 supertype and their respective correlation coefficients r investigated in this study: (A) Interface area vs. number of intermolecular hydrogen bonds; (B) Interface area vs. gap index; (C) Gap volume vs. number of intermolecular hydrogen bonds; and (D) Gap index vs. number of intermolecular hydrogen bonds.



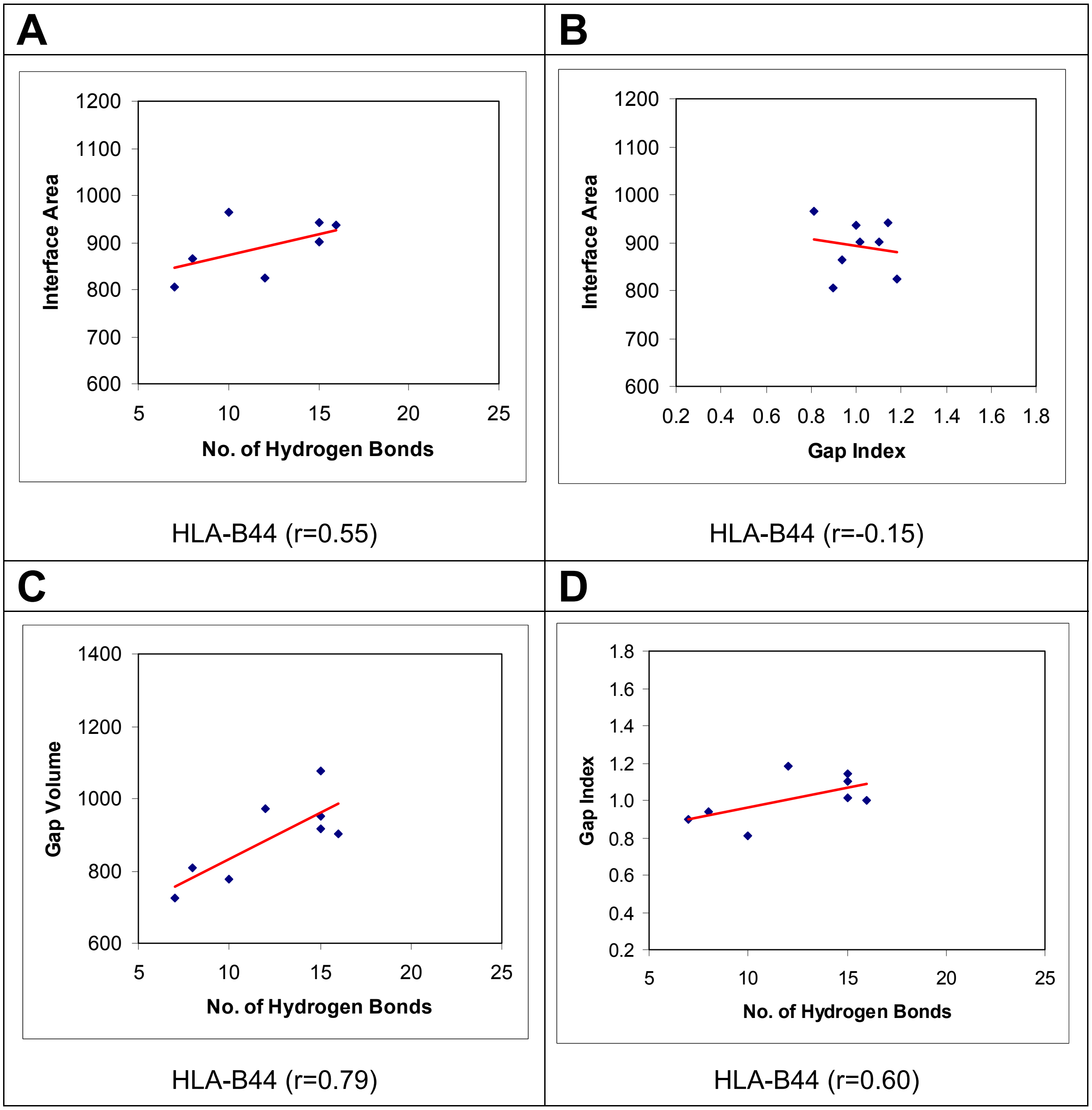
HLA-B27 (r=0.49)

HLA-B27 (r=-0.51)

HLA-B27 (r=-0.22)

HLA-B27 (r=-0.38)

Figure 13 Graphs of different structural interaction parameters for HLA-B27 supertype and their respective correlation coefficients r investigated in this study: (A) Interface area vs. number of intermolecular hydrogen bonds; (B) Interface area vs. gap index; (C) Gap volume vs. number of intermolecular hydrogen bonds; and (D) Gap index vs. number of intermolecular hydrogen bonds.



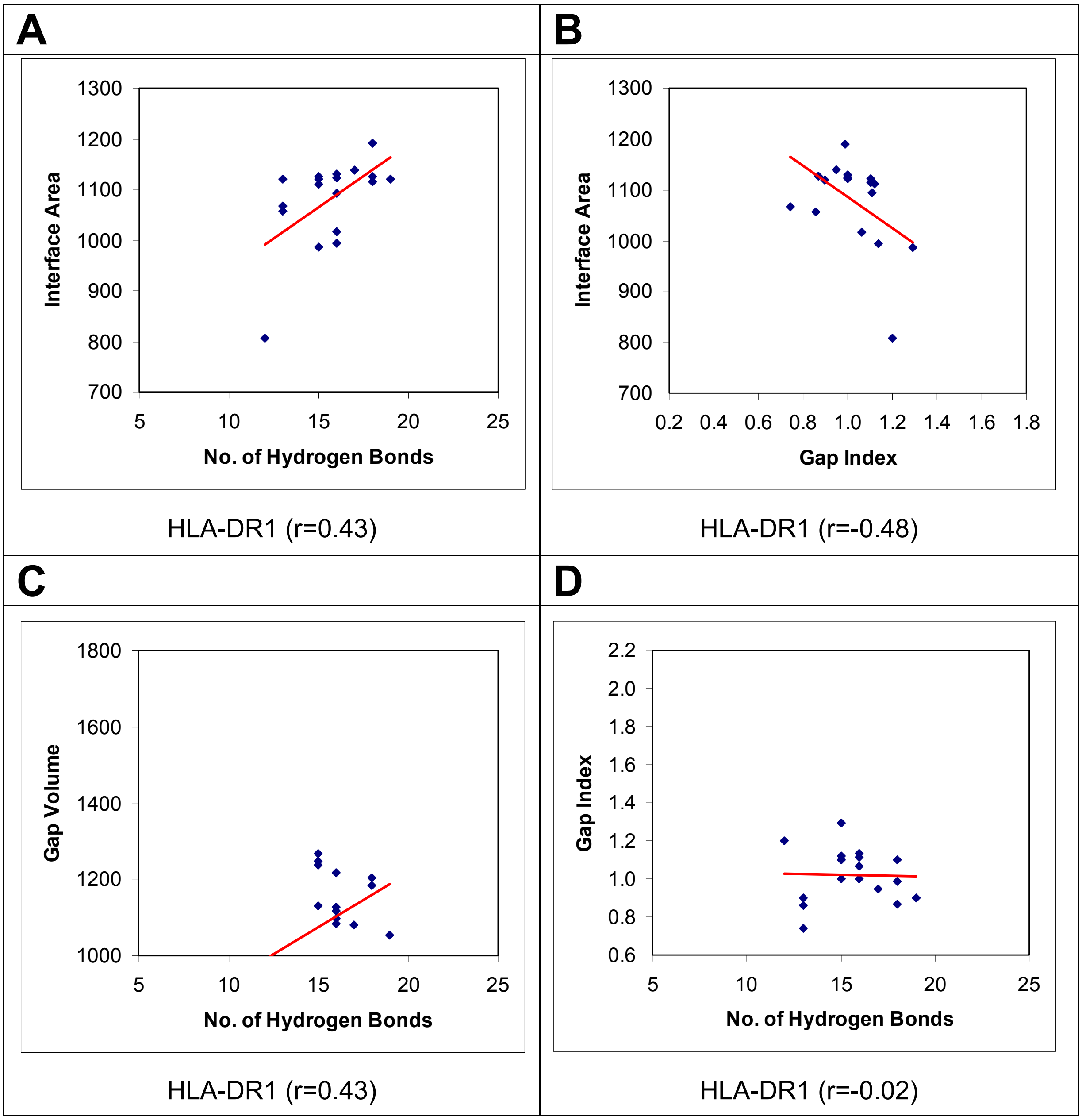
HLA-B44 (r=0.55)

HLA-B44 (r=-0.15)

HLA-B44 (r=0.79)

HLA-B44 (r=0.60)

Figure 14 Graphs of different structural interaction parameters for HLA-B44 supertype and their respective correlation coefficients r investigated in this study: (A) Interface area vs. number of intermolecular hydrogen bonds; (B) Interface area vs. gap index; (C) Gap volume vs. number of intermolecular hydrogen bonds; and (D) Gap index vs. number of intermolecular hydrogen bonds.



HLA-DR1 (r=0.43)

HLA-DR1 (r=-0.48)

HLA-DR1 (r=0.43)

HLA-DR1 (r=-0.02)

Figure 15 Graphs of different structural interaction parameters for HLA-DR1 supertype and their respective correlation coefficients r investigated in this study: (A) Interface area vs. number of intermolecular hydrogen bonds; (B) Interface area vs. gap index; (C) Gap volume vs. number of intermolecular hydrogen bonds; and (D) Gap index vs. number of intermolecular hydrogen bonds.

Table 3 Computed HLA structural interaction parameters investigated in this study. Three residues at the N- ("head") and C- ("tail") termini of the peptide within the MHC groove are compared. '#' represents the templates used for comparing Cα RMSDs of the various supertypes.

Class	Supertype	Allele	PDB ID	Interface Area	Gap Volume	Gap Index	No. of H-bonds	Peptide Length	Cα RMSD		
									Head	Tail	All
I	A2	A*0201	1AKJ	857.2	746.5	0.9	13	9	0.05	0.03	1.63
I	A2	A*0201	1AO7	883.3	1038.3	1.2	10	9	0.03	0.10	1.86
I	A2	A*0201	1B0G	860.2	441.3	0.5	12	9	0.05	0.09	1.44
I	A2	A*0201	1BD2	874.2	813.6	0.9	11	9	0.02	0.13	1.82
I	A2	A*0201	1DUY	717.7	1116.1	1.6	7	8	0.07	0.16	1.08
I	A2	A*0201	1DUZ	863.0	1069.6	1.2	11	9	0.05	0.13	1.22
I	A2	A*0201	1EEY	787.1	900.3	1.1	11	9	0.05	0.06	1.75
I	A2	A*0201	1EEZ	829.8	704.2	0.9	9	9	0.05	0.17	1.79
I	A2	A*0201	1HHG	765.8	1039.9	1.4	12	9	0.02	0.07	1.58
I	A2	A*0201	1HHH	918.4	530.4	0.6	11	10	0.04	0.16	2.07
I	A2	A*0201	1HHI	842.0	455.7	0.5	9	9	0.14	0.07	1.54
I	A2	A*0201	1HHJ	847.9	827.4	1.0	14	9	0.06	0.08	1.65
I	A2	A*0201	1HHK	865.5	1083.4	1.3	10	9	0.06	0.13	1.25
I	A2	A*0201	1I1F	850.9	800.8	0.9	11	9	0.05	0.07	1.70
I	A2	A*0201	1I1Y	877.9	745.0	0.9	13	9	0.06	0.03	1.77
I	A2	A*0201	1I4F#	820.1	877.1	1.1	15	10	0.00	0.00	0.00
I	A2	A*0201	1I7R	902.5	805.4	0.9	11	9	0.08	0.07	1.56
I	A2	A*0201	1I7T	847.9	591.4	0.7	9	9	0.11	0.07	1.44
I	A2	A*0201	1I7U	845.2	545.8	0.7	11	9	0.03	0.03	1.59
I	A2	A*0201	1IM3	855.0	954.9	1.1	11	9	0.04	0.14	1.17
I	A2	A*0201	1JF1	870.4	492.5	0.6	11	10	0.08	0.07	2.93
I	A2	A*0201	1JHT	781.0	588.8	0.8	12	9	0.06	0.04	1.93
I	A2	A*0201	1LP9	855.7	549.7	0.6	12	9	0.03	0.08	1.45

I	A2	A*0201	1OGA	856.0	539.5	0.6	12	9	0.06	0.10	1.45
I	A2	A*0201	1P7Q	810.8	648.1	0.8	7	9	0.05	0.03	1.74
I	A2	A*0201	1QR1	827.1	722.4	0.9	9	9	0.06	0.03	1.39
I	A2	A*0201	1QRN	871.5	921.6	1.1	10	9	0.05	0.12	1.84
I	A2	A*0201	1QSE	873.0	1097.7	1.3	11	9	0.04	0.18	1.82
I	A2	A*0201	1QSF	828.6	959.5	1.2	10	9	0.07	0.11	1.80
I	A2	A*0201	1S9W	904.9	933.9	1.0	13	9	0.05	0.06	2.03
I	A2	A*0201	1S9X	872.3	933.1	1.1	12	9	0.03	0.06	2.04
I	A2	A*0201	1S9Y	883.7	979.0	1.1	11	9	0.05	0.05	2.04
I	A2	A*0201	1TVB	829.2	866.8	1.0	12	9	0.01	0.08	2.15
I	A2	A*0201	1TVH	865.1	950.6	1.1	10	9	0.02	0.09	2.15
I	A2	A*0201	2CLR	896.5	911.4	1.0	10	10	0.05	0.08	1.78
I	B7	B*0801	1AGB[#]	820.8	881.7	1.1	15	8	0.00	0.00	0.00
I	B7	B*0801	1AGC	812.5	688.1	0.9	18	8	0.14	0.09	0.35
I	B7	B*0801	1AGD	819.7	816.1	1.0	16	8	0.14	0.03	0.32
I	B7	B*0801	1AGE	812.3	920.6	1.1	15	8	0.14	0.05	0.32
I	B7	B*0801	1AGF	860.0	765.4	0.9	14	8	0.07	0.03	0.32
I	B7	B*0801	1M05	1000.6	897.4	0.9	18	9	0.21	0.04	1.34
I	B7	B*0801	1MI5	1028.0	850.0	0.8	15	9	0.27	0.12	1.44
I	B7	B*3501	1A1N	857.9	670.2	0.8	11	8	0.15	0.17	1.06
I	B7	B*3501	1A9B	855.4	847.4	1.0	12	9	0.27	0.06	1.38
I	B7	B*3501	1A9E	882.6	779.3	0.9	12	9	0.10	0.06	1.39
I	B7	B*3501	1QEW	843.2	855.9	1.0	12	9	0.16	0.15	1.89
I	B7	B*3501	1XH3	927.0	1198.5	1.3	13	14	0.11	0.11	1.67
I	B27	B*1501	1XR8[#]	860.7	968.1	1.1	16	9	0.00	0.00	0.00
I	B27	B*2705	1HSA	691.8	1148.4	1.7	14	9	0.02	0.12	1.06
I	B27	B*2705	1OF2	1087.7	1015.5	0.9	17	9	0.04	0.08	0.67
I	B27	B*2705	1W0V	1007.2	898.9	0.9	16	9	0.09	0.08	0.91
I	B27	B*2705	1JGE	815.4	994.4	0.9	15	9	0.13	0.12	1.07
I	B27	B*2705	1OGT	1096.0	849.3	0.8	21	9	0.02	0.03	0.58
I	B27	B*2709	1UXW	1071.9	1116.8	1.0	18	9	0.04	0.09	1.13
I	B27	B*2709	1W0W	999.9	968.1	1.0	18	9	0.14	0.11	0.94
I	B27	B*2709	1JGD	979.7	994.4	1.0	23	10	0.06	0.09	1.56
I	B27	B*2709	1K5N	780.7	879.9	1.1	19	9	0.13	0.07	1.10

I	B44	B*4402	1M6O[#]	937.3	903.2	1.0	16	9	0.00	0.00	0.00
I	B44	B*4403	1SYS	941.9	1076.0	1.1	15	9	0.08	0.24	0.93
I	B44	B*4403	1N2R	901.0	949.9	1.1	15	9	0.01	0.05	0.41
I	B44	B*4405	1SYV	901.8	915.6	1.0	15	9	0.02	0.08	0.26
I	B44	B*5101	1E27	865.2	809.0	0.9	8	9	0.15	0.01	1.91
I	B44	B*5101	1E28	804.9	724.0	0.9	7	8	0.11	0.07	1.75
I	B44	B*5301	1A1M	823.9	971.3	1.2	12	9	0.05	0.24	1.23
I	B44	B*5301	1A1O	965.2	778.8	0.8	10	9	0.09	0.04	0.67
II	DR1	DRB1*0101	1AQD[#]	1190.3	1182.7	1.0	18	14	0.00	0.00	0.00
II	DR1	DRB1*0101	1DLH	1139.1	1081.7	1.0	17	13	0.15	0.13	0.61
II	DR1	DRB1*0101	1FYT	1125.6	979.5	0.9	18	13	0.27	0.02	0.53
II	DR1	DRB1*0101	1HXY	1110.6	1246.7	1.1	15	13	0.14	0.08	0.66
II	DR1	DRB1*0101	1JWM	1129.6	1117.6	1.0	16	13	0.21	0.06	0.62
II	DR1	DRB1*0101	1JWS	1122.9	1095.7	1.0	16	13	0.17	0.05	0.62
II	DR1	DRB1*0101	1JWU	1114.0	1203.6	1.1	18	13	0.17	0.05	0.62
II	DR1	DRB1*0101	1KG0	1120.8	1238.1	1.1	15	13	0.18	0.03	0.65
II	DR1	DRB1*0101	1KLG	1124.8	1130.0	1.0	15	15	0.12	0.13	0.64
II	DR1	DRB1*0101	1KLU	1093.4	1217.8	1.1	16	15	0.14	0.14	0.68
II	DR1	DRB1*0101	1LO5	1119.1	978.3	0.9	13	13	0.15	0.13	0.62
II	DR1	DRB1*0101	1R5I	1120.4	1054.9	0.9	19	13	0.08	0.09	0.73
II	DR1	DRB1*0101	1SJE	1017.3	1084.0	1.1	16	15	0.07	0.04	0.67
II	DR1	DRB1*0101	1SJH	993.0	1127.8	1.1	16	13	0.04	0.04	0.70
II	DR1	DRB1*0101	1T5W	1067.7	792.4	0.7	13	13	0.25	0.05	0.15
II	DR1	DRB1*0101	1T5X	1056.9	906.1	0.9	13	13	0.24	0.08	0.16
II	DR1	DRB1*1501	1BX2	985.7	1269.3	1.3	15	14	0.12	0.12	0.94

Three types of interaction patterns for class I supertypes can be identified in this study: (i) the number of intermolecular hydrogen bonds directly correlates with gap index (Figures 11D and 14D); (ii) the number of intermolecular hydrogen bonds does not correlate with gap index (Figure 12D); and (iii) the number of intermolecular hydrogen bonds inversely correlates with the gap index (Figure 13D). The mean number of intermolecular hydrogen bonds for these three groups are 11.2 ± 2.3, 14.3 ± 2.3, and 17.7 ± 2.6. Overall, the data indicate that there is a lack of any real correlationship in the scattergrams.

For the first group (A2 and B44 supertypes), the majority of intermolecular hydrogen bonds are concentrated at both ends of the binding groove (in pockets A, B and F). More hydrogen bonds are observed with decreasing geometric and electrostatic complementarity (i.e. increasing gap index; Figures 11D and 14D). A possible explanation for these observations is that the overall geometric and electrostatic complementarities of these complexes are low and extensive hydrogen bonding in the binding pockets at both ends of the binding grooves are necessary to stabilize the complexes for this supertype. For the second group (B7 supertype), no clear relationship between intermolecular hydrogen bonds with gap index is observed. In general, the interaction mechanism employed by this supertype may be degenerative and a combination of non-covalent interactions (hydrogen bonds, hydrophobic and ionic interactions) may be involved in peptide selection. However, it is not clear to what extent the different interactive forces contribute to the net stability of complex. For the third group (B27 supertype), the number of intermolecular hydrogen bonds increases with higher geometric and electrostatic complementarity (smaller gap index). This group also consists of the highest mean number of intermolecular hydrogen bonds. The results strongly indicate that the complexes formed by this group may be more stable, with higher overall geometric and electrostatic complementarity. High conservation of peptide termini residues is observed in all class I binding peptides. The mean Cα RMSD for N- and C- terminal residues is 0.08 Ǻ and 0.09 Ǻ respectively. This observation has been previously reported for HLA-Aw68 (Guo *et al*., 1992) and confirmed by this analysis. A similar highly conserved backbone conformation is observed at the ends of the core peptide fragments in the binding cleft of DR1 molecules. The Cα RMSD values of the N- and C-termini reveal their relatively fixed locations within the groove across both classes of peptide/MHC complexes.

Analysis for class II supertypes is currently restricted to DR1 due to the lack of sufficient crystallographic structures for other class II supertypes in the current Protein Databank (PDB). Although 6 DR4 crystallographic structures exist, majority of the ligands (PDB ID: 1D5M, 1D5X, 1D5Z, 1D6E) are primarily short inhibitor fragments, rendering the remaining dataset for this supertype too limited for statistically significant analysis. For DR1 supertype, another type of interaction pattern is observed. The number of intermolecular hydrogen bonds directly correlates with gap volume (Figure 15C) but no clear correlations with gap index can be seen (Figure 15D). It is possible that this supertype may employ a similar degenerative interaction mechanism as B7, and also engage the use of an extensive hydrogen bonding network to hold longer or less complementary peptides together.

The present analysis is difficult due to the limited number of peptide/HLA crystallographic structures in the current PDB. As the database grows, more alleles and supertypes will be analysed. Nonetheless, through the use of existing three-dimensional structures, the author has demonstrated that different HLA superfamilies employ the use of different binding mechanism for selectivity of antigenic peptides. By focusing solely on the use of experimental three-dimensional structures, this analysis is supported and verified by existing data. The present analysis suggests that although similar interactions exist between MHCs and their corresponding bound peptides, their relative exploitation by the groups and geometry of the binding site will vary between alleles.

3.4 Summary

- In this work, a new database for sequence-structure-function information on (TCR/) peptide/MHC interactions, termed MHC-Peptide Interaction Database version T (MPID-T) has been developed to facilitate the analysis of (TCR/) peptide/MHC structural interaction characteristics.
- MPID-T is a manually curated MySQL database containing experimentally determined structures of 187 peptide/MHC complexes and 16 TCR/peptide/MHC complexes available in the PDB and precomputed interaction parameters including solvent accessibility, number of intermolecular hydrogen bonds, gap volume and gap index.

- Structural visualization of the T cell receptor/peptide/MHC complex, peptide/MHC complex, MHC or the bound peptide can be performed using freely available graphics applications such as Chime or RasMol, while structural alignment (based on MHC class and peptide length) can be viewed using the Jmol molecular viewer or a Chime-compatible web browser client.
- Each MPID-T record is hyperlinked to external immunologic databases including IMGT/HLA, IMGT/3Dstructure-DB, SYFPEITHI, AntiJen, among others. Pre-computed schematic diagrams based on LIGPLOT program are provided to illustrate explicit peptide/MHC interactions. Consensus patterns among peptides of the same length or allele are generated using the program WebLogo. Other useful sources of information as referenced in Rammensee *et al*. (1999) are also provided under MHC resources on the MPID-T help page.
- Four interaction parameters (intermolecular hydrogen bonds, interface area, gap volume, and gap index) previously identified as being significant for the characterization of peptide/MHC interface were applied for analyzing the binding characteristics of 5 HLA supertypes (A2, B7, B27, B44, and DR1).
- For all supertypes investigated in this study, the gap index (geometric and electrostatic complementation between peptide and MHC) inversely correlates with increasing interface area. This implies that complexes with larger interface area have better geometric and electrostatic

complementarity, resulting in the formation of more intermolecular hydrogen bonds.

- This is the first report on the existence of different interaction patterns in HLA supertypes which were identified using computed structural interaction parameters. Three types of interaction patterns for class I supertypes have been identified: (i) the number of intermolecular hydrogen bonds directly correlates with gap volume and gap index; (ii) the number of intermolecular hydrogen bonds does not correlate with both gap volume and gap index; and (iii) the number of intermolecular hydrogen bonds inversely correlates with gap volume and gap index.
- The N- and C-terminal residues of class I peptides are highly conserved with mean Cα RMSD of 0.08 Ǻ and 0.09 Ǻ respectively. A similar highly conserved backbone conformation is observed at the ends of the core peptide fragments in the binding cleft of DR1 molecules with mean Cα RMSD of 0.08 Å and 0.09 Å for the two peptide termini, respectively.
- The present analysis suggests that the use of a standardized set of structural interaction rules may not be applicable for all HLA alleles as interaction characteristics vary across MHC supertypes.

Chapter 4: Modeling the structure of bound peptide ligands to MHC

4.1 Introduction

In recent years, protein structure prediction has been gaining prominence in the field of structural biology. A 3D model for a receptor-ligand complex of unknown structure can provide valuable insights in the study of structure-activity relationships. In the context of peptide/MHC complex, the availability of such models also allows detailed analysis of peptide/MHC interaction characteristics and the prediction of potential immunodominant epitopes at allele-specific level without the need of large experimental dataset for training. However, despite the many benefits of structure-based modeling, few peptide/MHC docking techniques have been developed due to higher complexity in development and longer computational time. To this end, the author reports the development of a highly accurate docking protocol for efficient and fast modeling of peptide/MHC complexes. The methodology presented here is applicable to the design of both sub-type specific vaccines as well as promiscuous peptide epitopes.

4.2 Implementation

4.2.1 Selection of probe residues

The main problem in docking simulation is to enumerate the large number of possible combinations for two molecules to interact within an enclosed sampling space. There are six degrees of global-rotational and translational freedom of one molecule relative to the other, as well as one internal dihedral rotation per rotational bond. Given four consecutive atoms A_{i-2}, A_{i-1}, A_i, *and* A_{i+1}, the dihedral angle is defined as the smallest angle between the planes π_1 and π_2, as shown in Figure 16. Variation of the dihedral angle is a consequence of rotation of the two outer bonds about the central bond. There are two freely rotatable backbone dihedral angles per amino acid residue in the protein chain: the phi-angle is a consequence of the rotation about the bond between N and Cα, and the psi-angle is a consequence of the rotation about the bond between Cα and C. The peptide bond between N of one residue and C of the adjacent residue is not rotatable. There are two backbone dihedrals per amino acid, but the number of side chain dihedrals varies with the length of the side chain. Its value ranges from 0, in the case of glycine, which has no sidechain, to 5 in the case of arginine. A full search on the conformational space increases with increasing molecule size and sampling space. As such, a key challenge in docking simulation is to minimize the conformational search space of ligand within the large sampling space enclosed by the receptor binding site.

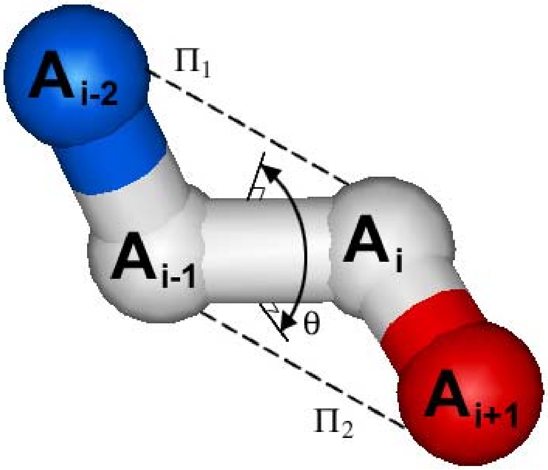


Figure 16: π_1 is the plane uniquely defined by the first three atoms A_{i-2}, A_{i-1}, and A_i. Similarly, π_2 is the plane uniquely defined by the last three atoms A_{i-1}, and A_i, and A_{i+1}. The dihedral angle, θ, is defined as the smallest angle between these two planes.

A possible approach is to identify suitable base or anchor fragments (referred to hereafter as the probes) for initiating docking simulations. A probe must satisfy two criteria: (i) the anchor must have sufficient contact with the receptor, and (ii) the structure of the anchor must be highly conserved. Probes that are too short in length will require the exploration of a larger search space and hence longer computational time, whilst probes that are too long may result in insufficient sampling of the receptor binding site. A systematic analysis of class I and class II peptide/MHC crystallographic structures (Chapter 5) revealed that the N- and C- termini residues of peptide binding registers are both highly conserved and in contact with the receptor binding pockets, thus offer a good starting point for docking simulation.

4.2.2 The peptide docking procedure

Beginning with the sequence of the ligand for which the structure is to be generated (referred to hereafter as the target peptide), and the availability of the MHC receptor structure, our docking protocol (Figure 17) for peptide binding registers consists of the following steps: (i) rigid docking of probe residues; (ii) loop closure of central residues by satisfaction of spatial constraints; followed by (iii) ab initio refinements of the binding register. An additional step (iv) extension of flanking residues is applied for the modeling of class II peptide ligands.

4.2.2.1 Rigid docking of probe residues

A fast soft-interaction energy function (Fernández-Recio *et al*., 2002) is adopted to dock each probe to the respective ends of the MHC binding groove. This is performed using an Internal Coordinate Mechanics (ICM; Abagyan and Maxim, 1999) global optimization algorithm; with flexible ligand interface side-chains and a grid map representation of the receptor energy localized to small cubic regions of 1.00Å radius from the backbone of each probe. Each probe performs a random walk within their respective grid map. At each step, the side-chain torsions were changed using a biased Monte Carlo procedure, which begins by randomly selecting a set of torsion angles in the probe and subsequently finding the local energy minimum about those angles. New conformations are adopted upon satisfaction of the Metropolis criteria with probability min(1,exp[-ΔG/RT]), where R is the universal gas constant and T is the absolute temperature of the simulation. Loose restraints were imposed on the positional variables of the ligand molecule to keep it close to the starting conformation. The stimulation temperature was set to 300K. The optimal energy function (Equation 3) used during stimulations consisted of the internal energy of the probe and the intermolecular energy based on the same optimized potential maps used in the docking step:

$$E = E_{Hvw} + E_{Cvw} + 2.16E_{el}^{solv} + 2.53E_{hb} + 4.35E_{hp} + 0.20E_{solv}$$

(*Equation 3*)

The internal energy included internal van der Waals interactions (hydrogen probe: E_{Hvw}; heavy atom probe: E_{Cvw}), electrostatic potential (E_{el}^{solv}), hydrogen bonding (E_{hb}), hydrophobicity (E_{hp}) and solvation energy (E_{solv}) calculated with ECEPP/3 parameters, and the Coulomb electrostatic energy with a distance-dependent dielectric constant (e=4r). The configurational entropy of side-chains and the surface-based solvation energy were included in the final energy to select the best-refined solutions.

4.2.2.2 Construction of loop

In this stage, an initial conformation of the central loop is generated by satisfaction of spatial constraints (Sali and Blundell, 1993) based on the allowed subspace for backbone dihedrals in accordance with the conformations of peptides docked into the ends of the binding groove. This is performed in three-steps: (i) Distance and dihedral angle restraints on the entire peptide sequence are derived from its alignment with the sequences of probes docked into the binding groove. (ii) The restraints on spatial features of the unknown center residues are derived by extrapolation from the known 3D structures of probes in the alignment, expressed as probability density functions. Stereochemical restraints include bond distances, bond angles, planarity of peptide groups and side-chain rings, chiralities of Cα atoms and side-chains, van der Waals contact distances and the bond lengths, bond angles and dihedral angles of cysteine disulfide bridges. (iii) Spatial restraints on the unknown center residues are satisfied by optimization of the molecular probability density function using variable target function technique that applies the conjugate gradients algorithm to positions of all non-hydrogen atoms.

4.2.2.3 Refinement of binding register

To improve the accuracy of the initial model, partial refinement was performed for both the ligand backbone and side-chain, using ICM biased Monte Carlo procedure (Abagyan and Maxim, 1999). Initial stages of refinements attempt to overcome the penalty derived from the initial rigid docking of terminal residues by introducing partial flexibility to the ligand backbone. Restraints were imposed upon the positional variables of the Cα atoms of probes to keep it close to the starting conformation. The energy function adopted for this refinement step is shown in Equation 4:

$$E = E_{vw} + E_{hbonds} + E_{torsions} + E_{electr} + E_{solv} + E_{entropy}$$

(*Equation 4*)

The internal energy included internal van der Waals interactions (E_{vw}), entropic energy ($E_{entropy}$), electrostatic potential (E_{electr}), hydrogen bonding (E_{hbonds}), torsion energy ($E_{torsions}$) and solvation energy (E_{solv}). Refinements of ligand and receptor side-chain torsions in the vicinity of 4.00 Å from the receptor were performed upon the final backbone structure.

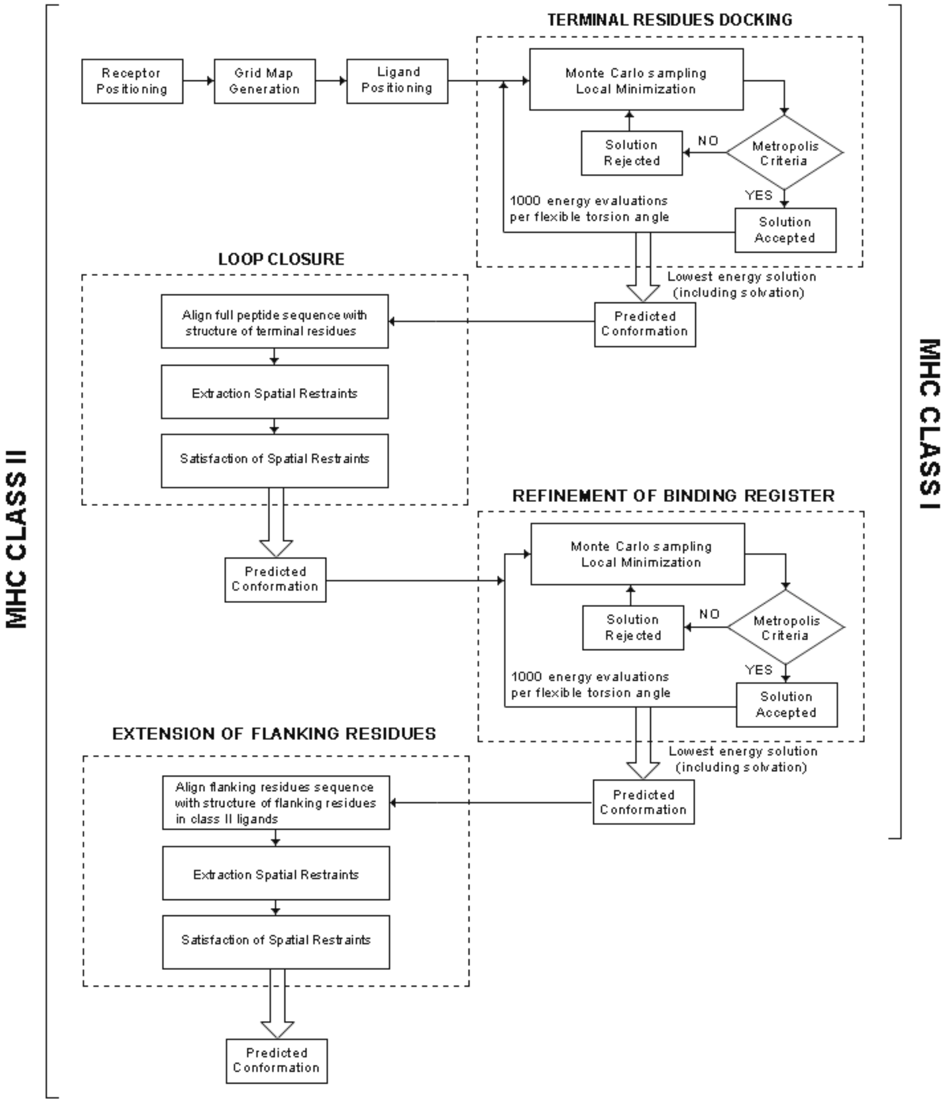

TERMINAL RESIDUES DOCKING
Receptor Positioning
Grid Map Generation
Ligand Positioning
Monte Carlo sampling Local Minimization
Solution Rejected
NO
Metropolis Criteria
YES
1000 energy evaluations per flexible torsion angle
Solution Accepted
LOOP CLOSURE
Lowest energy solution (including solvation)
Align full peptide sequence with structure of terminal residues
Predicted Conformation
Extraction Spatial Restraints
Satisfaction of Spatial Restraints
MHC CLASS I
MHC CLASS II
REFINEMENT OF BINDING REGISTER
Monte Carlo sampling Local Minimization
Predicted Conformation
Solution Rejected
NO
Metropolis Criteria
YES
1000 energy evaluations per flexible torsion angle
Solution Accepted
EXTENSION OF FLANKING RESIDUES
Lowest energy solution (including solvation)
Align flanking residues sequence with structure of flanking residues in class II ligands
Predicted Conformation
Extraction Spatial Restraints
Satisfaction of Spatial Restraints
Predicted Conformation

Figure 17 Flowchart of docking procedure used in this work.

4.2.2.4 Construction of flanking residues

At this stage, MHC class I ligand models have been fully constructed and the following task is applicable only to MHC class II ligands. Here, the only construction remaining is the flanking residues that extend out of the MHC class II binding groove. The conformations of the flanking peptide residues are generated by satisfying the spatial constraints in the allowed subspace for backbone dihedrals (Sali and Blundell, 1993), defined by the conformation of the bound core nonameric peptide docked into the binding groove. This is performed in three stages: (i) distance and dihedral angle restraints on the entire peptide sequence are derived from its alignment with the nonamer sequence in the binding groove; (ii) the restraints on spatial features of the flanking residues are derived by extrapolation from the known 3D structure of flanking residues in the alignment, expressed as probability density functions; and (iii) the spatial restraints on the flanking residues are then satisfied by optimization of the molecular probability density function using a variable target function technique that applies the conjugate gradients algorithm to positions of all non-hydrogen atoms.

4.3 Results

Evaluation of our docking procedure is performed systematically in the following three tests: (i) self-docking 40 test case complexes; (ii) cross-docking of 15 solved peptides into templates of appropriate alleles; and (iii) validation against existing techniques.

4.3.1 Self-docking bound peptides to MHC molecules

To validate our docking procedure, we first applied our technique to the rebuilding of 40 non-redundant MHC-peptide complexes by docking peptides extracted from peptide/MHC complexes back into their respective binding grooves. This initial experiment is an important first step for testing the capability of our technique to model peptides into their cognate MHC receptors. Peptides were separated from experimental structures and remodeled back into their own bound states. A correct docking result is defined as a complex with not more than 2.50 Å Cα RMSD from the known experimental structure. The RMSD for the near-native solution ranges from 0.09 Å (complex 1G7Q) to 1.53 Å (complex 1JF1, Figure 18). Our procedure generated 33 out of 40 non-redundant complexes (Table 4 and Figure 19) within a Cα RMSD of 1.00 Å.

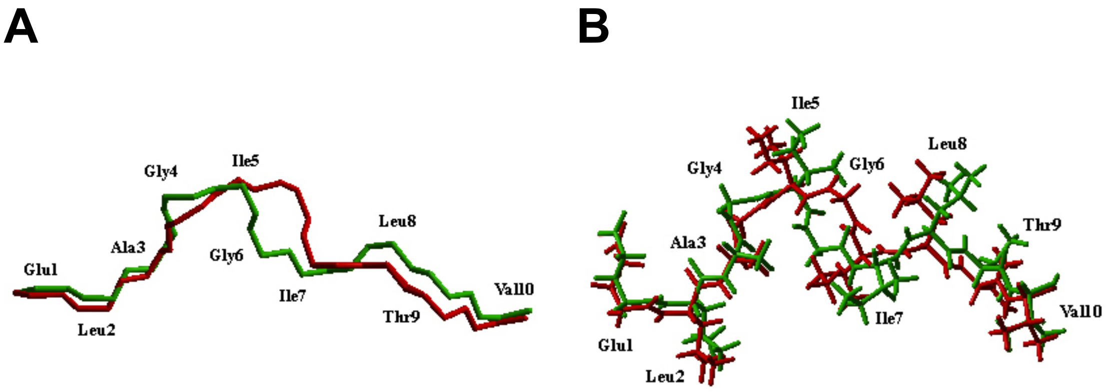


Figure 18 Comparison of the predicted and experimental structures of the ELAGIGILTV peptide in the 1JF1 complex (Table 2). The crystal structure (in red) and modeled structures (in green) are shown in (A) Cα trace representation, and (B) stick representation of all heavy atoms.

Table 4 Comparison of the position the bound peptide in the original crystal structure and after docking back into the MHC groove. RMSD values are calculated for the ligand interface Cα atoms of the lowest energy solution, superimposed onto the experimental structure.

Class	Allele	PDB	Res (Å)	Length	RMSD (Å)	Sequence
I	HLA-A*0201	1DUZ	1.80	9	0.33	LLFGYPVYV
I	HLA-A*0201	1HHG	2.60	9	0.46	TLTSCNTSV
I	HLA-A*0201	1HHJ	2.50	9	0.87	ILKEPVHGV
I	HLA-A*0201	1HHH	3.00	10	1.10	FLPSDFFPSV
I	HLA-A*0201	1I1Y	2.20	9	0.70	YLKEPVHGV
I	HLA-A*0201	1I4F	1.40	10	0.49	GVYDGREHTV
I	HLA-A*0201	1I7R	2.20	9	0.59	FAPGFFPYL
I	HLA-A*0201	1I7U	1.80	9	0.32	ALWGFVPVL
I	HLA-A*0201	1JF1	1.85	10	1.53	ELAGIGILTV
I	HLA-A*0201	1JHT	2.15	9	0.54	ALGIGILTV
I	HLA-A*0201	1OGA	1.40	9	0.32	GILGFVFTL
I	HLA-A*0201	1QRN	2.80	9	0.46	LLFGYAVYV
I	HLA-A*0201	1QSE	2.80	9	0.26	LLFGYPRYV
I	HLA-A*0201	1QSF	2.80	9	0.54	LLFGYPVAV
I	HLA-A*6801	1TMC	2.30	10	0.52	EVAPPEYHRK
I	HLA-B*0801	1AGD	2.05	8	0.28	GGKKKYKL
I	HLA-B*0801	1AGF	2.20	8	0.66	GGKKRYKL
I	HLA-B*3501	1A1N	2.00	8	0.10	VPLRPMTY
I	HLA-B*3501	1A9E	2.50	9	1.09	LPPLDITPY
I	HLA-B*5101	1E27	2.20	9	1.27	LPPVVAKEI
I	HLA-B*5301	1A1M	2.30	9	0.59	TPYDINQML
I	HLA-B*5301	1A1O	2.30	9	0.78	KPIVQYDNF
I	H2-Db	1JPF	2.18	11	1.14	SGVENPGGYCL
I	H2-Db	1JPG	2.20	9	0.33	FQPQNGQFI
I	H2-Dd	1BII	2.40	10	1.49	RGPGRAFVTI
I	H2-Kb	1FZM	1.80	8	0.32	RGYVYQGL
I	H2-Kb	1FZO	1.80	9	0.40	FAPGNYPAL
I	H2-Kb	1G7P	1.50	9	0.97	SRDHSRTPM
I	H2-Kb	1G7Q	1.60	8	0.09	SAPDTRPA
II	HLA-DR1	1AQD	2.45	10	0.63	DWRFLRGYHQ
II	HLA-DR1	1AQD	2.45	10	1.08	DWRFLRGYHQ
II	HLA-DR2	1BX2	2.60	10	0.60	VVHFFKNIVT
II	HLA-DR2	1BX2	2.60	10	0.81	VVHFFKNIVT
II	HLA-DR2	1FV1	1.90	10	0.47	HFFKNIVTPR
II	HLA-DR2	1FV1	1.90	10	0.58	HFFKNIVTPR
II	HLA-DR3	1A6A	2.75	10	0.38	KMRMATPLLM
II	HLA-DR4	1J8H	2.40	10	0.59	KYVKQNTLKL
II	HLA-DR4	2SEB	2.50	10	0.43	YMRADAAAGG
II	HLA I-Ak	1IAK	1.90	10	0.42	TDYGILQINS
II	HLA-DQ8	1JK8	2.40	10	0.21	VEALYLVCGE

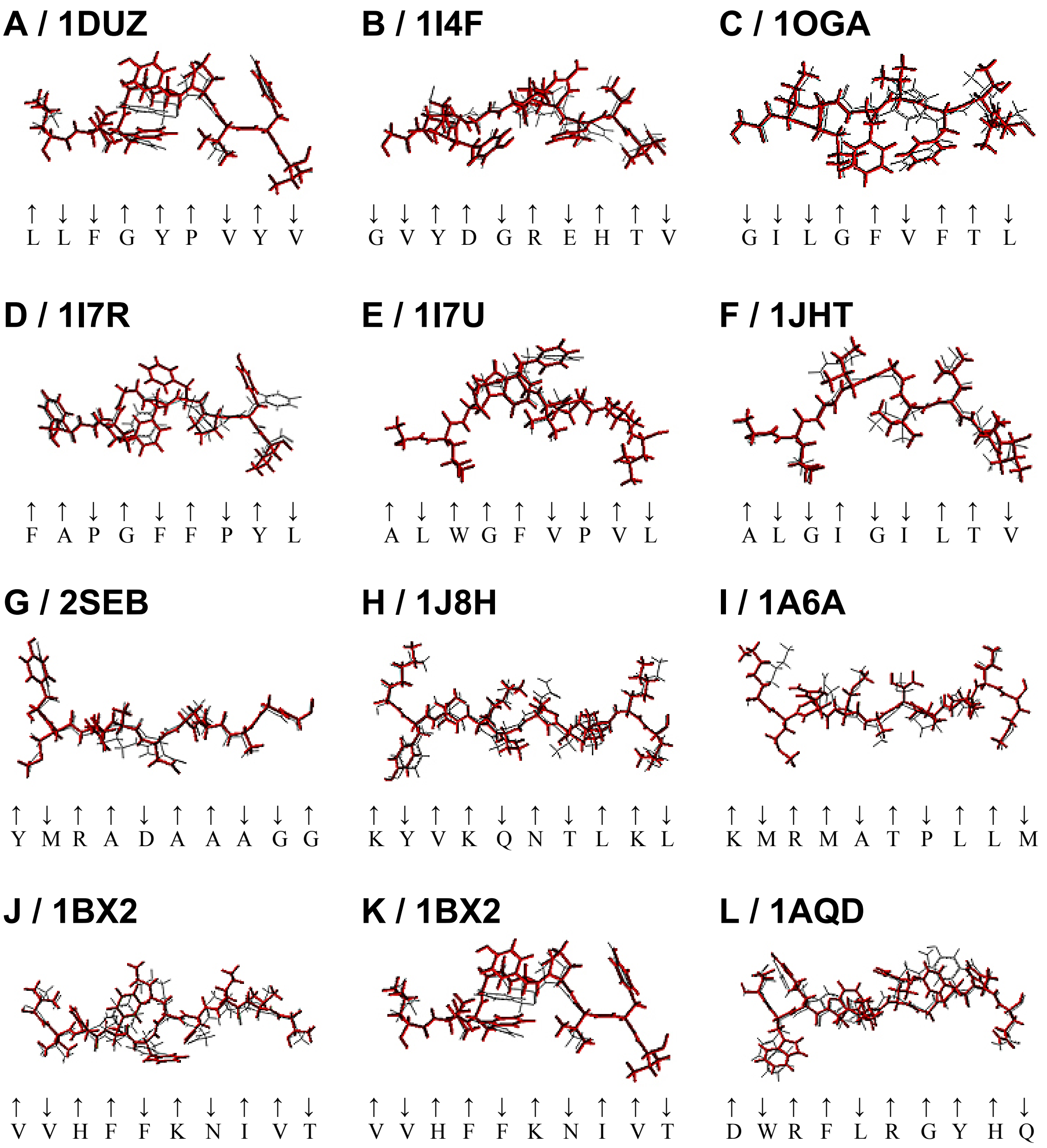


Figure 19 Representations of selected lowest energy solutions in the binding grooves obtained after redocking the peptides into the respective MHC grooves in the first benchmarking test (Table 4). Experimental peptide structures are represented as bold dark lines and remodeled structures as thin grey lines, showing all heavy atoms for MHC Class I (A-F) and class II (G-L) complexes. The relative orientations of the peptide side chains with respect to the floor of the binding groove are indicated by arrows pointing either up (away from the groove) or down (towards the groove).

Table 5 Comparison between modeled peptides and relevant crystal structures after docking onto a single template. RMSD values calculated for the ligand interface Cα atoms of the lowest energy solution superimposed onto the experimental PDB structure are listed.

Class	Allele	PDB	Length	Template	RMSD (Å)	Sequence
I	HLA-A*0201	1DUZ	9	1I4F	0.69	LLFGYPVYV
I	HLA-A*0201	1HHG	9	1I4F	0.58	TLTSCNTSV
I	HLA-A*0201	1HHJ	9	1I4F	0.73	ILKEPVHGV
I	HLA-A*0201	1HHH	10	1I4F	1.48	FLPSDFFPSV
I	HLA-A*0201	1I1Y	9	1I4F	0.77	YLKEPVHGV
I	HLA-A*0201	1I7R	9	1I4F	0.60	FAPGFFPYL
I	HLA-A*0201	1I7U	9	1I4F	0.70	ALWGFVPVL
I	HLA-A*0201	1JF1	10	1I4F	1.20	ELAGIGILTV
I	HLA-A*0201	1JHT	9	1I4F	1.09	ALGIGILTV
I	HLA-A*0201	1OGA	9	1I4F	0.38	GILGFVFTL
I	HLA-A*0201	1QRN	9	1I4F	0.81	LLFGYAVYV
I	HLA-A*0201	1QSE	9	1I4F	0.52	LLFGYPRYV
I	HLA-A*0201	1QSF	9	1I4F	0.57	LLFGYPVAV
II	HLA-DR2	1BX2	10	1FV1	1.22	VVHFFKNIVT
II	HLA-DR4	2SEB	10	1J8H	0.42	KYVKQNTLKL

Table 6 Benchmarking of our MHC-peptide procedure with previously published studies in MHC class I peptide modeling. *RMSD of peptide Cα atoms obtained in our work from self- and cross-docking respectively.

Peptide Sequence	Technique	Reference	*RMSD (Å)	
			Ref.	Current
TLTSCNTSV	Simulated Annealing	Rognan *et al*. (1999)	1.04	0.46, 0.58
FLPSDFFPSV	Simulated Annealing	Rognan *et al*. (1999)	1.59	1.10, 1.48
GILGFVFTL	Simulated Annealing	Rognan *et al*. (1999)	0.46	0.32, 0.38
ILKEPVHGV	Simulated Annealing	Rognan *et al*. (1999)	0.87	0.87, 0.73
LLFGYPVYV	Simulated Annealing	Rognan *et al*. (1999)	0.78	0.33, 0.69
RGYVYQGL	Combinatorial Algorithm	Desmet *et al*. (2000)	0.56	0.32, 0.66
FAPGNYPAL	Multiple copy Algorithm	Rosenfeld *et al*. (1993, 1995)	2.70	0.40, 0.90
GILGFVFTL	Multiple copy Algorithm	Rosenfeld *et al*. (1993, 1995)	1.40	0.32, 0.38
LLFGYPVYV	Combinatorial Algorithm	Sezerman *et al*. (1996)	1.40	0.33, 0.69
ILKGPVHGV	Combinatorial Algorithm	Sezerman *et al*. (1996)	1.30	0.87, 0.73
GILGFVFTL	Combinatorial Algorithm	Sezerman *et al*. (1996)	1.60	0.32, 0.38
TLTSCNTSV	Combinatorial Algorithm	Sezerman *et al*. (1996)	2.20	0.46, 0.58

This preliminary experiment establishes the validity of our approach, using the three-step proposed procedure. Encouraged by these results, we next apply our procedure to a more practical problem in allele-specific vaccine design, that is, the modeling of peptide/MHC complexes resulting from multiple peptides binding to a single MHC allele template.

4.3.2 Cross-docking peptides onto a single template

We next applied our technique to the modeling of 15 non-redundant peptides (13 class I and 2 class II) for which crystal structures are available into a single template. This stage of the testing is critical to determine the capability of our procedure to model unknown peptides onto available templates. Due to the deficiency of available class II crystal structures, only 2 class II peptides are tested in this stage. Our procedure constantly found a solution with RMSD below 1.48 Å. Table 5 shows the results obtained from this experiment.

4.3.3 Comparison with existing approaches

In order to determine the validity and accuracy of our procedure, we benchmark our technique with four previously published studies involving MHC class I peptide modeling as detailed in Table 6. As there was no previously reported accuracy for MHC class II peptide modeling, no benchmarking could be performed on the modeled MHC class II peptides. It is notable that validation process by Rognan *et al.* (1999), Desmet *et al.* (2000) and Sezerman *et al.* (1996) involved remodeling peptides back into their original crystal structure.

Using this criterion, our procedure is either comparable or outperforms the three earlier studies (Rognan *et al.*, 1999; Desmet *et al.*, 2000; Sezerman *et al.*, 1996) in terms of the Cα RMSD of the modeled peptides.

4.4 Discussion

Modeling the bound conformation of MHC-binding peptides is a complex problem in the field of immunology. In this chapter, a generic protocol for the modeling of both MHC class I and class II complexes has been developed. The proposed procedure forms a basis for the prediction of peptides that will bind to specific MHC alleles and hence vaccine design, based on computational immunological methods. To the best of the author's knowledge, the current study presents one of the most accurate peptide/MHC flexible docking techniques to date. The docking procedure has been assessed against a large dataset of non-redundant peptide/MHC complexes in which three-dimensional information is available. Out of 40 peptides considered in this study (Table 4), we have consistently obtained a Cα RMSD below 1.00 Å for 33 peptides by remodeling peptide-bound MHC structures.

The worst structure was generated from the remodeling of the bound peptide ELAGIGILTV from complex 1JF1 with Cα RMSD of 1.53 Å. The loop formed around residues 5 to 7 was erroneously predicted and this misplacement is a direct consequence of missing water molecules positioned around the loop in the template, which resulted in incorrect positioning of interacting residues. In the absence of explicit water molecules, the predicted conformation of our peptide is energetically more favorable than the crystal conformation. Nonetheless, our procedure can correctly predict the conformation of residues that extends into the binding cleft and identify essential contacts with the MHC receptor as shown in Figure 18. While water molecules and other common biological ions such as phosphate and chloride may mediate peptide/MHC interactions, they were left out in our preliminary experiments in order to determine the generic prediction capability of our docking protocol using a single template for each allele since the significance and contributions of these molecules varies between different peptides and the respective alleles. It is possible that for some peptide/MHC complexes, appropriate addition of mediating molecules or considerations of solvent effects may lead to an improvement in prediction accuracy.

The performance of our method, in terms of computational time, is highly efficient and requires approximately 11 minutes for the complete modeling of one peptide (with the first rigid-body docking step of ~3.5 minutes, loop closure of ~12 seconds and the final refinement step of ~7 minutes) on a 4-CPU SGI Origin 3200 workstation. Rapid flexible docking of target peptide into the receptor binding groove (with rigid backbone and flexible side chains) is possible by restraining the conformational spaces to be sampled in the early phase of our modeling protocol (please refer to *section 6.2.2.1 Rigid docking of probe* for details). Large scale modeling and scanning of potential MHC-binding sequences is possible through automation for all steps. Our docking procedure also proved to be capable of accurately modeling MHC-peptide complexes in the absence of essential anchor residues by exploiting the highly conserved backbone conformation of bound MHC class I and class II peptide termini.

4.5 Summary

- This work reports on the development of an efficient and fast docking protocol for modeling the bound conformation of peptide ligands to MHC class I and class II molecules. To the best of the author's knowledge, the current study presents one of the most accurate peptide/MHC flexible docking techniques to date.
- High prediction accuracy was obtained in three tests: (i) self-docking 40 test case complexes; (ii) cross-docking of 15 solved peptides into templates of appropriate alleles; and (iii) validation against existing techniques.
- The methodology reported here also proved to be capable of accurate modeling of MHC-peptide complexes in the absence of essential anchor residues by exploiting the highly conserved backbone conformation of bound MHC class I and class II peptide termini.
- This work demonstrates that structure-based predictive technique can be applied to the systematic functional analysis of MHC-binding peptides. This generic approach is applicable for the modeling of both MHC class I and class II complexes. The proposed procedure forms a basis for the prediction of peptides that will bind to specific MHC alleles and hence vaccine design, based on computational immunological methods.

Chapter 5: Analysis of PV associated and non-associated alleles

5.1 Introduction

Pemphigus Vulgaris (PV) is a potentially life-threatening form of autoimmune blistering skin disorder due to the loss of integrity of normal intercellular attachments within the epidermis and mucosal epithelium. The disease is characterized by the presence of pathogenic autoantibodies directed against a 130-kDa transmembrane glycoprotein, desmoglein-3 (Dsg3) (Amagai, 1994), within the desmosomes of the spinous layer of the skin. Although Dsg3 is thought to be important in maintaining cell-to-cell adhesion, there have been few in vivo models that confirm their actual function in the normal structure and function of hair (Koch *et al*., 1998). Strong association of PV to the major histocompatibility complex class II serotypes DR4 and DR6 has been reported in the literature (Ahmed *et al*., 1990, 1991; Scharf *et al*., 1989) with over 95% of PV patients possessing one or both of these alleles (Scharf *et al*., 1989). Direct nucleotide sequence analysis of DR4 and DR6 subtypes revealed that susceptibility to PV is strongly linked to DRB1*0402 and DQB1*0503 molecular subtypes, respectively (Scharf *et al*., 1989; Sinha *et al*., 1988).

This chapter aims to understand the functional correlation between MHC class II alleles and PV, from a structural interaction view point. Molecular modeling of ten PV associated and non-associated MHC class II receptors (DR4: DRB1*0401, *0402, *0404, *0406, DR6 (also classified now as DR14):

DRB1*1401, *1404, *1405, DQ2: DQB1*0201, *0202 and DQ5: DQB1*0503) were performed to explore the structural organization of the binding groove of these alleles. Nine previously identified epitopes, Dsg3 96-112, Dsg3 191-205, Dsg3 206-220, Dsg3 252-266, Dsg3 342-356, Dsg3 380-394, Dsg3 763-777, Dsg3 810-824 and Dsg3 963-977 (numbered in accordance with Swiss-Prot accession number P32926), capable of stimulating patient derived T cells, were selected. The binding of these peptides to the DR and DQ structural models were studied using the computational docking protocol discussed previously (Chapter 7). This is, to the author's knowledge, the first study of its kind, where structural principles have been used to discriminate between peptide binders and non-binders, for a number of disease-implicated and non-disease-implicated alleles. In the light shed by these atomic models, the binding specificities of each allele to the various Dsg3 peptides are discussed. The results obtained in the study are able to discriminate between PV associated and non-associated alleles, consistent with the experimental results obtained by Veldman *et al*. (2003) and Sinha *et al*. (unpublished results for Dsg3 342-356, 810-824 and 963-977). Insights into structural features behind the immune response provided by protective alleles for PV have also been obtained by our structural immunoinformatics approach.

5.2 Materials and Methods

5.2.1 Template search

In this study, ten PV associated, closely related non-associated and protective MHC class II alleles DRB1*0401, *0402, *0404, *0406, *1401, *1404, *1405, DQB1*0201, *0202, and *0503 were selected for analysis. MHC sequence data were obtained from the IMGT-HLA database (http://www.ebi.ac.uk/imgt/hla/). The α chain of all DR alleles investigated in this study is the DRA1*0101 sequence, with the β chain from the allele sub-type. To identify potential structural templates available in the PDB for model building, a sequence similarity search was performed using BLAST (Altschul *et al*., 1990) running on the servers at NCBI (www.ncbi.nlm.nih.gov/blast/) and the highest quality templates were selected among the returned results. Among these, the crystal structures of HLA-DR4 (PDB code 1D5Z) and HLA-DQ2 (PDB code 1S9V) were adopted as the structures of DRB1*0401 and DQB1*0201 respectively (100% sequence identity). The crystal structures of DRB1*0401 (PDB code 1D5Z), DQB1*0602 (PDB code 1UVQ) and DQB1*0201 (PDB code 1S9V) were selected as templates for all other DR subtypes, DQB1*0503 and DQB1*0202 respectively (Table 7).

5.2.2 Model building

The program MODELLER (Sali and Blundell, 1993) was employed for comparative modeling of both DRB1 (*0402, *0404, *0406, *1401, *1404, *1405) and DQB1 (*0202, *0503) subtypes. The models are constructed by optimally satisfying spatial constraints obtained from the alignment of the template structure with the target sequence and from the CHARMM-22 force field (MacKerell *et al*., 1998). The initial model was refined by assigning the rotameric states of essential side chains according to the corresponding crystal structure, followed by a short energy minimization (Abagyan *et al*., 1994) using the program Internal Coordinates Mechanics (ICM; Molsoft LLC, San Diego, CA) (Abagyan *et al*., 1999).

5.2.3 Peptide set

Nine previously identified epitopes Dsg3 96-112, 191-205, Dsg3 206-220, Dsg3 252-266, Dsg3 342-356, Dsg3 380-394, Dsg3 763-777, Dsg3 810-824 and Dsg3 963-977 that elicited primary proliferative T cell response in PV patients (Sinha *et al*., 1988, 1990; Veldman *et al*., 2003; Wucherpfennig *et al*., 1995; Hertl *et al*., 1998) were selected for modeling studies. T cell response to eight of these peptides (Dsg3 191-205, Dsg3 206-220, Dsg3 252-266, Dsg3 342-356, Dsg3 380-394, Dsg3 763-777, Dsg3 810-824 and Dsg3 963-977) has been reported in patients carrying DRB1*0402. Dsg3 96-112 has been reported to elicit T cell response in patients with DQB1*0503 but lacking DRB1*0402 (Veldman *et al*., 2003). Of these Dsg3 191-205, Dsg3 342-356, Dsg3 810-824 and Dsg3 963-977 were shown to directly bind to DRB1*0402 by competitive binding assays (Sinha *et al*., personal communications). Briefly, soluble HLA DRA1*0101/DRB1*0402 were purified by DR-specific affinity chromatography and incubated with different concentrations of experimental peptides (0-40 μM) in the presence of biotinylated class II-associated invariant-chain peptide (CLIP) (1 μM) for 2 hours. The MHC-peptide complexes were then captured on a 96-well plate coated with anti-HLA-DR (L243) (BD Pharmingen, San Diego, CA). The CLIP bound to the MHC molecules was directly assayed using Europium (Eu)-labeled streptavidin (Perkin Elmer, Boston, MA). The relative binding of peptides was subsequently determined by measuring the displacement of the CLIP at different peptide concentrations.

Table 7 Sequence and structural similarity between the eight (DRB1*0402, *0404, *0406, *1401, *1404, *1405, DQB1*0202, and *0503) MHC structural models and their corresponding template structures (1D5Z: DRB1*0401, 1S9V: DQB1*0201, 1UVQ: DQB1*0602). Positives represent a measure of sequence similarity, accounting for identical and conservatively substituted residues. Root mean square deviations (RMSD) values in Å are shown for the Cα atoms of both MHC chains and for the residues comprising the different peptide-binding pockets.

Allele	Template	Sequence Identity	Positives	Cα RMSD (Å)					
				α & β chains	Pockets				
					P1	P4	P6	P7	P9
DRB1*0402	1D5Z	97.9%	99.0%	0.35	0.12	0.06	0.07	0.10	0.09
DRB1*0404	1D5Z	99.0%	99.5%	0.31	0.15	0.10	0.06	0.07	0.18
DRB1*0406	1D5Z	97.9%	98.4%	0.32	0.11	0.15	0.07	0.11	0.22
DRB1*1401	1D5Z	94.1%	97.3%	0.25	0.11	0.09	0.02	0.09	0.18
DRB1*1404	1D5Z	85.8%	89.5%	0.29	0.12	0.10	0.02	0.06	0.22
DRB1*1405	1D5Z	81.0%	83.2%	0.24	0.11	0.07	0.02	0.08	0.07
DQB1*0202	1S9V	98.0%	99.0%	0.57	0.16	0.09	0.04	0.15	0.05
DQB1*0503	1UVQ	93.0%	96.0%	0.39	0.03	0.07	0.01	0.10	0.06

5.2.4 Peptide docking

Analysis of binding motifs (www.syfpeithi.de) (Rammensee *et al*., 1999) and available crystal structures suggested a core region of nine amino acids as essential for binding. To represent the possibility that any core peptide sequences can be recognized by the binding groove of MHC class II alleles, a sliding window input of size nine as illustrated in Figure 20 was applied to generate all possible combinations of core nonamer peptides from each Dsg3 peptide. This method can eliminate any bias in selecting core peptides based on sequence patterns alone. Each core peptide fragment is docked into the binding groove using the docking protocol described previously (Chapter 6). For each ligand, the best solution is obtained based on the following criteria: pattern of hydrogen bonding to the MHC molecule, pattern of hydrophobic burial of peptide side chains, and the absence of atomic clashes or repulsive contacts.

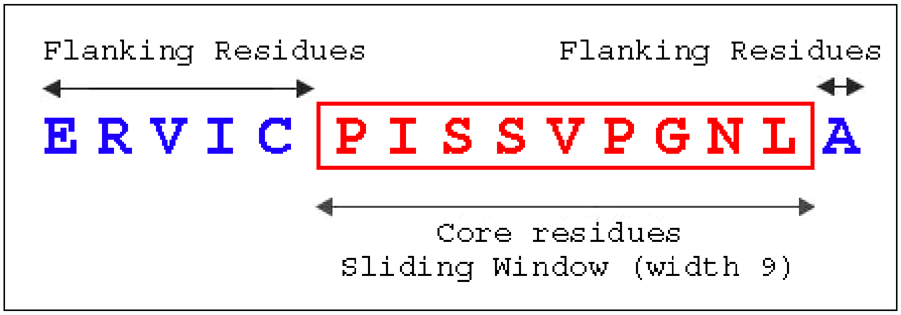


Figure 20 Sliding window of width 9 applied to identify core residues of Dsg3 963-977 to be modeled into binding groove.

5.2.5 Definition of contact residues

In this study, peptide/MHC residues were considered to be in contact if at least one pair of their non-hydrogen (“heavy”) atoms was found to be within 4.00 Å radius (Fischer and Marquesee, 2000). Intra-peptide interactions and intra-MHC interactions were not considered as they have minor influence on peptide/protein backbone structure. Any atom in the peptide and any atom in the MHC were considered to be experiencing atomic clash if their separation is below 2.00 Å (Samudrala and Moult, 1997) for non-hydrogen atoms and below 1.60 Å for atoms participating in hydrogen bonds (Wallace *et al*., 1995; Samanta *et al*., 2002).

5.2.6 Definition of binding pockets for MHC class II alleles

Interactions between side-chains of bound peptide ligands and polymorphic cavities (or anchor “pockets”) in the binding site of MHC class II alleles are important in determining the peptide binding affinity and sequence specificity of MHC molecules and are defined according to the work of Stern *et al*. (Stern and Wiley, 1994; Murthy and Stern, 1997).

5.3 Results and Discussion

5.3.1 Alleles comparisons

5.3.1.1 DR4 PV

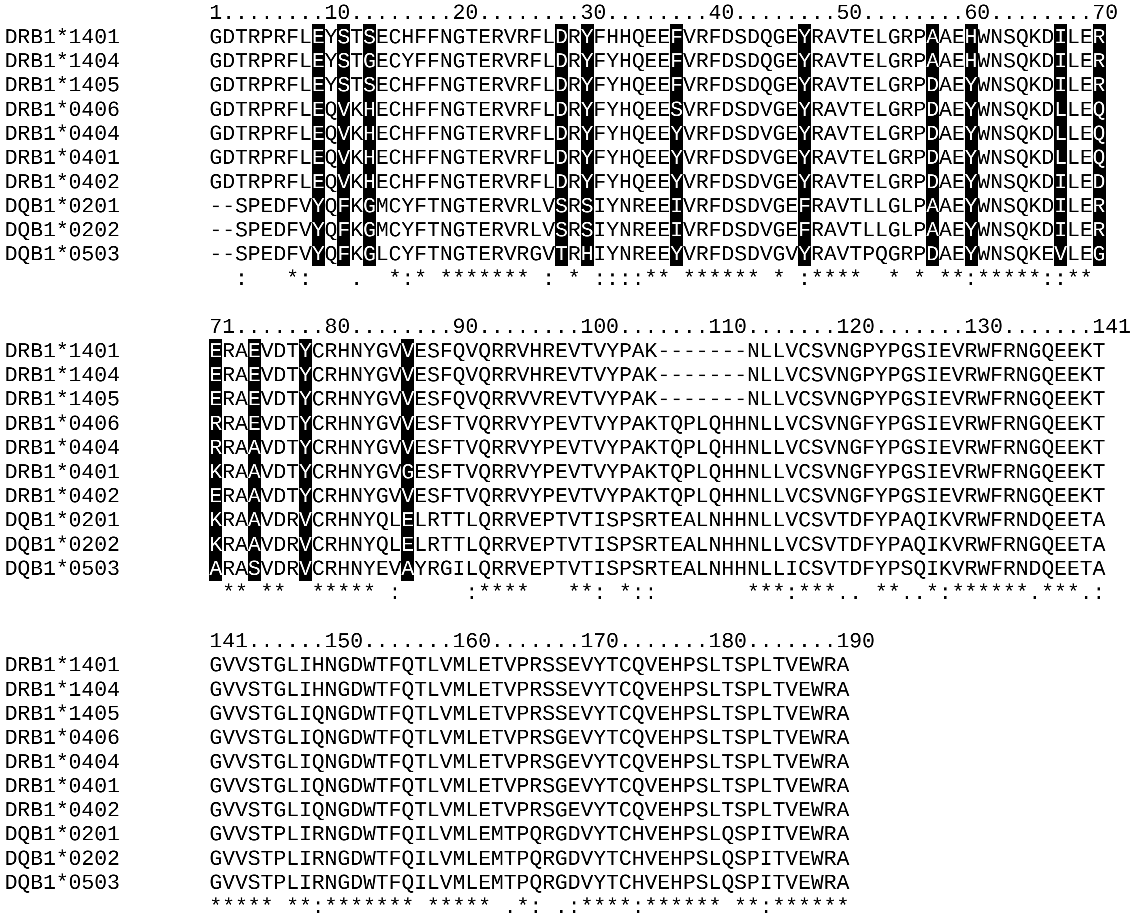
The sequence identity between the DR4 alleles (excluding DRB1*0401) with their corresponding templates ranges from 97.9 to 99.0%, and the sequence similarity (representing identical and conservatively substituted residues) was between 98.4 and 99.5% (Table 7). All five important peptide-binding pockets 1, 4, 6, 7 and 9 show extremely high structural conservation at the Cα positions, suggesting that any peptide discrimination leading to epitope selection between the alleles is mainly due to the size and nature of the side chains of the pocket residues. In order to further isolate the true disease-relevant allele within a haplotype, we compared specific residues in the polymorphic pockets regarded as important in conferring specificity for antigen presentation (Figure 21). Pocket 1, characterized by a Val/Gly β86 dimorphism, is the deepest cavity and thus, the most important anchor for peptide binding (Wucherpfennig *et al*., 1995). In addition, the functional specificity of DR4 molecules is also affected by polymorphisms at position β70, β71, β74, which contribute to pocket 4. Two negatively charged residues at position β70 and β71 that were previously suggested to influence peptide selectivity in PV patients (Hertl *et al*., 1998) could be found in DRB1*0402 (Asp β70 and Glu β71) but a positively charge Arg/Lys β71 was found in DRB1*0404, *0406 and *0401. Amino acid polymorphism can also be observed at position β11 of pocket 6, β71 of pocket 7 and β37 of pocket 9 respectively.

5.3.1.2 DR6 PV

Study of individual allele frequencies in DR6 PV patients revealed that the relevant disease susceptibility allele is DQB1*0503 instead of DR6 alleles (Sinha *et al*., personal communications). DQB1*0503 and the DR6 PV non-associated alleles investigated in this study show a significant degree of overlap in alignment, with 14 amino acid differences in areas of the binding cleft that could affect peptide binding. Clear differences in the amino acid sequences are observed at residue β86 of pocket 1, residues β13, β70, β71, β74, β78 of pocket 4, residue β11 of pocket 6, residues β28, β30, β67, β71 of pocket 7 and residues β9, β37, β57, β60 of pocket 9. Similar to the DR4 alleles, all five important peptide-binding pockets 1, 4, 6, 7 and 9 in DBQ1*0503 and DR6 alleles demonstrate exceptionally high structural conservation at the Cα positions. A significant difference is that DQB1*0503 contains a negatively charged Asp β57 that differs from the uncharged Ala β57 found in non-PV associated DRB1*1401 and *1404. Also, at positions β70 and β71, DQB1*0503 does not contain negatively charged residues identified in DRB1*0402 that are critical for binding of self-antigens in DR4 PV patients. Instead, these positions were replaced by two small neutral hydrophobic residues (Gly β70 and Ala β71), suggesting that DRB1*0402 and DQB1*0503 may recognize different sets of PV epitopes under the influence of a different balance of intermolecular forces. Positions β70 and β74 show charge reversal, in the non-PV associated DRB1*1401, *1404 and *1405 alleles while the negative charge at β71 alone is conserved, compared to DRB1*0402, making pocket 4 the single dominant factor discriminating between PV non-association and susceptibility.

5.3.1.3 PV protective and susceptible alleles

Differences in the amino acid sequences are observed at residue β86 of pocket 1, residue β70, β71 of pocket 4, residues β28, β30, β47, β71 of pocket 7 and residues β37, β57 of pocket 9. Both protective alleles (DQB1*0201 and DQB1*0202) do not contain negatively charged residues at position β70 (pocket 4) and β71 (pocket 7). Instead, these positions were replaced by two large and positively charged amino acids (Arg β70 and Lys β71). The functional specificities of PV protective and susceptible alleles are also affected by clear structural differences in the Cα positions of both α and β chains (Cα RMSD > 0.57 Å) indicating that any differences in peptide discrimination between the alleles is due to a combination of both the backbone conformation as well as the size and nature of the side chains of the pocket residues.



```
                1........10........20........30........40........50........60........70
DRB1*1401       GDTRPRFLEYSTSECHFFNGTERVRFLDRYFHHQEEFVRFDSDQGEYRAVTELGRPAAEHWNSQKDILER
DRB1*1404       GDTRPRFLEYSTGECYFFNGTERVRFLDRYFYHQEEFVRFDSDQGEYRAVTELGRPAAEHWNSQKDILER
DRB1*1405       GDTRPRFLEYSTSECHFFNGTERVRFLDRYFYHQEEFVRFDSDQGEYRAVTELGRPDAEYWNSQKDILER
DRB1*0406       GDTRPRFLEQVKHECHFFNGTERVRFLDRYFYHQEESVRFDSDVGEYRAVTELGRPDAEYWNSQKDLLEQ
DRB1*0404       GDTRPRFLEQVKHECHFFNGTERVRFLDRYFYHQEEYVRFDSDVGEYRAVTELGRPDAEYWNSQKDLLEQ
DRB1*0401       GDTRPRFLEQVKHECHFFNGTERVRFLDRYFYHQEEYVRFDSDVGEYRAVTELGRPDAEYWNSQKDLLEQ
DRB1*0402       GDTRPRFLEQVKHECHFFNGTERVRFLDRYFYHQEEYVRFDSDVGEYRAVTELGRPDAEYWNSQKDILED
DQB1*0201       --SPEDFVYQFKGMCYFTNGTERVRLVSRSIYNREEIVRFDSDVGEFRAVTLLGLPAAEYWNSQKDILER
DQB1*0202       --SPEDFVYQFKGMCYFTNGTERVRLVSRSIYNREEIVRFDSDVGEFRAVTLLGLPAAEYWNSQKDILER
DQB1*0503       --SPEDFVYQFKGLCYFTNGTERVRGVTRHIYNREEYVRFDSDVGVYRAVTPQGRPDAEYWNSQKEVLEG
                  :   *:   .  *:* ******* : * ::::** ****** * :****  * * **:*****::**

                71.......80........90........100.......110.......120.......130.......141
DRB1*1401       ERAEVDTYCRHNYGVVESFQVQRRVHREVTVYPAK-------NLLVCSVNGPYPGSIEVRWFRNGQEEKT
DRB1*1404       ERAEVDTYCRHNYGVVESFQVQRRVHREVTVYPAK-------NLLVCSVNGPYPGSIEVRWFRNGQEEKT
DRB1*1405       ERAEVDTYCRHNYGVVESFQVQRRVVREVTVYPAK-------NLLVCSVNGPYPGSIEVRWFRNGQEEKT
DRB1*0406       RRAEVDTYCRHNYGVVESFTVQRRVYPEVTVYPAKTQPLQHHNLLVCSVNGFYPGSIEVRWFRNGQEEKT
DRB1*0404       RRAAVDTYCRHNYGVVESFTVQRRVYPEVTVYPAKTQPLQHHNLLVCSVNGFYPGSIEVRWFRNGQEEKT
DRB1*0401       KRAAVDTYCRHNYGVGESFTVQRRVYPEVTVYPAKTQPLQHHNLLVCSVNGFYPGSIEVRWFRNGQEEKT
DRB1*0402       ERAAVDTYCRHNYGVVESFTVQRRVYPEVTVYPAKTQPLQHHNLLVCSVNGFYPGSIEVRWFRNGQEEKT
DQB1*0201       KRAAVDRVCRHNYQLELRTTLQRRVEPTVTISPSRTEALNHHNLLVCSVTDFYPAQIKVRWFRNDQEETA
DQB1*0202       KRAAVDRVCRHNYQLELRTTLQRRVEPTVTISPSRTEALNHHNLLVCSVTDFYPAQIKVRWFRNGQEETA
DQB1*0503       ARASVDRVCRHNYEVAYRGILQRRVEPTVTISPSRTEALNHHNLLICSVTDFYPSQIKVRWFRNDQEETA
                 ** **  ***** :     :****   **: *::       ***:***.. **..*:******.***.:

                141......150.......160.......170.......180.......190
DRB1*1401       GVVSTGLIHNGDWTFQTLVMLETVPRSSEVYTCQVEHPSLTSPLTVEWRA
DRB1*1404       GVVSTGLIHNGDWTFQTLVMLETVPRSSEVYTCQVEHPSLTSPLTVEWRA
DRB1*1405       GVVSTGLIQNGDWTFQTLVMLETVPRSSEVYTCQVEHPSLTSPLTVEWRA
DRB1*0406       GVVSTGLIQNGDWTFQTLVMLETVPRSGEVYTCQVEHPSLTSPLTVEWRA
DRB1*0404       GVVSTGLIQNGDWTFQTLVMLETVPRSGEVYTCQVEHPSLTSPLTVEWRA
DRB1*0401       GVVSTGLIQNGDWTFQTLVMLETVPRSGEVYTCQVEHPSLTSPLTVEWRA
DRB1*0402       GVVSTGLIQNGDWTFQTLVMLETVPRSGEVYTCQVEHPSLTSPLTVEWRA
DQB1*0201       GVVSTPLIRNGDWTFQILVMLEMTPQRGDVYTCHVEHPSLQSPITVEWRA
DQB1*0202       GVVSTPLIRNGDWTFQILVMLEMTPQRGDVYTCHVEHPSLQSPITVEWRA
DQB1*0503       GVVSTPLIRNGDWTFQILVMLEMTPQRGDVYTCHVEHPSLQSPITVEWRA
                ***** **:******* ***** .*: .:****:****** **:******
```

Figure 21 Multiple sequence alignment of the β chains of DR and DQ alleles. Pocket residues are shaded in black.

5.3.2 Epitope comparisons

5.3.2.1 DR4 PV

Eight previously identified stimulatory Dsg3 epitopes (Dsg3 191-205, Dsg3 206-220, Dsg3 252-266, Dsg3 342-356, Dsg3 380-394, Dsg3 763-777, Dsg3 810-824 and Dsg3 963-977) for DRB1*0402 were docked into the binding groove of all

DR4 (DRB1*0401, *0402, *0404, *0406) alleles investigated in this study. Analysis of these Dsg3 peptide-bound alleles revealed that only one peptide conformation can fit perfectly into the binding cleft of DRB1*0402, and atomic clashes of these Dsg3 peptides are obtained for all other DR4 subtypes investigated in this study. Two epitopes (Dsg3 342-356 and Dsg3 810-824) have small residues (Ser/Cys) in pocket 1, suggesting that small residues at anchor positions may also result in high affinity binding with DR4 PV molecules, an observation previously documented for the influenza-associated I-A^d allele of mice (Scott *et al*., 1998). This finding supports the association of DRB1*0402 with PV whereas other DR4 subtypes are non-associated, with the exception of DRB1*0406 that is reported to be associated in the Japanese population (Yamashina *et al*., 1998). As such, there is a possibility of the existence of other peptides relevant in the Japanese populations that bind to *0406 but are yet to be determined.

5.3.2.2 DR6 PV

Dsg3 96-112, a recently identified epitope in DR6 PV patients (Veldman *et al*., 2003), fits perfectly into the binding groove of DQB1*0503 with two identified core sequences at residues 101-109 and residues 102-110. The identified 101-109 core has four intermolecular hydrogen bonds compared to seven intermolecular hydrogen bonds in the core of 102-110. Perfect fitting of Dsg3 206-220, Dsg3 252-266, Dsg3 342-356, Dsg3 810-824 and Dsg3 963-977 into the binding groove of DQB1*0503 is also obtained. Atomic clashes are obtained for Dsg3

191-205, Dsg3 380-394 and Dsg3 763-777 as well as all DR6 alleles investigated in this study. The proportion of DRB1*1401, *1404 and *1405 has been reported to be increased in PV probably due to linkage disequilibrium. The lack of binding of all stimulatory peptides investigated in this study to these alleles indicates that the HLA association in DR6 PV patients is more likely at the DQB1 locus (DQB1*0503 allele) and not the linked DRB1 loci (DRB1*1401, *1404 and *1405). Our data supports the notion that the reported associations of this disease with DRB1*1401, *1404, *1405 are due to linkage disequilibrium with the true disease associated allele (DQB1*0503).

5.3.2.3 PV susceptible alleles

Our docking simulations reveal strong evidence that DRB1*0402 and DQB1*0503 can bind to different sets of PV epitopes by recognizing different core peptide sequences in the binding groove (Table 8). Three PV epitopes (Dsg3 191-205, Dsg3 380-394 and Dsg3 763-77) can only bind to DRB1*0402, four PV epitopes (Dsg3 206-220, Dsg3 252-266, Dsg3 342-356 and Dsg3 810-824) can bind to both alleles with different core peptide sequences (Figure 22), one PV epitope (Dsg3 963-977) can bind to both alleles with the same core peptide sequence, and one PV epitope (Dsg3 96-112) can only bind to DQB1*0503. DRB1*0402 and DQB1*0503 may recognize the same Dsg3 epitope at two unique sets of core sequences (which may be in close proximity) within the epitope itself. These findings are completely in accord with experimental data (Veldman *et al*., 2003).

5.3.2.4 PV protective alleles

Our simulation results indicate that DQB1*0201 and DQB1*0202 can bind to multiple core sequences for the majority of PV epitopes investigated in this study. DQB1*0201 can bind one epitope (Dsg3 963-977) at two core regions, one epitope (Dsg3 206-220) at three core regions, three epitopes (Dsg3 191-205, 252-266 and 342-356) at four core regions, and two epitopes (Dsg3 96-112 and 810-824) at five core regions. DQB1*0202 can bind two epitopes (Dsg3 96-112 and 963-977) at three core regions, two epitopes (Dsg3 342-356 and 810-824) at four core regions and one epitope (Dsg3 252-266) at five core regions. In contrast, the majority of PV epitopes (with the exception of Dsg3 96-112 and 252-266) can bind to PV susceptible alleles DRB1*0402 and DQB1*0503 at a single core. This finding lends support to the hypothesis that the protective alleles DQB1*0201, *0202 may be capable of binding to most peptides with greater affinity than PV susceptible alleles, allowing for efficient deletion of autoreactive T cells (Gebe *et al*., 2002).

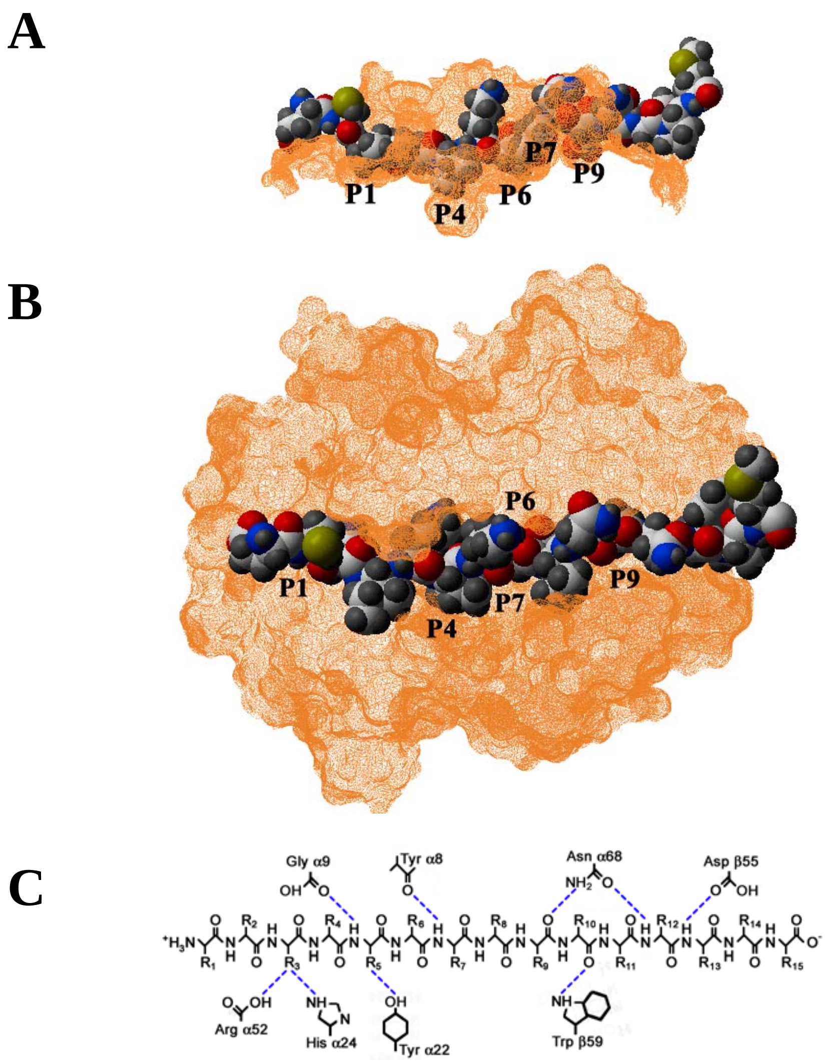


Figure 22 Dsg3 252-266 peptide (ECNIKVKDVNDNFPM) docked into the binding grooves of DQB1*0503 without any atomic clashes. The relevant binding register is underlined. (A) Side view of the molecular surface of the DQB1*0503 peptide-binding site is shown in orange with the binding pockets labeled and the Dsg3 252-266 peptide displayed as a CPK model. (B) As *b*, except from a top view. (C) Hydrogen bonds between Dsg3 252-266 and DQB1*0503 indicated by dashed lines.

Table 8 Preferred core residues for PV associated alleles. Best fitting binding registers in the binding groove are underlined.

No.	Residues	Allele	Core peptide sequences
I	96-112	DRB1*0402	-
		DQB1*0503	PFGIFVVDKNTGDINIT
			PFGIFVVDKNTGDINIT
II	191-205	DRB1*0402	NSKIAFKIVSQEPAG
		DQB1*0503	–
III	206-220	DRB1*0402	TPMFLLSRNTGEVRT
		DQB1*0503	TPMFLLSRNTGEVRT
IV	252-266	DRB1*0402	ECNIKVKDVNDNFPM
		DQB1*0503	ECNIKVKDVNDNFPM
			ECNIKVKDVNDNFPM
V	342-356	DRB1*0402	SVKLSIAVKNKAEFH
		DQB1*0503	SVKLSIAVKNKAEFH
VI	380-394	DRB1*0402	GIAFRPASKTFTVQK
		DQB1*0503	–
VII	763-777	DRB1*0402	SGTMRTRHSTGGTNK
		DQB1*0503	–
VIII	810-824	DRB1*0402	NDCLLIYDNEGADAT
		DQB1*0503	NDCLLIYDNEGADAT
IX	963-977	DRB1*0402	ERVICPISSVPGNLA
		DQB1*0503	ERVICPISSVPGNLA

5.3.3 Role of flanking residues in peptide selection

Our data demonstrates that the conformations of flanking peptide residues that extend beyond the binding groove are critical to peptide selection in MHC class II alleles. The core sequences of Dsg3 963-977 fit perfectly within the binding grooves of non-associated alleles DRB1*0401, *0404, and *1404 but poor contacts to the respective alleles at Phe α50 are obtained when the conformation of the N-terminal flanking residue Ile4 is taken into account. These results suggest that binding is determined by both the core and flanking segments while considering the overall interactions between each peptide and the respective alleles.

Table 9 Comparison of core peptides (numbering according to Table 8) from structural docking in the different binding pockets with the sequence-based binding motifs. '+' indicates compliance of amino acid residues within the core (bold underlined) with the respective binding motifs defined by the groups of [a]Veldman *et al*. (2003) and [b]Sinha (1990, personal communications)

No.	Residues	Peptide Sequence and positions in the bound conformation for DRB1*0402	Core peptide residue positions as defined by binding motifs							
			p1		p4		p6		p9	
		1 2 3 4 5 6 7 8 9	V[a]	S[b]	V	S	V	S	V	S
II	191-205	NSKIA **F K I V S Q E P A** G	+			+		+		+
III	206-220	TPM **F L L S R N T G E** VRT	+				+	+		+
IV	252-266	ECNI **K V K D V N D N F** PM					+	+		
V	342-356	SVKL **S I A V K N K A E** FH				+	+	+		+
VI	380-394	GIA **F R P A S K T F T** VQK	+			+		+		+
VII	763-777	SGT **M R T R H S T G G** TNK	+	+	+	+	+	+		+
VIII	810-824	ND **C L L I Y D N E G** ADAT				+		+		+
IX	963-977	ERVICP **I S S V P G N L A**	+	+		+				+

5.3.4 Sequence Motifs

Sequence-based epitope prediction relies on the identification of sequence motifs from available experimental data. The correlation of core peptide residues with binding motifs previously defined by Veldman *et al*. (2003) and Sinha *et al*. (unpublished results) is shown in Table 9, to understand to what extent sequence-based approaches will be valid with specific reference to PV. The sequence conservation observed here is too low to warrant the generation of a consensus sequence pattern.

Peptide VII (Dsg3 763-777) agrees well with the motifs from Veldman *et al*. (2003) and Sinha *et al*. (personal communications), while all other peptides show low to moderate compliance. Of the four positions compared, peptide IV (Dsg3 252-266) shows agreement only at position p6. For Dsg3 342-356 peptide, the core nonamer identified by our models is 346-354, which is register-shifted by one residue from the core of 347-355 reported by Veldman *et al*. (2003), and 345-353 identified by Sinha *et al*. (personal communications), for the binding groove of *0402. This shift is critical as residues p1 and p4 identified by us do not fit well into both binding motifs. Our modeling studies suggest that peptide position p4 need not be positively charged as indicated by Veldman *et al*. (2003), supporting the existence of a more degenerate motif by Sinha *et al*. (personal communications) at this position. In addition, p1 also appears to be more degenerate than previously suggested (Veldman *et al*., 2003), showing a preference for hydrophobic and large residues but can accommodate residues of other sizes as well. Hence for generating sequence patterns to design peptides for vaccine design, structural information is important (Schirle *et al*., 2001) and the exact peptide in the binding groove identified by our docking protocol will be most useful here.

5.3.5 Disease progression in PV

T cell response to a number of epitopes among PV patients has been reported in several studies (Sinha *et al*., 1988, 1990; Boeckmann *et al*., 2003; Wucherpfennig *et al*., 1995; Hertl *et al*., 1998). There may be disease heterogeneity, meaning that clinically similar but distinct phenotypes could operate by alternate pathways, each with a different initial immunodominant epitope(s). The differential T cell reactivities among individual patients to individual peptides may also be a function of the disease stage or severity and correlate with mechanisms of disease progression. While there may be a limited set of epitopes present in patients in the early stages of the disease, epitope spreading can occur during disease progression, resulting in reactivity to previously innocuous epitopes. In addition, reactivities to multiple epitopes within individual patients were detected in two cases (Dsg3 191-205 and 342-356 for PV107; Dsg3 191-205, 810-824 and 963-977 for PV117). Autoantibodies against desmoglein 1 have also been reported in severe disease (Harman *et al*., 2000). One other incidence of multiple T cell reactivities within a PV patient has been previously reported (Wucherpfennig *et al*., 1995). These findings, together with our simulation results, lend further credence to the hypothesis that no single epitope is responsible for both disease initiation and propagation and are consistent with the expected and observed ability to generate multiple peptide/MHC complexes from a single target autoantigen.

5.4 Summary

- In this work, docking simulations at the binding site of PV associated and non-associated DR and DQ alleles have been performed to analyze the structural aspects of binding and allele-specificity for nine previously identified Dsg3 epitopes.

- The author has demonstrated the existence of best-fit core residues at different positions of each peptide (excepting Dsg3 96-112) into the binding groove of DRB1*0402 with no observed atomic clash penalties or bad contacts. In contrast, atomic clashes are experienced in all other PV non-associated DR4 alleles. This discrimination supports existing hypothesis with regards to the crucial role that DRB1*0402 plays in selecting specific self-peptides in DR4 PV.
- This study indicates that DRB1*0402 and DQB1*0503 do not necessarily share the same core residues. It is possible that DRB1*0402, DQB1*0503 and all other PV non-associated alleles may have different sets of binding specificities.
- This study also indicates that perfect fitting of the core nonameric peptide residues within the binding groove of MHC class II alleles may not guarantee perfect fitting of the entire peptide, and flanking residues outside the binding groove may play a critical part in peptide selection.
- Comparison of binding registers with existing binding motifs indicates that sequence-based methods are currently insufficient for the design of PV epitopes as there are both register shifts in the suggested motifs as well as polymorphism observed in the core residues in the binding groove.
- The present analysis supports the hypothesis (Gebe *et al*., 2002) that the alleles DQB1*0201 and *0202 play a protective role by binding Dsg3 peptides with greater affinity than the susceptible alleles, facilitating efficient deletion of autoreactive T cells.

- The current analysis supports existing evidence that no single epitope may be responsible for both disease initiation and propagation in PV, and it is valuable to identify all Dsg3 peptides that bind to the PV susceptible alleles.

Chapter 6: Functional prediction of MHC class II binding peptides

6.1 Introduction

MHC class II molecules play a critical role in immune responses. They bind short antigenic peptide fragments and present them on the surface of antigen-presenting cells for recognition by the CD4$^+$ helper T cells. T cell recognition of the peptide/MHC complex initiates a cascade of immunological events necessary for initiation and regulation of immune responses. These events are necessary for normal immune responses but may also be involved in the pathogenesis of autoimmune disorders (Klein *et al.*, 2000; Flynn *et al.*, 2004) and hypersensitivity reactions (Neeno *et al*., 1996; Krco *et al*., 2000).

The HLA-DQ allele, DQ3.2*β* (DQA1*0301/DQB1*0302), commonly known as DQ8, is present in approximately 20% of the human population (Middleton *et al.*, 2003). DQ3.2*β* is of particular interest in the study of allergenicity and autoimmunity because of its association to house dust mite allergy (Neeno *et al*., 1996; Krco *et al*., 2000) and several human autoimmune disorders, including celiac disease (CD) (Sollid and Thorsby, 1993), insulin-dependent diabetes mellitus (IDDM) (Nepom and Kwok, 1998; Erlich *et al.*, 1993), IDDM-associated periodontal disease (Faustman *et al.*, 1991), and autoimmune polyendocrine syndrome type II (APS-II) (Robles *et al.*, 2002). Some 70% of IDDM patients (Kwok *et al.*, 1989) have DQ3.2*β*. Improved understanding of peptide binding to this molecule is important for elucidating the role of DQ3.2*β* in both autoimmunity and allergies. Peptide-binding studies are invaluable for designing vaccines and immunotherapies for controlling allergic or autoimmune responses.

Computational methods for the identification of peptides that bind to HLA-DR molecules are relatively advanced (Brusic *et al.*, 2004), while methods for prediction of peptide binding to HLA-DQ molecules have encountered limited success due to the paucity of peptides as training data for sequence-based techniques. Computational strategies for DQ3.2*β* binding peptides using sequence motifs (Godkin *et al.*, 1997, 1998; Rammensee *et al.*, 1999; Moustakas *et al.*, 2000) have been used with varying degrees of success (Harfouch-Hammoud *et al.*, 1999) but an effective model for large-scale screening is still currently lacking. Up to now, few prediction techniques for HLA-DQ molecules have been developed using three-dimensional models as the dual issues of docking and scoring must be addressed (Ranganathan *et al.,* 2005).

In the previous chapter (Chapter 5), a new technique for rapid and accurate docking of binding registers to class I and class II alleles has been developed. This approach has been successfully applied to discriminate between alleles implicated in the autoimmune disorder, pemphigus vulgaris from non-disease implicated and protective alleles (Chapter 7). However, despite the accuracy of our docking experiments, these results are qualitative rather than quantitative, as energy-based scoring was not considered. The earlier model, therefore, cannot be used for effective discrimination of peptide binding affinities (strong, moderate and weak binders from non-binders). This chapter reports the development of a scoring function to complement the docking protocol to effectively identify MHC class II epitopes, with the correct binding register. An analysis on the binding patterns of DQ3.2*β* peptides was performed to understand MHC class II binding characteristics. The results of these analyzes would be important for understanding the principles of self/non-self discrimination and strategies for the design of epitope-based vaccines.

6.2 Materials and Methods

6.2.1 Data

6.2.1.1 Crystallographic data

The coordinates of DQ3.2*β* was extracted from the crystal structure of DQ3.2*β*–insulin B9-23 complex, with PDB code 1JK8 (Lee *et al.*, 2001). The structure was relaxed by conjugate gradient minimization, using the Internal Coordinate Mechanics (ICM) 3.0 package (Abagyan *et al.*, 1994a).

6.2.1.2 Experimental binding data

Two sets of data are used in this study: (i) peptides with experimental IC_{50} values from biochemical studies and (ii) peptides with experimental T cell proliferation values from functional studies.

Dataset I comprises 127 peptides (Table 10) with experimentally determined IC_{50} values (70 high-affinity, 13 medium-affinity and 23 low-affinity binders and 21 non-binders) derived from biochemical studies (Godkin *et al.*, 1998; Sidney *et al.*, 2002; Suri *et al.*, 2005). Largely for discussion

Table 10 DQ3.2*β* specific peptides with experimentally determined IC_{50} values used in this study. For peptides with experimentally determined binding registers (#1-#87), the nonamer in the binding groove is underlined in bold font.

No.	Category	Description	Peptide	IC_{50} (nM)	Reference
1	Training Set	Thyroid per 632-645Y	IDV**WLGGLAENF**LPY	39	Sidney *et al.* 2002
2	Training Set	Thyroid per 632-645Y analog	IDV**DLGGLAENF**LPY	20	Sidney *et al.* 2002
3	Training Set	Thyroid per 632-645Y analog	IDV**YLGGLAENF**LPY	52	Sidney *et al.* 2002
4	Training Set	Thyroid per 632-645Y analog	IDV**SLGGLAENF**LPY	72	Sidney *et al.* 2002
5	Training Set	Thyroid per 632-645Y analog	IDV**LLGGLAENF**LPY	119	Sidney *et al.* 2002
6	Training Set	Thyroid per 632-645Y analog	IDV**KLGGLAENF**LPY	2028	Sidney *et al.* 2002
7	Training Set	Thyroid per 632-645Y analog	IDV**WSGGLAENF**LPY	36	Sidney *et al.* 2002
8	Training Set	Thyroid per 632-645Y analog	IDV**WVGGLAENF**LPY	44	Sidney *et al.* 2002
9	Training Set	Thyroid per 632-645Y analog	IDV**WYGGLAENF**LPY	63	Sidney *et al.* 2002
10	Training Set	Thyroid per 632-645Y analog	IDV**WDGGLAENF**LPY	83	Sidney *et al.* 2002
11	Training Set	Thyroid per 632-645Y analog	IDV**WKGGLAENF**LPY	97	Sidney *et al.* 2002
12	Training Set	Thyroid per 632-645Y analog	IDV**WLDGLAENF**LPY	100	Sidney *et al.* 2002
13	Training Set	Thyroid per 632-645Y analog	IDV**WLSGLAENF**LPY	105	Sidney *et al.* 2002
14	Training Set	Thyroid per 632-645Y analog	IDV**WLYGLAENF**LPY	126	Sidney *et al.* 2002
15	Training Set	Thyroid per 632-645Y analog	IDV**WLLGLAENF**LPY	130	Sidney *et al.* 2002
16	Training Set	Thyroid per 632-645Y analog	IDV**WLKGLAENF**LPY	325	Sidney *et al.* 2002
17	Training Set	Thyroid per 632-645Y analog	IDV**WLGLLAENF**LPY	51	Sidney *et al.* 2002
18	Training Set	Thyroid per 632-645Y analog	IDV**WLGSLAENF**LPY	78	Sidney *et al.* 2002
19	Training Set	Thyroid per 632-645Y analog	IDV**WLGDLAENF**LPY	105	Sidney *et al.* 2002
20	Training Set	Thyroid per 632-645Y analog	IDV**WLGYLAENF**LPY	325	Sidney *et al.* 2002
21	Training Set	Thyroid per 632-645Y analog	IDV**WLGGVAENF**LPY	93	Sidney *et al.* 2002
22	Training Set	Thyroid per 632-645Y analog	IDV**WLGGYAENF**LPY	139	Sidney *et al.* 2002
23	Training Set	Thyroid per 632-645Y analog	IDV**WLGGKAENF**LPY	177	Sidney *et al.* 2002
24	Training Set	Thyroid per 632-645Y analog	IDV**WLGGSAENF**LPY	217	Sidney *et al.* 2002
25	Training Set	Thyroid per 632-645Y analog	IDV**WLGGLDENF**LPY	177	Sidney *et al.* 2002
26	Training Set	Thyroid per 632-645Y analog	IDV**WLGGLYENF**LPY	195	Sidney *et al.* 2002
27	Training Set	Thyroid per 632-645Y analog	IDV**WLGGLSENF**LPY	390	Sidney *et al.* 2002
28	Training Set	Thyroid per 632-645Y analog	IDV**WLGGLAYNF**LPY	100	Sidney *et al.* 2002

29	Training Set	Thyroid per 632-645Y analog	IDV**WLGGLALNF**LPY	130	Sidney *et al.* 2002
30	Training Set	Thyroid per 632-645Y analog	IDV**WLGGLASNF**LPY	355	Sidney *et al.* 2002
31	Training Set	Thyroid per 632-645Y analog	IDV**WLGGLAESF**LPY	18	Sidney *et al.* 2002
32	Training Set	Thyroid per 632-645Y analog	IDV**WLGGLAELF**LPY	23	Sidney *et al.* 2002
33	Training Set	Thyroid per 632-645Y analog	IDV**WLGGLAEYF**LPY	25	Sidney *et al.* 2002
34	Training Set	Thyroid per 632-645Y analog	IDV**WLGGLAEKF**LPY	31	Sidney *et al.* 2002
35	Training Set	Thyroid per 632-645Y analog	IDV**WLGGLAEDF**LPY	34	Sidney *et al.* 2002
36	Training Set	Thyroid per 632-645Y analog	IDV**WLGGLAEQF**LPY	35	Sidney *et al.* 2002
37	Training Set	Thyroid per 632-645Y analog	IDV**WLGGLAEND**LPY	17	Sidney *et al.* 2002
38	Training Set	Thyroid per 632-645Y analog	IDV**WLGGLAENV**LPY	23	Sidney *et al.* 2002
39	Training Set	Thyroid per 632-645Y analog	IDV**WLGGLAENY**LPY	30	Sidney *et al.* 2002
40	Training Set	Thyroid per 632-645Y analog	IDV**WLGGLAENS**LPY	35	Sidney *et al.* 2002
41	Training Set	Thyroid per 632-645Y analog	IDV**WLGGLAENK**LPY	1677	Sidney *et al.* 2002
42	Training Set	Thyroid per 632-645Y analog	IDV**WLGGLAENF**DPY	25	Sidney *et al.* 2002
43	Training Set	Thyroid per 632-645Y analog	IDV**WLGGLAENF**SPY	54	Sidney *et al.* 2002
44	Training Set	Thyroid per 632-645Y analog	IDV**WLGGLAENF**YPY	58	Sidney *et al.* 2002
45	Training Set	Thyroid per 632-645Y analog	IDV**WLGGLAENF**KPY	75	Sidney *et al.* 2002
46	Training Set	Thyroid per 632-645Y analog	IDV**WLGGLAENF**VPY	77	Sidney *et al.* 2002
47	Training Set	Thyroid per 632-645Y analog	IDV**WLGGLAENF**LDY	42	Sidney *et al.* 2002
48	Training Set	Thyroid per 632-645Y analog	IDV**WLGGLAENF**LYY	108	Sidney *et al.* 2002
49	Training Set	Thyroid per 632-645Y analog	IDV**WLGGLAENF**LLY	139	Sidney *et al.* 2002
50	Training Set	Thyroid per 632-645Y analog	IDV**WLGGLAENF**LSY	195	Sidney *et al.* 2002
51	Training Set	Thyroid per 632-645Y analog	IDV**WLGGLAENF**LKY	279	Sidney *et al.* 2002
52	Training Set	Thyroid per 632-645Y analog	IDV**WLGGLAENF**LPD	26	Sidney *et al.* 2002
53	Training Set	Thyroid per 632-645Y analog	IDV**WLGGLAENF**LPL	32	Sidney *et al.* 2002
54	Training Set	Thyroid per 632-645Y analog	IDV**WLGGLAENF**LPF	50	Sidney *et al.* 2002
55	Training Set	Thyroid per 632-645Y analog	IDV**WLGGLAENF**LPK	66	Sidney *et al.* 2002
56	Training Set	Thyroid per 632-645Y analog	IDV**WLGGLAENF**LPS	70	Sidney *et al.* 2002
57	Test Set 1	E25B protein 112-126	YQTI**EENIKIFEE**DA	800	Suri *et al*. 2005
58	Test Set 1	E25B protein 112-126 analog	YQTI**EENIKIFKE**DA	1000	Suri *et al*. 2005
59	Test Set 1	E25B protein 112-126 analog	YQTI**EENIKIFEE**KA	1700	Suri *et al*. 2005
60	Test Set 1	E25B protein 112-126 analog	YQTI**EENIKIFEA**DA	1800	Suri *et al*. 2005
61	Test Set 1	E25B protein 112-126 analog	YQTI**EENIKIFEA**AA	2500	Suri *et al*. 2005

62	Test Set 1	E25B protein 112-126 analog	YQTI**EENIKIFAA**AA	1700	Suri *et al*. 2005
63	Test Set 1	E25B protein 112-126 analog	YQTI**KENIKIFEE**DA	3800	Suri *et al*. 2005
64	Test Set 1	TRAIL receptor 2 364-380	GRFT**YQNAAAQPE**TGPG	1700	Suri *et al*. 2005
65	Test Set 1	TRAIL receptor 2 364-380 analog	GRFT**YQNAAAQPA**TGPG	1000	Suri *et al*. 2005
66	Test Set 1	TRAIL receptor 2 364-380 analog	GRFT**AQNAAAQPE**TGPG	1700	Suri *et al*. 2005
67	Test Set 1	TRAIL receptor 2 364-380 analog	GRFT**KQNAAAQPE**TGPG	3700	Suri *et al*. 2005
68	Test Set 1	TRAIL receptor 2 364-380 analog	GRFT**AQNAAAQPA**TGPG	3100	Suri *et al*. 2005
69	Test Set 1	Nicastrin 65-78	ISG**DTGVIHVVE**KE	1000	Suri *et al*. 2005
70	Test Set 1	Nicastrin 65-78 analog	ISG**KTGVIHVVE**KE	N.B.	Suri *et al*. 2005
71	Test Set 1	Nicastrin 65-78 analog	ISG**KTGVIHVVK**KE	N.B.	Suri *et al*. 2005
72	Test Set 1	Nicastrin 65-78 analog	ISG**DTGVIHVVA**KE	4300	Suri *et al*. 2005
73	Test Set 1	Nicastrin 65-78 analog	ISG**ATGVIHVVE**KE	2300	Suri *et al*. 2005
74	Test Set 1	Superoxide dimutase 1 90-103	AGK**DGVANVSIE**DR	2000	Suri *et al*. 2005
75	Test Set 1	Superoxide dimutase 1 90-103 analog	AGK**AGVANVSIE**DR	1800	Suri *et al*. 2005
76	Test Set 1	Superoxide dimutase 1 90-103 analog	AGK**DGVANASIE**DR	2800	Suri *et al*. 2005
77	Test Set 1	Superoxide dimutase 1 90-103 analog	AGK**DGVANVSIK**DR	N.B.	Suri *et al*. 2005
78	Test Set 1	Superoxide dimutase 1 90-103 analog	AGK**KGVANVSIK**DR	N.B.	Suri *et al*. 2005
79	Test Set 1	Superoxide dimutase 1 90-103 analog	AGK**DGVANKSIE**DR	N.B.	Suri *et al*. 2005
80	Test Set 1	MHC II Eα 51-65	FDG**DEIFHVDIE**KSE	1000	Suri *et al*. 2005
81	Test Set 1	MHC II Eα 51-65 analog	FDG**DEIFHVDIK**KSE	N.B.	Suri *et al*. 2005
82	Test Set 1	MHC II Eα 51-65 analog	FDG**KEIFHVDIK**KSE	N.B.	Suri *et al*. 2005
83	Test Set 1	MHC II Eα 51-65 analog	FDG**DEIFHKDIE**KSE	N.B.	Suri *et al*. 2005
84	Test Set 1	MHC II Eα 51-65 analog	FDG**KEIFHVDIE**KSE	2800	Suri *et al*. 2005
85	Test Set 1	MHC II Eα 51-65 analog	FDG**AEIFHVDIE**KSE	2000	Suri *et al*. 2005
86	Test Set 1	MHC II Eα 51-65 analog	FDG**DEIAHVDIE**KSE	3300	Suri *et al*. 2005
87	Test Set 1	MHC II Eα 51-65 analog	FDG**DEIFHADIE**KSE	3100	Suri *et al*. 2005
88	Test Set 1	A-gliadin 49-63	FPSQQPYLQLQPFPQ	20	Godkin *et al*. 1998
89	Test Set 1	A-gliadin 207-221	YPLGQGSFRPSQQNP	100	Godkin *et al*. 1998
90	Test Set 1	A-gliadin 77-91	SFPPQQPYPQPQPQY	370	Godkin *et al*. 1998
91	Test Set 1	A-gliadin 30-44	FPGQQQQFPPQQPYP	600	Godkin *et al*. 1998
92	Test Set 1	A-gliadin 196-210	PSSQFQQPLQQYPLG	10000	Godkin *et al*. 1998
93	Test Set 1	A-gliadin 41-55	QPYPQPQPFPSQQPY	1120	Godkin *et al*. 1998
94	Test Set 1	A-gliadin 56-70	LQLQPFPQPQPFPPL	20	Godkin *et al*. 1998
95	Test Set 1	A-gliadin 227-241	VQPQQQLPQFEIRNL	73	Godkin *et al*. 1998

96	Test Set 1	A-gliadin 34-48	QQQFPPQQPYPQPQP	10000	Godkin *et al*. 1998
97	Test Set 1	A-gliadin 201-215	QQPLQQYPLGQGSFR	2180	Godkin *et al*. 1998
98	Test Set 1	HSV	DMTPADALDDFDL	173	Sidney *et al.* 2002
99	Test Set 1	CD20 249–262 analog	EEDIEIIPIQEEEY	21	Sidney *et al.* 2002
100	Test Set 1	34P3A	IARAKMFPAVAEK	541	Sidney *et al.* 2002
101	Test Set 1	HA 255–271Y	FESTGNLIAPEYGFKISY	62	Sidney *et al.* 2002
102	Test Set 1	GAD 101–115	CDGERPTLAFLQDVM	69	Sidney *et al.* 2002
103	Test Set 1	FceR 104–122	SQDLELSWNLNGLQADLSS	123	Sidney *et al.* 2002
104	Test Set 1	Pf ABRA 487–506	DSNIMNSINNVMDEIDFFEK	171	Sidney *et al.* 2002
105	Test Set 1	p21 51–66; C out	LLDILDTAGLEEYSAMRD	202	Sidney *et al.* 2002
106	Test Set 1	Lamba repressor 12–24	LEDARRLKAIYEK	717	Sidney *et al.* 2002
107	Test Set 1	GAD65 253–265	IARFKMFPEVKEK	3712	Sidney *et al.* 2002
108	Test Set 1	Artificial sequence	AAAAAVAAEAY	48	Sidney *et al.* 2002
109	Test Set 1	OVA 267-276 Y	LTEWTSSNVMEERY	62	Sidney *et al.* 2002
110	Test Set 1	IA-2 499-509	GVAGLLVALAV	95	Sidney *et al.* 2002
111	Test Set 1	MHC Ia 46-63	EPRAPWIEQEGPEYW	519	Sidney *et al.* 2002
112	Test Set 1	VP16	PPLYATGRLSQAQLMPSPPM	538	Sidney *et al.* 2002
113	Test Set 1	IA-2 499–509	MSSGSFINISV	2470	Sidney *et al.* 2002
114	Test Set 1	Artificial sequence (ROIV)	YAHAAHAAHAAHAAHAA	2924	Sidney *et al.* 2002
115	Test Set 1	Lol p1 101–120	APYHFDLSGHAFGSMAKKGE	3602	Sidney *et al.* 2002
116	Test Set 1	CLIP 95-102	KPVSKMRMATPLLMQALP	650	Sidney *et al.* 2002
117	Test Set 1	FceR 104–122 analog	SQDLELSWNLNGLQAY	118	Sidney *et al.* 2002
118	Test Set 1	MHC Ia 51–63 analog	YPFIEQEGPEFFDQE	1156	Sidney *et al.* 2002
119	Test Set 1	B2m 91–104	TPTEKDEYCARVNH	> 10000	Sidney *et al.* 2002
120	Test Set 1	ML LSR2 5–17	GVTYEIDLTNKN	> 10000	Sidney *et al.* 2002
121	Test Set 1	Insulin B 5–15	FVNQHLCGSHLVEAL	> 10000	Sidney *et al.* 2002
122	Test Set 1	Artificial sequence	YARFQSQTTLKQKT	> 10000	Sidney *et al.* 2002
123	Test Set 1	Artificial sequence	YARFQRQTTLKAAA	> 10000	Sidney *et al.* 2002
124	Test Set 1	Pf cp 379–396 truncated analog	IEKKIAKMEKASY	> 10000	Sidney *et al.* 2002
125	Test Set 1	CLIP 96-114	KLPKPPKPVSKMRMATPLL	> 10000	Sidney *et al.* 2002
126	Test Set 1	Pf MSP-1 250-271	FGYRKPLDNIKDNVGKMEDYIKK	> 10000	Sidney *et al.* 2002
127	Test Set 1	DQa1 0501 16-30	YQSYGPSGQYTHEFD	> 10000	Sidney *et al.* 2002

purposes, peptides are classified into their experimental IC_{50} values (high-affinity binders: $IC_{50} \leq 500$ nM, medium-affinity binders: 500 nM < $IC_{50} \leq 1500$ nM, low-affinity binders: 1500 < $IC_{50} \leq 5000$ nM and non-binders: 5000 < IC_{50}). In this dataset, 87 binding peptides had experimentally determined binding registers.

Dataset II consists of 12 Dermatophagoides pternnyssinus (*Der p*) peptides with experimental T cell proliferation values from functional studies (Krco *et al*., 2000; Neeno *et al*., 1996), with seven peptides eliciting DQ3.2*β*-restricted T cell proliferation.

6.2.2 Model

6.2.2.1 Peptide docking

In this study, an overlapping sliding window of size nine is applied to each peptide to generate all combinations of nonameric core-regions to be modeled into the binding groove of DQ3.2*β.* Docking was performed using an extension of the protocol as described in Chapter 6: (i) pseudo-Brownian rigid body docking of peptide fragments to the ends of the binding groove, (ii) central loop closure by satisfaction of spatial constraints, (iii) refinement of the backbone and side-chain atoms of the core recognition residues and receptor contact regions and (iv) extension of flanking peptide residues by satisfaction of spatial constraints. The conformations of the flanking peptide residues are generated by satisfying the spatial constraints in the allowed subspace for backbone dihedrals (Sali and Blundell, 1993), defined by the conformation of the bound core nonameric peptide docked into the binding groove. In brief, this is performed in three stages:

(i) distance and dihedral angle restraints on the entire peptide sequence are derived from its alignment with the nonamer sequence in the binding groove; (ii) the restraints on spatial features of the flanking residues are derived by extrapolation from the known 3D structure of flanking residues (PDB code 1JK8) in the alignment, expressed as probability density functions; and (iii) the spatial restraints on the flanking residues are then satisfied by optimization of the molecular probability density function using a variable target function technique that applies the conjugate gradients algorithm to positions of all non-hydrogen atoms.

6.2.2.2 Empirical free energy functions

The scoring function presented in the study is based on the free energy potential in ICM3.0 package (Abagyan and Totrov, 1999). The binding free energy is computed as the difference between the energy of the solvated complex and the sum of the energy of the solvated receptor and that of the peptide ligand. The reference state chosen for the peptide is the fully relaxed conformation of the free peptide in water (Schapira *et al.,* 1999). In all binding energy calculations, the protein and the ligand are separated after docking and their relaxed energies computed, following energy minimization in water. The binding free energy function (ΔG_{bind}) is expressed as

$$\Delta G_{\mathrm{bind}} = \alpha\Delta G_{H} + \beta\Delta G_{S} + \gamma\Delta G_{EL} + C$$

(*Equation 5*)

Here, ΔG_H is the hydrophobic energy computed as the product of solvent accessible surface area (determined by rolling a sphere of 1.40 Å radius along the surface of the molecule) by the surface tension. ΔG_S refers to the entropic contribution from the protein side-chains computed from the maximal burial entropies for each type of amino acid and their relative accessibilities. ΔG_{EL} denotes the electrostatic term composed of coulombic interactions between receptor and ligand and the desolvation of partial charges transferred from an aqueous medium to a protein core environment, and is determined by the numeric solution of the Poisson equation using an implementation of the boundary element algorithm (Zauhar and Morgan, 1985; Bharadwaj *et al.*, 1995; Schapira *et al.*, 1999). An additional constant term *C* (or *K*; Rognan *et al.,* 1999) accounts for entropy change in the system due to the decrease of free molecular concentration and the loss of rotational/translational degrees of freedom upon binding (Schapira *et al.*, 1999). In theory, *C* represents physical parameters which are independent on the data set used and there are great variations in its value among various research groups (reviewed in Janin, 1995). The coefficients (α, β, γ) assigned to each energy term were optimized in this study, to obtain the best separation of binders and non-binders in the peptide-DQ3.2β model. This partitioning scheme has been successfully adopted as a framework in many earlier studies (Krystek *et al.*, 1993; Weng *et al.*, 1996; Novotny *et al.*, 1997; Froloff *et al.*, 1997; Schapira *et al.*, 1999) and consists of the most significant potentials contributing to protein-protein, protein-ligand and protein-peptide interactions.

6.2.2.3 Optimization of the scoring function

Reported IC_{50} values, representing the concentration of ligand required to saturate half of the available binding sites of the protein (Bock and Gough, 2002), were assumed to be similar to equilibrium dissociation constants K_d as the concentration of the ligand in the unbound state is much lower than the equilibrium dissociation constant K_d of the ligand in the binding assay, so that $\Delta G_{bind} \approx -RT \ln (IC_{50})$ (Rognan *et al.,* 1999). ΔG_{bind} is usually reported in units of pK_d ($-\log_{10}(K_d)$), where 1 pK_d = -1.364 kcal/mol (Wang *et al.,* 2002) or -5.708 kJ/mol at 298.15 K. To improve the discriminative power of the scoring function, the coefficients of the different energy terms were recalibrated using standard least-square multivariate regression analyses of the training set (Wang *et al*., 2002). This step was followed by 10-fold cross-validation (Bock and Gough, 2002) to assess to quality of the scoring function.. In *k*-fold cross-validation, *k* random, (approximately) equal-sized, disjoint partitions of the sample data are constructed, and a given model is trained on (*k*-1) partitions and tested on the excluded partition. The results are averaged after *k* such experiments, and the observed error rate may be taken as an estimate of the error rate expected upon generalization to new data. The predictive power of the models was assessed by the cross-validation coefficient q^2 and the standard error of prediction s_{press}. The robustness of the predictive model was further evaluated using evolutionary regression analysis (Wang *et al*., 2002), with different subsets representing 5-fold, 4-fold, 3-fold and 2-fold cross-validation.

6.2.3 Training, testing and validation

Peptide data obtained from biochemical studies with experimental IC_{50} values was divided into training and test datasets. Training of the DQ3.2*β* prediction model was performed by sampling (i) the bound conformations of binding peptides with experimentally determined registers that can be recognized by MHC, and (ii) the best conformations of non-binding peptides without any preferred register in the binding groove. The training set comprised 56 binding conformations with known registers and 30 non-binding conformations generated from 3 non-binding peptides (from Dataset I) without any binding registers. Two external sets of test data were used: (i) Test set 1: 68 peptides (the rest of Dataset I) with experimental IC_{50} values (16 high-affinity binders, 13 medium affinity binders, 21 low affinity binders and 18 non-binders) from biochemical studies and (ii) Test set 2: all peptides from Dataset II, with known T cell proliferation values.

The predictive performance of our model was assessed using sensitivity (SE), specificity (SP) and receiver operating characteristic (ROC) analysis as described previously (Brusic *et al.*, 2002). SE=TP/(TP+FN) and SP=TN/(TN+FP), indicate percentages of correctly predicted binders and non-binders, respectively. TP (true positives) stands for experimental binders with at least one predicted binding register and TN (true negatives) for experimental non-binders with no predicted binding register. FN (false negatives) denotes experimental binders predicted as non-binders and FP (false positives) represents experimental non-binders predicted as binders. The accuracy of our predictions was assessed by

ROC analysis where the ROC curve is generated by plotting SE as a function of (1-SP) for various classification thresholds. The area under the ROC curve (A_{ROC}) provides a measure of overall prediction accuracy, A_{ROC}<70% for poor, A_{ROC}>80% for good and A_{ROC}>90% for excellent predictions (Brusic *et al.*, 2002). We consider values of SP≥80% useful in practice and assessed SE for three values of SP (80%, 90% and 95%).

6.3 Results

The accuracy of the DQ3.2*β* prediction model was evaluated using (i) peptides with experimental IC_{50} values obtained from biochemical studies with experimental IC_{50} values and (ii) peptides with T cell proliferation values obtained from functional studies.

Three threshold binding energy values were used to evaluate the accuracy of the DQ3.2*β* prediction model on Test set 1 – LMH (low-, medium-, high-affinity binders; A_{ROC}=0.88); MH (medium- and high-affinity binders; A_{ROC}=0.93) and H (high-affinity binders only; A_{ROC}=0.93). The results indicate that, overall, three-dimensional models are suitable for discriminating class II binding ligands from the background with good accuracy (A_{ROC}>0.80). The accuracy of our model relies on the scoring function derived from the training dataset of experimentally determined binders with known binding registers and non-binders with no binding registers. A scoring function based on the default ICM coefficients (α=β=γ=1; C=0) resulted in poor correlation (r^2=0.43, s=2.91 kJ/mol) to experimental data when tested with the novel peptide-DQ3.2*β* system. The discriminative power of our model improved significantly with better correlation (r^2=0.89, s=4.77 kJ/mol) after recalibration of the scoring function (Equation 5) by fitting to the training data, using multiple linear regression. The optimal scoring function, after 10-fold cross-validation (q^2=0.85, s_{press}=2.20 kJ/mol) is:

$$\Delta G_{\mathrm{bind}} = 1.55\Delta G_{\mathrm{H}} + 4.08\Delta G_{\mathrm{S}} - 0.23\Delta G_{\mathrm{EL}} - 7.12$$

(*Equation 6*)

The training set of 86 complexes in the current study is too large for the leave-one-out cross-validation done by Rognan *et al*. (1999) on training datasets of five and 37 MHC-peptide complexes. At the same time, it is smaller than the training set of 200 complexes used by Wang *et al.* (2002) or the 2617 protein-ligand complexes studies by Bock and Gough (2002) for extensive cross-validation analyses. The higher standard error in the training set (s=4.77 kJ/mol=1.13 kcal/mol=0.84 pK_{d}) than the standard error after 10-fold cross-validation (s_{press}=2.20 kJ/mol=0.52 kcal/mol=0.39 pK_{d}), is attributable to the noise in binding energy values in the complete training set spanning three orders of magnitude as illustrated in Table 10 compared to the subsets in 10-fold cross-validation, with a subset size (N) of 78 or 91% of the training set. These values are lower than error values (1.47-1.62 pK_{d} or 8.36-9.22 kJ/mol) reported by Wang *et al.* (2002), for a training set of 200 protein-ligand complexes, after several rounds of evolutionary regression analysis. Using a similar but limited evolutionary regression analysis approach, the robustness of our predictive model has been estimated for 5-fold (N=69, q^2=0.89, s_{press}=2.47 kJ/mol=0.43 pK_{d}), 4-fold (N=65, q^2=0.86, s_{press}=2.70 kJ/mol=0.47 pK_{d}), 3-fold (N=57, q^2=0.87, s_{press}=2.50 kJ/mol=0.44 pK_{d}) and 2-fold (N=43, q^2=0.83 s_{press}=3.29 kJ/mol=0.58 pK_{d}) cross-validation. The results indicate that despite a very slight increase in the error value for the 2-fold cross-validation, the cross-validation coefficient q^2 and the standard error of prediction s_{press} are stable, with mean values of q^2=0.86 and s_{press}=2.63 kJ/mol=0.46 pK_{d}, and respective standard deviation values of 0.02 and 0.41 kJ/mol=0.07 pK_{d}. This iterative regression procedure thus validates the internal consistency of the scoring function in the current model, rendering it suitable for predictions on the test datasets.

The sensitivity of our prediction model was determined on Test set 1 for three decision thresholds (Table 11) that define levels of specificities suitable for practical applications (Brusic *et al.*, 2002). SP=0.80 offers high-sensitivity predictions, whereas SP=0.95 results in very few false positives but fewer true positives. The prediction results for our model were in accordance with expected binding patterns of DQ3.2β peptides and provided a sensitivity of 90% (SP=0.80). The sensitivity values decrease with higher levels of specificity (SP=0.90, SE=0.84 or SP=0.95, SE=0.81), while still correctly predicting more than half of the high-affinity binders in the worst case scenario (high-binders alone, SP=0.95, SE=0.63). The efficacy of our model in detecting binding registers was then evaluated with experimentally determined registers. Our external test data comprised 23 peptides from Test set 1, with known binding energy for each register (Suri *et al.*, 2005). At a threshold of -30.82 kJ/mol

(SP=0.80, SE=0.75), our model accurately detected 87% (20/23) of the experimentally determined binding registers. We also correctly predicted the only experimentally determined register (4-12) for *Der p* 2 1-20 (Krco *et al.*, 2000), from Test set 2.

Next, the predictive performance of the optimized model was tested on the functional dataset of 12 peptides (Test set 2) with experimental T cell proliferation values using the decision thresholds defined above. The top five predictions (*Der p* 2 61-80, 51-70, 110-129, 101-120, 91-110) are experimental positives (Table 12) with binding energy values of -34.52 kJ/mol or less (predicted high-binders for SE = 0.63, SP = 0.95). This is in agreement with existing studies that high-affinity binders have a greater chance of stimulating T cell proliferation (Deng *et al*., 1997; Keogh *et al*. 2001) and this knowledge is crucial for peptide vaccine design. Peptide *Der p* 2 31-50, ranked #6, is a predicted high-affinity binder at threshold -33.59 kJ/mol (SE = 0.63, SP = 0.95). It is possible that *Der p* 2 31-50 is either a high-affinity binder that failed to stimulate T cell proliferation (Deng *et al*., 1997; Keogh *et al*., 2001) or is a false positive in the prediction. At this cut-off, correct predictions number 5/7 (71%), with one false positive (14%) and two false negatives (28%). Peptide *Der p* 2 41-60, ranked #12 in our prediction, is possibly an outlier, as it failed to stimulate detectable T cell response in study of Neeno *et al.* (1996), despite a similar reported T cell stimulatory propensity as *Der p* 2 91-110; and deletion experiments confirm the criticality of only residues 55-70 in the region 41-70 (Table 2 in Krco *et al*., 2000). For T cell proliferation pr-

Table 11 Sensitivity values and binding energy thresholds for prediction of peptide binding to DQ3.2*β* for the specificity levels of 0.80, 0.90 and 0.95.

Specificity (SP)	Group	Sensitivity (SE)	Binding Energy Threshold (kJ/mol)
SP = 0.80	LMH	0.90	-28.70
	MH	0.85	-29.10
	H	0.75	-30.82
SP = 0.90	LMH	0.84	-29.10
	MH	0.77	-30.50
	H	0.75	-32.74
SP = 0.95	LMH	0.81	-29.93
	MH	0.73	-32.12
	H	0.63	-33.59

Table 12 Predicted Dermatophagoides pternnyssinus (*Der p 2)* allergenic peptide sequences to DQ3.2*β*. The top 5 predictions are experimentally positive. '–' indicates non-immunostimulatory in the relevant experiments. '*' indicates peptides that elicit T cell responses in one or both experimental studies. Values <3500 are not considered significant (Neeno *et al.*, 1996). The experimentally determined binding register is shown in underlined bold type.

Peptide	Sequence	T cell proliferation (Δ cpm)		Predicted Binding Energy (kJ/mol)
		Krco *et al.* 2000	Neeno *et al.* 1996	
Der p 2 61-80*	LEVDVPGIDPNACHYMKCPL	24,381	5,455	-38.97
Der p 2 51-70*	KIEIKASIDGLEVDVPGIDP	49,958	<3,500	-38.60
Der p 2 111-129*	MGDDGVLACAIATHAKIRD	–	8,839	-37.00
Der p 2 101-120*	SENVVVTVKVMGDDGVLACA	54,256	<3,500	-35.26
Der p 2 91-110*	TWNVPKIAPKSENVVVTVKV	47,711	8,409	-34.52
Der p 2 31-50	RGKPFQLEAVFEAVQNTKTA	–	–	-33.96
Der p 2 1-20*	DQV**DVKDCANHE**IKKVLVPG	36,389	–	-31.44
Der p 2 11-30	HEIKKVLVPGCHGSEPCIIN	–	–	-31.40
Der p 2 81-100	VKGQQYDIKYTWNVPKIAPK	–	<3,500	-31.35
Der p 2 21-40	CHGSEPCIIHRGKPFQLEAV	–	–	-31.33
Der p 2 71-90	HACHYMKCPLVKGQYDIDKY	–	<3,500	-30.71
Der p 2 41-60*	FEAVQNTKTAKIEIKASIDG	46,871	–	-26.49

edictions, the current model is suitable to screen for high-affinity binders at SP=0.95.

6.4 Discussion

6.4.1 Detection of epitopes that do not conform to binding motifs

Consensus peptide-binding motifs for identifying potential immunodominant epitopes within autoantigenic proteins have been developed for many HLA class II molecules. However, earlier studies (Harfouch-Hammoud *et al*. 1999) reveal that these motifs do not correlate with binding to a specific allele. In Test set 1, 63 out of 68 binding peptide sequences have amino acid residues that do not conform to available DQ3.2*β* binding motifs (Godkin *et al*., 1998; Rammensee *et al*., 1999) considering all relevant positions (P1, P4, P6, P7, P9). Table 13 lists 17 LMH predictions from this dataset. A-gliadin 49-63 (#10), MHC Ia 46–63 (#14) and VP16 (#15) are classified negatives using existing DQ3.2*β* binding motifs. However, using our scoring function these T cell epitopes are easily identified, with A-gliadin 49-63 as a high affinity binder and the MHC Ia 46–63 and VP16 as medium affinity binders. This reaffirms our earlier observation that binding motifs may be inadequate for defining T cell epitopes and many other factors including the physicochemical composition of the peptide, (affecting the overall stability of the peptide/MHC complex) have to be considered in prediction systems for HLA-binding peptides.

6.4.2 Detection of multiple registers in experimental binders

Our results support the existence of multiple registers with different nonameric core regions within a candidate binding peptide that serve as recognition sites for MHC class II molecules (Figure 23). In particular, the results indicate that several binding registers (with different nonameric core recognition regions) exist within an MHC class II binding peptide, facilitating binding to DQ3.2*β* in several different conformations. 58% of binding peptides in Test set 1 exhibit two or more registers that can be docked to DQ3.2*β* with favorable binding energy values. Multiple registers occur predominantly in medium- and high-affinity binders, suggesting that recognition using flexible fitting may play a critical role in binding to MHC class II alleles as well as in T cell recognition and this knowledge should be taken into consideration in vaccine design. For example, two conformations of the high affinity binding peptide Pf ABRA 487–506 showed ∆G values less than the decision threshold -33.59 kJ/mol (SP = 0.95, SE = 0.63), with the 496–504 register (shown in Table 13) being the preferred binding mode.

It is possible that the open binding groove of DQ3.2*β* (and other class II alleles) accommodates peptides with differing pocket specificities and can recognize multiple regions within a single candidate peptide. Whilst not all binding registers may elicit T cell response, the existence of multiple registers within a candidate peptide (especially for high-affinity binders) can facilitate binding to a particular allele, enhancing T cell recognition, with the highest binding affinity register acting as the primary recognition region. It is also possible that a peptide may initially bind in one register and migrate laterally into another.

Table 13 Analysis of DQ3.2*β* binding motifs. Predicted binding registers of the top 17 DQ3.2*β* ligands/epitopes (Test set 1; SP = 0.95, SE = 0.81) from the conformation for the lowest predicted binding energy (BE) values. Residues that conform to existing DQ3.2*β* peptide binding motifs (Godkin *et al.*, 1998; Rammensee *et al.*, 1999) are underlined. Predicted high-affinity, moderate affinity and low affinity binders are ligands 1-10, 11-16 and 17 respectively.

	Position	Source	BE (kJ/mol)	IC50 (nM)	Reference
	1 4 67 9				
Binding Motif	**T D RR Q** **S V VV N** **W M DD G** **K A AA D** **E I II P** **D YY R** **Q E** **F** **L** **M**				
Ligands / Epitopes	LQLQP**F**PQPQPFP**P**L	A-gliadin 56-70	-41.01	20	Godkin *et al*., 1998
	DMTPAD**A**L**DD**F**D**L	HSV	-40.53	173	Sidney *et al*., 2002
	AAAA**A**V**AA**EAY	Artificial sequence	-39.98	48	Sidney *et al*., 2002
	GVAGLL**VA**LAV	IA-2 499-509	-36.16	95	Sidney *et al*., 2002
	DSNIMNSINNVM**D**E**ID**FFEK	Pf ABRA 487–506	-36.01	171	Sidney *et al*., 2002
	FES**T**GNLI**A**PEYGFKISY	HA 255–271Y	-35.70	62	Sidney *et al*., 2002
	YP**F**IEQEGPEFFDQE	MHC Ia 51–63 analog	-35.34	1156	Sidney *et al*., 2002
	LLDI**L**DT**A**GLEEYSAMRD	p21 51–66; C out	-35.27	202	Sidney *et al*., 2002
	QPYPQPQPFPSQQ**P**Y	A-gliadin 41-55	-35.26	1120	Godkin *et al*., 1998
	FPSQ**Q**PYLQLQPFPQ	A-gliadin 49-63	-33.93	20	Godkin *et al*., 1998
	CDGERPTL**A**FLQDVM	GAD 101–115	-33.57	69	Sidney *et al*., 2002
	S**F**PPQQP**Y**P**Q**PQPQY	A-gliadin 77-91	-33.35	370	Godkin *et al*., 1998
	SQDLEL**S**WNLNGLQADLSS	FceR 104–122	-32.89	123	Sidney *et al*., 2002
	EPRAPW**I**EQEGPEYW	MHC Ia 46-63	-32.89	519	Sidney *et al*., 2002
	PPLYATGRLSQAQLMPSPPM	VP16	-32.59	538	Sidney *et al*., 2002
	SQDLEL**S**WNLNGLQAY	FceR 104–122 analog	-32.49	118	Sidney *et al*., 2002
	IARAKMFPAVAEK	34P3A	-31.91	541	Sidney *et al*., 2002

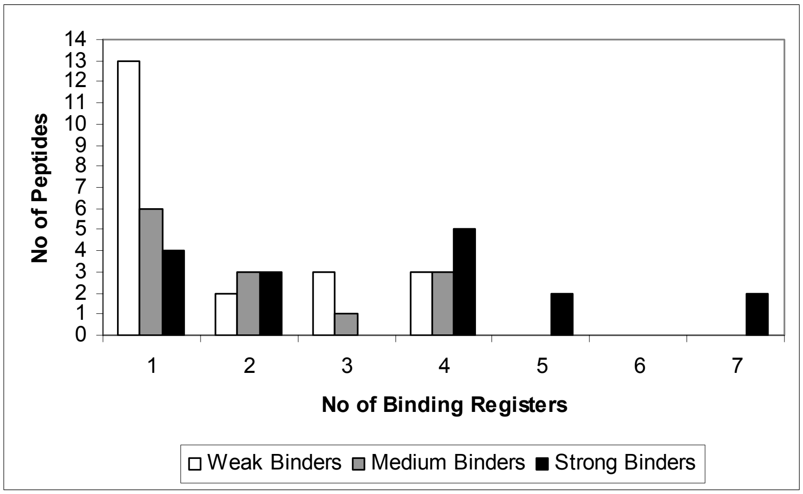


Figure 23 Number of binding registers within predicted DQ3.2*β* binding peptides for Test set 1 (SE = 0.81, SP = 0.95). Medium- and high-affinity binders represent the highest proportion of binding peptides predicted to contain more than one binding register.

Peptide vaccine development is advancing rapidly with recent successes in malaria (Lopez *et al*., 2001) and anti-tumor vaccines (Knutson *et al*., 2001). A key research area is to identify allele-specific candidate T cell epitopes suitable for designing vaccines and immunotherapies to control allergic or autoimmune responses. The task of identifying candidate class II binding ligands is a challenging process due to the open binding groove that can potentially accommodate multiple binding registers (Li *et al.*, 2000; Seamons *et al.*, 2003) and has wholly occupied the energies of researchers. A polynomial derived scoring matrix for DRB1*0401 of Southwood *et al.* (1998) and a genetic algorithm (Brusic *et al*., 1998) are excellent approaches. However, the nonameric core regions used for training predictive models were often preselected based on existing binding motifs, usually extracted from multiple sequence alignment and not experimentally validated. Such methodologies exclude the prediction of other binding registers within a candidate class II binding ligand capable of eliciting a strong T cell response. Moreover, the possibility of the existence of multiple binding registers, particularly for high-affinity binders, suggests that all possible nonameric core regions within a candidate binding ligand must be carefully examined. For training computational models, the utilization of experimentally validated binding registers is preferred.

Recently, Sinha *et al*. (unpublished results) discovered that DRB1*0402-specific binding motifs are insufficient for the design of pemphigus vulgaris epitopes, due to the presence of register shifts as well as polymorphisms in the binding register. With increasing evidence suggesting the inadequacy of binding motifs in defining class II T cell epitopes, the current approach of predictive model building and virtual screening for vaccine candidates is independent of sequence motifs and takes into account the presence of multiple registers within class II ligands. In a wider context, the methodology presented here might be helpful in defining peptide antigenicity in a wide spectrum of human diseases including cancer pathologies as well as a range of diverse autoimmune disorders such as insulin-dependent diabetes mellitus, multiple sclerosis, rheumatoid arthritis, and pemphigus vulgaris. In this study, we have illustrated that it is possible to efficiently discriminate between categories of binders from non-binders and predict the binding register of class II ligands with good accuracy. Our docking methodology, combined with a sensitive scoring function, provides a set of sensitive and specific computational tools to facilitate systematic screening of peptides for immunotherapeutic applications.

6.5 Summary

- In this work, a scoring function has been developed as an extension of the docking protocol developed in Chapter 6 for functional prediction of MHC-binding peptides.
- High accuracy of predictions was obtained by validation with experimental biochemical and functional data. This approach successfully identified peptide binders which lack conserved binding motifs.
- The present analysis reveal the possible existence of multiple binding registers within a candidate class II binding peptide, suggesting that recognition via flexible fitting may play a critical role in binding to class II alleles.

Chapter 7: Analysis of T cell epitope repertoire in Dsg3

7.1 Introduction

Pemphigus vulgaris (PV) is a severe autoimmune blistering skin disorder due to loss of integrity of normal intercellular attachments within the epidermis and mucosal epithelium. Strong association of PV to the major histocompatibility complex (MHC) class II alleles DRB1*0402 and DQB1*0503 have been reported in the literature (Carcassi *et al*., 1996; Delgado *et al*., 1997; Loiseau *et al*., 2000; Miyagawa *et al*., 1997, 1999; Nizeki *et al*., 1991; Scharf *et al*., 1988; Sinha *et al*., 1988) with over 95% of PV patients possessing one or both of these alleles (Scharf *et al*., 1988; Sinha *et al*., 1988). The target antigen of PV, desmoglein (Dsg) 3, is a 130-kDa transmembrane glycoprotein that belongs to the cadherin superfamily of cell adhesion molecules (Amagai *et al*., 1991). In the early stage of disease, patients demonstrate autoimmunity only to Dsg3 and develop mucosal blisters; while at the later stage, patients exhibit non-cross-reactive immunity to both Dsg3 and Dsg1 (Salato *et al*., 2005). Despite several reports of T cell specificities for Dsg3 (Veldman *et al*., 2004; Hertl *et al*., 1998; Salato *et al*., 2005; Wucherpfennig *et al*., 1995; Riechers *et al*., 1999), much remains unknown with regards to the role of T cells in the pathogenesis of PV.

Bioinformatic tools are now commonly used to facilitate T cell epitope discovery (Schirle *et al*., 2001; Yu *et al*., 2002, Srinivasan *et al*., 2004). Computational methods for predicting MHC-binding peptides include procedures based on sequence motifs (Wucherpfennig *et al*., 1995), quantitative matrices

(Parker *et al*., 1994; Davenport *et al*., 1995; Gulukota *et al*., 1997), decision trees (Savoie *et al*., 1999; Segal *et al*., 2001), artificial neural networks (Brusic *et al*., 1994, 1998), hidden Markov models (Mamitsuka, 1998) and support vector machines (Dönnes and Elofsson, 2002; Bhasin and Raghava, 2004; Bozic *et al*., 2005). However, despite recent advances in sequence-based predictive techniques, effective models for DRB1*0402 and DQB1*0503 have been lacking, mainly due to the paucity of sufficient peptides as training data (Tong *et al*., 2006a) as well as the presence of register shifts and polymorphisms in the binding registers. To date, few prediction techniques for MHC class II molecules have been developed using three-dimensional models since the dual issues of model quality and discriminative technique are still to be addressed (Ranganathan *et al.,* 2005).

Our strategy for prediction of T cell epitopes involves three-dimensional modeling of peptide/MHC complexes using a hybrid docking approach that integrates the strength of Monte Carlo simulations and homology modeling (Tong *et al*., 2004, 2006a,b). In an earlier study, we have successfully discriminated disease-implicated from non-disease implicated and protective alleles in PV based on structural interaction rules (Tong *et al*., 2006a). A complementary scoring function has now been developed for effective identification of DRB1*0402 and DQB1*0503 epitopes. We investigated the T cell epitope repertoire of the entire Dsg3 glycoprotein and show the existence of multiple extracellular and intracellular specificities within the Dsg3 self-antigen. Further analysis reveal that DRB1*0402 and DQB1*0503 share similar specificities by binding peptides at different core recognition regions. These data impact our understanding of the mechanism of HLA mediated control of disease.

7.2 Materials and Methods

7.2.1 Template search

MHC sequence data were obtained from IMGT-HLA database (http://www.ebi.ac.uk/imgt/hla/) (Robinson *et al*., 2003). To identify potential structural templates available in the Protein Data Bank (PDB) (Berman *et al*., 2000) for model building, a sequence similarity search was performed using BLAST (Altschul *et al*., 1990) running on the servers at NCBI (www.ncbi.nlm.nih.gov/blast/) and the highest quality templates were selected among the returned results. The crystal structures of the highly conserved DRB1*0401 (99% similarity; PDB code 1D5Z) and DQB1*0602 (96% similarity; PDB code 1UVQ) were selected as templates for DRB1*0402 and DQB1*0503 respectively.

7.2.2 Model building

The program MODELLER (Sali and Blundell, 1993) was employed for comparative modeling of both DRB1*0402 and DQB1*0503. The models are constructed by optimally satisfying spatial constraints obtained from the alignment of the template structure with the target sequence and from the CHARMM-22 force field (MacKerell *et al*., 1998). The structures were relaxed by conjugate gradient minimization, using the Internal Coordinate Mechanics (ICM) 3.0 package (Abagyan *et al.*, 1994).

7.2.3 Experimental binding data

Two sets of data are used in this study: (i) peptides with experimental IC_{50} values from biochemical studies and (ii) peptides with experimental T cell proliferation values/responses from functional studies.

Dataset I (Table 14) comprises 59 DRB1*0402-specific peptides derived from biochemical studies (20 high-affinity, 11 medium-affinity and 13 low-affinity binders and 15 non-binders). Peptides are classified based on their experimental IC_{50} values (high-affinity binders: $IC_{50} \leq 500$ nM, medium-affinity binders: 500 nM $< IC_{50} \leq 1500$ nM, low-affinity binders: $1500 < IC_{50} \leq 5000$ nM and non-binders: $5000 < IC_{50}$).

Table 14 DRB1*0402-specific peptides with experimental IC_{50} values used in this study.

No.	Allele	Category	Description	Peptide	IC_{50} (nM)	Reference
1	DRB1*0402	Training Set	Dsg3 342-356	LNSKIAFKIVSQEPA	2600	Sinha *et al.* (unpublished)
2	DRB1*0402	Training Set	Dsg3 786-800	RNPIAKITSDYQATQ	4700	Sinha *et al.* (unpublished)
3	DRB1*0402	Training Set	Dsg3 810-824	PFGIFVVDKNTGDIN	>40000	Sinha *et al.* (unpublished)
4	DRB1*0402	Training Set	Dsg3 67-81	SVKLSIAVKNKAEFH	>40000	Sinha *et al.* (unpublished)
5	DRB1*0402	Training Set	Dsg3 846-860	MNFLDSYFSQKAFAC	8900	Sinha *et al.* (unpublished)
6	DRB1*0402	Training Set	Dsg3 963-977	NDCLLIYDNEGADAT	>40000	Sinha *et al.* (unpublished)
7	DRB1*0402	Training Set	Dsg3 96-110	LDSLGPKFKKLAEIS	>40000	Sinha *et al.* (unpublished)
8	DRB1*0402	Training Set	Dsg3 191-205	ERVICPISSVPGNLA	2700	Sinha *et al.* (unpublished)
9	DRB1*0402	Test Set	HADP analogue	AAVAAAKAAAAAA	17	Marshall *et al*. (1994)
10	DRB1*0402	Test Set	HADP analogue	AAWAAAKAAAAAA	2200	Marshall *et al*. (1994)
11	DRB1*0402	Test Set	HADP 7.18	AAYAAAKAAALAA	7000	Marshall *et al*. (1994)
12	DRB1*0402	Test Set	Myoglobin 110-122	AIIHVLHSRHPGD	1.9	Marshall *et al*. (1994)
13	DRB1*0402	Test Set	Myoglobin 67-79	TVLTALGAILKKK	640	Marshall *et al*. (1994)
14	DRB1*0402	Test Set	Myelin BP 90-102	HFFKNIVTPRTPA	61	Marshall *et al*. (1994)
15	DRB1*0402	Test Set	Tetanus toxoid 828-840	MQYIKANSKFIGI	900	Marshall *et al*. (1994)
16	DRB1*0402	Test Set	Pertussis Toxin 31-43	NVLDHLTGRSSQV	340	Marshall *et al*. (1994)

17	DRB1*0402	Test Set	Hemagglutinin 103-115	PDYASLRSLVASS	18	Marshall *et al*. (1994)
18	DRB1*0402	Test Set	Hemagglutinin 307-319	PKYVKQNTLKLAT	110	Marshall *et al*. (1994)
19	DRB1*0402	Test Set	*M. leprae* 65 kDa 416-428	TLLQAAPALDKLK	870	Marshall *et al*. (1994)
20	DRB1*0402	Test Set	HADP analogue	AAFAAAKAAAAAA	47	Marshall *et al*. (1994)
21	DRB1*0402	Test Set	HADP analogue	AALAAAKAAAAAA	2	Marshall *et al*. (1994)
22	DRB1*0402	Test Set	HADP analogue	AASAASKAAAAAA	60000	Marshall *et al*. (1994)
23	DRB1*0402	Test Set	HADP 7.20	AAYAAAKAAAVAA	13000	Marshall *et al*. (1994)
24	DRB1*0402	Test Set	HADP 7.21	AAYAAAKAAASAA	4500	Marshall *et al*. (1994)
25	DRB1*0402	Test Set	HADP 7.44	AAYAAAKAEAAAA	3400	Marshall *et al*. (1994)
26	DRB1*0402	Test Set	HADP 18.7	AAYAAAKAAAAAA	1600	Marshall *et al*. (1994)
27	DRB1*0402	Test Set	HADP 7.25	AAYAAAKAAAGAA	5700	Marshall *et al*. (1994)
28	DRB1*0402	Test Set	HADP 7.45	AAYAAAKALAAAA	60	Marshall *et al*. (1994)
29	DRB1*0402	Test Set	HADP 7.50	AAYAAFKAAAAAA	460	Marshall *et al*. (1994)
30	DRB1*0402	Test Set	HADP 7.27	AAYAAKKAAAAAA	850	Marshall *et al*. (1994)
31	DRB1*0402	Test Set	HADP 7.30	AAYAALKAAAAAA	1900	Marshall *et al*. (1994)
32	DRB1*0402	Test Set	HADP 7.39	AAYAAQKAAAAAA	2700	Marshall *et al*. (1994)
33	DRB1*0402	Test Set	HADP 7.29	AAYAASKAAAAAA	5200	Marshall *et al*. (1994)
34	DRB1*0402	Test Set	Flu NP 383-395	SRYWAIRTRSGGI	13	Marshall *et al*. (1994)
35	DRB1*0402	Test Set	Matrix 18-30	GPLKAEIAQRLED	25000	Marshall *et al*. (1994)
36	DRB1*0402	Test Set	Tetanus toxoid 591-603	KIYSYFPSVISKV	9.2	Marshall *et al*. (1994)
37	DRB1*0402	Test Set	Haemagglutinin 23-35	GTLVKTITDDQIE	1200	Marshall *et al*. (1994)
38	DRB1*0402	Test Set	HADP 7.23	AAYAAAKAAARAA	4200	Marshall *et al*. (1994)
39	DRB1*0402	Test Set	HADP 7.46	AAYAAAKAFAAAA	200	Marshall *et al*. (1994)
40	DRB1*0402	Test Set	HADP 7.43	AAYAAAKAKAAAA	830	Marshall *et al*. (1994)
41	DRB1*0402	Test Set	HA Y307-319	YPKFVKQNTLKAA	2200	Harfouch-Hammoud *et al*. (1999)
42	DRB1*0402	Test Set	Hsp65 189-201 analogue	EGMRFAKGYISGY	1000	Hammer *et al*. (1995)
43	DRB1*0402	Test Set	Designer peptide	GFKYAAAAAA	6000	Hammer *et al*. (1995)
44	DRB1*0402	Test Set	Designer peptide	GFKAAARAAA	9509	Hammer *et al*. (1995)
45	DRB1*0402	Test Set	Designer peptide	GFKAAAHAAA	60000	Hammer *et al*. (1995)
46	DRB1*0402	Test Set	HLA-B	GRLLRGHNQFAYDGK	5	Kirschmann *et al*. (1995)
47	DRB1*0402	Test Set	Synthetic peptide	DTQFVRFDSDAASQR	600	Kirschmann *et al*. (1995)
48	DRB1*0402	Test Set	Apoliopoprotein	TPDFIVPLTDLRIPS	70	Kirschmann *et al*. (1995)
49	DRB1*0402	Test Set	Actin peptide	YPIEHGIVTNWDDM	4000	Kirschmann *et al*. (1995)
50	DRB1*0402	Test Set	Synthetic peptide	EEFVVEFDLPGIK	100	Kirschmann *et al*. (1995)
51	DRB1*0402	Test Set	Synthetic peptide	AEFVVEFDLPGIK	1000	Kirschmann *et al*. (1995)
52	DRB1*0402	Test Set	Synthetic peptide	EAFVVEFDLPGIK	1000	Kirschmann *et al*. (1995)
53	DRB1*0402	Test Set	Synthetic peptide	EEFAVEFDLPGIK	250	Kirschmann *et al*. (1995)
54	DRB1*0402	Test Set	Synthetic peptide	EEFVAEFDLPGIK	250	Kirschmann *et al*. (1995)
55	DRB1*0402	Test Set	Synthetic peptide	EEFVVAFDLPGIK	100	Kirschmann *et al*. (1995)
56	DRB1*0402	Test Set	Synthetic peptide	EEFVVEADLPGIK	1000	Kirschmann *et al*. (1995)
57	DRB1*0402	Test Set	Synthetic peptide	EEFVVEFALPGIK	400	Kirschmann *et al*. (1995)
58	DRB1*0402	Test Set	Synthetic peptide	EEFVVEFDAPGIK	5500	Kirschmann *et al*. (1995)
59	DRB1*0402	Test Set	Synthetic peptide	EEFVVEFDLAGIK	7000	Kirschmann *et al*. (1995)

Table 15 Dsg3 peptides with experimental T cell proliferation values/responses used in this study.

No.	Allele	Category	Description	Peptide	Reference
1	DRB1*0402	Test Set	Dsg3 342-358	SVKLSIAVKNKAEFHQS	Veldman *et al*. (2004)
2	DRB1*0402	Test Set	Dsg3 376-392	NVREGIAFRPASKTFTV	Veldman *et al*. (2004)
3	DRB1*0402	Test Set	Dsg3 205-221	GTPMFLLSRNTGEVRTL	Veldman *et al*. (2004)
4	DRB1*0402	Test Set	Dsg3 380-396	GIAFRPASKTFTVQKGI	Riechers *et al*. (1999)
5	DRB1*0402	Test Set	Dsg3 190-204	LNSKIAFKIVSQEPA	Wucherpfennig *et al*. (1995)
6	DRB1*0402	Test Set	Dsg3 189-205	HLNSKIAFKIVSQEPAG	Veldman *et al*. (2004)
7	DRB1*0402	Test Set	Dsg3 512-526	SARTLNNRYTGPYTF	Wucherpfennig *et al*. (1995)
8	DRB1*0402	Test Set	Dsg3 78-94	QATQKITYRISGVGIDQ	Wucherpfennig *et al*. (1995)
9	DRB1*0402	Test Set	Dsg3 78-93	QATQKITYRISGVGID	Veldman *et al*. (2004)
10	DRB1*0402	Test Set	Dsg3 206-220	TPMFLLSRNTGEVRT	Wucherpfennig *et al*. (1995)
11	DRB1*0402	Test Set	Dsg3 210-226	LLSRNTGEVRTLTNSL	Veldman *et al*. (2004)
12	DRB1*0402	Test Set	Dsg3 251-265	CECNIKVKDVNDNFP	Wucherpfennig *et al*. (1995)
13	DRB1*0402	Test Set	Dsg3 250-266	QCECNIKVKDVNDNFPM	Veldman *et al*. (2004)
14	DRB1*0402	Test Set	Dsg3 483-499	VRVPDFNDNCPTAVLEK	Veldman *et al*. (2004)
15	DRB1*0402	Test Set	Dsg3 762-776	QSGTMRTRHSTGGTN	Wucherpfennig *et al*. (1995)
16	DRB1*0402	Test Set	Dsg3 161-177	IFMGEIEENSASNSLVM	Hertl *et al*. (1998)
17	DRB1*0402	Test Set	Dsg3 96-112	PFGIFVVDKNTGDINIT	Veldman *et al*. (2004)
18	DRB1*0402	Test Set	Dsg3 97-111	FGIFVVDKNTGDINI	Wucherpfennig *et al*. (1995)
19	DQB1*0503	Training Set	Dsg3 342-358	LNSKIAFKIVSQEPA	Veldman *et al*. (2004)
20	DQB1*0503	Training Set	Dsg3 376-392	RNPIAKITSDYQATQ	Veldman *et al*. (2004)
21	DQB1*0503	Training Set	Dsg3 205-221	PFGIFVVDKNTGDIN	Veldman *et al*. (2004)
22	DQB1*0503	Training Set	Dsg3 250-266	SVKLSIAVKNKAEFH	Veldman *et al*. (2004)
23	DQB1*0503	Training Set	Dsg3 96-112	MNFLDSYFSQKAFAC	Veldman *et al*. (2004)
24	DQB1*0503	Training Set	Dsg3 512-526	NDCLLIYDNEGADAT	Wucherpfennig *et al*. (1995)
25	DQB1*0503	Training Set	Dsg3 97-111	LDSLGPKFKKLAEIS	Wucherpfennig *et al*. (1995)
26	DQB1*0503	Training Set	Dsg3 78-93	ERVICPISSVPGNLA	Wucherpfennig *et al*. (1995)
27	DQB1*0503	Test Set	Dsg3 189-205	HLNSKIAFKIVSQEPAG	Veldman *et al*. (2004)
28	DQB1*0503	Test Set	Dsg3 762-786	QSGTMRTRHSTGGTN	Wucherpfennig *et al*. (1995)
29	DQB1*0503	Test Set	Dsg3 190-204	LNSKIAFKIVSQEPA	Wucherpfennig *et al*. (1995)
30	DQB1*0503	Test Set	Dsg3 206-220	TPMFLLSRNTGEVRT	Wucherpfennig *et al*. (1995)
31	DQB1*0503	Test Set	Dsg3 251-265	CECNIKVKDVNDNFP	Wucherpfennig *et al*. (1995)

Dataset II (Table 15) consists of 18 DRB1*0402-specific Dsg3 peptides and 13 DQB1*0503-specific Dsg3 peptides with T cell proliferation values/responses (Veldman *et al*., 2004; Hertl *et al*., 1998; Wucherpfennig *et al*., 1995).

7.2.4 Peptide docking

A sliding window of size nine was applied to each peptide to generate all possible overlapping nonameric core-regions that can be modeled into the binding grooves of DRB1*0402 and DQB1*0503*.* Docking was performed using an extension of the protocol, illustrated in Chapter 6.

7.2.5 Empirical free energy functions

The scoring function presented in this study is based on the free energy potential in ICM3.0 package (Abagyan and Totrov, 1999). The binding free energy function is partitioned into three terms (Tong *et al*., 2006) expressed by the equation:

$$\Delta G = \alpha\Delta G_H + \beta\Delta G_S + \gamma\Delta G_{EL} + C. \qquad (1)$$

ΔG_H is the hydrophobic energy computed as the product of solvent accessible surface area (determined by rolling a sphere of 1.40Å radius along the surface of the molecule) by the surface tension. ΔG_S refers to the entropic contribution from the protein side-chains computed from the maximal burial entropies for each type of amino acid and their relative accessibilities. ΔG_{EL} denotes the electrostatic term composed of coulombic interactions between receptor and ligand and the desolvation of partial charges transferred from an aqueous medium to a protein core environment, and is determined by the numeric solution of the Poisson equation using an implementation of the boundary element algorithm (Zauhar and Morgan, 1985; Bharadwaj *et al.*, 1995; Schapira *et al*., 1999). The constant term *C* accounts for entropy change in the system due to the decrease of free molecular concentration and the loss of rotational/translational degrees of freedom upon binding (reviewed in Janin, 1995).

7.2.6 Training, testing and validation

Two computational models are trained in this study – one model for the prediction of peptide binding to DRB1*0402 and the other for DQB1*0503.

DRB1*0402-specific peptide data derived from biochemical studies with experimental IC_{50} values was divided into training and test sets. The training set comprised 8 (5 binding and 3 non-binding) Dsg3 sequences with experimentally determined binding registers (from Dataset I). Two external sets of test data were used: (i) Test set 1: 51 peptides with experimental IC_{50} values (20 high-affinity binders, 11 medium affinity binders, 9 low affinity binders and 11 non-binders) from biochemical studies, and (ii) Test set 2: all DRB1*0402-specific Dsg3 peptides from Dataset II, with known T cell proliferation values.

DQB1*0503 prediction model was trained using Dsg3 peptide data from functional studies in the absence of relevant biochemical data. The training set comprised 8 (5 stimulatory and 3 non-stimulatory) sequences from Dataset II. For each peptide sequence, T cell proliferation value (Wucherpfennig *et al*., 1995) is mapped to a theoretical IC_{50} value in accordance with expected binding patterns of Dsg3 binding peptides (Sinha *et al*., personal communications). The performance of the prediction model was subsequently evaluated on an external set of 5 peptides with known T cell proliferation values.

Coefficients (α, β, γ) and the constant term *C* in Equation 1 were derived using standard least-square multivariate regression analyses of the training set, followed by leave-one-out analysis to assess the quality of the scoring function (Rognan *et al*., 1999). For each model, the entire procedure is repeated 8 times to reduce noise in all computations, the results averaged and the observed error rate is used to estimate the expected error rate upon generalization to new data.

The optimal scoring function selected from each cross-validation analysis was further assessed using sensitivity (SE), specificity (SP) and receiver operating characteristic (ROC) analysis (Tong *et al.*, 2006). SE=TP/(TP+FN) and SP=TN/(TN+FP), indicate percentages of correctly predicted binders and non-binders, respectively. TP (true positives) represents experimental binders with at least one predicted binding register and TN (true negatives) for experimental non-binders with no predicted binding register. FN (false negatives) denotes experimental binders predicted as non-binders and FP (false positives) represents experimental non-binders predicted as binders. The accuracy of our predictions was assessed by ROC analysis where the ROC curve is generated by plotting SE as a function of (1-SP) for various classification thresholds. The area under the ROC curve (A_{ROC}) provides a measure of overall prediction accuracy, A_{ROC}<70% for poor, A_{ROC}>80% for good and A_{ROC}>90% for excellent predictions (Tong *et al.*, 2006b). In this study, we assessed SE for three values of SP (80%, 90% and 95%) that are considered useful in practice. All regression and validation results, including correlation coefficient (r^2), standard deviation (*s*), cross-validation coefficient (q^2), standard error of prediction (s_{press}), SE, SP, A_{ROC}, coefficients and constant terms in the scoring function, are recorded.

7.3 Results and Discussion

7.3.1 DRB1*0402 predictive model

The DRB1*0402 (r^2=0.90, s=1.20 kJ/mol, q^2=0.82, s_{press}=1.61 kJ/mol) model shows excellent predictivity. The accuracy of the prediction model was further evaluated using (i) peptides with experimental IC_{50} values obtained from biochemical studies and (ii) Dsg3 peptides with T cell proliferation values obtained from functional studies.

Three threshold binding energy values (Table 16) that define levels of specificities suitable for practical applications (Tong *et al.*, 2006b) were used to evaluate the accuracy of the DRB1*0402 prediction model on the biochemical dataset (Test set 1) – LMH (low-, medium-, high-affinity binders; A_{ROC}=0.93); MH (medium- and high-affinity binders; A_{ROC}=0.86) and H (high-affinity binders only; A_{ROC}=0.81). The results indicate that, overall, our DRB1*0402 peptide-binding models are highly accurate (A_{ROC}≥0.81). SP=0.80 offers high-sensitivity predictions, whereas SP=0.95 results in very few false positives but fewer true positives. The prediction results for our model were consistent with expected binding patterns of DRB1*0402 peptides and provided a sensitivity of 70% (SP=0.95) for DRB1*0402-binding peptides.

Next, the predictive performance of the DRB1*0402 model was tested on the functional dataset of 18 peptides (Test set 2) with experimental T cell proliferation values/responses using the decision thresholds defined above. All experimental positives (Table 17) are predicted with binding energy values of -26.94 kJ/mol or less (SE = 0.70, SP = 0.95). Dsg3 512-526 (ranked #7) and Dsg3 78-93 (ranked #9) are predicted binders at threshold -26.94 kJ/mol (SE = 0.70, SP = 0.95) that did not stimulate T cell responses in the relevant experiments (Veldman *et al*., 2004; Wucherpfennig *et al*., 1995). A noteworthy observation is that peptide Dsg3 78-93 (Wucherpfennig *et al*., 1995) is fully contained within Dsg3 78-94 (ranked #8), an experimental true positive identified in an independent study (Veldman *et al*., 2004). It is possible that these peptides are binders that led to clonal deletion or anergy of the peptide-specific T cells or are false positives in the prediction.

Table 16 Sensitivity values and binding energy thresholds for DRB1*0402 peptide-binding model at specificity levels 0.80, 0.90 and 0.95.

Specificity (SP) Level	Group	Sensitivity (SE)	Binding Energy Threshold (kJ/mol)
SP = 0.80	LMH	0.78	-25.55
	MH	0.81	-25.79
	H	0.65	-26.64
SP = 0.90	LMH	0.75	-25.79
	MH	0.52	-26.94
	H	0.30	-28.83
SP = 0.95	LMH	0.70	-26.94
	MH	0.42	-27.72
	H	0.25	-30.57

Table 17 Predicted Dsg3 peptide sequences to DRB1*0402. The top 5 predictions are experimentally positive. ‘.’ indicates non-immunostimulatory in the relevant experiments.

Rank	Peptide	Sequence	Predicted BE (kJ/mol)	References
1	Dsg3 342-358	SVKLSIAVKNKAEFHQS	-31.46	Veldman *et al*. (2004)
2	Dsg3 376-392	NVREGIAFRPASKTFTV	-30.47	Veldman *et al*. (2004)
3	Dsg3 205-221	GTPMFLLSRNTGEVRTL	-30.44	Veldman *et al*. (2004)
4	Dsg3 380-396	GIAFRPASKTFTVQKGI	-29.97	Riechers *et al*. (1999)
5	Dsg3 190-204	LNSKIAFKIVSQEPA	-29.74	Wucherpfennig *et al*. (1995)
6	Dsg3 189-205	HLNSKIAFKIVSQEPAG	-29.24	Veldman *et al*. (2004)
7*	Dsg3 512-526	SARTLNNRYTGPYTF	-29.21	Wucherpfennig *et al*. (1995)
8	Dsg3 78-94	QATQKITYRISGVGIDQ	-28.37	Wucherpfennig *et al*. (1995)
9*	Dsg3 78-93	QATQKITYRISGVGID	-28.30	Veldman *et al*. (2004)
10	Dsg3 206-220	TPMFLLSRNTGEVRT	-28.02	Wucherpfennig *et al*. (1995)
11	Dsg3 210-226	LLSRNTGEVRTLTNSL	-27.98	Veldman *et al*. (2004)
12	Dsg3 251-265	CECNIKVKDVNDNFP	-27.88	Wucherpfennig *et al*. (1995)
13	Dsg3 250-266	QCECNIKVKDVNDNFPM	-27.68	Veldman *et al*. (2004)
14	Dsg3 483-499	VRVPDFNDNCPTAVLEK	-27.48	Veldman *et al*. (2004)
15	Dsg3 762-776	QSGTMRTRHSTGGTN	-27.28	Wucherpfennig *et al*. (1995)
16	Dsg3 161-177	IFMGEIEENSASNSLVM	-27.09	Hertl *et al*. (1998)
17	Dsg3 96-112	PFGIFVVDKNTGDINIT	-27.09	Veldman *et al*. (2004)
18*	Dsg3 97-111	FGIFVVDKNTGDINI	-26.58	Wucherpfennig *et al*. (1995)

7.3.2 DQB1*0503 predictive model

High predictivity is achieved for the DQB1*0503 (r^2=0.95, s=1.20 kJ/mol) prediction model. The DQB1*0503 model outperforms the prediction models undertaken by Rognan *et al*. (1999) on training datasets of 5 HLA-A*0204 (r^2=0.85, s_{press}=2.40 kJ/mol) and 37 HLA-2K^k (r^2=0.78, s_{press}=3.16 kJ/mol) peptide sequences. The cross-validation coefficient q^2 and the standard error of prediction s_{press} are stable, with q^2=0.75 and s_{press}=2.15 kJ/mol. This iterative regression procedure validates the internal consistency of the scoring function in the current model, rendering it suitable for predictions on the test dataset obtained from functional studies.

The accuracy of the DQB1*0503 prediction model was assessed on a dataset of 5 (4 stimulatory and 1 non-stimulatory peptides) Dsg3 peptides with known T cell proliferation values (Table 3). All DQB1*0503-specific Dsg3 stimulatory peptides can be effectively discriminated from the background at the prediction threshold -26.65 kJ/mol.

Table 18 Predicted Dsg3 peptide sequences to DQB1*0503. The top 4 predictions are experimentally positive. ‘.’ indicates non-immunostimulatory in the relevant experiments.

Rank	Peptide	Sequence	Predicted BE (kJ/mol)	References
1	Dsg3 206-220	`TPMFLLSRNTGEVRT`	-30.53	Wucherpfennig *et al*. (1995)
2	Dsg3 189-205	`HLNSKIAFKIVSQEPAG`	-29.10	Wucherpfennig *et al*. (1995)
3	Dsg3 190-204	`LNSKIAFKIVSQEPA`	-26.88	Wucherpfennig *et al*. (1995)
4	Dsg3 251-265	`CECNIKVKDVNDNFP`	-26.65	Wucherpfennig *et al*. (1995)
5∗	Dsg3 762-786	`QSGTMRTRHSTGGTNKDYADGAISM`	-22.42	Wucherpfennig *et al*. (1995)

7.3.3 Disease Progression in PV

A variety of studies have demonstrated that a limited set of epitopes may be present in early disease, and intra-molecular epitope spreading may occur during disease transition at the B-cell level (Salato *et al*., 2005). Our data support the existence of multiple immunodominant T cell epitopes that may be responsible for both disease initiation and propagation (refer Figures 23 and 24). These findings are in line with T cell proliferation data obtained from DR4 and DR6 PV patients (Sinha *et al*., 2006; Chow *et al*., 2006; Tong *et al*., 2006a; Wucherpfennig et al., 1995). Our analysis showed that the potential Dsg3 T cell epitope repertoire is evenly distributed throughout all 5 extracellular domains (ECDs) as well as the majority of the transmembrane region (Figures 23 and 24).

Many of these epitopes may not be generated via antigen processing events, or are subdominant or “cryptic”, which are not recognized during the initial immune response. It is possible that immune responses may be developed against these secondary epitopes at a later stage as a result of intracellular epitope spreading.

7.3.4 DRB1*0402 and DQB1*0503 cross reactivity

An in-depth analysis was performed to investigate the extent of overlap in the Dsg3 peptide-binding repertoires of DRB1*0402 and DQB1*0503. A panel of 936 15mer Dsg3 sequences were generated using an overlapping sliding window of size 15 across the entire Dsg3 glycoprotein and modeled into the binding grooves of both DRB1*0402 and DQB1*0503 (refer *Peptide Docking*).

Both the DRB1*0402 and DQB1*0503 alleles are particularly efficient in binding Dsg3-derived peptides. Furthermore, a significant level of cross-reactivity was observed between DRB1*0402 and DQB1*0503. Of the 936 overlapping 15mer peptides derived from the entire Dsg3 glycoprotein investigated in this study, 539 (57%) were predicted high-affinity binders to both alleles at threshold -26.64 kJ/mol. The computer simulation results are shown in Figures 23 and 24. It was noteworthy that three previously defined immunoreactive segments of the Dsg3 extracellular domains (ECD) (Dsg3 145-192, 240-303, and 570-614) (Lin *et al*., 1997) were also predicted by our models at this specific threshold. These observations are of particular interest in that both DRB1*0402 and DQB1*0503 are strongly linked to PV (Lee *et al*., 2006), indicating that common or overlapping dominant epitopes may be responsible for inducing disease in DR4 and DR6 patients respectively. The inability of some Dsg3 peptides to be recognized by autoreactive cells may be at the level of T cell recognition rather than the level of epitope selection by MHC molecules.

7.3.5 DRB1*0402 and DQB1*0503 peptide binding specificities

The basis for the high degree of cross-reactivities between DRB1*0402 and DQB1*0503 was subjected to further analysis. Our data support the existence of multiple binding registers within a candidate binding peptide that serve as recognition sites for DRB1*0402 and DQB1*0503. Of 936 Dsg3-derived peptides, 614 were predicted high-affinity binders with 76% displaying 2 or more registers that can be docked into the binding groove of DRB1*0402 (Figure 4). Similar results are obtained for DQB1*0503, with 673 predicted high-affinity binders and 57% exhibiting 2 or more binding registers (Figure 26). DRB1*0402 and DQB1*0503 predicted consensus binding sequences number 539. A striking aspect of this analysis is that DRB1*0402 and DQB1*0503 were predicted to bind a large portion of these peptides (354/539 or 66%) at different binding registers. For example, the consensus binding peptide Dsg 205-221 showed ΔG values less than the decision threshold -26.64 kJ/mol, with the 209-217 and 207-215 registers being the preferred binding modes for DRB1*0402 and DQB1*0503 respectively. We propose that DRB1*0402 and DQB1*0503 share similar specificities by binding peptides at different binding registers.

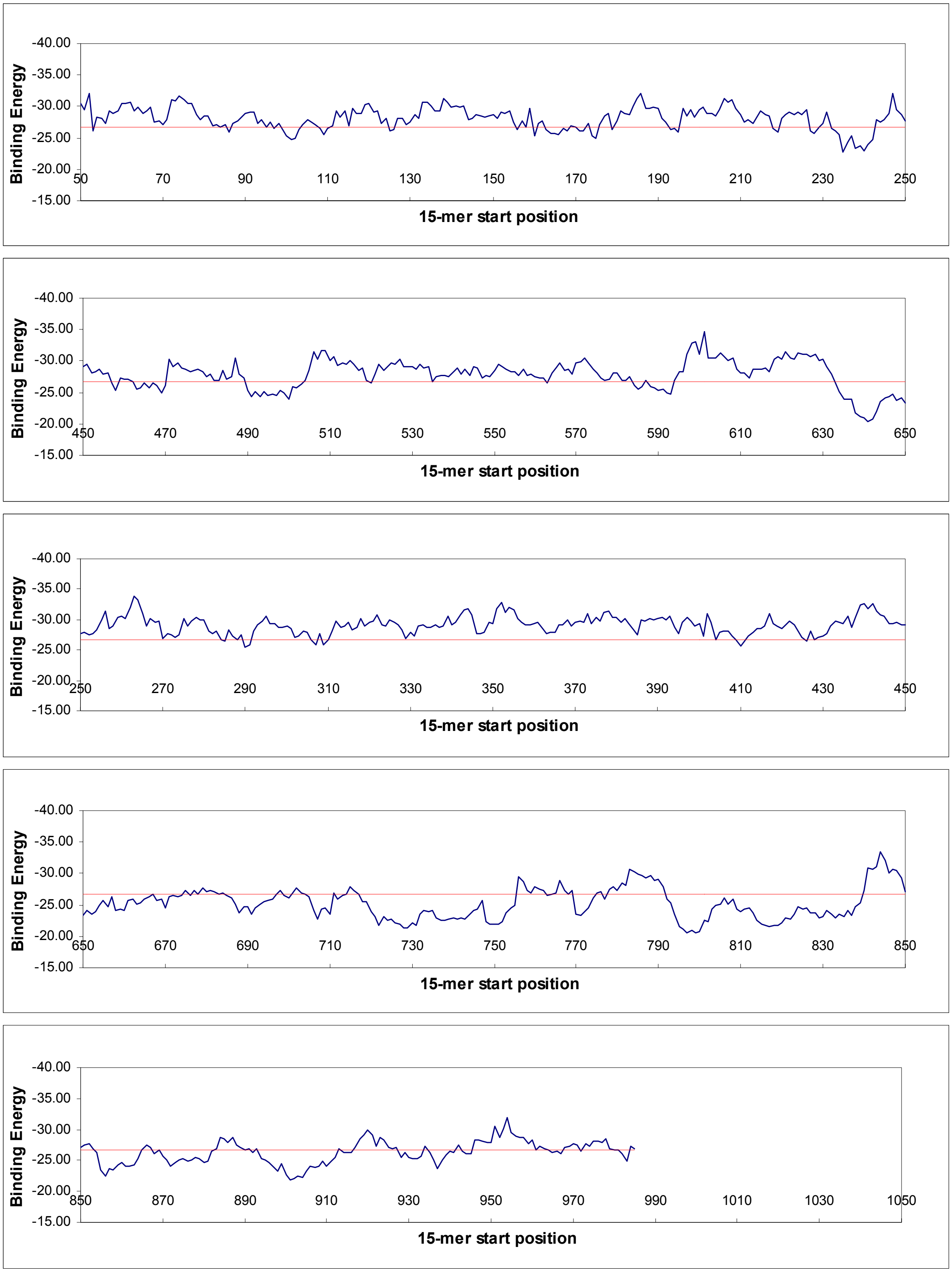


Figure 24 Predicted DRB1*0402-specific T cell epitope repertoire within Dsg3 glycoprotein at threshold of -26.64 kJ/mol (red line).

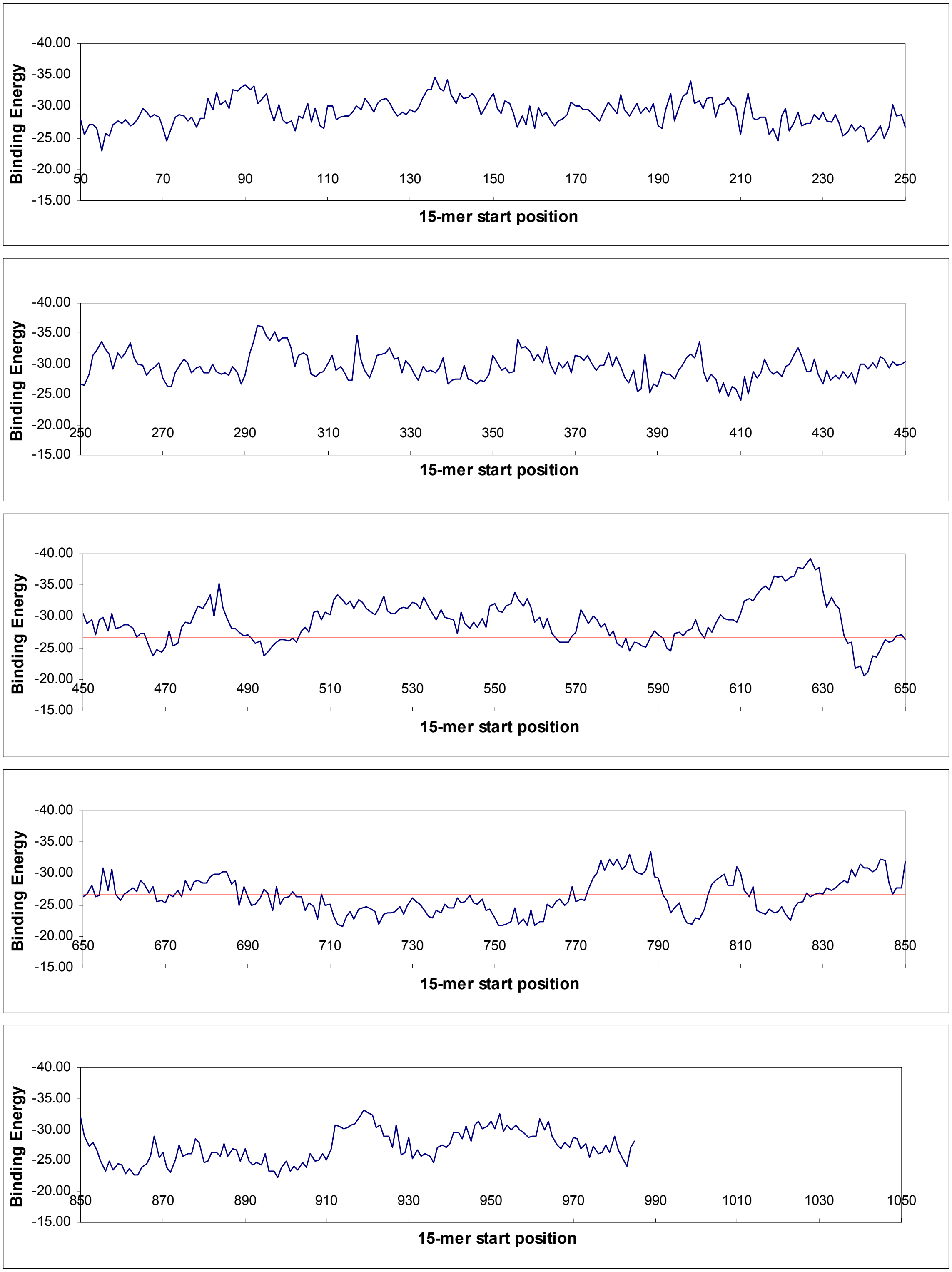


Figure 25 Predicted DQB1*0503-specific T cell epitope repertoire within Dsg3 glycoprotein at threshold of -26.64 kJ/mol (red line).

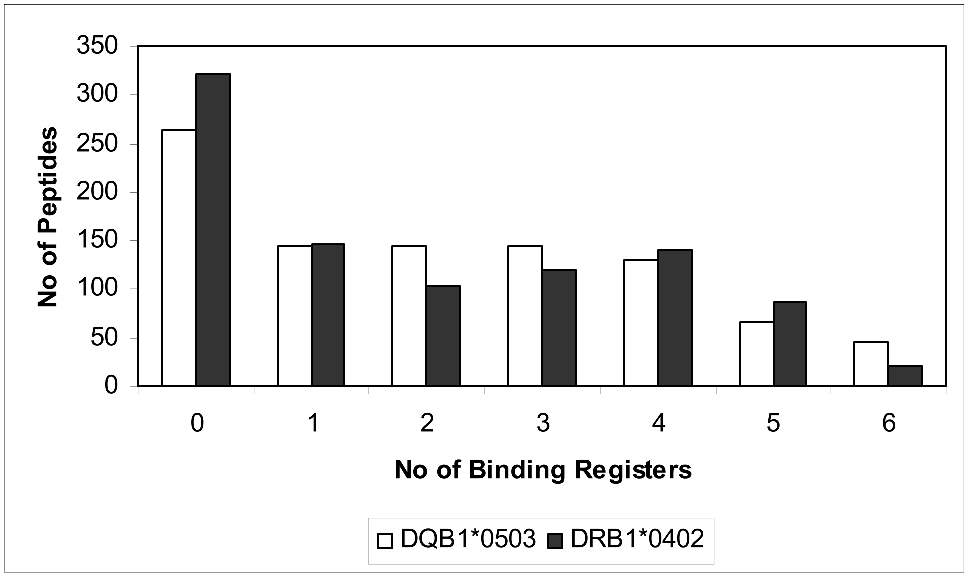


Figure 26 Number of predicted binding registers within Dsg3 peptides to DRB1*0402 and DQB1*0503.

7.5 Summary

- In this work, a scoring function has been developed as an extension of the docking protocol developed in Chapter 6 for functional prediction of peptides binding to DRB1*0402 and DQB1*0503.
- High accuracy of predictions was obtained by validation with experimental binding and non-binding peptide sequences.
- High degree of cross-reactivities between DRB1*0402 and DQB1*0503 was obtained. Of the 936 15mer peptides generated from the entire Dsg3 glycoprotein, 539 (57%) were predicted high-affinity binders to both alleles at threshold -26.64 kJ/mol. The results also indicate that DRB1*0402 and

DQB1*0503 share similar specificities by binding peptides at different binding registers.

- The present analysis reveals the possible existence of multiple shared immunodominant epitopes within the Dsg3 glycoprotein, suggesting that no single epitope is responsible for both disease initiation and propagation. The inability of some Dsg3 peptides to be recognized by autoreactive cells may be at the level of T cell recognition rather than the level of epitope selection by MHC molecules.
- These observations are of particular interest in that both DRB1*0402 and DQB1*0503 are strongly linked to PV, indicating that common or overlapping dominant epitopes may be responsible for inducing disease in DR4 and DR6 patients respectively. The inability of some Dsg3 peptides to be recognized by autoreactive cells may be at the level of T cell recognition rather than the level of epitope selection by MHC molecules.

Chapter 8: Conclusion

This work began with an investigation of the use of bioinformatic-based approaches for the study of peptide/MHC interactions. Through the systematic application of bioinformatic resources and tools, the author has delved into the interaction characteristics of peptide/MHC complexes and its significance in the pathology of the autoimmune disorder PV. By the completion of this project, the author has advanced the knowledge of the use of bioinformatics in immunological research. The author now summarizes his conclusions.

8.1 Database of (TCR/) peptide/MHC interaction parameters

The interaction of peptide/MHC complexes with TCRs on the surface of T cells is responsible for T cell activation and stimulation of the adaptive immune response. In this context, in-depth understanding of the structural principles involved in the selection of specific antigenic peptides by different MHC alleles and subsequent recognition by TCRs is an important step towards effective development of epitope-based vaccines.

Here, the author has described the development of a new database termed MHC-Peptide Interaction Database version T (MPID-T) to facilitate the analysis of TCR/peptide/MHC and peptide/MHC structural interaction characteristics. With a large repository of computed (TCR/) peptide/MHC interaction parameters, a high-level perspective of the general interaction patterns can be inferred. This present database relates to current knowledge of the field and with the availability of more information in the future, it will provide a better understanding of the structural principles involved in the selection of antigenic peptides by the different MHC alleles. This generic classification approach may also be applied to the analysis of other families of ligand/receptor complexes.

8.2 Analysis of MHC supertype interaction characteristics

With the rapid growth of immunological data, there is a need to develop computational strategies for the classification of MHC alleles into supertypes to support research in the development of new generation peptide vaccines with wide population coverage (Doytchinova *et al.*, 2004b, 2005). The analysis of MHC supertype interaction characteristics could draw more accurate inferences about MHC binding specificities and enhance our understanding of the relationship among different MHC alleles.

This work reports the first analysis of HLA supertype interaction characteristics using four interaction parameters (interface area, intermolecular hydrogen bonds, gap volume, and gap index). The present analysis is difficult due to the limited number of peptide/HLA crystallographic structures in the current PDB. Nonetheless, we have demonstrated that different HLA alleles employ the use of different binding mechanism for selectivity of antigenic peptides in a supertype dependent manner. By focusing solely on the use of experimental three-dimensional structures, our analysis is supported and verified by existing data. *In silico* analysis of HLA supertype interaction characteristics opens the way for more in-depth understanding of the binding mechanism involved in peptide selection and better characterization of HLA supertypes. This systematic approach for analyzing receptor/ligand interactions can serve as a model for other MHC supertypes and receptor/ligand systems where 3D information is available.

8.3 Development of generic peptide/MHC docking protocol

In recent years, bioinformatic tools modeling the immune system network are playing an increasingly important role in advancing epitope-based vaccine research. At the present time, experimental data for the majority of MHC alleles do not exist and there is a great need for computational strategies that are not constrained by the limited availability of large training datasets.

This work reports on the development of an efficient and fast docking protocol for modeling the bound conformation of peptide ligands to MHC class I and class II molecules without the need for large training datasets. High prediction accuracy was obtained in three independent experiments: (i) self-docking 40 test case complexes; (ii) cross-docking of 15 solved peptides into the templates of appropriate alleles; and (iii) validation against existing techniques.

8.4 Analysis of PV related and non-related alleles

PV is a severe autoimmune blistering skin disorder due to loss of integrity of normal intercellular attachments within the epidermis and mucosal epithelium. To date, much remains unknown with regards to the functional correlation between MHC class II alleles and PV.

In this work, the author describes the structural analysis of ten PV associated, non-associated and protective MHC class II receptors. Nine previously identified epitopes capable of stimulating patient derived T cells, were docked into the binding groove of each model to analyze the structural aspects of allele-specific binding. This study has addressed three important issues with regards to the pathology of PV: (i) DRB1*0402 and DQB1*0503 have different binding specificities and play a crucial role in DR4 and DR6 PV respectively; (ii) DQB1*0201 and *0202 play a protective role by binding Dsg3 peptides with greater affinity than the susceptible alleles, facilitating efficient deletion of autoreactive T cells; and (iii) no single epitope may be responsible for both disease initiation and propagation in PV.

8.5 Development of functional prediction technique

A new approach for predicting the binding affinities of MHC-binding peptides (Chapter 6) was developed to complement to the peptide/MHC docking protocol (Chapter 4). This approach has been successfully applied to screen peptide binders which lack conserved binding motifs. A systematic attempt to analyze MHC class II binding and non-binding peptides reveals the possibility of multiple registers within a candidate class II binding peptide. These results suggest that recognition via flexible fitting may play a critical role in binding to class II alleles. It is possible that the existence of multiple registers within a candidate peptide can facilitate binding to a particular allele, with the highest binding affinity register acting as the primary recognition region.

8.6 Analysis of T cell epitope repertoire in Dsg3

Knowledge of the nature of peptide binding to PV-implicated class II alleles has advanced rapidly in the last years through experimental (Wucherpfennig *et al*., 1995; Sinha *et al*., 1988, 1990; 2006; Veldman *et al*., 2004) and computational studies (Tong *et al*., 2006a). Because of the paucity of experimental peptide binding data, screening of PV epitopes has been based primarily on sequence motifs (Wucherpfennig *et al*., 1995; Veldman *et al*., 2004). However, Sinha *et al*. (unpublished results) recently discovered that DRB1*0402-specific binding motifs are insufficient for the design of PV epitopes, due to the presence of register shifts as well as polymorphisms in the binding register. With increasing evidence suggesting the inadequacy of binding motifs in defining class II T cell epitopes, the current approach of predictive model building and virtual screening for T cell epitope candidates is independent of sequence motifs with excellent predictivity trained using a limited dataset.

The first report of a high degree of cross-reactivity between DRB1*0402 and DQB1*0503 provides new insights into the pathology of PV, suggesting the possible existence of multiple shared immunodominant epitopes within the Dsg3 glycoprotein. Both disease-implicated alleles DRB1*0402 and DQB1*0503 share similar specificities by binding peptides at different core recognition regions. These observations are of particular interest in that both DRB1*0402 and

DQB1*0503 are strongly linked to PV (Lee *et al*., 2006), indicating that common or overlapping dominant epitopes may be responsible for inducing disease in DR4 and DR6 patients respectively. The inability of some Dsg3 peptides to be recognized by autoreactive cells may be at the level of T cell recognition rather than the level of epitope selection by MHC molecules as a result of clonal deletion or anergic response.

8.7 Overall conclusions

This thesis reports pioneering work in the field of immunoinformatics through the use of 3D structural models. In summary, the following conclusions can be drawn:

- The extremely high polymorphism of HLA alleles (Williams, 2001) has been a confounding factor in the study of HLA peptide binding specificities. Sequence-structure-function information is critical in understanding the principles governing peptide/MHC recognition and binding. In this context, the author introduced the use of structural interaction information to analyze high-level relationships hidden within peptide/HLA crystallographic structures and demonstrated the existence of different interaction characteristics among different MHC supertypes (Chapter 3). The result of this analysis paves the way for more accurate inferences about MHC binding specificities.
- Through the systematic application of structural models for the analysis of PV-implicated and non-implicated alleles (Chapter 5), the author has

addressed three important issues with regards to the pathology of PV: (i) DRB1*0402 and DQB1*0503 have different binding specificities and play a crucial role in DR4 and DR6 PV respectively; (ii) DQB1*0201 and *0202 play a protective role by binding Dsg3 peptides with greater affinity than the susceptible alleles, facilitating efficient deletion of autoreactive T cells; and (iii) no single epitope may be responsible for both disease initiation and propagation in PV.

- Through systematic analysis of MHC class II binding and non-binding peptides, the author addressed the issue of degeneracy in peptide binding to MHC class II molecules by providing evidence on the possible existence of multiple registers within a candidate class II binding peptide (Chapter 6). This discovery provides new insights into the binding specificities of class II alleles, suggesting that recognition via flexible fitting may play a critical role in binding to class II alleles. Whilst not all binding registers may elicit T cell response, it is possible that the existence of multiple registers within a candidate peptide (especially for high-affinity binders) may facilitate binding to a particular allele, with the highest binding affinity register acting as the primary recognition region.
- High level of cross-reactivities were detected between DRB1*0402 and DQB1*0503, suggesting the possible existence of multiple shared immunodominant epitopes within the Dsg3 self-antigen. It was noteworthy that both disease-implicated alleles share similar

specificities by binding peptides at different core recognition regions. These observations indicate that common or overlapping dominant epitopes may be responsible for inducing disease in DR4 and DR6 patients respectively. The inability of some Dsg3 peptides to be recognized by autoreactive cells may be at the level of T cell recognition rather than the level of epitope selection by MHC molecules as a result of clonal deletion or anergic response.

8.8 Future directions

The work done in Chapter 3 paves the way for further development that will facilitate the extraction of high-level relationships hidden within TCR/peptide/MHC interaction data by mapping the TCR footprint on the MHC and its bound peptide. Preliminary work in this area has begun with the computation of interaction parameters for 16 TCR/peptide/MHC complexes. Future developments will also include computed data on additional structural parameters characterizing the TCR/peptide/MHC and peptide/MHC interaction region.

The analysis covered in Chapter 3 has revealed the striking observation that peptide/HLA structural interaction patterns vary among different alleles and may be grouped in a supertype dependent manner. In general, the interaction patterns (gap index, gap volume, interface area, and the number of intermolecular hydrogen bonds) of peptide/MHC complexes are conserved at the supertype level but not across different superfamilies. While some studies showed excellent results when applied to specific sets of alleles, the results presented here suggest that the use of a standardized set of structural interaction rules or free energy scoring functions to discriminate binding peptides may not be applicable for all MHC alleles as interaction characteristics vary across MHC supertypes.

Although the current methodology focuses on the use of existing crystallographic data for analysis, the work may be extended to theoretical models for alleles without experimental structures. Such analysis will prove useful as the majority of MHC alleles have not been crystallized and much remains unknown with regards to the binding mechanisms underlying peptide/MHC interactions. The classification of MHC alleles into supertypes may be formulated according to peptide/MHC interaction characteristics and serve as an alternative to HLA supertype analysis using either sequence or receptor structure information alone. This will allow the finer selection of representative molecules that can effectively cover the HLA specificity space.

The docking protocol covered in Chapters 4 to 7 provides a rapid way for accurate modeling of peptides binding to MHC. While water molecules and other common biological ions such as phosphate and chloride may mediate peptide/MHC interactions, they were left out in the current experiments due to complexities in modeling and contributions of these molecules vary between different peptides and the respective alleles. While the lack of placement of water molecules at the peptide/MHC interface has been used to account for errors in many predictions (Rognan *et al*., 1999), a recent study that incorporate water molecules in docking simulations to predict the bound conformation of peptides binding to an array of MHC class I molecules led to poor binding prediction with average RMSD of modeled peptides between 1.50 to 2.47 Å from the original crystallographic structures (Bui *et al*., 2006). Future developments will explore strategies for accurate mapping of conserved water molecules and other ions into the MHC binding groove.

The analysis covered in Chapters 5 and 7 serve as an essential preliminary step towards better understanding of the pathology of PV by focusing on its main (and sometimes sole) antigen Dsg3. A similar approach may be applied for the analysis of Dsg1 which may occur at the later stage of disease. This will also provide valuable insights into disease pathology and facilitates the fine profiling of the PV-associated T cell epitope repertoire in disease-implicated alleles.

Bibiliography

1. Abagyan, R., Totrov, M. and Kuznetsov, D. (1994) ICM – a new method for protein modeling and design: Applications to docking and structure prediction from the distorted native conformation. *J. Comp. Chem.* **15**: 488-506.
2. Abagyan, R. and Maxim, T. (1999) Ab initio folding of peptides by the optimal-bias Monte Carlo minimization procedure. *J. Comput. Phys.* **151**: 402-421.
3. Abagyan, R. and Totrov, M. (2001) High-throughput docking for lead generation. *Curr. Opin. Chem. Biol.* **5**: 375-382.
4. Adams, H. P. and Koziol, J. A. (1995) Prediction of binding to MHC class I molecules. *J. Immunol. Methods* **185**: 181-190.
5. Ahmed, A. R., Yunis, E. J., Khatri, K., Wagner, R., Notani, G., Awdeh, Z. and Alper, C. A. (1990) Major histocompatibility complex haplotype studies in Ashkenazi Jewish patients with pemphigus vulgaris. *Proc. Natl. Acad. Sci. USA* **87**: 7658-7662.
6. Ahmed, A. R., Wagner, R., Khatri, K., Notani, G., Awdeh, Z., Alper, C. A. and Yunis, E. J. (1991) Major histocompatibility complex haplotypes and class II genes in non-Jewish patients with pemphigus vulgaris. *Proc. Natl. Acad. Sci. USA* **88**: 5056-5060.
7. Akutsu, T. and Sim, K. L. (1999) Protein threading based on multiple protein structure alignment. *Genome Inform.* **10**: 23-29

8. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. and Lipman, D. J. (1990) Basic local alignment search tool. *J. Mol. Biol.* **215**: 403-410.

9. Altuvia, Y., Schueler, O. and Margalit, H. (1995) Ranking potential binding peptides to MHC molecules by a computational threading approach. *J. Mol. Biol.* **249**: 244-250.

10. 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**: 1-11.

11. Amagai, M. (1994) Autoantibodies against cell adhesion molecules in pemphigus. *J. Dermatol.* **21**: 833-837.

12. Androlewicz, M. J., Ortmann, B., van Endert, P., Spies, T. and Cresswell, P. (1994) Characteristics of peptide and major histocompatibility complex class I/ß2-microglobulin binding to the transporters associated with antigen processing (TAP1 and TAP2). *Proc. Natl. Acad. Sci. USA*. **91**: 12716-12720.

13. Antes, I., Siu, S. W. and Lengauer, T. (2006) DynaPred: A structure and sequence based method for the prediction of MHC class I binding peptide sequences and conformations. *Bioinformatics* **22**: e16-e24.

14. Angelini, G., Bonamonte, D., Lin, M. S., Lucchese, A., Mittelman, A., Serpico, R., Simone, S., Sinha, A. A. and Kanduc, D. (2005) Characterization of polyclonal antibodies raised against a linear peptide determinant of desmoglein-3. J*. Exp. Ther. Oncol.* **5**: 1-7.

15. Bairoch, A. and Apweiler, R. (1998) The SWISS-PROT protein sequence databank and its supplement TrEMBL in 1998. *Nucleic Acids Res.* **26**: 38-42.

16. Bairoch, A., Boeckmann, B., Ferro, S. and Gasteiger, E. (2004) Swiss-Prot: juggling between evolution and stability. *Brief Bioinform.* **5**: 39-55.

17. Bateman, A., Coin, L., Durbin, R., Finn, R. D., Hollich, V., Griffiths-Jones, S., Khanna, A., Marshall, M., Moxon, S., Sonnhammer, E. L., Studholme, D. J., Yeats, C. and Eddy, S. R. (2004) The Pfam protein families database. *Nucleic Acids Res.* **32**: D138-141.

18. Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. and Bourne, P. E. (2000) The Protein Data Bank. *Nucleic Acids Res*. **28**: 235-242.

19. Bernstein, F. C., Koetzle, T. F., Williams, G. J., Meyer, E. E., Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T. and Tasumi, M. (1977) The Protein Data Bank: A computer-based archival file for macromolecular structures. *J. Mol. Biol.* **112**: 535-542.

20. Betancourt, M. R. 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**: 361-369.

21. Bharadwaj, A., Windemuth, A., Sridharan, S., Honig, B. and Nicholls, A. (1995) The fast multipole boundary element method for molecular electrostatics: an optimal approach for large system. *J. Comput. Chem*. **16**: 898-910.

22. Bhasin, M., Singh, H. and Raghava, G. P. S. (2003) MHCBN: a comprehensive database of MHC binding and non-binding peptides. *Bioinformatics* **19**: 665-666.

23. Bhasin, M. and Raghava, G. P. S. (2004a) SVM based method for predicting HLA-DRB1*0401 binding peptides in an antigen sequence. *Bioinformatics* **20**: 421-423.

24. Bhasin, M. and Raghava, G. P. S. (2004b) Prediction of CTL epitopes using QM, SVM and ANN techniques. *Vaccine* **22**: 3195-3204.

25. Bielekova, B. and Martin, R. (2001) Antigen-specific immunomodulation via altered peptide ligands. *J. Mol. Med*. **79**: 552-565.

26. Blythe, M. J., Doytchinova, I. A. and Flower, D. R. (2002) JenPep: a database of quantitative functional peptide data for immunology. *Bioinformatics* **18**: 434-439.

27. Bock, J. R. and Gough, D. A. (2002) A new method to estimate ligand-receptor energetics. *Mol. Cell. Proteomics* **1**: 904-910.

28. Bodmer, J. G., Marsh, S. G. E., Albert, E. D. Bodmer, W. F., Bontrop, R. E., Dupont, B., Erlich, H. A., Hansen, J. A., Mach, B., Mayr, W. R., Parham, P., Petersdorf, E. W., Sasazuki, T., Schreuder, G. M., Strominger, J. L., Svejgaard, A. and Terasaki, P. I. (1999) Nomenclature for factors of the HLA system. *Tissue Antigens* **53**: 407-446.

29. Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M. C., Estreicher, A., Gasteiger, E., Martin, M. J., Michoud, K., O'Donovan, C., Phan, I., Pilbout

S. and Schneider, M. (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res*. **31**: 365-370.

30. Bourdette, D. N., Edmonds, E., Smith, C., Bowen, J. D., Guttmann, C. R., Nagy, Z. P., Simon, J., Whitham, R., Lovera, J., Yadav, V., Mass, M., Spencer, L., Culbertson, N., Bartholomew, R. M., Theofan, G., Milano, J., Offner, H. and Vandenbark, A. A. (2005) A highly immunogenic trivalent T cell receptor peptide vaccine for multiple sclerosis. *Mult*. *Scler.* **11**: 552-561.

31. Bower, M. J., Cohen, F. E. and Dunbrack, R. L., Jr. (1997) Prediction of protein side-chain rotamers from a backbone-dependent rotamer library: a new homology modeling tool. *J. Mol. Biol.* **267**: 1268-1282.

32. Bozic, I., Zhang, G. and Brusic, V. (2005) Predictive vaccinology: optimisation of predictions using support vector machine classifiers. *IDEAL 2005*: 375-381.

33. Brown, J. H., Jardetzky, T. S., Gorga, J. C., Stern, L. J., Urban, R. G., Strominger, J. L. and Wiley, D. C. (1993) Three-dimensional structure of the human class II histocompatibility antigen HLA-DR1. *Nature* **364**: 33-39.

34. Brusic, V., Rudy, G. and Harrison, L.C. (1994) Prediction of MHC binding peptides using artificial neural networks. *In Stonier, R. J. and Yu, X. S., (eds), Complex Systems: Mechanism of Adaptation. IOS Press, Amsterdam*: 253-260.

35. Brusic, V., Schonbach, C., Takiguchi, M., Ciesielski, V. and Harrison, L. C. (1997) Application of genetic search in derivation of matrix models of

peptide binding to MHC molecules. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **5**: 75-83.

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

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

38. Brusic, V., van Endert, P., Zeleznikow, J., Daniel, S., Hammer, J. and Petrovsky, N. (1999) A neural network model approach to the study of human TAP transporter. *In Silico Biol*. **1**: 109-121.

39. Brusic, V., Petrovsky, N., Zhang, G. and Bajic, V.B. (2002) Prediction of promiscuous peptides that bind HLA class I molecules. *Immunol. Cell Biol.*, **80**: 280-285.

40. Bui, H-.H., Schiewe, A. J., von Grafenstein, H., and Haworth, I. S. (2006) Structural prediction of peptides binding to MHC class I molecules. Proteins, **63**: 43-52.

41. Buus, S. (1999) Description and prediction of peptide-MHC binding: the 'human MHC project'. *Curr. Opin. Immunol*. **11**: 209-213.

42. Caflisch, A., Niederer, P. and Anliker, M. (1992) Monte Carlo docking of oligopeptides to proteins. *Proteins* **13**: 223-230.

43. Cavasotto, C. N., Orry, A. J. W. and Abagyan, R. (2003) Structure-based identification of binding sites, native ligands and potential inhibitors for G-protein coupled receptors. *Proteins*. **51**: 423-433.

44. Chen, W., Khilko, S., Fecondo, J., Margulies, D. H. 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**: 1471-1483.

45. Chicz, R. M., Urban, R. G., Gorga, J. C., Vignali, D. A., Lane, W. S. and Strominger, J. L. (1993) Specificity and promiscuity among naturally processed peptides bound to HLA-DR alleles. *J. Exp*. *Med*. **178**: 27-47.

46. Chothia, C. and Janin, J. (1975) Principles of protein-protein recognition. *Nature* **28**: 705-708.

47. Chow, S., Rizzo, C., Ravitskiy, L. and Sinha, A. A. (2005) The role of T cells in cutaneous autoimmune disease. *Autoimmunity* **38**: 303-317.

48. Collins, E. J., Garboczi, D. N., Karpusas, M. N. and Wiley, D. C. (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*. **92**: 1218-1221.

49. Corradin, G. and Demotz, S. (1997) Peptide-MHC complexes assembled following multiple pathways: an opportunity for the design of vaccines and therapeutic molecules. *Hum. Immunol*. **54**: 137-147.

50. Cresswell P. (1994) Assembly, transport, and function of MHC class II molecules. *Annu. Rev. Immunol*. **12**: 259-293.

51. Crooks, G. E., Hon, G., Chandonia, J. M. and Brenner, S. E. (2004) WebLogo: A sequence logo generator, *Genome Res*. **14**: 1188-1190.

52. D'Amaro, J., Houbiers, J. G. A., Drijfhout, J. W., Brandt, R. M. P., Schipper, R., Bavinck, J. N. B., Melief, C. J. M. and Kast, W. M. (1995) A computer program for predicting possible cytotoxic T lymphocyte epitopes based on HLA class I peptide-binding motifs. *Hum. Immunol.* **43**: 13-18.

53. Davenport, M. P., Ho Shon, I. A. P. and Hill, A. V. S. (1995) An empirical method for the prediction of T cell epitopes. *Immunogenetics* **42**: 392-397.

54. Deng, Y., Yewdell, J. W., Eisenlohr, L. C. and Bennink, J.R. (1997) MHC affinity, peptide liberation, T cell repertoire, and immunodominance all contribute to the paucity of MHC class I-restricted peptides recognized by antiviral CTL. *J. Immunol*., **158**: 1507-1515.

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

56. Doytchinova, I. A. and Flower, D. R. (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**: 3572-3581.

57. Doytchinova, I. A., Blythe, M. J. and Flower, D. R. (2002) Additive method for the prediction of protein-peptide binding affinity. Application to the MHC Class 1 molecule HLA-A*0201. *J. Proteome Res*. **1**: 263-272.

58. Doytchinova, I. A. and Flower, D. R. (2003) Towards the *in silico* identification of class II restricted T cell epitopes: a partial least squares iterative self-consistent algorithm for affinity prediction. *Bioinformatics* **19**: 2263-2270.

59. Doytchinova, I. A., Guan, P. and Flower, D. R. (2004a) Quantitative structure-activity relationships and the prediction of MHC supermotifs. *Methods* **34**: 444-453.

60. Doytchinova, I. A., Guan, P. and Flower, D. R. (2004b) Identifying human MHC supertypes using bioinformatic methods. *J. Immunol.* **172**: 4314-4323.

61. Doytchinova, I. A., Walshe, V., Borrow, P. and Flower, D. R. (2005) Towards the chemometric dissection of peptide-HLA-A*0201 binding affinity: comparison of local and global QSAR models. *J. Comput. Aided. Mol*. Des. **19**: 203-212.

62. Doytchinova, I. A. and Flower, D. R. (2005) *In silico* identification of supertypes for class II MHCs. *J. Immunol.* **174**: 7085-7095.

63. Duda, R. O., Hart, P. E. and Stork, D. G. (2001) *Pattern Classification. Wiley-Interscience*.

64. Durbin, R., Eddy, S., Krogh, A. and Mitchison, G. (1998) Biological Sequence Analysis: *Probabilistic models of proteins and nucleic acids. Cambridge University Press*: 51-68.

65. Erlich, H. A., Zeidler, A., Chang, J., Shaw, S., Raffel, L. J., Klitz, W., Beshkov, Y., Costin, G., Pressman, S., Bugawan, T. and Rotter, J. I. (1993) HLA class II alleles and susceptibility and resistance to insulin dependent diabetes mellitus in Mexican-American families. *Nat. Genet*. **3**: 358-364.

66. Evavold, B. D., Sloan-Lancaster, J. and Allen, P. M. (1993) Tickling the TCR: selective T cell functions stimulated by altered peptide ligands. *Immunol Today* **14**: 602-609.

67. Falk, K., Rötzschke, O., Deres, K., Metzger, J., Jung, G. and Rammensee, H. G. (1991a) Identification of naturally processed viral nonapeptides allows their quantification in infected cells and suggests an allele-specific T cell epitope forecast. *J. Exp. Med*. **174**: 425-434.

68. Falk, K., Rötzschke, O., Stevanovic, S., Jung, G. and Rammensee, H. G. (1991b) Allele-specific motifs revealed by sequencing of self-peptides eluted from MHC molecules. *Nature* **351**: 290-296.

69. Faustman, D., Li, X. P., Lin, H. Y., Fu, Y. E., Eisenbarth, G., Avruch, J. and Guo, J. (1991) Linkage of faulty major histocompatibility complex class I to autoimmune diabetes. *Science* **254**: 1756-1761.

70. Feltkamp M. C., Vierboom M. P., Kast W. M. and Melief C. J. (1994) Efficient MHC class I-peptide binding is required but does not ensure MHC class I-restricted immunogenicity. *Mol. Immunol*. **31**: 1391-401.

71. Fernández-Recio, J., Totrov, M., and Abagyan, R. (2002) Soft protein-protein docking in internal coordinates. *Protein Sci.* **11**: 280-291.

72. Ferrari, G., Kostyu, D. D., Cox, J., Dawson, D. V., Flores, J., Weinhold, K. J. and Osmanov, S. (2000) Identification of highly conserved and broadly cross-reactive HIV type 1 cytotoxic T lymphocyte epitopes as candidate immunogens for inclusion in Mycobacterium bovis BCG-vectored HIV vaccines. *AIDS Res. Hum*. *Retroviruses* **16**: 1433-1443.

73. Fickel, S. and del Carpio, C. A. (2000) A QSAR study for modeling MHC class-I binding oligo-peptides. *Genome Informatics* **11**: 436-437.

74. Fischer, K. F. and Marquesee, S. (2000) A rapid test for identification of autonomous folding units in proteins. *J. Mol. Biol.* **302**: 701-712.

75. Flynn, J.C., McCormick, D. J., Brusic, V., Wan, Q., Panos, J. C., Giraldo, A. A., David, C. S. and Kong, Y. C. (2004) Pathogenic human thyroglobulin peptides in HLA-DR3 transgenic mouse model of autoimmune thyroiditis. *Cell. Immunol.*, **229**, 79-85.

76. Froloff, N., Windemuth, A. and Honig, B. (1997) On the calculation of binding free energies using continuum methods: application to MHC class I protein-peptide interactions. *Protein Sci*. **6**: 1293-1301.

77. Garrett T. P. J., Saper M. A., Bjorkman P. J., Strominger J. L. and Wiley D. C. (1989) Specificity pockets for the side chains of peptide antigens in HLA-Aw68. *Nature* **342**: 692-696.

78. Gebe, J. A., Swanson, E. and Kwok, W. W. (2002) HLA class II peptide-binding and autoimmunity. *Tissue Antigens* **59**: 78-87.

79. Godkin, A. J., Davenport, M. P., Willis, A., Jewell, D. P. and Hill, A. V. (1998) Use of complete eluted peptide sequence data from HLA-DR and -DQ molecules to predict T cell epitopes, and the influence of the nonbinding terminal regions of ligands in epitope selection. *J. Immunol*., **161**: 850-858.

80. Gorer P. A. J. (1937) *Pathol. Bacteriol*. **44**: 691.

81. Govindarajan, K. R., Kangueane, P., Tan, T. W. And Ranganathan, S. (2003) MPID: MHC-Peptide Interaction Database for sequence-structure-

function information on peptides binding to MHC molecules. *Bioinformatics* **19**: 309-310.

82. Guan, P., Doytchinova, I. A., Zygouri, C. and Flower, D. R. (2003) MHCPred: bringing a quantitative dimension to the online prediction of MHC binding. *Appl. Bioinformatics* **2**: 63-66.

83. Guan, P., Doytchinova, I.A. and Flower, D.R. (2003) HLA-A3 supermotif defined by quantitative structure–activity relationship analysis. *Protein Eng.* **16**: 11-18.

84. Guan, P., Doytchinova, I. A. and Flower, D. R. (2003) A comparative molecular similarity indices (CoMSIA) study of peptide binding to the HLA-A3 superfamily. *Bioorg. Med. Chem*. **11**: 2307-2311.

85. 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**: 1258-1267.

86. Guo H. C., Jardetzky T. S., Garrett T. P., Lane W. S., Strominger J. L. and Wiley D. C. (1992) Different length peptides bind to HLA-Aw68 similarly at their ends but bulge out in the middle. *Nature* **360**: 364-366.

87. Hammer, G. E., Gonzalez, F., Champsaur, M., Cado, D. and Shastri, N. (2005) The aminopeptidase ERAAP shapes the peptide repertoire displayed by major histocompatibility complex class I molecules. *Nature Immunol*. **7**: 103-112.

88. Hammer, J., Bono, E., Gallazzi, F., Belunis, C., Nagy, Z. and Sinigaglia, F. (1994) Precise prediction of major histocompatibility complex class II-

peptide interaction based on peptide side chain scanning. *J. Exp. Med.* **180**: 2353-2358.

89. Hammer, J., Gallazzi, F., Bono, E., Karr, R. W., Guenot, J., Valsasnini, P., Nagy, Z. A. and Sinigaglia, F. (1995) Peptide binding specificity of HLA-DR4 molecules: correlation with rheumatoid arthritis association. *J. Exp. Med.* **181**: 1847-1855.

90. Hammerstrom, D. (1993) Neural networks at work. *IEEE Spectrum* **30**: 26-32.

91. Han, L. Y., Cai, C. Z., Ji, Z. L., Cao, Z. W., Cui, J. and Chen, Y. Z. (2004) Predicting functional family of novel enzymes irrespective of sequence similarity: a statistical learning approach. *Nucleic Acids Res.* **32**: 6437-6444.

92. Hanada, K., Yewdell, J. W. and Yang, J. C. (2004) Immune recognition of a human renal cancer antigen through post-translational protein splicing. *Nature* **427**: 252-256.

93. Harfouch-Hammoud E, Walk T, Otto H, Jung G, Bach, J. F., van Endert, P. M. and Caillat-Zucman, S. (1999) Identification of peptides from autoantigens GAD65 and IA-2 that bind to HLA class II molecules predisposing to or protecting from type 1 diabetes. *Diabetes* **48**: 1937-1947.

94. Harman, K. E., Gratian, M. J., Bhogal, B. S., Challacombe, S. J. and Black, M. M. (2000) A study of desmoglein 1 autoantibodies in pemphigus vulgaris: Racial differences in frequency and the association with a more severe phenotype. *Br. J. Dermatol.* **143**: 343–348.

95. Haselden, B. M., Kay, A. B. and Larche, M. (2000) Peptide-mediated immune responses in specific immunotherapy. *Int. Arch. Allergy Immunol*. **122**: 229-237.

96. Hattotuwagama, C. K., Doytchinova, I. A. and Flower, D. R. (2005) In silico prediction of peptide binding affinity to class I mouse major histocompatibility complexes: a comparative molecular similarity index analysis (CoMSIA) study. *J. Chem. Inf. Model*. **45**: 1415-1423.

97. Hertl, M., Amagai, M., Sundaram, H., Stanley, J., Ishii, K. and Katz, S. I. (1998) Recognition of desmoglein 3 by autoreactive T cells in pemphigus vulgaris patients and normals. *J. Invest. Dermatol*. **110**: 62-66.

98. Hertl, M., Karr, R. W., Amagai, M. and Katz, S. I. (1998) Heterogeneous MHC II restriction pattern of autoreactive desmoglein 3 specific T cell responses in pemphigus vulgaris patients and normals. *J. Invest. Dermatol*. **110**: 388-392.

99. Hubbard, S. J. and Thornton, J. M. (1993) 'NACCESS', Computer Program, Department of Biochemistry and Molecular Biology, University College London, UK.

100. Hunt, D. F., Henderson, R. A., Shabanowitz, J., Sakaguchi, K., Michel, H., Sevilir, N., Cox, A. L., Appella, E. and Engelhard, V. H. (1992) Characterization of peptides bound to the class I MHC molecule HLA-A2.1 by mass spectrometry. *Science* **255**: 1261-1263.

101. Jardetzky, T. S., Lane, W. S., Robinson, R. A., Madden, D. R. and Wiley, D. C. (1991) Identification of self peptides bound to purified HLA-B27. *Nature* **353**: 326-329.

102. Jameson, S. C. and Bevan, M. J. (1992) Dissection of major histocompatibility complex (MHC) and T cell receptor contact residues in a Kb-restricted ovalbumin peptide and an assessment of the predictive power of MHC-binding motifs. *Eur. J. Immunol.* **22**: 2663-2667.

103. Janin, J. (1995) Protein-protein recognition. *Prog. Biophys. Mol. Biol*. **64**: 145-166.

104. Joachims, T. (2001) *Learning To Classify Text Using Support Vector Machines – Methods, Theory, Algorithms. Kluwer Academic Publishers.*

105. Jones, S., and Thornton, J. M. (1996) Principles of protein-protein interactions. *Proc. Natl. Acad. Sci. USA* **93**: 13-20.

106. Kaas, Q., Ruiz, M. and Lefranc, M. P. (2004) IMGT/3Dstructure-DB and IMGT/StructuralQuery, a database and a tool for immunoglobulin, T cell receptor and MHC structural data. *Nucleic Acids Res*. **32**: D208-210.

107. Kangueane, P., Sakharkar, M. K., Lim, K. S., Hao, H., Lin, K., Ren, E. C. and Kolatkar, P. R. (2000) Knowledge-based grouping of modeled HLA peptide complexes. *Hum. Immunol.* **61**: 460-466.

108. Kangueane, P., Sakharkar, M. K., Kolatkar, P. R. and Ren, E. C. (2001) Towards the MHC-peptide combinatorics. *Hum. Immunol.* **62**: 539-556.

109. Karplus, K. (1995) Evaluating regularizers for estimating distributions of amino acids. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **3**: 188-196.

110. Karpenko, O., Shi, J. and Dai, Y. (2005) Prediction of MHC class II binders using the ant colony search strategy. *Artif. Intell. Med.* **35**: 147-156.

111. Keogh, E., Fikes, J., Southwood, S., Celis, E., Chesnut, R. and Sette, A. (2001) Identification of new epitopes from four different tumor-associated antigens: recognition of naturally processed epitopes correlates with HLA-A*0201 binding affinity. *J. Immunol*., **167**: 787-796.

112. Kirschmann, D. A., Duffin, K. L., Smith, C. E., Welply, J. K., Howard, S. C. Schwartz, B. D. and Woulfe, S. L. (1995) Naturally processed peptides from rheumatoid arthritis associated and non-associated HLA-DR alleles. *J. Immunol*. **155**: 5655-5662.

113. Koch, P. J., Mahoney, M. G., Cotsarelis, G., Rothenberger, K., Lavker, R. M. and Stanley, J. R. (1998) Desmoglein 3 anchors telogen hair in the follicle. *J. Cell Sci*. **111**, 2529-2537.

114. Krystek, S., Stouch, T. and Novotny, J. (1993) Affinity and specificity of serine endopeptidase-protein inhibitor interactions. Empirical free energy calculations based on X-ray crystallographic structures. *J. Mol. Biol*. **234**: 661-679.

115. Klein, J. (1986) *Natural history of the major histocompatibility complex. Wiley & Sons, Inc.*

116. Klein,L., Klugmann, M., Nave, K. A., Tuohy, V. K. and Kyewski, B. (2000) Shaping of the autoreactive T cell repertoire by a splice variant of self protein expressed in thymic epithelial cells. *Nat. Med*., **6**: 56-61.

117. Knutson, K. L., Schiffman, K. and Disis, M. L. (2001) Immunization with a HER-2/neu helper peptide vaccine generates HER-2/neu CD8 T cell immunity in cancer patients. *J. Clin. Invest*. **107**: 477-484.

118. Krco, C. J., Harders, J., Chapoval, S. and David, C. S. (2000) Immune response of HLA-DQ transgenic mice to house dust mite allergen p2: identification of HLA-DQ restricted minimal epitopes and critical residues. *Clin. Immunol*. **97**: 154-161.

119. Krystek,S., Stouch, T. and Novotny, J. (1993) Affinity and specificity of serine endopeptidase-protein inhibitor interactions. Empirical free energy calculations based on X-ray crystallographic structures. *J. Mol. Biol.* **234**: 661-679.

120. Kwok, W. W., Lotshaw, C., Milner, E. C., Knitter-Jack, N. and Nepom, G. T. (1989) Mutational analysis of the HLA-DQ3.2 insulin-dependent diabetes mellitus susceptibility gene. *Proc. Natl. Acad. Sci. USA* **86**: 1027-1030.

121. Laskowski, R. A. (1991) SURFNET computer program (Department of Biochemistry and Molecular Biology, University College, London, England).

122. Lee, E., Lendas, K. A., Chow, S., Pirani, Y., Gordon, D. Dionisio, R. Nguyen, D. Spizuoco, A. Fotino, M. Zhang, Y. and Sinha, A. A. (2006) Disease relevant HLA class II alleles isolated by genotypic, haplotypic, and sequence analysis in north American Caucasians with pemphigus vulgaris. *Hum Immunol*. **67**: 125-139.

123. Lee, K. H., Wucherpfennig, K. W., Wiley, D. C. (2001) Structure of a human insulin peptide-HLA-DQ8 complex and susceptibility to type 1 diabetes. Nat Immunol 2: 501-507.

124. Lefranc, M. P. and Lefranc, G. (2001) The T cell receptor FactsBook, Academic Press.

125. Levitsky, V., Zhang, Q. -J., Levitskaya, J. and Masucci, M. G. (1996) The lifespan of MHC-peptide complexes influences the efficiency of presentation and immunogenicity of two class I restricted cytotoxic T lymphocyte epitopes in the Epstein-Barr virus nuclear antigen-4. *J. Exp. Med.* **183**: 915-926.

126. Li, Y., Li, H., Martin, R. and Mariuzza, R. A. (2000) Structural basis for the binding of an immunodominant peptide from myelin basic protein in different registers by two HLA-DR2 proteins. *J. Mol. Biol*. **304**: 177-188.

127. Lipford, G. B., Hoffman, M., Wagner, H. and Heeg, K. (1993) Primary in vivo responses to ovalbumin. Probing the predictive value of the Kb binding motif. *J. Immunol.* **4**: 1212-1222.

128. Lim, J. S., Kim, S., Lee, H. G., Lee, K. Y., Kwon, T. J. and Kim, K. (1996) Selection of peptides that bind to the HLA-A2.1 molecule by molecular modelling. *Mol. Immunol.* **33**: 221-230.

129. Lin, M. S., Swartz, S. J., Lopez, A., Ding, X., Fernandez-Vina, M. A., Stastny, P., Fairley, J. A. and Diaz, L. A. (1997) Development and characterization of desmoglein-3 specific T cells from patients with pemphigus vulgaris. *J. Clin. Invest*. **99**: 31-40.

130. 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**: 229-243.

131. 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**: 675-679.

132. Lopez, J.A., Weilenman, C., Audran, R., Roggero, M. A., Bonelo, A., Tiercy, J. M., Spertini, F. and Corradin, G. (2001) A synthetic malaria vaccine elicits a potent CD8 (+) and CD4 (+) T lymphocyte immune response in humans. Implications for vaccination strategies. *Eur. J. Immunol*. **31**: 1989-1998.

133. Lund, O., Nielsen, M., Kesmir, C., Petersen, A. G., Lundegaard, C., Worning, P., Sylvester-Hvid, C., Lamberth, K., Roder, G., Justesen, S., Buus, S. and Brunak, S. (2004) Definition of supertypes for HLA molecules using clustering of specificity matrices. *Immunogenetics* **12**: 797-810.

134. MacKerell, A. D., Bashford, D., Bellott, M., Dunbrack, R. L., Evanseck, J. D., Field, M. J., Fischer, S., Gao, J., Guo, H., Ha, S., *et al*. (1998) All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J. Phys. Chem.* ***B*****102**: 3586-3617.

135. Madden, D. R., Gorga, J. C., Strominger, J. L. and Wiley, D. C. (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**: 1035-1048.

136. Madden, D. R., Garboczi, D. N. and Wiley, D. C. (1993) The antigenic identity of peptide-MHC complexes: a comparison of the conformations of five viral peptides presented by HLA-A2. *Cell* **75**: 693-708.

137. Mallios, R. R. (1999) Class II MHC quantitative binding motifs derived from a large molecular database with a versatile iterative stepwise discriminant analysis meta-algorithm. *Bioinformatics* **15**: 432-439.

138. Mallios, R. R. (2001) Predicting class II MHC/peptide multi-level binding with an iterative stepwise discriminant analysis meta-algorithm. *Bioinformatics* **17**: 942-948.

139. Mamitsuka, H. (1998) Predicting peptides that bind to MHC molecules using supervised learning of hidden Markov models. *Proteins* **33**: 460-474.

140. Marshall, K. W., Liu, A. F., Canales, J., Perahia, B., Jorgensen, B. Gantzos, R. D., Aguilar, B. Devaux, B. and Rothbard, J. B. (1994) Role of the polymorphic residues in HLA-DR molecules in allele-specific binding of peptide ligands. *J. Immunol.* **152**: 4946-4957.

141. Martin, W., Sbai, H. and de Groot, A. S. (2003) Bioinformatics tools for identifying class I-restricted epitopes. *Methods* **29**: 289-298.

142. May, A. C. and Johnson, M, S. (1995) Improved genetic algorithm-based structure comparisons: Pairwise and multiple superpositions. *Protein Eng*. **8**: 873-82.

143. Mazza, G., Housset, D., Piras, C., Gregoire, C., Lin, S. Y., Fontecilla-Camps, J. C. and Malissen, B. (1998) Glimpses at the recognition of

peptide/MHC complexes by T cell antigen receptors. *Immunol. Rev*. **163**: 187-196.

144. McDonald, I. K. and Thornton, J. M. (1994) Satisfying hydrogen bonding potential in proteins. *J. Mol. Biol.* **238**: 777-793.

145. McSparron, H., Blythe, M. J., Zygouri, C., Doytchinova, I. A. and Flower, D. R. (2003) JenPep: a novel computational information resource for immunobiology and vaccinology. *J. Chem. Inf. Comput. Sci.* **43**: 1276-1287.

146. Meister, G. E., Roberts, C. G. P., Berzofsky, J. A. and de Groot, A. S. (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**: 581-591.

147. Michielin, O., Luescher, I. and Karplus, M. (2000) Modeling of the TCR-MHC-peptide complex. *J. Mol. Biol.* **300**: 1205-1235.

148. Michielin, O. and Karplus, M. (2002) Binding free energy differences in a TCR-peptide-MHC complex induced by a peptide mutation: a stimulation analysis. *J. Mol. Biol*. **324**: 547-569.

149. Middleton, D., Menchaca, L., Rood, H. and Komerofsky, R. (2003) New allele frequency database: http:// www.allelefrequencies.net. *Tissue Antigens*, **61**, 403-407.

150. Milik, M., Sauer, D., Brunmark, A. P., Yuan, L., Vitiello, A., Jackson, M. R., Peterson, P. A., Skolnick, J. and Glass, C. A. (1998) Application of an

artificial neural network to predict specific class I MHC binding peptide sequences. *Nat. Biotechnol.* **16**: 753-756.

151. Miyazawa, S. and Jernigan, R. L. (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**: 623-644.

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

153. Moesta, A. K., Stevanovic, S., Rammensee, H. G., Zhang, J., Rasmussen, J., Kanduc, D., Miele, W. R., Steinman, L. and Sinha, A. A. (2003) Reverse immunology identifies T cell epitopes mapping to intra- and extracellular regions of Dsg3 in pemphigus vulgaris. Submitted.

154. Momburg, F. and Hämmerling, G. (1998) Generation and TAP-mediated transport of peptides for major histocompatibility complex class I molecules. *Adv. Immunol*. **68**: 191-256.

155. Morrison, R. T. and Boyd, R. N. (1992) *Organic Chemistry Sixth Edition. Prentice Hall.*

156. Moustakas, A. K., van de Wal, Y., Routsias, J., Kooy, Y. M., van Veelen, P., Drijfhout, J. W., Koning, F. and Papadopoulos, G. K. (2000) Structure of celiac disease-associated HLA-DQ8 and non-associated HLA-DQ9 alleles in complex with two disease-specific epitopes. *Int. Immunol*. **12**: 1157-1166.

157. Mulder, N. J., Apweiler, R., Attwood, T. K., Bairoch, A., Barrell, D., Bateman, A., Binns, D., Biswas, M., Bradley, P., Bork, P., Bucher, P., Copley, R. R., Courcelle, E., Das, U., Durbin, R., Falquet, L., Fleischmann, W., Griffiths-Jones, S., Haft, D., Harte, N., Hulo, N., Kahn, D., Kanapin, A., Krestyaninova, M., Lopez, R., Letunic, I., Lonsdale, D., Silventoinen, V., Orchard, S. E., Pagni, M., Peyruc, D., Ponting, C. P., Selengut, J. D., Servant, F., Sigrist, C. J., Vaughan, R. and Zdobnov, E. M. (2003) The InterPro Database, 2003 brings increased coverage and new features. *Nucleic Acids Res.* **31**: 315-318.

158. Murthy, V. L. and Stern, L. J. (1997) The class II MHC protein HLA-DR1 in complex with an endogenous peptide: implications for the structural basis of the specificity of peptide binding. *Structure* **5**: 1385-1396.

159. Neeno, T., Krco, C. J., Harders, J., Baisch, J., Cheng, S. and David, C. S. (1996) HLA-DQ8 transgenic mice lacking endogenous class II molecules respond to house dust allergens: identification of antigenic epitopes. *J. Immunol*. **156**: 3191-3195.

160. Needleman, S. B. and Wunsch, C. D. (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.* **48**: 443-453.

161. Nepom, G. T. and Kwok, W. W. (1998) Molecular basis for HLA-DQ associations with IDDM. *Diabetes* **47**: 1177-1184.

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

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

164. 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.

165. Novotny, J., Bruccoleri, R. E., Davis, M. and Sharp, K. A. (1997) Empirical free energy calculations: a blind test and further improvements to the method. *J. Mol. Biol*. **268**: 401-411.

166. Ortmann, B., Androlewicz, M. J. and Cresswell, P. (1994) MHC class I/beta 2-microglobulin complexes associate with TAP transporters before peptide binding. *Nature* **368**: 864-867.

167. Palmowski, M. J., Gileadi, U., Salio, M., Gallimore, A., Millrain, M., James, E., Addey, C., Scott, D., Dyson, J., Simpson, E. and Cerundolo, V. (2006) Role of immunoproteasomes in cross-presentation. *J. Immunol.* **177**: 983-990.

168. Parker, K. C., Bednarek, M. A. and Coligan, J. E. (1994) Scheme for ranking potential HLA-A2 binding peptides based on independent binding of individual peptide side-chains. *J. Immunol.* **152**: 163-175.

169. Peters, P. J., Raposo, G., Neefjes, J. J., Oorschot, V., Leijendekker, R. L., Geuze, H. J. and Ploegh, H. L. (1995) Major histocompatibility complex

class II compartments in human B lymphoblastoid cells are distinct from early endosomes. *J. Exp. Med*. **182**: 325-334.

170. Rabiner, L. R. (1989) A tutorial on Hidden Markov Models and selected applications in speech recognition. *Proc. IEEE* **77**: 257-286.

171. Rajapakse, M., Schmidt, B. and Brusic, V. (2006) Multi-objective evolutionary algorithm for discovering peptide binding motifs. *Lecture Notes in Computer Science.* In Press.

172. Rajapakse, M., Wyse, L., Schmidt, B. and Brusic, V. (2005) Deriving matrix of peptide-MHC interactions in diabetic mouse by genetic algorithm. *IDEAL 2005*: 440-447.

173. Rammensee, H. G., Falk, K. and Rotzschke, O. (1993) Peptides naturally presented by MHC class I molecules. *Annu. Rev. Immunol*. **11**: 213-244.

174. Rammensee, H. G., Friede, T. and Stevanović, S. (1995) MHC ligands and peptide motifs: first listing. *Immunogenetics* **41**: 178-228.

175. Rammensee, H., Bachmann, J., Emmerich, N. P., Bachor, O. A. and Stevanović, S. (1999) SYFPEITHI: database for MHC ligands and peptide motifs. *Immunogenetics* **50**: 213-219.

176. Reche, P. A., Glutting, J. P. and Reinherz, E. L. (2002) Prediction of MHC class I binding peptides using profile motifs. *Hum. Immunol.* **63**: 701-709.

177. Riechers, R., Grotzinger, J. and Hertl, M. (1999) HLA class II restriction of autoreactive T cell responses in pemphigus vulgaris: review of the literature and potential applications for the development of a specific immunotherapy. *Autoimmunity* **30**: 183-196.

178. Roberts, J. D., Niedzwiecki, D., Carson, W. E., Chapman, P. B., Gajewski, T. F., Ernstoff, M. S., Hodi, F. S., Shea, C., Leong, S. P., Johnson, J., Zhang, D., Houghton, A., Haluska, F. G, Cancer and Leukemia Group B. (2006) Phase 2 study of the g209-2M melanoma peptide vaccine and low-dose interleukin-2 in advanced melanoma: Cancer and Leukemia Group B 509901. *J. Immunother.* **29**: 95-101.

179. Robinson, J., Waller, M. J., Parham, P., Bodmer, J. G. and Marsh, S. G. E. (2001) IMGT/HLA Database - a sequence database for the human major histocompatibility complex. *Nucleic Acids Res* **29**: 210-213.

180. Robles, D. T., Fain, P. R. Gottlieb, P. A. and Eissenbarth, G. S. (2002) The genetics of autoimmune polyendocrine syndrome type II. *Endocrinol. Metab. Clin. North. Am.* **31**: 353-368.

181. Roetzschke, O., Falk, K., Stefanovic, S., Jung, G., Walden, P. and Rammensee, H. G. (1991) Exact prediction of a natural T cell epitope. *Eur. J. Immunol.* **21** : 2891-2894.

182. Rognan, D., Laumoeller, S. L., 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**: 4650-4658.

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

184. Rosenfeld, R., Zheng, Q., Vajda, S. and Delisi, C. (1993) Computing the structure of bound peptides: Application to antigen recognition by class I major histocompatibility complex receptors. *J. Mol. Biol.* **234**: 515-521.

185. Rost, B. and Sander, C. (1994) Conservation and prediction of solvent accessibility in protein families. *Proteins* **20**: 216-226.

186. Rudensky, A. Y., Maric, M., Eastman, S., Shoemaker, L., DeRoos, P. C. and Blum, J. S. (1994) Intracellular assembly and transport of endogenous peptide-MHC class II complexes. *Immunity* **1**: 585-594.

187. Ruppert, J., Sidney, J., Celis, E., Kubo, R. T., Grey, H. M. and Sette, A. (1993) Prominent role of secondary anchor residues in peptide binding to HLA-A2.1 molecules. *Cell* **74**: 929-937.

188. Saitou, N. and Nei, M. (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* **4**: 406-425.

189. Salato, V. K., Hacker-Foegen, M. K., Lazarova, Z., Fairley, J. A. and Lin, M. S. (2005) Role of intracellular epitope spreading in pemphigus vulgaris. *Clin. Immunol*. **116**: 54-64.

190. Sali, A. and Blundell, T. L. (1993) Comparative protein modeling by satisfaction of spatial restraints. *J. Mol. Biol.* **234**: 774-815.

191. Samanta, U., Bahadur, R. P. and Chakrabarti, P. (2002) Quantifying the accessible surface area of protein residues in their local environment. *Prot. Eng.* **15**: 659-667.

192. Samudrala, R. and Moult, J. (1997) Handling context-sensitivity in protein structures using graph theory: bona fide prediction. *Proteins* **Suppl 1**: 43-49.

193. Sanchez, R. and Sali, S. (1997) Evaluation of comparative protein structure modeling by MODELLER-3. *Proteins* **Suppl**., 50-58.

194. Savoie, C. J., Kamikawaji, N., Sasazuki, T. and Kuhara, S. (1999) Use of BONSAI decision trees for the identification of potential MHC class I peptide epitope motifs. *Pac Symp Biocomput.*: 182-189.

195. Schafer, J. R., Jesdale, B. M., George, J. A., Kouttab, N. M. and de Groot, A. S. (1998) Prediction of well-conserved HIV-1 ligands using a matrix-based algorithm, EpiMatrix. *Vaccine* **16**: 1880-1884.

196. Schapira, M., Totrov, M. and Abagyan, R. (1999) Prediction of the binding energy for small molecules, peptides and proteins. *J. Mol. Recognit*. **12**: 177-190.

197. Scharf, S. J., Freidmann, A., Steinman, L., Brautbar, C. and Erlich, H. A. (1989) Specific HLA-DQB and HLA-DRB1 alleles confer susceptibility to pemphigus vulgaris. *Proc. Natl. Acad. Sci. USA* **86**: 6215-6219.

198. Schirle, M., Weinschenk, T. and Stevanović, S. (2001) Combining computer algorithms with experimental approaches permits the rapid and accurate identification of T cell epitopes from defined antigens. *J. Immunol. Methods* **257**: 1-16.

199. Schönbach, C., Koh, L. Y., Sheng, X., Wong, L. and Brusic, V. (2000) FIMM, a database of functional molecular immunology. *Nucleic Acids Res*. **28**: 1 222-224

200. 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**: 549-564.

201. 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**: 1838-1846.

202. Scott, C. A., Peterson, P. A., Teyton, L. and Wilson, I. A. (1998) Crystal structures of two I-A^{d}- peptide complexes reveal that high affinity can be achieved without large anchor residues. *Immunity* **8**: 319-329.

203. Seamons, A., Sutton, J., Bai, D., Baird, E., Bonn, N., Kafsack, B. F., Shabanowitz, J., Hunt, D. F., Beeson, C. and Goverman, J. (2003) Competition between two MHC binding registers in a single peptide processed from myelin basic protein influences tolerance and susceptibility to autoimmunity. *J. Exp. Med*. **197**: 1391-1397.

204. Segal, M. R., Cummings, M. P. and Hubbard, A. E. (2001) Relating amino acid sequence to phenotype: analysis of peptide-binding data. *Biometrics* **57**: 632-642.

205. Sette, A., Sidney, J., Oseroff, C., del Guercio, M. F., Southwood, S., Arrhenious, T., Powell, M. F., Colon, S. M., Gaeta, F. C. and Grey, H. M. (1993) HLA DR4w4-binding motifs illustrate the biochemical basis of

degeneracy and specificity in peptide-DR interactions. *J. Immunol*. **151**: 3163–3170.

206. Sette, A., Vitiello, A., Reherman, B., Fowler, P., Nayersina, R., Kast, W. M., Melief, C. J., Oseroff, C., Yuan, L. and Ruppert, J. (1994) The relationship between class I binding affinity and immunogenecity of potential cytotoxic T cell epitopes. *J. Immunol*. **153**: 5586-5592.

207. Sette, A., Livingstone, B., McKinney, D., Appella, E., Fikes, J., Sidney, J., Newman, M. and Chesnut, R. (2001) The development of multi-epitope vaccines: epitope identification, vaccine design and clinical evaluation. *Biologicals* **29**: 271-276.

208. 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**: 443-451.

209. Sezerman, U., Vajda, S. and DeLisi, C. (1996) Free energy mapping of class I MHC molecules and structural determination of bound peptides. *Protein Sci.* **5**: 1272-1281.

210. Sidney, J., Oseroff, C., del Guercio, M. F., Southwood, S., Krieger, J. I., Ishioka, G. Y., Sakaguchi, K., Appella, E. and Sette, A. (1994) Definition of a DQ3.1-specific binding motif. *J. Immunol.* **152**: 4516-4525.

211. Sidney, J., del Guercio, M. F., Southwood, S. and Sette, A. (2002) The HLA molecules DQA1*0501/B1*0201 and DQA1*0301/B1*0302 share an

extensive overlap in peptide binding specificity. *J. Immunol.*, **169**, 5098-5108.

212. Singh, R. R. (2000) The potential use of peptides and vaccination to treat systemic lupus erythematosus. *Curr. Opin. Rheumatol.* **12**: 399-406.

213. Sinha, A.A., Lopez, M.T. and McDevitt, H.O. (1990) Autoimmune diseases: the failure of self-tolerance. *Science* **248**: 1380-1388.

214. Sinha, A. A., Brautbar, C., Szafer, F., Friedmann, A., Tzfoni, E., Todd, J. A., Steinman, L. and McDevitt, H. O. (1988) A newly characterized HLA DQ beta allele associated with pemphigus vulgaris. *Science* **239**: 1026-1029.

215. Smith, T. F. and Waterman, M. S. (1981) Identification of common molecular subsequences. *J. Mol. Biol*. **147**: 195-197.

216. Sollid, L. M. and Thorsby, E. (1993) HLA susceptibility genes in celiac disease: genetic mapping and role in pathogenesis. *Gastroenterology* **105**: 910-922.

217. Southwood, S., Sidney, J., Kondo, A., del Guercio, M. F., Appella, E., Hoffman, S., Kubo, R. T., Chesnut, R. W., Grey, H. M. and Sette, A. (1998) Several common HLA-DR types share largely overlapping peptide binding repertoires. *J. Immunol.* **160**: 3363-3373.

218. Spiliotis, E. T., Manley, H., Osorio, M., Zuniga, M. C. and Edidin, M. (2000) Selective export of MHC class I molecules from the ER after their dissociation from TAP. *Immunity* **6**: 841-851.

219. Srinivasan, K. N., Zhang, G. L., Khan, A. M., August, J. T. and Brusic, V. (2004) Prediction of class I T cell epitopes: evidence of presence of

immunological hot spots inside antigens. *Bioinformatics* **20 Suppl 1**: i297-i302.

220. Stern, L. J., Brown, J. H., Jardetzky, T. S., Gorga, J. C., Urban, R. G., Strominger, J. L. and Wiley, D. C. (1994) Crystal structure of the human class II MHC protein HLA-DR1 complexed with an influenza virus peptide. *Nature* **368**: 215-221.

221. Stern, L. J. and Wiley, D. C. (1994) Antigenic peptide binding by class I and class II histocompatibility proteins. *Structure* **2**: 245-251.

222. Sturniolo, T., Bono, E., Ding, J., Raddrizzani, L., Tuereci, O., Sahin, U., Braxenthaler, M., Gallazzi, F., Protti, M. P., 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**: 555-561.

223. Suh, W. K., Cohen-Doyle, M. F., Fruh, K., Wang, K., Peterson, K. A. and Williams, D. B. (1994) Interaction of MHC class I molecules with the transporter associated with antigen processing. *Science* **264**: 1322-1326.

224. Suri, A., Walters, J. J., Gross, M. L. and Unanue, E. R. (2005) Natural peptides selected by diabetogenic DQ8 and murine I-A^{g7} molecules show common sequence specificity. *J. Clin. Invest*. **115**: 2268-2276.

225. Swindells, M. B. and Thornton, J. M. (1991) *Curr. Opin. Struct. Biol*. **1**: 219.

226. Todd, J.A., Acha-Orbea, H., Bell, J.I., Chao, N., Fronek, Z., Jacob, C.O., McDermott, M., Sinha, A.A., Timmerman, L., Steinman, L., *et al.* (1988) A

molecular basis for MHC class II-associated autoimmunity. *Science.* **240**: 1003-1009.

227. Toseland, C. P., Clayton, D. J., McSparron, H., Hemsley, S. L., Blythe, M. J., Paine, K., Doytchinova, I. A., Guan, P., Hattotuwagama, C. K. and Flower, D. (2005) AntiJen: a quantitative immunology database integrating functional thermodynamic, kinetic, biophysical, and cellular data. *Immunome Res*. **1**: 4.

228. Totrov, M. and Abagyan, R. (2001) Protein-ligand docking as an energy optimization problem. In *Drug-receptor thermodynamics: Introduction and experimental applications*. Raffa, R. B., ed. New York: John Wiley & Sons, 603-624.

229. Totrov, M. and Abagyan, R. (1999) Derivation of sensitive discrimination potential for virtual ligand screening. *Proceedings of the Third Annual International Conference on Computational Molecular Biology*. Lyon, France: ACM Press, 37-38.

230. Tzakos, A. G., Fuchs, P., van Nuland, N. A., Troqanis, A., Tselios, T., Deraos, S., Matsoukas, J., Gerothanassis, I. P. and Bonvin, A. M. (2004) *Eur. J. Biochem*. **271**: 3399-3413.

231. Uebel S. and Tampe R. (1999) Specificity of the proteosome and the TAP transporter. *Curr. Opin. Immunol*. **11**: 203-208.

232. Vapnik, V. (1998) Statistical Learning Theory. *Wiley, Chichester, G.B.*

233. Veldman, C. M., Gebhard, K. L., Uter, W., Wassmuth, R., Grötzinger, J., Schultz, E. and Hertl, M. (2003) T cell recognition of Desmoglein 3 peptides

in patients with Pemphigus Vulgaris and healthy individuals. *J. Immunol*. **172**: 3883-3892.

234. Wang, E., Phan, G. Q. and Marincola, F. M. (2001) T cell-directed cancer vaccines: the melanoma model. *Expert Opin. Biol. Ther.* **1**: 277-290.

235. Wang, R. N., Lai, L. N. and Wang, S. N. (2002) Further development and validation of empirical scoring functions for structure-based binding affinity prediction. *J. Comput. Aided Mol. Des*. **16**: 11-26.

236. Wallace A. C., Laskowski, R. A. and Thornton, J. M. (1995) LIGPLOT: A program to generate schematic diagrams of protein-ligand interactions. *Prot. Eng.* **8**: 127-134.

237. Wan, S., Coveney, P. and Flower, D. R. (2004) Large-scale molecular dynamics simulations of HLA-A*0201 complexed with a tumor-specific antigenic peptide: can the alpha3 and beta2m domains be neglected? *J. Comput. Chem*. **25**: 1803-1813.

238. Weng, Z., Vajda, S. and Delisi, C. (1996) Prediction of protein complexes using empirical free energy functions. *Protein Sci*. **5**: 614-626.

239. Williams, T. M. (2001) ‘Human leukocyte antigen gene polymorphism and the histocompatibility laboratory’, *J. Mol. Diagn.* **3**: 98-104.

240. Wucherpfennig, K. W., Yu, B., Bhol, K., Monos, D. S., Argyris, E., Karr, R. W., Ahmed, A. R. and Strominger, J. L. (1995) Structural basis for major histocompatibility complex (MHC)-linked susceptibility to autoimmunity: charged residues of a single MHC binding pocket confer selective

presentation of self-peptides in pemphigus vulgaris. *Proc. Natl. Acad. Sci. USA* **92**: 11935-11939.

241. Yamashina, Y., Miyagawa, S., Kawatsu, T., Iida, T., Higashimine, I., Shirai, T. and Kaneshige, T. (1998) Polymorphisms of HLA class II genes in Japanese patients with pemphigus vulgaris. *Tissue Antigens* **52**: 74-77.

242. Yewdell J. W., Reits E. and Neefjes J. (2003) Quantitating the MHC class I antigen processing pathway. *Nature Rev. Immunol*. **3**: 952-961.

243. Yu, K., Petrovsky, N., Schonbach, C., Koh, J. Y. and Brusic, V. (2002) Methods for prediction of peptide binding to MHC molecules: a comparative study. *Mol. Med.* **8**: 137-148.

244. Zacharias, M. and Springer, S. (2004) Conformational flexibility of the MHC class I α_1 – α_2 domain in peptide bound and free states: a molecular dynamics simulation study. *Biophysical J*. **87**: 2203-2214.

245. Zauhar, R. J. and Morgan, R. S. (1985) A new method for computing the macromolecular electric potential. *J. Mol. Biol*. **20**: 815-820.

246. Zhang, Q. J., Gavioli, R., Klein, G. and Masucci, M. G. (1993) An HLA-A11-specific motif in nonamer peptides derived from viral and cellular proteins. *Proc. Natl. Acad. Sci. USA* **90**: 2217-2221.

247. Zhao, Y., Pinilla, C., Valmori, D., Martin, R. and Simon, R. (2003) Application of support vector machines for T cell epitopes prediction. *Bioinformatics* **19**: 1978-1984.

248. Zhang, G. L., Khan, A. M., Srinivasan, K. N., August, J. T. and Brusic, V. (2005) MULTIPRED: a computational system for prediction of promiscuous HLA binding peptides. *Nucleic Acids Res.* **33**: W172-W179.

249. Zhihua, L., Yuzhang, W., Bo, Z., Bing, N. and Li, W. (2004) Toward the quantitative prediction of T-cell epitopes: QSAR studies on peptides having affinity with the class I MHC molecular HLA-A*0201. *J. Comput. Biol*. **11**: 683-694.

Publications

1. J. C. Tong, T. W. Tan, S. Ranganathan. 2006. Methods and Protocols for Predicting Immunogenic Epitopes. *Briefings in Bioinformatics*. In press. Accepted 14/09/06.
2. Tong, J. C., Kong, L., Tan, T. W. and Ranganathan, S. (2006) MPID-T: database for sequence-structure-function information on TCR/peptide/MHC interactions. *Applied Bioinformatics* **5**: 111-114.
3. Tong, J. C., Tan, T. W. and Ranganathan, S. (2004) Modeling the structure of bound peptide ligands to major histocompatibility complex. *Protein Sci*. **13**: 2523-2532.
4. Ranganathan, S. and Tong, J. C. (2006) A Practical Guide to Structure-based Prediction of MHC Binding Peptides, In D. R. Flower. (ed.), *Methods in Molecular Biology: Immunoinformatics: Predicting Immunogenicity in silico*. Humana Press.
5. Ranganathan, S. Tong, J. C. and Tan, T. W. Structural Immunoinformatics. Immunoinformatics: Opportunities and Challenges of Bridging Immunology with Computer and Information Sciences. Kluwer Plenum. In press.
6. Tong, J. C., Bramson, J., Kanduc, D., Chow, S., Sinha, A. A. and Ranganathan, S. (2006) Modeling the bound conformation of pemphigus vulgaris-associated peptides to MHC class II DR and DQ Alleles. *Immunome Res.* **2**: 1.

7. J. C. Tong, Zhang, G. L., Tan, T. W., August, J. T., Brusic, V. and Ranganathan, S. (2006) Prediction of HLA-DQ3.2□ ligands: Evidence of multiple registers in class II binding peptides. *Bioinformatics* **22**: 1232-1238.

8. K. H. Choo, J. C. Tong, L. Zhang. (2004) Recent applications of hidden Markov models in computational biology – A Review. *Genomics, Proteomics & Bioinformatics* **2**: 84-96.

9. L. Kong, B. T. K. Lee, J. C. Tong, T. W. Tan, S. Ranganathan, 2004. SDPMOD: an automated comparative modeling server for small disulphide-bonded proteins. *Nucleic Acids Res.* **32**: W356–W359.

10. J. C. Tong, J. Bramson, D. Kanduc, A. A. Sinha, S. Ranganathan. 2006. Prediction of desmoglein-3 peptides reveals multiple shared T-cell epitopes in HLA DR4- and DR6- associated pemphigus vulgaris. *BMC Bioinformatics* **7** Suppl 5: S7.