Class I HLA supertype and supermotif definition by chemometric approaches

by

Pingping Guan

A thesis submitted for the degree of Doctor of Philosophy at the University of London

October 2004

The Edward Jenner Institute for Vaccine Research

University College London

ProQuest LLC
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106-1346

Abstract

Activation of cytotoxic T cells in human requires specific binding of antigenic peptides to human leukocyte antigen (HLA) molecules. HLA is the most polymorphic protein in the human body, currently 1814 different alleles collected in the HLA sequence database at the European Bioinformatics Institute. Most of the HLA molecules recognise different peptides. Also, some peptides can be recognised by several of HLA molecules. In the present project, all available class I HLA alleles are classified into supertypes. Super – binding motifs for peptides binding to some supertypes are defined where binding data are available.

A variety of chemometric techniques are used in the project, including 2D and 3D QSAR techniques and different variable selection methods like SIMCA, GOLPE and genetic algorithm. Principal component analysis combined with molecular interaction fields calculation by the program GRID is used in the class I HLA classification.

This thesis defines an HLA-A3 supermotif using two QSAR methods: the 3D-QSAR method CoMSIA, and a recently developed 2D-QSAR method, which is named the additive method. Four alleles with high phenotype frequency were included in the study: HLA-A*0301, HLA-A*1101, HLA-A*3101 and HLA-A*6801. An A*0201 binding motif is also defined using amino acid descriptors and variable selection methods. Novel peptides have been designed according to the motifs and the binding affinity is tested experimentally. The results of the additive method are used in the online server, MHCPred, to predict binding affinity of unknown peptides. In HLA classification, the HLA-A, B and C molecules are classified into supertypes separately. A total of eight supertypes are observed for class I HLA, including A2, A3, A24, B7, B27, B44, C1 and C4 supertype. Using the HLA classification, any newly discovered class I HLA molecule can be grouped into a supertype easily, thus simplifying the experimental function characterisation process.

Acknowledgements

I would like to thank my supervisor Dr. Darren R. Flower for giving me the opportunity to work in his group, also thank him for his support and encouragement during my PhD.

I would also thank for people in the Bioinformatics lab for their invaluable help throughout my work. Thanks to Darren and Dr. Irini Doytchinova for many helpful discussions and suggestions for my project. Thanks to Valerie Walshe for helping me with the laboratory assays.

Contents

List of figures

Chapter 5

Chapter 6

List of tables

Amino acid abbreviations

The table below lists the name of the 20 amino acids, their three letter and one letter code which are used in the text.

Amino acids	*Three letter code*	*One letter code*
Non-polar		
Glycine	Gly	G
Alanine	Ala	A
Valine	Val	V
Leucine	Leu	L
Isoleucine	Ile	I
Methionine	Met	M
Phenylalanine	Phe	F
Tryptophan	Trp	W
Proline	Pro	P
Polar		
Serine	Ser	S
Threonine	Thr	T
Cysteine	Cys	C
Tyrosine	Tyr	Y
Asparagine	Asn	N
Glutamine	Gln	Q
Electrically charged		
Aspartic acid	Asp	D
Glutamic acid	Glu	E
Lysine	Lys	K
Arginine	Arg	R
Histidine	His	H

Chapter 1
Introduction

1.1 The major histocompatibility complex

1.1.1 Overview

Major histocompatibility complex (MHC) molecules are polymorphic membrane glycoproteins (Zinkernagel, 1986). Human MHC is also called human leukocyte antigen, often abbreviated as HLA (Clark and Forman, 1984). There are two classes of HLA, class I and class II. Class I HLA is present on most nucleated cells, in particular the surfaces of lymphocytes, which have 1000 to 10000 HLA molecules per cell (Goust, 1993). Class II HLA is mostly expressed on antigen presenting cells (APC). An APC is a cell that has the ability to present antigen to helper T cells to activate an immune response. Examples of APCs are macrophages, dendritic cells, B cells and thymic epithelium cells.

One of the principal functions of the immune system is to recognise and eliminate foreign antigens in the body, such as viruses, bacteria and parasites (Janeway, 2001). Extracellular antigens can be recognised and destroyed by macrophages, T cells and B cells. The intracellular antigens, however, do not have direct contact with the immune system, and they have to be eliminated with the help of MHC proteins. The MHC proteins take up degraded intracellular antigenic fragments and present them to T cells to induce an immune system response (Ljunggren and Thorpe, 1996; Madnaka and Yvonne Jones, 1999; Rau *et al.*, 2001).

Class I HLA molecules mainly bind to 8 – 11, but up to 15 amino acids long intracellular peptide fragments and present them to the cytotoxic T cells. Both self and viral peptides are degraded in the cell by proteasomes, and class I HLAs are able to bind to both. Under normal conditions, HLA molecules bind to fragments of self proteins degraded in the cell. In infected cells, HLA molecules bind to fragments of degraded foreign proteins (Haeney, 1995). Activated cytotoxic T cells release proteins such as perforin and granzymes to induce cell lysis. Fas ligand is expressed on the surfaces of cytotoxic T cells and this is recognised by Fas receptors on infected cells to induce apoptosis.

Class II HLA binds to 9 – 25 amino acids long peptides from extracellular bacteria and viruses ingested by APCs and present these peptides to helper T cells to activate humoral and cellular immunity in the host (Madden *et al.*, 1993). Upon binding, T cells are stimulated and proliferate into either Th1 or Th2 helper cells. Th1 cells secrete interferon gamma receptor (IFN-γ) and express CD40 ligand or Fas ligand on their surface. CD40 ligands bind to and activate cells with CD40 on their surfaces. Fas ligand sends death signals to cells with Fas receptor on their surfaces. Th2 cells, on the other hand, activate B cells. Th2 cells secrete cytokins interleukin (Il) -4 and Il-5, which are B cell growth factors. Th2 cells also express CD40 ligand, which can bind to CD40 receptor on B cells and stimulate B cell proliferation.

Most peptides binding to class I and class II HLA alleles contain binding motifs, which is a combination of preferred residues at specific positions of the peptide. Most binding motifs are allele-specific. However, some alleles have similar

binding motifs, and some peptides can be recognised by more than one allele. HLA alleles can be classified into superfamilies according to their similar binding motifs. The discovery of cross-reactive peptides is very important in epitope based vaccination, where epitopes restricted to the superfamilies can be used in a vaccine that is effective in the global population.

HLA molecules play an important role in the immune system. The binding of peptides to HLA molecules, and their subsequent presentation to T cells, is the initial step of the host adaptive immune system defence against infectious agents. Recent experimental evidence suggests that MHC molecules can interact with the natural killer (NK) cell receptors (Valiante and Parham, 1996; Yokoyama *et al.*, 1995). Also the HLA is important in non-self tissue transplantation, in which the donor and the host's HLA type must match to avoid tissue rejection. At present, the interaction between peptides and HLA alleles is not fully understood. The present thesis focuses on class I HLA molecules, therefore in the following sections several aspects of class I HLA molecules are explored: MHC genetics, class I HLA structure, peptide-MHC interactions, laboratory and *in silico* techniques used to identify epitopes, protein degradation and MHC presentation pathway, HLA superfamilies and the application of epitopes in vaccine design.

1.1.2 HLA nomenclature

Since its initial discovery, 1814 different HLA alleles have been identified (Robinson *et al.*, 2003). A nomenclature system is used by the WHO Nomenclature Committee to name alleles (Bodmer *et al.*, 1990a; Bodmer *et al.*, 1990b; Marsh, 2003; Marsh, 2004). Allele names used in the thesis adopt the

following style: HLA-locus*allele. For example, HLA-A*0101 means the 0101 allele that is encoded by the A locus. HLA-DRB1*0701 is the 0701 allele encoded by the DRB locus, etc. As many alleles can be encoded by the same locus, the allele number can vary from 0101 to 8001 excluding the mutants. The same applies to other class I and II alleles. There are also mutants of the alleles in the IMGT/HLA database containing silent mutations that do not affect the protein sequences. Two digits are added to the allele name to distinguish the mutants. For example, HLA-A*020101 is an allele that is a mutant of the HLA-A*0201. Mutants are not included in the present studies. Full information about HLA nomenclature can be found on the HLA informatics group web page URL: http://www.anthonynolan.org.uk/HIG/.

1.1.3 MHC genetics

The MHC genes are located on chromosome 17 in mice and on the short arm of chromosome 6 in human (Koeller and Ozato, 1986). Mouse MHC is about 1.5 centimorgans long and is divided into three loci: H-2K, D and L. Human MHC is about 2 centimorgans long and consists of about 4000 kilobases of DNA (Goust, 1993). There are six loci encoding class I and II HLA alleles. The first three loci are HLA-DP, HLA-DR and HLA-DQ, which encode class II HLA alleles. The other three loci, HLA-A, HLA-B and HLA-C, are the three major loci encoding the class I HLA alleles (Fig. 1.1). Six minor loci have also been identified, which are HLA-E, HLA-F, HLA-G and HLA-H for class I and HLA-DN and HLA-DO for class II. Each class I loci is composed of eight exons divided by seven introns. Exon 1 encodes the leading sequence, which is cleaved during post-translation modification. The three extracellular domains α1 α2 α3 are encoded by exons 2

to 4. Exon 5 encodes the transmembrane helices and exons 6 to 8 encode the cytoplasmic tail (Fig. 1.1) (Maloy, 1987). Genes on the minor loci produce so-called non-classical MHC proteins, which are not involved in CD4 and CD8 T cell activation. Experimental evidence shows that HLA-E and G can contact natural killer cells and may inhibit natural killer cell induced cell lysis (Borrego *et al.*, 1998; Braud *et al.*, 1998; Lopez-Botet and Bellon, 1999; Marchal-Bras-Goncalves *et al.*, 2001; Matsunami *et al.*, 2000a; Matsunami *et al.*, 2000b; O'Callaghan, 2000; Pazmany *et al.*, 1996; Posch *et al.*, 1998; Rouas-Freiss *et al.*, 1997; Sasaki *et al.*, 1999).

The DP, DR and DQ loci of class II MHC exist in pairs, which may come from gene duplication (Koskimies and Eklund, 1997). Each pair encodes one α and one β chain of the class II MHC protein (Janeway, 2001). Each major locus of class I MHC encodes a single polypeptide (α chain), which binds to β2-microglobulin in the ER and forms the HLA complex (Bouvier and Wiley, 1998a). β2-microglobulin is encoded separately on chromosome 15 (Haeney, 1995). Both class I and II loci are highly polymorphic and are able to express, within a population, hundreds of different alleles (Caiozzo *et al.*, 2000). The most polymorphic region is found in the α chains of both class I and II MHC and β chains of class II MHC (Koeller and Ozato, 1986).

Some of the proteins in the MHC assembly process are also encoded in the MHC region, such as TAP and LMP, both genes are arranged in pairs as the class II genes and the arrangement may be the result of gene duplication (Beck *et al.*, 1992). The function of the TAP protein is to transport peptides into the ER, and

LMP proteins are subunits of the proteasome used for digesting proteins (Janeway, 2001). Some non-MHC proteins are encoded between the class I and class II regions, such as proteins of the complement system C2, C4 and factor B (Goust, 1993). These proteins have limited polymorphism and participate in the innate immune system responses.

Closely linked genes that are inherited together, like class I and class II MHCs, are also known as haplotypes (Zhao *et al.*, 2003b). An individual inherits one haplotype from the mother and one from the father, each containing three class I (HLA-A, B and C) and class II (HLA-DP, DR and DQ) loci (Rhodes and Trowsdale, 1999). Therefore an individual will have a maximum of six different class I specificities (Goust, 1993). The situation is more complicated for class II HLA. Class II alleles consists of two chains named α and β, and an individual can have one α gene from one parent and one β gene from another (Janeway, 2001). Hence, an individual may have a maximum of 12 different combinations. Occasionally a crossover occurs between the parental chromosomes and this generates new haplotypes, with mixed specificity from each locus, that also contributes to the HLA heterogenecity in the population (Koskimies and Eklund, 1997).

Alleles that are specific for each locus can be recognised by serotyping or mixed leukocyte reaction (MLR). In serotyping, the patient haplotype is obtained by adding monoclonal antibodies to the serum (Festenstein and Ollier, 1987; Geffrotin *et al.*, 1984; Welsh, 1989). MLR is often used to match class II HLA phenotypes by mixing lymphocytes from the two patients and testing whether the

lymphocytes are stimulated to proliferate. If there is lymphocyte proliferation, then the phenotypes from the two individuals do not match (Mohler and Streilein, 1989; Yoshizawa and Yano, 1984). With the advance in molecular genetics, the use of polymerase chain reaction (PCR) is now more common and is gradually replacing the traditional biochemical assays (Collins *et al.*, 2003; Konnai *et al.*, 2003; Middleton, 1999; Welsh and Bunce, 1999; Westerdahl *et al.*, 2004; Zheng *et al.*, 1999)

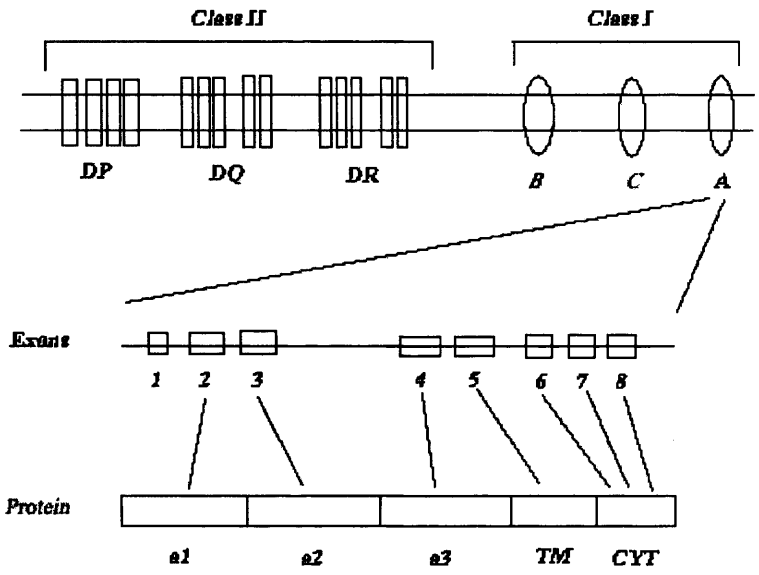


Figure 1.1. The arrangement of the HLA genes. Class I HLA is encoded on chromosome 6 and there are six loci: HLA-A, B and C that encode class I HLA alleles, and DP, DQ and DR that encode class II alleles. Class I HLA protein sequences are encoded on 8 exons. Exon 1 encodes the leading sequence which is cleaved during post-translation modification. α1 α2 α3 domains are encoded on exons 2, 3 and 4, respectively. Exon 5 encodes the transmembrane helices (TM) and exon 6, 7 and 8 encode the cytoplasmic tail (CYT) (Duran and Pease, 1986).

1.1.4 Class I HLA structure

MHC Class I molecules were first discovered in 1916 as cell surface antigens that caused transplant rejection in mice (Little and Tyzzer, 1916), hence MHC molecules were initially called transplantation antigens. Twenty years later, Gorer identified MHC molecules in mice, which was later named the H-2 antigen (Gorer, 1936; Gorer, 1937). Human MHC was discovered later by the finding that blood taken from pregnant women contained antibodies that agglutinated leukocytes, i.e, the antibodies targeted the leukocytes (van Rood *et al.*, 1958). The same situation was also found in people receiving blood transfusion even when the blood type was matched. H-2 molecules were purified for the first time in 1966 by Nathenson and Davies using gel filtration and ion exchange chromatography techniques. The molecular weight of the molecules was found to be 45,000 kD (Nathenson and Davis, 1966). The H-Kb protein was sequenced nearly fifteen years later by Coligan and his colleagues using a radiochemical assay (Coligan *et al.*, 1981; Uehara *et al.*, 1981a; Uehara *et al.*, 1981b). Later on, the three-dimensional structures of several MHC molecules were crystallised, such as HLA- Aw68(Garrett *et al.*, 1989), A2 (Saper *et al.*, 1991), B27 (Madden *et al.*, 1991a), H-2Db (Young *et al.*, 1994) and H-Kb (Ghendler *et al.*, 1998). The MHC crystal structures are produced by removing the hydrophobic transmembrane region and cytoplasmic tail to make the molecules soluble or express mutated MHC molecules as soluble molecules (Graff *et al.*, 1970; Mann *et al.*, 1968). These structures are used as the prototypes for the study of all MHC molecules.

The major part of a class I HLA molecule consists of a transmembrane heavy chain of 44 KD (Ploegh *et al.*, 1981). The extracellular part of the heavy chain is divided into 3 domains α1, α2 and α3 (Krensky and Clayberger, 1996), each about 90 amino acids long. These domains have been defined by sequence analysis and do not correspond to the two structural domains apparent in the extracellular part of MHC crystal structures. We use this nomenclature to remain consistent with immunological literature. Sequence alignment of human, murine, pig and rabbit MHC molecules suggested that the three α domains are conserved between species, but the transmembrane and cytopolasmic tail vary greatly in different species (Maloy, 1987). The HLA-A2 structure solved by Bjorkman revealed that the α2 and α3 domains are connected to a short helix consisting of residue 177 to 181 (Bjorkman *et al.*, 1987a). The α3 domain is linked to a 25 amino acids long transmembrane region followed by a short intracellular cytoplasmic tail of 30-35 amino acids at the C terminal (Shields *et al.*, 1999). The heavy chain is non-covalently attached to a 12 KD protein named β2-microglobulin (β_2m) and forms the complete MHC complex (Willcox *et al.*, 2003).

Solved human and mouse MHC crystal structures revealed that the peptide binding site is formed by an eight-strand antiparallel β sheet with two α helices running parallel to each other over the top of the β sheet (Bjorkman *et al.*, 1987a). This three-dimensional arrangement creates a long groove between the helices which is the peptide binding site (Bjorkman and Parham, 1990; Jones, 1997; Takiguchi, 1994), the β sheet forms the 'floor' of the binding site. (figure 1.2, figure 1.3). The binding of peptide to the MHC is confirmed by the observed

extra electron density within the binding site of the crystallised MHC proteins (Saper *et al.*, 1991).

The α3 domain and the β_2m have similar structures (Saper *et al.*, 1991). Sequence analysis showed that the α3 domain and β_2m are both immunoglobulin type domains (Gussow *et al.*, 1987; Nathenson *et al.*, 1986). Both the α3 domain and the β_2m consist of two antiparallel β sheets linked by a disulphide bond (Gussow *et al.*, 1987), one of the β sheets has four strands and the other has three. Conformational changes in β_2m can induce changes in the structure of the MHC complex and alter CTL recognition (Bjorkman *et al.*, 1987b). The α1 and α2 domain are polymorphic, while the β_2m, α3 domain, the transmembrane and cytoplasmic regions are more conserved (Clark and Forman, 1984).

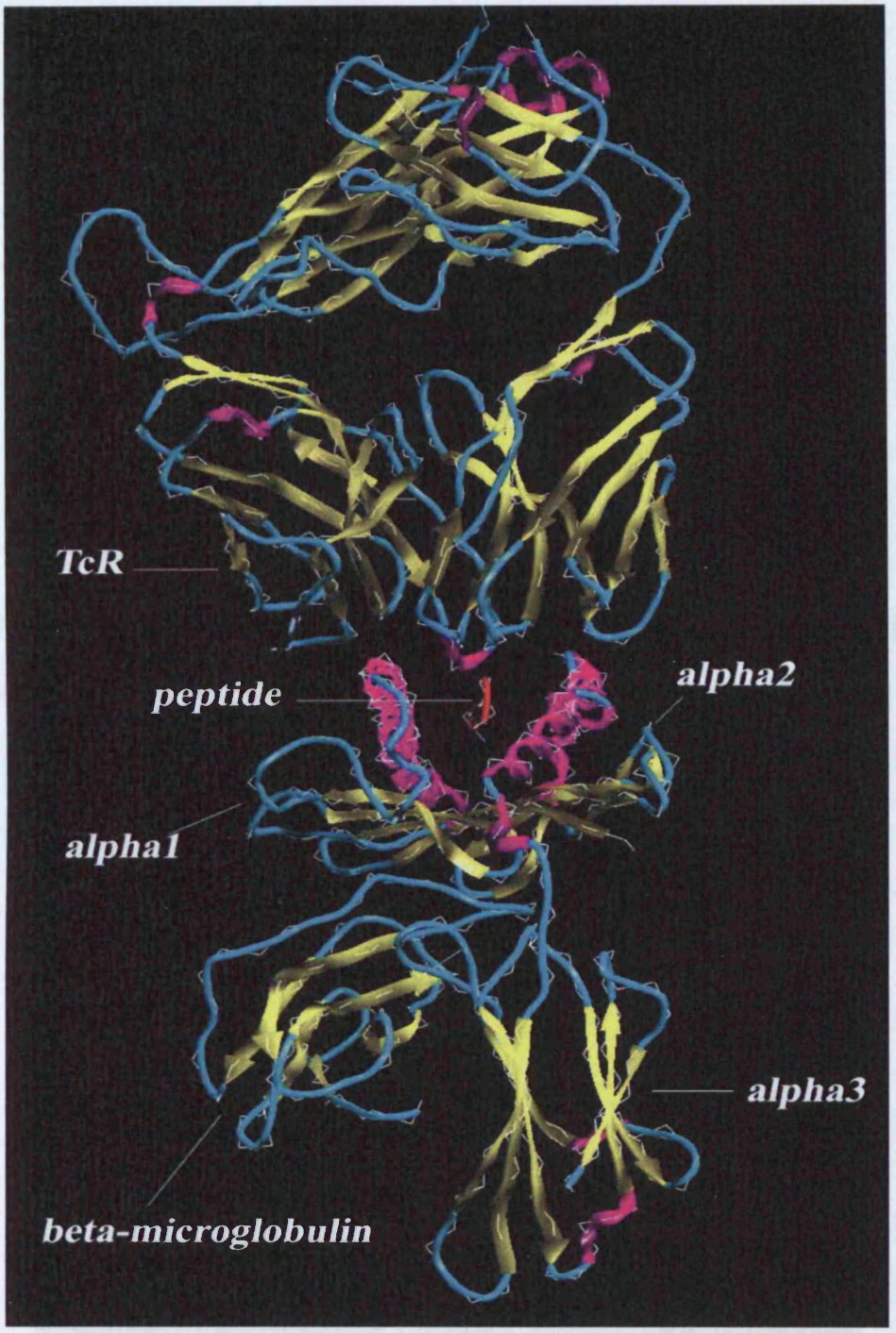


Figure 1.2. The crystal structure of human MHC class I allele A*0201 complexed with human T cell receptor B7. A viral peptide Tax is bound inside the peptide binding groove (Ding *et al.*, 1998).

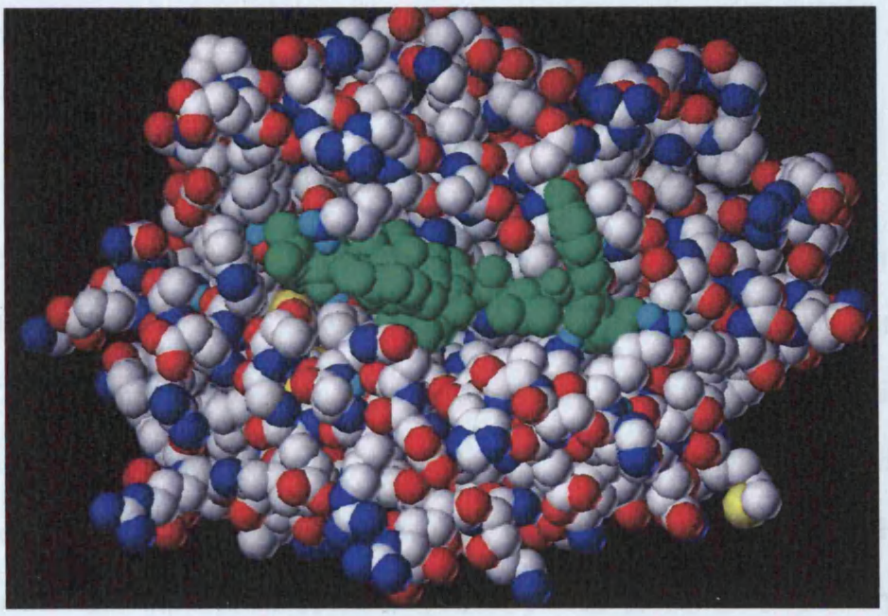


Figure 1.3. A spacefill graph of the peptide bound to the HLA-A*0201 binding site. The peptide is in green.

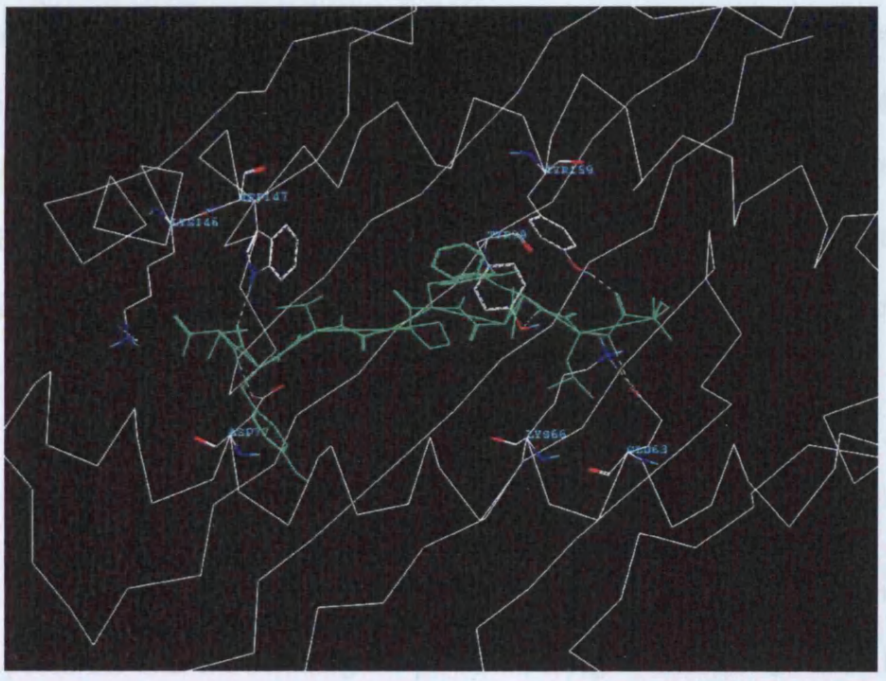


Figure 1.4. The major force for stabilising peptide is the hydrogen bond between the main chain atoms of the peptide and the A*0201 molecule. The hydrogen bonds are shown as dotted lines in the graph.

1.1.5 Peptide-MHC binding

1.1.5.1 The binding site

The peptide binding site is about 25-30Å long and 10Å wide (Bjorkman *et al.*, 1987a). The two ends of the class I HLA binding site are closed by the contacts of the side chains between the two helices (Saper *et al.*, 1991). The binding site is situated at the top of the MHC molecule and faces away from the cell, which facilitates recognition by T cell receptors (TCR) (Fig. 1.2). Most of the HLA residues that face the binding site are polymorphic (Saper *et al.*, 1991). In crystal structures, the N and C termini of the peptide are buried in the binding site and form hydrogen bonds with the residues of the HLA molecule (Hillig *et al.*, 2001). These hydrogen bonds are the main stabilisation force securing peptides in the binding site (Bouvier and Wiley, 1994) (Fig 1.4). It has been observed in thermal denaturation studies that substitution of the N and C terminal residues of the peptide accelerates the denaturation rate of the complex (Bouvier *et al.*, 1998b; Khan *et al.*, 2000). The centre of the peptide is more exposed to the solvent and has contact with the T cell receptor in some of the MHC-TCR structures (Ding *et al.*, 1998; Garboczi *et al.*, 1996).

1.1.5.2 T cell epitopes vs. MHC ligands

MHC molecules do not have the ability to distinguish between self and non-self peptides. As both self and antigenic proteins are degraded in the cytosol, a large proportion of MHC molecules bind to peptides from self proteins. Normally T cells that recognise self peptides are deleted by negative selection during maturation in the thymus (Janeway, 2001), therefore mature T cells only

recognise foreign epitopes. A T cell epitope is a linear peptide that can both bind to MHC molecules and induce T cell mediated immune response (Madden, 1995). T cell-epitope recognition and subsequent immune responses play an important role in host immune defence, changes in this process can lead to serious problems. Some viruses escape immune recognition through mutations that change their protein sequences, such as HIV (Letvin and Walker, 2001). Also in the case of autoimmune diseases, T cells recognise self peptides and destroy tissues in the host. An important application of epitopes is their use in epitope based vaccines, in which engineered epitope strings are injected into the body so that they can be recognised by T cells and induce immune response in the recipient. Details of epitope based vaccines are discussed in section 1.7.

1.1.5.3 Binding pockets and binding motifs

The optimal length of peptide binding to class I HLA alleles is nine residues but the binding of peptides with 8 - 15 residues have also been found (Rammensee *et al.*, 1995). In immunology, the residues of peptides are often denoted as P1, P2 ...P8, P9....., starting from the N terminus. Peptides with 8 or 9 residues are bound to MHC in an extended conformation (Madden *et al.*, 1991a) (Figure 1.5a, 1.5b). Structural studies showed that nonamer peptides bind to class I HLA molecules with a similar structural conformation in the binding site (Madden, 1995). The statistical pair-wise study of the structure of peptide in the MHC binding site and in native proteins confirmed that the structures of peptides in their native proteins vary from extended to helical, but this did not affect the peptide conformation in the binding site (Schueler-Furman *et al.*, 2001). In the study, there was a slight preference for an extended conformation of residue 8

and 9 in the native structures, which may facilitate the binding of the anchor residue in the pocket F.

The crystal structure of the A*0201/peptide complex defined six binding pockets in the binding site, termed A to F (Saper *et al.*, 1991) (Figure 1.6, 1.7). Binding pockets accommodate side chains of peptides and are important in both stabilising the peptide-MHC complex and determining peptide specificity. Pocket B, C, D and E are situated between the helices and the β sheet (Petrone and Garcia, 2004) (Fig. 1.6). The two pockets A and F, located at the two ends of the binding groove, are involved in the interactions with the side chains of the amino and carboxyl termini of the peptide, respectively (Carreno *et al.*, 1993). This interaction provides a major source of binding energy (Fremont *et al.*, 1992; Madden *et al.*, 1991b; Madden *et al.*, 1992; Matsumura *et al.*, 1992a; Silver *et al.*, 1992) and also decides the orientation of the peptide. Study of the HLA-A2/peptide complex identified 14 water molecules in the binding site (Petrone and Garcia, 2004). The water molecules are scattered in the binding pockets of the binding site, the water molecules both form the hydrogen bonds between the MHC and the peptide and fill the empty spaces in the binding groove.

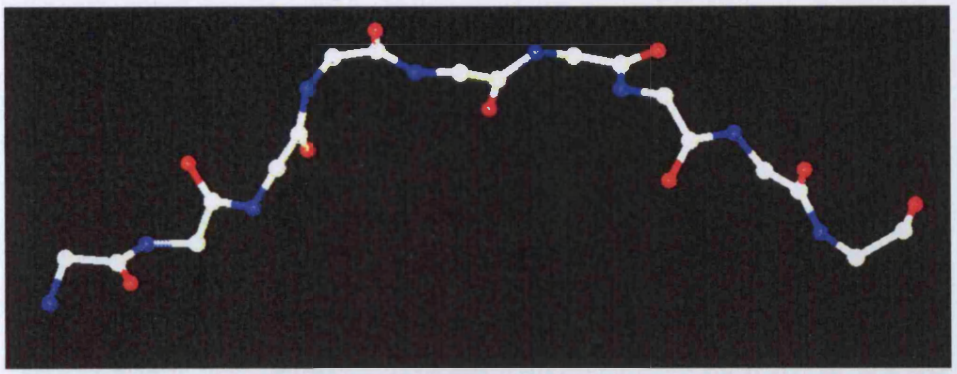
Binding pockets also help determine the peptide specificity of the MHC molecules (Johansen *et al.*, 1997; Schafroth and Floudas, 2004). The peptide specificity of H-2Kb was altered when the residues in the Pocket C Val9, Val97 and Ser99 were changed to bulkier amino acids from A*0201 Phe9, Arg97 and Tyr99 (Johansen *et al.*, 1997). Analysis of peptides eluted from MHC complexes showed that peptides bound to the same MHC molecule often contain the same

or chemically related amino acids at certain positions (Madden, 1995). These positions are called anchor positions (Deres *et al.*, 1993). The combination of anchor positions is named the peptide binding motif (Sette *et al.*, 2001). Most class I HLA motifs contain P2 and the carboxyl terminal residue (Sette and Sidney, 1999). The P2 residue for the class I HLA molecules is usually aliphatic: leucine, isoleucine, valine and methionine are the most commonly found anchor residues for alleles such as A*0101, A*0201, A*0202, etc (Altfeld *et al.*, 2001; Cerny *et al.*, 1995; Kurokohchi *et al.*, 1996; Yoon *et al.*, 1998). Basic amino acids (arginine, lysine) and aromatic amino acids (tyrosine, phenylalanine) have also been observed for alleles such as H-2Kd, H-2Kk, and A*24 (Burrows *et al.*, 1996; Jiang *et al.*, 2002) (Parker et al., 1995). In the crystal structure of HLA-A2/peptide complex (Hillig *et al.*, 2001), the P2 residue interacts with the side chains of the amino acids lining the pocket B and the P2 specificity is largely directed by the residues in the pocket (Altuvia *et al.*, 1997).

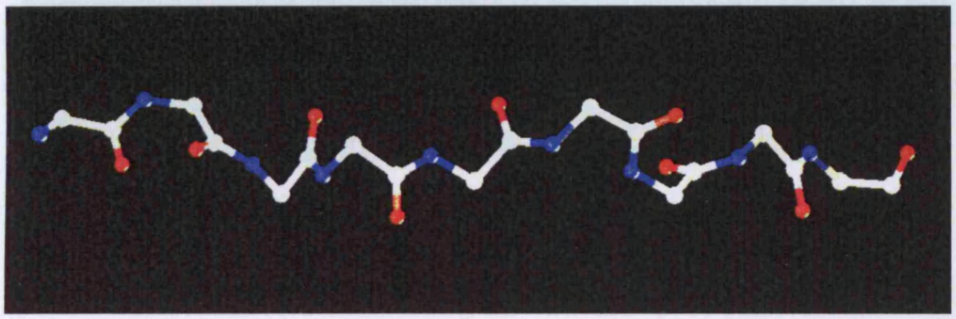
Another common anchor position is the C terminal residue (Rammensee *et al.*, 1995). Substitution of the carboxyl terminus significantly lowered the binding affinity of the peptide (Elvin *et al.*, 1991; Fahnestock *et al.*, 1994; Parker *et al.*, 1992a; Rohren *et al.*, 1993; Wettstein *et al.*, 1993). The nature of the residue is influenced by the residues in pocket F, in particular amino acid 116 of the MHC molecule, which is situated at the bottom of the pocket F (Saper *et al.*, 1991). For example, A*0201 has tyrosine at position 116 and the peptides bound to A*0201 usually have aliphatic amino acids at the C termini (Madden *et al.*, 1993).

Apart from anchor residues, other positions have also been identified to be important in high affinity binding (Falk *et al.*, 1991b). These positions are termed secondary anchor positions (Jameson and Bevan, 1992; Ruppert *et al.*, 1993). Substitution of secondary anchor amino acids also reduces the binding affinity of the peptide (Boehncke *et al.*, 1993). An example is the peptide stabilisation assay by Chujoh et al., in which the binding affinity of the mutated peptides to HLA-A*1101 were tested (Chujoh *et al.*, 1998). Peptides with the anchor residues (hydrophobic residue at P2, and lys at C terminus) were able to form a stable complex with HLA-A*1101. Peptides with mutated amino acids at P9 (lys was substituted to Asp, Glu and Thr) were unable to bind to the MHC molecule, but those with arginine at P9 were still able to form the peptide-MHC complex, although with less stability. Similar results were found for the other anchor positions. Peptides with Val, leu and Ile at P2 had higher affinity.

However, the binding motif can not fully explain the interactions between a peptide and an MHC molecule, as peptides with the same binding motif have different affinities, which indicates that the binding motif is not the only factor in peptide-MHC binding. In some experiments, H-2Db and H-2Kk binding peptides have been competitively inhibited by peptides that do not have the required motifs (Bodmer *et al.*, 1989). The interactions between anchor residues and the peptide binding site provide a stabilisation force to secure the peptide in the binding site. However, interactions between other residues and the binding site are also important and can influence binding affinity. In this thesis, the contributions of peptide residues to binding are further explored by two QSAR techniques: CoMSIA and the additive method.



(a)



(b)

Figure 1.5. Backbone structure of peptide TLTSCNTSV in the HLA-A*0201 binding site (Madden *et al.*, 1993). Peptide is bound to the binding site in an extended conformation. (a) Side view. (b) View from the top of the binding site.

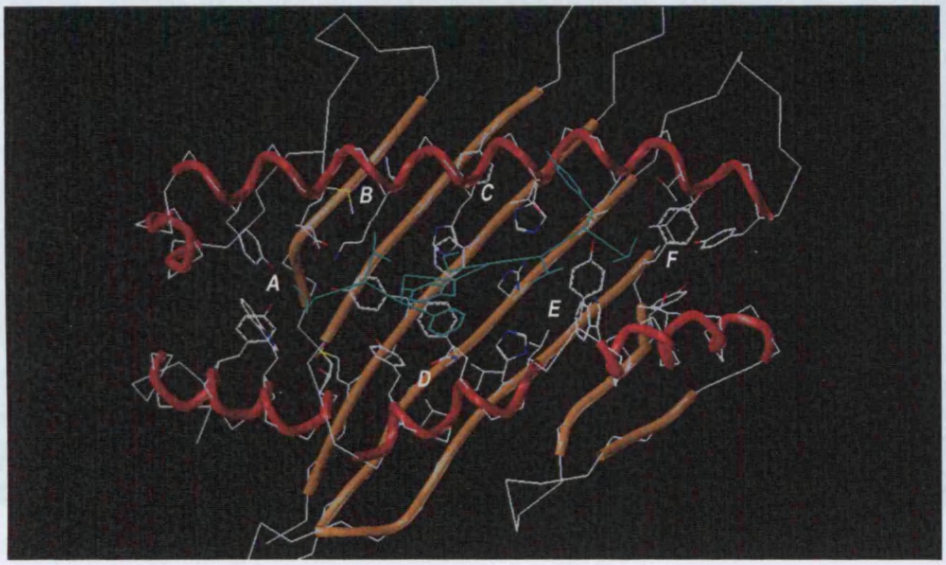


Figure 1.6. Peptide is bound in the binding cleft of A*0201 MHC molecule. For simplicity, only the α1 and α2 domains are shown. The α helices are shown in red, and β sheet is in brown. The peptide is in green. Six binding pockets, A to F, are indicated in the graph.

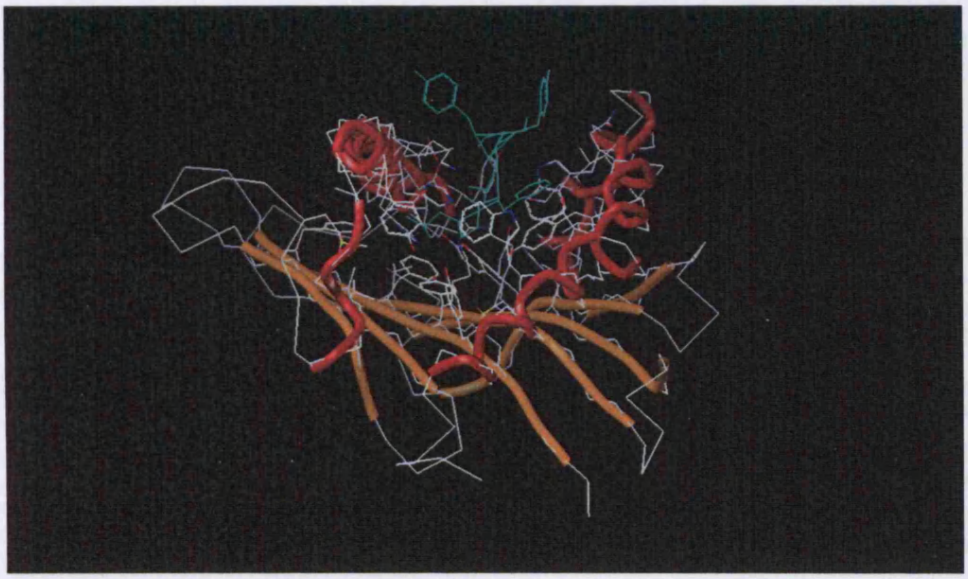


Figure 1.7. A side view of the A*0201 binding cleft. . For simplicity, only α1 and α2 domains are shown. The α helices are shown in red, and β sheet is in brown. The peptide is in green. The peptide binds to the MHC in an extended conformation. Most of the residues in the peptide interact with the binding cleft and stabilize the peptide inside the cleft, while side chains of other positions (position 5 in this graph) point towards the outside of the binding cleft and interact with the T cell receptor.

1.2 Techniques used in identifying MHC binders

Identification of peptides binding to MHC molecules is an important aspect of vaccinology. Techniques that have been used to identify eptiopes can be divided into two categories, experimental and computational.

1.2.1 Experimental methods

1.2.1.1 Pool sequencing

Traditionally, T cell epitopes are identified using pool sequencing (Falk *et al.*, 1991b), in which the peptides bound to the MHC molecules are separated and sequenced (Buus *et al.*, 1988). In the experiment carried out by Chicz et al., the MHC-peptide complexes were purified by immunoaffinity purification, and the purified complexes were then exposed to acidic conditions to elute bound peptides. The sequences of the peptides were analysed by Edman degradation. The percentage of peptides recovered from the experiment was estimated to be 70% - 80% of those bound to the recovered MHC molecules (Chicz *et al.*, 1993). Pool sequencing is able to identify the position of the primary anchor residues (Banks *et al.*, 1993; Cossins *et al.*, 1993; Falk *et al.*, 1991a; Pamer *et al.*, 1991). The process is good for examining peptide specificity of a particular MHC molecule and has been used for identifying tumour associated antigens presented by MHC class I molecules to specific CTLs (Traversari *et al.*, 1992).

1.2.1.2 Mass spectrometry

In mass spectrometry, individual peptides are separated by reverse-phase high-performance liquid chromatography (HPLC) and are sequenced by mass spectroscopy using collision activated dissociation analysis (Flad *et al.*, 2003; Gallego *et al.*, 2001; Hunt *et al.*, 1992; Hunt *et al.*, 1986; Papac *et al.*, 1994; Parker *et al.*, 1996; Peter and Tomer, 2001; Rotzschke *et al.*, 1991; Zhao and Chalt, 1994). Pascolo used mass spectrometry to test peptide binders predicted by computer programs. The peptide fragments are synthesised and incubated with the cells. A*0201 MHC-peptide complexes are isolated by adding HLA-A2 specific antibodies. The complexes are then fractionated by HPLC and analysed by mass spectrometry (Pascolo *et al.*, 2001). Mass spectrometry has also been used in analysing protein digestion by proteasomes (Emmerich *et al.*, 2000; Nussbaum *et al.*, 1998). The limitation of the technique is the amount of peptides required for reverse HPLC isolation.

1.2.1.3 Peptide binding studies

The peptide-binding assay has been developed to study the physicochemical properties of the interactions between peptides and MHC molecules (Joyce and Nathenson, 1994). Testing sets of overlapping peptides generated from a given protein sequence is a popular way of identifying both B cell and T cell epitopes (Goulder *et al.*, 1997; Hosoyama *et al.*, 1996; Schneider *et al.*, 1998; Schol *et al.*, 1998). The assay includes direct binding and the quantitative measurement of radio- or fluorescence-labelled peptides bound to the MHC molecules (Cerottini and Luescher, 1991; Chen and Parham, 1989; Christinck *et al.*, 1991; Kast and Melief, 1991; Levitsky *et al.*, 2000; Mendez-Samperio and Jimenez-Zamudio,

1991; Schumacher *et al.*, 1990; Stuber *et al.*, 1995; Wauben *et al.*, 1997). The MHC molecules can be either expressed on the cell surface or soluble (Fahnestock *et al.*, 1994; Matsumura *et al.*, 1992b; Ojcius *et al.*, 1992). Multiple steps of isolation and washing are required to isolate the bound peptides from the free ones (Manfredi *et al.*, 1993; Nag *et al.*, 1994). Binding affinity can be expressed as BL_{50} (concentration of peptides required for 50% of maximal binding) (Marshall *et al.*, 1994; van der Burg *et al.*, 1996), IC_{50} (concentration of peptides that is required for 50% inhibition of the standard peptide by the test peptide) (Chen *et al.*, 1994; Kono *et al.*, 1998; Kubo *et al.*, 1994; Rongcun *et al.*, 1999) or EC_{50} (plasma concentration needed for obtaining 50% of the maximum effect in vivo) values (Holt *et al.*, 2004; Monneret *et al.*, 2004; Suzuki *et al.*, 2004).

Peptide binding assays are the most common way of identifying T cell epitopes and measuring peptide binding affinities. Several databases have been set up to store peptide binding affinity data, such as MHCPEP (Brusic *et al.*, 1998a), MHCBN (Bhasin *et al.*, 2003) and AntiJen (Blythe *et al.*, 2002; McSparron *et al.*, 2003).

1.2.2 *In silico* methods

The experimental protocols for T cell epitope identification are both labour intensive and time consuming. Furthermore, they are often limited by the low quantity of peptides or the weak association between the peptide and the MHC molecules. Alternative non-experimental methods have been developed to

accelerate the epitope identification process. Non-experimental methods can be divided into two categories: the sequence approach and the structure approach.

1.2.2.1 The sequence approach

1.2.2.1.1 Motif search

The most commonly used method of epitope prediction is the use of motif patterns (Joyce and Nathenson, 1994; Pamer *et al.*, 1991; Sette *et al.*, 1989a; Suhrbier *et al.*, 1993). The concept of motif based algorithms is similar to that of the PROSITE database (Hulo *et al.*, 2004). PROSITE finds sequence regions in the test protein which are similar to conserved domains in protein families (Mondal *et al.*, 2003). The motif based prediction uses the motifs available in the literature and searches the input sequence against a library of known motifs (Rammensee *et al.*, 1999). Altuvia et al. identified class II mouse MHC binding motifs by studying binders and non-binders in the literature (Altuvia *et al.*, 1994). HLA-DR epitopes from *Plasmodium falciparum* have been identified using motif searching (Doolan *et al.*, 2000). A motif based program, EPIPREDICT, has been applied to predict class II epitopes associated with celiac disease (Jung *et al.*, 2001).

D'Amaro et al. developed a computer program MOTIF, which contains a collection of available A*0201 motifs (D'Amaro *et al.*, 1995). The program divides the query protein sequence into nonamers. Each position of the nonamer is compared with the motif, and the corresponding coefficient is added to the

total coefficient value. A higher coefficient indicates higher binding affinity. In the validation test, the program predicted 27 possible binders, 18 of which were identified as binders in an *in vitro* binding assay. The accuracy of the program was 61%. Another program EpiMer has been developed by the researchers of the TB/HIV laboratory in Brown University (De Groot *et al.*, 2001) and has been used to predict HIV epitopes (Meister *et al.*, 1995).

One of the most well known T cell epitope prediction algorithms, SYFPEITHI, is based on motif pattern searching (Dick *et al.*, 1998). Each residue in the input peptide is evaluated using the motif library. Peptides with predicted binding affinities less than 500nM are identified as potential epitopes (Rammensee *et al.*, 1999). Many epitope predictions are made by identifying the peptides containing the correct motifs first, followed by laboratory testing of their affinities (Amicosante *et al.*, 2002; Cossins *et al.*, 1993; Dong *et al.*, 2003; Hansson *et al.*, 2003; Liu *et al.*, 2004b; Neumann *et al.*, 2004; Pelte *et al.*, 2004; Suhrbier *et al.*, 1993; Ullenhag *et al.*, 2004; Wagner *et al.*, 2003; Zehbe *et al.*, 2003).

Motif based algorithms usually have a predictivity of 60-70%, as not all MHC binders contain the defined motif (Nussbaum *et al.*, 2003). In some cases, the correlation between predicted and experimental high binders was poor. In the study carried out by Anderson, the binding affinities of oncogenic and viral peptides were tested experimentally, and the results were compared with the predictions from SYFPEITHI and BIMAS. It was found that the algorithm predicted many false positives. Also some high binders were predicted to be non-binders by the algorithm (Andersen *et al.*, 2000)

1.2.2.1.2 Scoring matrix

Peptide binding motifs only define amino acids at certain positions and do not include information for the other positions, which are also important in binding (Margalit and Altuvia, 2003). Scoring matrix methods are similar to an expanded binding motif with coefficients for amino acid at each position of the peptide (Gulukota *et al.*, 1997). The method is based on the assumption that each amino acid contributes independently to the binding of the peptide, and the contribution is the same for a particular amino acid in different sequences (Brusic *et al.*, 1998b). Some studies also take into account interactions between amino acid side chains (Peters *et al.*, 2003; Segal *et al.*, 2001). Cano and Fan generated matrices for HLA-A and B alleles by the mathematical analysis of known MHC-peptide complexes (Cano and Fan, 2001). Another class I HLA prediction model was generated by defining the interactions between the peptide and residues of the MHC molecules in crystal structures. These residues formed the virtual binding pockets and binding of other peptides can be predicted by predicting the interactions between the peptide and residues in the virtual pockets (Zhao *et al.*, 2003a).

Matrix based predictions have been applied to class II alleles. Southwood et al. used the results from peptide binding studies to generate the DRB1*0401 model (Southwood *et al.*, 1998). Hammer et al. constructed a matrix for class II MHC alleles by analysing the side chains of peptides in the training set and used it in their predictions (Hammer *et al.*, 1994). Side chain scanning of combinatorial libraries has also been used to identify high affinity ligands (Dooley and Houghten, 1993; Pinilla *et al.*, 1992). The proteasomal cleavage prediction

program MAPPP is also based on quantitative matrices (Hakenberg *et al.*, 2003). Similar algorithms have been applied to predict linear B cell epitopes. Alix calculated the molecular properties such as hydrophilicity, side chain flexibility and surface accessibility for each of the twenty amino acids and used these scales to predict potential epitope regions in protein sequences (Alix, 1999).

The online prediction service BIMAS is probably the best known T cell epitope prediction algorithm based on quantitative matrices. BIMAS has been used in many experiments to identify potential epitopes (Hansson *et al.*, 2003; Lu and Celis, 2000; Ullenhag *et al.*, 2004; Vonderheide *et al.*, 1999). Developed by Park et al., BIMAS identifies potential epitopes by their predicted half-life dissociate rate of MHC-peptide complex (Parker *et al.*, 1992a; Parker *et al.*, 1992b; Silver *et al.*, 1991). The half-life dissociate rate is measured by the rate at which radiolabeled β2m dissociates from the MHC-peptide complex at 37°C (Parker *et al.*, 1992b). However, only the A*0201 model in BIMAS is based on half-time dissociate binding studies by Park et al., while models of other alleles are derived from motifs published in literature. Laboratory tests showed that BIMAS and SYFPEITHI are good at predicting epitopes within known T cell targets, but are less efficient in screening random proteins, that is, proteins that are not known to be T cell targets (Pelte *et al.*, 2004).

Another matrix based algorithm EpiMatrix has been developed De Groot and colleagues. The program has been used to identify HIV-1 antigens (Meister *et al.*, 1995; Schafer *et al.*, 1998). Two similar algorithms have also been developed. One is named ClustiMer, which searches for cross-presentation of peptides to

HLA superfamilies. The other program, Conservatrix, was generated to search for conserved regions across the isolates of the pathogen (Sbai *et al.*, 2001). Matrices have been generated using synthesised peptide libraries that reflect binding strengths of different amino acids at different positions (Lauemoller *et al.*, 2001). Pascolo et al. have used similar programs to identify tumour antigens encoded by the MAGE-A1 gene. The results were verified by mass spectrometry (Pascolo *et al.*, 2001). Sometimes models are obtained by aligning the known peptides and calculating the frequency of amino acid at each position, such as the position specific scoring matrices (PSSM) developed by Reche et al. (Reche *et al.*, 2002). Another example is the Gibbs sampling approach for class I and class II epitopes by Nielsen (Nielsen *et al.*, 2004).

A variation of the quantitative matrix algorithms, virtual matrices, has been generated by Sturniolo et al. Virtual matrices model the interactions of each amino acid and the binding pockets of the MHC binding site. Virtual matrices containing pocket specific binding information and can be applied to other alleles by MHC sequence comparison, while quantitative matrices have to be determined for individual alleles separately (Sturniolo *et al.*, 1999). A commercial program, TEPITOPE, is based on virtual matrices for the prediction of HLA-DR alleles. TEPITOPE has been applied to predict epitopes in tumour antigen MAGE-3 (Cochlovius *et al.*, 2000; Manici *et al.*, 1999). A free web based application, ProPred, has been developed by Singh and Raghava (Singh and Raghava, 2001), using the HLA-DR matrices from the pocket profile database maintained by Sturniolo.

Motif based predictions only consider the motif residues, while scoring matrices take into account the interactions between each residue of the peptide and the MHC binding site. The disadvantage of this approach is that a new matrix has to be generated every time new data is added. Also, the quality of the prediction is dependent on the number of peptides in the training set. Brusic et al. claimed that 150 peptides were required to derive an allele specific matrix with acceptable prediction accuracy and the ideal number of peptides in the training set was 600 (Brusic *et al.*, 1997). In real situations, many alleles have only 50 or less peptides verified experimentally and some alleles have none. Therefore it is difficult to generate matrices for all the alleles.

1.2.2.1.3 Artificial neural network

Artificial neural networks (ANN) are good at dealing with nonlinear data (Beale and Jackson, 1990). ANN has been applied to solve many biological problems including asthma (Tomita *et al.*, 2004), heart disease (Stefaniak *et al.*, 2004), drug solubility (Jouyban *et al.*, 2004) and peptide prediction and analysis of MHC haplotypes (Bellgard *et al.*, 1998). Because the length of the peptides varies, the training data used to build an ANN model is usually aligned by the anchor residues. The task is simple for MHC class I alleles as the length differences among the peptides is small, but it is more complicated for class II alleles where the length variation is much larger.

In ANN prediction, peptides with known binding affinities are used as the training set to train the ANN. The peptides are aligned in a matrix (Brusic *et al.*, 1998b). The ANN contains computer nodes (elements) that can extract and

remember the patterns in the matrix and recognise them in the test set. ANN has been applied to predict A*0201 peptide binding affinities using 552 A*0201 nonamers and 486 decamers as the training set. The network achieved a correct prediction rate of 0.78 (Adams and Koziol, 1995). Brusic and colleagues used ANN to predict MHC class I and II epitopes with an accuracy of 50-60% (Honeyman *et al.*, 1998). Other applications of ANN include class I MHC binding peptides prediction by Milik et al. (Milik *et al.*, 1998) and HLA-DR models generated by Bisset and Fierz (Bisset and Fierz, 1993). Two online proteasome cleavage site prediction services, NetChop and ProPrac, also use ANN models in their prediction (Kesmir *et al.*, 2002; Nussbaum *et al.*, 2001).

Some applications combine ANN with other algorithms to build predictive models (Gulukota *et al.*, 1997). The other algorithms are usually used in the selection of training data set which the ANN used to generate models, it is especially useful for alleles that have not been studied extensively. For example, Brusic et al. used an evolutionary algorithm and an ANN to predict HLA class II binding peptides (Brusic *et al.*, 1998b). The evolutionary algorithm generated alignment matrices using the training set and the aligned peptides were used to produce the final model. In another study, Buus et al. used previously defined peptide specificity matrices to scan the SWISS-PROT database and identified peptides that are potential binders for the A*0204 allele data set (Buus *et al.*, 2003).

1.2.2.1.4 Hidden Markov model

The hidden Markov model (HMM) is one of the probabilistic discrete dynamic system models. It uses a set of defined states to describe possible states of the modelled system (Mamitsuka, 1998). Some of the states can be observed while some can not, therefore they are 'hidden'. When dealing with biological problems, HMM usually generates a series of states that are in sequential order. Any state in a HMM is dependent on the previous ones. The probability of moving from one state to the next can be calculated. A variation of the HMM model, profile HMM (Eddy, 1998), has been applied to proteomics, such as prediction of coiled-coil domains (Delorenzi and Speed, 2002), transmembrane regions within protein sequences (Liu *et al.*, 2003; Martelli *et al.*, 2002) and protein homology analysis (Qian and Goldstein, 2004). HMM is extensively used in protein sequence alignments (Krogh *et al.*, 1994), examples of HMM applications are the online protein domain family identification service Pfam and the database SMART maintained at the EBI (Bateman and Haft, 2002). HMM is also used in genomics to study gene splicing (Cawley and Pachter, 2003), analysis of phelogenetic trees (Jojic *et al.*, 2004) and identifying genes in prokaryotic genomes (Azad and Borodovsky, 2004).

HMM has been applied in peptide predictions. Profile HMM has been used to build models for signal peptide predictions (Zhang and Wood, 2003). Mamitsuka built HMM models of A*0201, DR1 and DR1 alleles. The training data set was taken from the MHCPEP database and the model had high level of sensitivity (>90%) (Mamitsuka, 1998). Based on Mamitsuka's approach, Udaka et al. used a committee based HMM model to predict peptides binding to class I MHC alleles

(Udaka *et al.*, 2002). Brusic used HMM to predict peptides binding to the HLA-A2 family (Brusic *et al.*, 2002). Only amino acids inside the binding site that contact the bound peptide were included in his study. An HMM model was built for each allele. Each model was trained using a peptide set that includes peptides binding to all other HLA-A2 alleles. For example, an A*0201 model was trained using peptides bound to A*0202, A*0203, A*0204, A*0205, A*0206, A*0207, A*A0209 and A*0214 allele. The test for each model includes all peptides that are known to bind the allele, with both T cell ligands and epitopes included. The accuracy of the prediction is measured by ROC analysis. The A*0201, A*0204 and A*0205 models have high predictivity (Aroc > 0.9), while the predictivity of some of the models is low, such as A*0202.

Schonback and his colleagues compared the performance of the machine learning methods (Schonbach *et al.*, 2000). In his work, scoring matrices, ANN and HMM models were used to screen more than 500 sequences of HIV-1, -2 protein (Gag, Env and Pol) for A*0201 and B*3501 epitopes. It was also found that the ANN model for A*0201 showed high accuracy, and the HMM model was good at predicting B*3501 peptides. Subsequent experiments showed that about 26% of epitopes were correctly predicted by both scoring matrices and ANN models. In the same experiment, scoring matrices techniques showed better performance than HMM model and predicted more epitopes.

1.2.2.1.5 Support vector machines

Support vector machines (SVM) were developed by Vapnik in the 1970s (Vapnik, 1998) and were originally used in pattern recognition and data

classification (Ding and Dubchak, 2001). SVM belongs to the group of kernel based methods (Scholkopf *et al.*, 1999). SVM classifies the data by creating a hyperplane in the space where the data is, and separates the data by calculating the distance between the data points and the plane (Fig 1.8). An implementation of SVM, the software SVM^{light} developed by Thorsten Joachims is widely used in SVM applications. SVMs have been used in many areas such as study of radio frequency fields (Maby *et al.*, 2004), enzymes and protein binding sites, analysing digital images (Cai *et al.*, 2004; Chen *et al.*, 2004) and fluorescence spectra (Lin *et al.*, 2004), etc.

SVM has been used to predict eukaryotic protein subcellular locations (Bhasin and Raghava, 2004b), subfamilies of G-proteins (Bhasin and Raghava, 2004c), membrane proteins (Wang *et al.*, 2004), study gene functions (Vinayagam *et al.*, 2004) and DNA arrays (Williams *et al.*, 2004), classifying nuclear receptors (Bhasin and Raghava, 2004d) and predicting T cell epitopes (Bhasin and Raghava, 2004f; Bhasin and Raghava, 2004e). Users of SVM claim that the method is especially suitable for multivariate data when the number of objects is small compared to the number of variables. Also SVM can avoid over-fitting the training data which often limits other machine learning methods (Zhao *et al.*, 2003c). SVM has been applied to generate models for 26 class I HLA molecules and given satisfactory results (Donnes and Elofsson, 2002; Zhao *et al.*, 2003c). Machine learning methods have been used together with scoring matrices to predict T cell epitopes. Bhasin and Raghava used quantitative matrices, SVM and ANN to predict T cell epitopes, the accuracy of the three methods was over 70% (Bhasin and Raghava, 2004a).

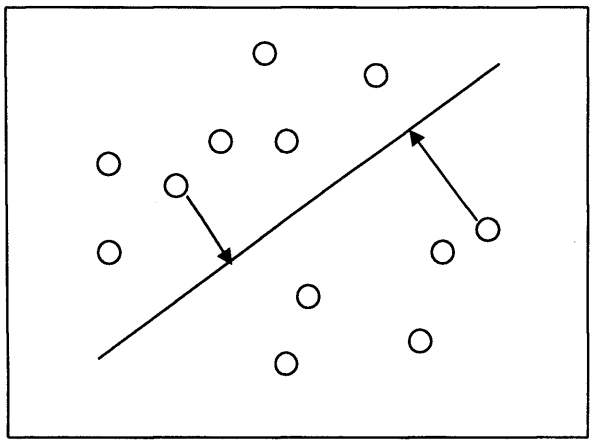


Figure. 1.8 A schematic diagram of the SVM theory. SVM creates a hyperplane in the space where the data is, and separates the data according to the distance between each data point and the plane. For example, data points with positive values are separated from data points with negative values.

SVM has been applied with other algorithms. Liu et al. built quantitative structure-property relationship (QSPR) models of 35 amino acids (Liu *et al.*, 2004a). The descriptors used in the analysis were selected by genetic algorithm coupled with partial least squares (GA-PLS), and were used as the input of the SVM calculation to predict activities (the isoelectric point) of the amino acids. The model had a prediction correlation coefficient of 0.970 with the root-mean-square error of 0.238. SVM has also been used with statistical methods such as least squares to generate QSPR models (Thissen *et al.*, 2004) in other studies. SVM generated models had good predictivity compared with models generated by multiple linear regression and neural networks (Xue *et al.*, 2004a; Xue *et al.*, 2004b; Xue *et al.*, 2004c; Yao *et al.*, 2004).

1.2.2.2 Structural approach

1.2.2.2.1 Threading

Adrian studied the atomic interactions at the interface between the crystallised MHC molecules and peptides and found that interactions between MHC side chains and peptide backbones contribute significantly to binding (Adrian *et al.*, 2002). His study also showed the importance of having correct peptide backbone conformations in prediction. Altuvia et al. used the threading technique to predict T cell epitopes. In their study, the input peptide is threaded through the backbone of the peptide template, which is taken from the crystal structure of peptides binding to the MHC binding site (Altuvia *et al.*, 1995; Margalit and Altuvia, 2003). The binding energy is calculated for each residue of the peptide, that is, the sum of the interaction energy between the residue and all the surrounding

MHC residues (Altuvia *et al.*, 1997). The binding energy of the peptide equals the sum of binding energy of each amino acid, a lower binding energy indicates high affinity binding (Schueler-Furman *et al.*, 2001; Schueler-Furman *et al.*, 1998). A drawback of the technique is that although the anchor residues of the template and the input peptide are super-imposable and some of the side chains were in the same orientation, there are still some side chains that point in different directions. Therefore further side chain modelling is required for precise measurement of peptide-MHC interactions (Schueler-Furman *et al.*, 2000).

1.2.2.2.2 Binding energy and molecular dynamics

A recent development in epitope prediction is the calculation of the binding free energy. Free energy is the difference between the free energy of the free peptide and the free energy of the peptide bound to the MHC (Sezerman *et al.*, 1996; Zhang *et al.*, 1997). Good binders can be found by directly comparing the free energy between two peptides, by estimating energy values using scoring functions, or by molecular dynamics (MD) simulations (Meng *et al.*, 2000). Examples of MD simulations include studying the binding of synthetic peptides (Scapozza, 1995), MHC-peptide complexes (Mata *et al.*, 1998; Rognan *et al.*, 1994), the contribution of water molecules to peptide binding (Petrone and Garcia, 2004), the interactions between A2 peptides and the binding site (Meng *et al.*, 2000; Pohlmann *et al.*, 2004), peptide dissociation (Binz *et al.*, 2003), and interaction between TCR and MHC-peptide complex (Michielin and Karplus, 2002). Rognan et al. simulated the binding of six peptides to B*2705 proteins and suggested the importance of secondary anchor residues in binding (Rognan *et al.*, 1994). MD simulation has been used in A*0201 peptide binding prediction

(Lim *et al.*, 1996), in which the binding of peptides to MHC was modelled by MD using crystal structures as templates and good binders were validated by binding experiments. In another study, MD simulation was used to identify properties of each position of the peptide binding to A*0201 and the results used to generate a quantitative matrix for prediction (Zeng *et al.*, 2001). Similar MD simulations were applied to search for the anchor amino acid preferences of the A*0217 molecule (Toh *et al.*, 2000). MD simulations have also been used to study peptides bound to DRB1 alleles (Androulakis *et al.*, 1997). Davies et al. built class II MHC peptide prediction models using a simulated annealing approach, in which the global energy minimum of existing crystallised class II MHC-peptide complex was obtained by increasing the temperature of the complex steeply and then gradually removing kinetic energy. After annealing, the interaction energy between the MHC binding site and the peptide was calculated and used to predict the binding affinities of other peptides (Davies *et al.*, 2003).

Another often used method in calculating the binding energy is the partitioning approach. Schapira obtained the ligand binding energy by calculating the difference between energy of the solvated complex and that of the solvated receptor and the ligand. The forms of energy considered were the hydrophobic and electrostatic forces, and the energy difference between the solvated and the reference state (Schapira *et al.*, 1999).

Most of the MHC-peptide interaction predictions use energy scoring functions. The advantage of this method over other structural methods is that it can give a

better description of the relationship between the peptide and the surrounding MHC residues (Logean *et al.*, 2001). Sezerman et al. generated free energy maps of the class I MHC binding sites using the electrostatic energy, the solvation free energy and the side-chain conformational entropy (Sezerman *et al.*, 1996). Froloff et al. calculated the binding energy of eight class I MHC-peptide complexes based on electrostatic and non-electrostatic interactions (Froloff *et al.*, 1997). Schapira et al. divided the total binding energy into three terms: the entropic, electrostatic and hydrophobic potentials and predicted the binding energy of small protein complexes (Schapira *et al.*, 1999).

Free energy calculation has been applied to predict HLA-A*0201 epitopes (Rognan *et al.*, 1999). In Rognan's experiment, the total free energy comes from five sources: the contribution of hydrogen bonds between the peptide and the MHC molecule, the interaction of lipophilic atoms, the loss of entropy from freezing rotational bonds upon binding, the negative contribution from the contact between the lipophilic and polar atoms, and finally, the energy required for transferring the atom from one continuum dielectric to another, such as from vacuum to water. In another experiment, Rognan used a new method (Fresno) to predict the free energy of peptides. Five HLA-A*0201 restricted peptides with both crystal structure and experimental binding affinities known were used in the training set. The model was used to predict the binding energy of 26 peptides to a HLA-A*0201 related allele, A*0204. The performance test showed that it was more accurate when a crystal structure of the MHC molecule was available. The algorithm has also been applied to estimate the binding energy of A*0201 and B*2705 peptides using existing crystallised structures as templates (Logean *et al.*,

2001). Later, Fresno was incorporated into the computational algorithm EpiDock, which builds the structure of a MHC-peptide complex by homology modelling and the scoring function Fresno was used to calculate the binding energy (Logean and Rognan, 2002).

However, a wide application of the binding energy method is hampered by the difficulty in predicting absolute binding free energy, also the intensive calculation it requires is a problem for wider applications, such as internet implementation.

1.2.2.2.3 Peptide docking and library screening

In the recent years many techniques used in the pharmaceutical industry have been used in biological research, such as combinatorial library screening and docking. Davenport et al. generated class II MHC models by scoring each amino acid according to the abundance of the amino acid at each position in the library in relation to the peptide binding affinity (Davenport *et al.*, 1995). New peptides binding to DRB1*0101 have also been designed according to the amino acid specificities of the peptide library (Fleckenstein *et al.*, 1996). Library screening has also been applied to other MHC alleles. Stryhn et al. analysed amino acid specificity of peptides binding to class I MHC alleles using peptide libraries (Stryhn *et al.*, 1996). Stevens used random peptide libraries to study the preferred peptide length of mice MHC alleles (Stevens *et al.*, 1998). Udaka et al. characterised specificities of peptides binding to H-KB, Db and Ld alleles by positional scanning of combinatorial peptide libraries (Udaka *et al.*, 2000; Udaka *et al.*, 1995). The frequencies of amino acids with different chemical properties

were obtained through scanning and the values were stored in a scoring matrix and were used in epitope prediction. Their prediction was 80% accurate within the test set. Similar studies has been carried out by Sung et al. (Sung *et al.*, 2002) and Nino-Vasquez et al. (Nino-Vasquez *et al.*, 2004).

Protein docking is often used in ligand design in the pharmaceutical industry (Vajda and Camacho, 2004). It has recently been applied to design peptides binding to MHC molecules. Early attempts focused on class I peptides (Rosenfeld *et al.*, 1995; Sezerman *et al.*, 1993). In the study by Zeng et al., chemical functional groups were used, each representing different properties (polar, non-polar, charged and so on). The chemical groups were docked into the peptide binding site to find the best property/residue favoured at each position of the peptide, generating estimated high binders for the HLA allele (Zeng *et al.*, 2001). In another study, Del Carpio used a genetic algorithm with peptide profile analysis to find the optimised matrix table for A2 and A24 alleles (Del Carpio *et al.*, 2002). Predicted good binders enter the second phase of analysis, where their structures were modelled and the peptides were docked into the peptide binding site. The MHC-peptide binding interface was obtained and the electrostatic and hydrophobic energy was calculated. In the study, the predicted good binders relate well with their experimental affinity.

Docking has been applied to class II MHC binding peptides to identify anchor residues and solvent exposed residues in long peptide fragments (Tzakos *et al.*, 2004). TCR structures have been docked to the MHC complex in order to study TCR-MHC interactions (Buslepp *et al.*, 2003; Wu *et al.*, 2002b). Tong et al.

developed a new docking technique which involved three steps: 1. docking of peptide terminal residues to the binding site. 2. loop closure of the remaining peptide backbone. 3. refinement of the backbone and side chains. The method was reported to be more accurate in studying class I and II MHC molecules than existing methods (Tong *et al.*, 2004). Liu et al. also took into account the MHC molecule flexibility in docking experiments (Liu *et al.*, 2004c). However, it is difficult to apply docking on a wider scale such as online prediction services, since it is CPU intensive and can only analyse a few peptides at a time. Moreover, the prediction is dependent on the resolution of the peptide binding site structure, and the accuracy of the peptide structure prediction.

1.3 MHC-TCR interaction

T cells recognise epitopes through interactions between T cell receptors (TCR) and the MHC-peptide complex. TCR molecules are membrane bound glycoproteins. Most TCR molecules consist of two polypeptide chains α and β (de la Hera *et al.*, 1991). A small percentage of TCR molecules have γ and δ chains. The γδ T cells are expressed predominantly in the skin, intestinal epithelium and pulmonary epithelium (Hampl *et al.*, 1999). The function of γδ T cells is different from αβ T cells. The exact function of γδ T cells are not clear, but it is known that they are able to recognise antigens directly (Mukasa *et al.*, 1999). In this section, only αβ T cells will be discussed. The αβ T cells can be divided into two subsets depending on which of the two glycoproteins, CD8 and CD4, is expressed on their surface. The CD8 and CD4 T cells identify antigenic fragments presented by class I and II MHC molecules, respectively. TCR is associated with the CD3 complex, which consists of four invariant polypeptide

chains two ε, one δ and one γ (Feito *et al.*, 2002). CD3 is synthesised co-ordinately with the TCR. The function of CD3 is to help transport TCR to the cell surface and sends activating signals to the T cell when the TCR recognises MHC-peptide complexes (Gouaillard *et al.*, 2001).

There are several TCR crystal structures available (Bentley *et al.*, 1995; Fields *et al.*, 1995; Garboczi *et al.*, 1996; Garcia *et al.*, 1996a). The structure of a TCR binding to the A*0201/Tax peptide complex is shown in figure 1.9. The TCR is structurally similar to the immunoglobulin Fab fragments (Garcia *et al.*, 1996a), each polypeptide chain has a variable (V) region at the N-terminus and a constant (C) region at the C-terminus (Wilson and Garcia, 1997). The TCR proteins are produced by gene rearrangement as are immunoglobulins (Arden *et al.*, 1995). The α chain is formed by the rearrangement of the variable (V) to the joining (J) segment, and the β chain is produced by the rearrangement of the variable (V), diversity (D) and joining (J) genes (Krangel *et al.*, 2000). The rearranged genes are attached to the constant (C) gene to form the complete α and β chains. The contact site between the TCR and the MHC complex is in the V region, formed by peptide loops (figure 1.10). There are four hypervariable regions on α and β chains, three of which (CDR1, CDR2, CDR3) resemble the complementarity-determining regions (CDRs) of immunoglobulins (Garboczi *et al.*, 1996). These hypervariable regions form the contact site between the TCR and the peptide-MHC complex. CDR3 is the most variable and is considered to be responsible for TCR specificity. CDR3 has contact with P5 to P8 of the peptide in the crystal structure (Garcia *et al.*, 1996a). Mutations on the CDR3 loops can abolish MHC-peptide recognition (Engel and Hedrick, 1988). The crystal structure of the TCR

recognising H-2Kb complex shows that CDR1 loops contact the N terminal residues of the peptide, and CDR2 loops cover the C terminus of the peptide (Garcia *et al.*, 1996a). The same binding orientation is found in other TCR-A2 complexes (Garboczi *et al.*, 1996). The constant region comprises the transmembrane region and the cytoplasmic tail that stabilises the TCR on the membrane (Wilson and Garcia, 1997). The two chains are linked by a disulphide bridge at the hinge region connecting the C region and the transmembrane region. Crystal structures of TCR showed that the V regions are similar to the Fab part of the antibody and adopt the immunoglobulin fold with two β sheets packed tightly against each other. (Wilson and Garcia, 1997). The crystallised structure of a TCR complex with class I MHC shows that the TCR-MHC binding surface is not parallel, but is about 20-30° towards diagonal (Garcia *et al.*, 1996a) (figure 1.11). A hydrophobic pocket was formed above the binding site between residue 93-104 of the α chain and 95-107 of the β chain, which could accommodate a side chain of the peptide (Garcia *et al.*, 1996a).

TCR recognises class I MHC-peptide complex with the help of several co-receptors (Bjorkman *et al.*, 1987b). The co-receptors are invariant and are involved in the interactions between TCR and MHC-peptide complex, the so-called immunological synapse (IS) (Creusot *et al.*, 2002; Dustin, 2002). An important co-receptor is the leukocyte function-associated antigen (LFA) -1, which recognises the intercellular adhesion molecule (ICAM) -1 (Goldstein *et al.*, 2000). In one experiment, MHC and ICAM-1 were put on a planar bilayer and the reactions between the TCR and MHC were monitored (Grakoui *et al.*, 1999). The binding of LFA-1 and ICAM-1 was the stop signal for T cells, after which

the T cell stopped migrating and attached itself to the MHC membrane bilayer, which marked the start of the IS (Dustin *et al.*, 1997). The nature of the signal is not clear but may involve chemokines. The engaged TCRs were then translocated to the centre of the interaction area after several minutes, surrounded by bound LFA-1/ICAM-1 (Dustin, 2002). The situation could be stable for hours during which TCR recognised the peptide presented by the MHC molecule and transduced signals to trigger T cell activation.

Figure 1.9. The crystal structure of TCR complexed with HLA-A*0201 and viral peptide Tax (Garboczi *et al.*, 1996). The TCR α chain is in red, β chain is in green. The HLA α chain is in yellow, β2-microglobulin is in orange. The peptide is in purple.

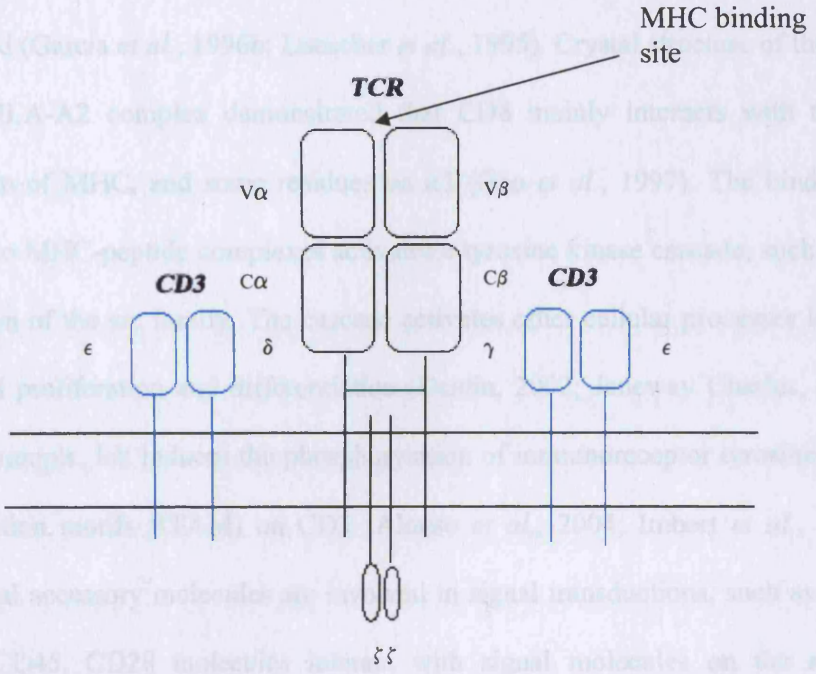


Figure 1.10. The structure of TCR and co-receptor CD3 molecules. The TCR consists of two variable domains Vα and Vβ which also contain the MHC binding site, and two constant domains Cα and Cβ. A disulphide bond (in green) stabilises the TCR structure. The ζ domains are involved in intracellular signal transmission. The CD3 complexes are co-expressed with the TCR molecules. CD3 helps transport the TCR molecules to the cell surface and sends activation signals to the T cells when the TCR recognises the MHC complex.

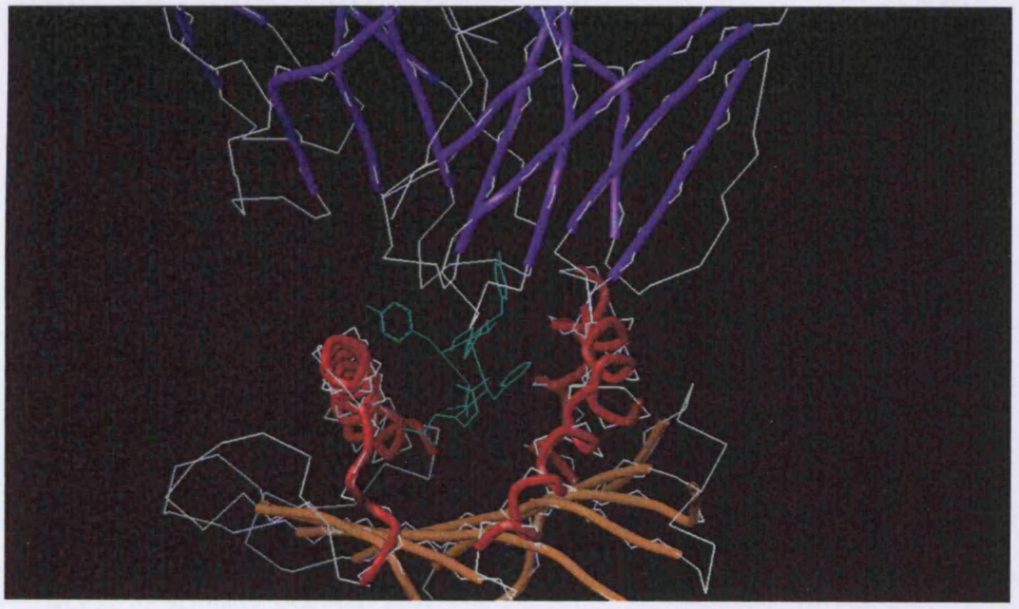
Stable binding of TCR to class I MHC-peptide complexes requires the help of CD8 molecules (Davis *et al.*, 1998). It has been demonstrated that the presence of CD8 on live T lymphotypes increased the affinity of TCR/MHC complex by 10-fold (Garcia *et al.*, 1996b; Luescher *et al.*, 1995). Crystal structure of the CD8 and HLA-A2 complex demonstrated that CD8 mainly interacts with the α3 domain of MHC, and some residues on α2 (Gao *et al.*, 1997). The binding of TCR to MHC-peptide complexes activates a tyrosine kinase cascade, such as lck and fyn of the src family. The cascade activates other cellular processes leading to cell proliferation and differentiation (Dustin, 2002; Janeway Charles, 2001). For example, lck induces the phosphorylation of immunoreceptor tyrosine-based activation motifs (ITAM) on CD3 (Alonso *et al.*, 2004; Imbert *et al.*, 1996). Several accessory molecules are involved in signal transductions, such as CD28 and CD45. CD28 molecules interact with signal molecules on the antigen presenting cell and activates the T cell (Beecham *et al.*, 2000; Hombach *et al.*, 2001). CD45 catalyses and activates tyrosine protein kinases by dephosphorylation (Koretzky *et al.*, 1993; Ross *et al.*, 1994; Stone *et al.*, 1997). For antigen presenting cells that express strong co-stimulatory molecules on their surfaces, such as dendritic cells, the bound T cell can be directly activated, producing Il-2 and other cytokines, which in turn activates the T cell itself to proliferate and differentiate into cytotoxic CD8 T cells, killing other infected cells (Dai *et al.*, 2000). The secretion of IL-2 has been shown to suppress the expression of TCR, CD3 and CD8 to avoid activation of non-specific T cells (Kambayashi *et al.*, 2001). For antigen presenting cells that only weakly express co-stimulatory molecules, the T cell will be activated if CD4 T cells bind to the same cell. The CD4 T cells either express co-stimulatory molecules to activate

CD8 T cells, or they produce Il-2, which is required for CD8 T cell proliferation. The signals that turn off T cell activation are not well understood. Possible agents are phosphatases and attachment of ubiquitin and degradation of the TCR α chain by the proteasome (Dittel *et al.*, 1999; Liu *et al.*, 2000).

1.4 Antigen degradation, transport and recognition

Class I HLA molecules recognise degraded intracellular protein fragments. Intracellular peptide fragments are from two sources: self-peptides and antigenic peptides. Intracellular proteins are degraded at a fast rate, including some newly synthesised proteins, producing large amounts of short peptides. On the other hand, external agents such as viruses invade the body and produce viral proteins, which are degraded by the host in a similar way to self-proteins. Intracellular protein degradation is carried out by a complex called the proteasome. After peptides are generated, they are translocated into the ER lumen by the transporter associated with antigen processing (TAP). TAP also has the ability to interact with peptide-free class I HLA molecules in the ER. After peptides associate with class I HLA molecules, the HLA molecules are released from TAP and are transferred to the cell surface.

a



b

Figure 1.11. The HLA class I allele A*0201 binds to the viral peptide (Tax) and presents it to the T cell receptor (Garboczi *et al.*, 1996). The viral peptide is in light green. a. A detailed image showing the interaction between the peptide-MHC complex and the TCR. The peptide is coloured in green. The MHC α helix is shown in red, and β sheet is in brown. The TCR is in purple. b. Side view of TCR binding to peptide-MHC complex. T cell receptor consists of two polypeptide chains α and β, which are coloured red and green in the graph. T cell receptor binds diagonally cross the peptide-binding groove. The Vα domain contacts the amino terminus of the peptide, while the Vβ domain contacts the carboxy-terminus of the peptide and some surrounding MHC residues.

1.4.1 Peptide generation

A small polypeptide called ubiquitin attaches to the protein before it can be recognised and degraded by the proteasome. ATP and ubiquitin-activating enzymes are required for attachment (Townsend *et al.*, 1990). Ubiquitin is attached to the protein by binding to an exposed lysine residue on the protein.

Several different forms of the proteasome are found in the cell. The important ones are the 20S core proteasome, the ATP-stimulated 26S proteasome and the immunoproteasome (Song and Harding, 1996). The most basic form is the 20S proteasome. The mammalian 20S proteasome is a large protein complex consisting of 28 copies of α and β sub-units, 14 each (Lowe *et al.*, 1995), and is responsible for protein degradation in both the cytosol and nucleus (Baumeister, 1998). The crystallised structure of the proteasome reveals that the sub-units are arranged in four rings stacked on top of each other, forming a cylindrical structure (figure 1.12). The sub-units are arranged in α7β7β7α7 order. The active sites are found on the inner surface of the cylinder on the β subunit (Unno *et al.*, 2002). The α rings form the gate through which unfolded polypeptides enter, they are also the binding site of proteasome activator 28 (PA28) (Sun *et al.*, 2002; Yamano *et al.*, 2002). The structure of proteasome is similar to the bacterial chaperonin GroEL, both have a cylinder like structure and both have active sites inside the cylinder (Chen and Sigler, 1999). However, the access to the inner compartment of the proteasome is guarded by a 19S regulator, allowing only completely unfolded and ubiquitin-bound protein to enter (Kloetzel and Ossendorp, 2004).

The exact mechanism of proteasome cleavage is unclear. The position of the peptide in the protein and the adjacent sequences influence proteasome cleavage. In one experiment, the nonamer murine cytomegalovirus epitope failed to be cleaved when inserted into the hepatitis B virus protein, but was recognised when a poly-alanine peptide was inserted next to it (Del Val *et al.*, 1991). It has been discovered that the active sites have different specificities for the P1 residue of the peptide. Some of the established activities are: trypsin like property (recognises and cleaves basic residues), chymotrypsin like activity (cleaves after hydrophobic residues) and peptidyl-glutamyl-peptide hydrolyzing activity (cleaves acidic residues) (Dick *et al.*, 1998; Heinermeyer *et al.*, 1997; Nussbaum *et al.*, 1998). The mammalian proteasome also has specificities for cleavage after small neutral amino acids and after branched-chain amino acids. Statistical analysis of naturally cleaved peptides found that up to five residues flanking the N terminal and the residue on either side of the C-terminal are also related in proteasome cleavage (Altuvia and Margalit, 2000; Bergmann *et al.*, 1996; Bergmann *et al.*, 1994; Holzhutter *et al.*, 1999; Nussbaum *et al.*, 1998; Shastri *et al.*, 1995; Vijh *et al.*, 1998; Yellen-Shaw *et al.*, 1997). Bioinformatians have used the proteasome cleavage patterns to predict potential T cell binding peptides. Several prediction algorithms are available online, such as PAProC (http://www.paproc.de) (Kuttler *et al.*, 2000; Nussbaum *et al.*, 2001), MAPPP (http://www.mpiib-berlin.mpg.de/) (Holzhutter *et al.*, 1999) and NetChop (http://www.cbs.dtu.dk/services/NetChop/) (Kesmir *et al.*, 2002). PAProcC predicts both human and yeast proteasome cleavage sites. MAPPP predicts proteasome cleavage using statistical analysis using existing motifs. NetChop is

based on artificial neural network approach using known peptides as the training set.

Sources of proteins for proteasome degradation include self and antigenic proteins. Apart from being involved in the antigen presentation pathway, proteasomes can act as a quality control system for self-proteins. Non-functional proteins, or defective ribosomal products (DRiP), are defective proteins due to errors in translation. These proteins constitute a large part of newly synthesised proteins and are rapidly degraded by the proteasome. Incorrectly folded or assembled proteins are also degraded by proteasomes. For antigenic proteins, the proteasome seems to favour oxidised proteins as substrates (Teoh and Davies, 2004). The hypothesis is supported by the finding that 70-80% of oxidised intracellular proteins are degraded by proteasomes, and the fact that the 20S proteasome preferred hydrophobic groups of residues on the surface of partially denatured oxidised proteins.

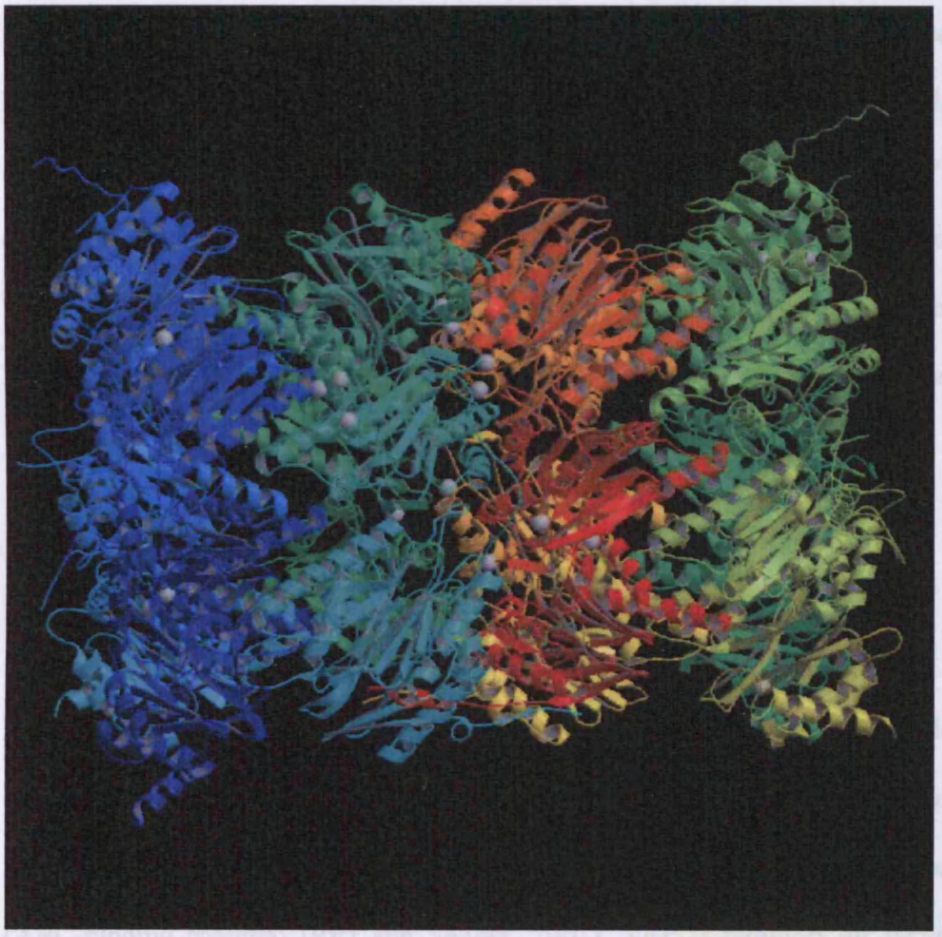


Figure 1.12. The three-dimensional structure of mammalian 20S proteasome at 2.75Å resolution (Unno *et al.*, 2002).

Peptides cleaved by the proteasome are 3 to 25 amino acids long, while most class I MHC epitopes are 8 to 11 amino acids long. It is estimated that only 15% of peptides degraded by the proteasomes are of the appropriate length for class I MHC binding. 70% of peptides are too short and 15% are too long. Long peptides can be trimmed to the correct size by various cellular peptidases. Peptides are digested in the cytosol by several peptidases such as leucine aminopeptidase (LAP) (Beninga *et al.*, 1998), tripeptidyl peptidase II (TPPII) (Tomkinson, 1999), thimet oligopeptidease (TOP) (Saric *et al.*, 2001), bleomycin hydrolase (BH) and puromycin-sensitive aminopeptidease (PSA) (Stoltze *et al.*, 2000). Inside the ER, longer peptides can be degraded by the aminopeptidases ERAAP or ERAP1 (Serwold *et al.*, 2002; York *et al.*, 2002).

1.4.2 Peptide translocation and class I MHC assembly

Degraded peptides bind to TAP and are transported to the ER. The translocation process consumes ATP. TAP protein is a heterodimer consisting of TAP1 and TAP2, both are part of a family of transporters known as the ABC transporters (ABC-binding cassette) (Ritz and Seliger, 2001). TAP is required for peptide translocation. Some viruses escape immune recognition by inhibition of TAP, such as Epstein-Barr virus , human cytomegalovirus (CMV) and herpes simplex virus (Ambagala *et al.*, 2003; Ambagala *et al.*, 2004; Khanna *et al.*, 1996; Koppers-Lalic *et al.*, 2003). The immediate early gene ICP47 of the herpes virus produces a short 88 amino acid polypeptide, which has high affinity for TAP and completely blocks the binding of peptides (Ambagala *et al.*, 2004). CMV virus inhibits TAP through binding of the viral protein US6 to the TAP region inside

the ER lumen. Another viral protein US3 binds newly synthesised MHC proteins and prevents MHC proteins from being transported to the cell surface (Ahn *et al.*, 1997; Bauer and Tampe, 2002; Hewitt *et al.*, 2001; Ulbrecht *et al.*, 2003).

TAP1 and TAP2 associate in the ER and form the TAP heterodimer. The central region of the TAP protein is likely to be the binding site as polymorphic residues in rat TAP2 have been shown to contact the peptide and influence peptide selection and transport (Elliott, 1997; Nijenhuis *et al.*, 1996). Binding of peptides to TAP proteins does not require ATP, while to transport peptide across the ER membrane consumes ATP.

Newly synthesised class I MHC molecules are unstable. They are retained in the ER in a partially folded state (Bouvier *et al.*, 1998b). A series of chaperone proteins are needed for the MHC proteins to fold completely. Newly synthesised class I MHC α chain is associated with a chaperone protein named calnexin, which is a membrane protein and holds MHC molecules inside the ER (Suh *et al.*, 1996). Some evidence suggests that calnexin is not an absolute requirement for class I HLA assembly as HLA molecules are expressed in calnexin-negative cell lines (Balow *et al.*, 1995; Scott and Dawson, 1995). The immunoglobulin binding protein (BiP) has a similar function to calnexin and may replace calnexin in the cell. Upon binding to β_2-microglobulin, the MHC molecule is released from calnexin and binds to a complex consisting of two proteins. One is calreticulin, which is also a chaperone and has a similar function to calnexin. The other protein is tapasin. After the binding of peptide, tapasin and calreticulin dissociate from the MHC molecule; the MHC molecule is now folded completely.

Some viruses escape the immune system by interfering with the MHC transport process. Adenovirus expresses the viral E19 protein that can associate with the class I MHC molecule and retains it inside the ER (Andersson *et al.*, 1987; Paabo *et al.*, 1989). Human cytomegalovirus (HCMV) synthesises two viral proteins US2 and US11, which stimulate the MHC heavy chains to be transported from the ER into the cytosol where they are rapidly degraded by proteasomes (Wiertz *et al.*, 1997).

TAP is able to associate with MHC proteins to facilitate peptide binding, although the interactions with TAP are not essential for peptide binding to the MHC proteins and the reason for the TAP-MHC attachment is unknown (Carreno *et al.*, 1995). Mullbacher postulated that the MHC molecules were more actively involved in peptide generation and binding (Mullbacher, 1997). It has been observed that TAP is able to transport large polypeptides into the ER (Higgins, 1992). Mullbacher suggested that the MHC molecule non-covalently attaches to TAP in the ER, the polypeptide slides through TAP, move along the MHC binding groove until the anchor residues occupy the binding pockets of the MHC. The MHC molecule then acts as a peptidase and removes the two ends of the polypeptide (Falk *et al.*, 1990). At present, there is no clear evidence to prove that MHC molecules have catalytic activities and more research is required to validate the hypothesis. After binding to peptides, the MHC protein leaves the ER and is transported to the cell surface (Townsend *et al.*, 1989). The peptide binding process is considered as the rate limiting step of MHC protein assembly as only a fraction of the peptides are able to bind to MHC.

1.5 HLA superfamily classification

HLA is one of the most polymorphic proteins in mammals. The IMGT/HLA database stores over 1800 different HLA class I and II alleles (Robinson *et al.*, 2003). The common way of finding whether a specific peptide will bind to one MHC allele is through binding assays. Many HLA alleles have been demonstrated to bind peptides with similar anchor residues (Sidney *et al.*, 1995). The experimental research process will be greatly shortened if there is a set of rules to group HLA alleles with similar specificities together. Several research groups have tried to classify HLA alleles (Cano *et al.*, 1998; Chelvanayagam, 1996; Lawlor *et al.*, 1991; Lund *et al.*, 2004; Sette and Sidney, 1998; Sidney *et al.*, 1996). The classification reduces the experimental workload as it is not necessary to study each HLA individually and it makes the design of epitope based vaccines and other immunological treatment targeted at multiple alleles more efficient.

1.5.1 Evolutionary analysis

Sequence alignment was often used in early classification. An early attempt to classify MHC molecules was from the evolutionary studies (Lawlor *et al.*, 1991). As chimpanzee and gorillas are the most closely related to human species and possibly share a common ancestor 7-10 million years ago, Lawlor compared the sequences of 14 gorilla class I MHC alleles with HLA-A, B and C alleles in human and MHC in chimpanzees. Sequences of human, gorilla and chimpanzee MHC alleles are similar but not identical, as most of the polymorphic residues appear in the same region. Also genes at A, B and C locus of gorilla and

chimpanzee MHCs are similar to HLA-A, HLA-B and HLA-C, respectively. Phylogenetic trees were generated for A, B and C genes and it was found that HLA-A alleles were divided into five families: A2, A3, A9, A10 and A19. Two divergent groups of HLA-C alleles were found, one containing Cw*0701 and Cw*0702, the other with Cw*0101-Cw*0601 and Cw*1201. HLA-B is the most polymorphic locus in the human HLA genes and no consensus group was found in the study. Based on Lawlor's research, Jakobsen et al. aligned DNA and protein sequences of the HLA-A alleles. The DNA alignment showed that family signatures are not focused on one region but are distributed throughout the sequence. The protein sequence alignment revealed that three positions in the binding site, 62, 97 and 114 were important in classifying alleles within the families (Jakobsen *et al.*, 1998).

Another HLA classification based on evolutionary analysis was done by McKenzie et al. in 1999 (McKenzie *et al.*, 1999). In their study, phylogenetic trees were constructed using three methods: maximum parsimony, distance-based minimum evolution and maximum likelihood. Different classifications were carried out, based on either whole protein/nucleotide sequence, sequence of the binding site, or sequence excluding the binding site. Two clusters were found for HLA-A class: one with A1, A3, A9, A11, A36, A*8001 and some of the A19 and the other with A2, A10, A28, A4301 and the other A19 members. HLA-B and HLA-C did not form any consistent clusters.

1.5.2 Structural analysis

The binding of peptides to MHC molecules is influenced by the interactions between side chains of peptides and the binding pockets in the peptide binding site. Kurata and Berzofsky studied the binding of peptide analogs to the MHC binding site and interactions with the TCR. It was identified that the same peptide can bind to I-Ed molecule in more than one conformation. Moreover, the change in peptide conformation did not affect the recognition by T cells, indicating that the TCR may interact with different positions of the peptide in different conformations (Kurata and Berzofsky, 1990). Similarly, Gopalakrishnan and Roques simulated the interactions between the peptide HLA-A2 170-180 and the H-2Kd binding site using the program AMBER. They found that the binding orientation of the peptide may be dependent on the sequence and structure of the peptide and may be allele specific (Gopalakrishnan and Roques, 1992).

In 1996, Chelvanayagam studied binding pockets and grouped HLA molecules according to the amino acid composition in each pocket (Chelvanayagam, 1996). HLA molecules within one group have the same amino acids or amino acids with similar chemical properties in a particular binding pocket and are expected to bind to the same peptide residue. The classification was used to classify HLA molecules that have not been studied experimentally and also predict their binding motif. Although classified separately, groups of HLA-B and C molecules share the same binding specificity with HLA-A if they have the same amino acids in the binding site. The drawback to this form of classification is that since the classification is done according to the residues surrounding one position of

the peptide, for a nonamer peptide, the HLA alleles are classified nine times and the same allele is often found in different groups in different classifications. A similar study has been carried out by Zhang et al., in which the binding pockets of class I MHC are classified into families by modelling the structures of MHC-peptide complexes using crystal structures as templates. Five families were defined according to specificities in the pocket B, and three families were defined based on specificities inside pocket D. Three more families were also defined for alleles with a joint specificity of pocket C and D (Zhang *et al.*, 1998).

1.5.3 Geometrical similarity matrix

Cano et al. clustered the HLA-A and HLA-B alleles by constructing similarity matrices (Cano *et al.*, 1998). MHC molecules were compared in a geometric space, where each amino acid occupied one dimension. The similarities among chemical properties of the twenty amino acids such as polarity and charges were compared and the results were stored in an amino acid similarity matrix. Another reference matrix, the binding affinity matrix was generated by calculating the stability of each amino acid side-chain at each position of the peptide. The similarity among MHC alleles was measured using both experimental peptide elution data and by comparing the alleles using the similarity matrix. The method identified three clusters. Cluster 1 includes HLA-A3, HLA-A11, HLA-31, and HLA-33. Cluster 2 includes HLA-B7, HLA-B35, HLA-B51, HLA-B53 and HLA-B54. Cluster 3 includes HLA-A29, HLA-B44 and HLA-B61.

1.5.4 Sequence and binding motif approach

The most common way of classifying HLA molecules is to group those with similar binding motifs. Class I HLA molecules have been classified into superfamilies by Sette and Sidney using motif based approaches. Sidney et al. defined four supertypes by examining reported cross-reactive epitopes, from which MHC alleles that can be possibly grouped into one supertype were identified (Sidney *et al.*, 1996). They then compare the binding pockets for the anchor residues, pocket B and F. Experimentally confirmed binding motifs of the alleles were also examined, and those with similar motifs are grouped into one supertype (Sidney *et al.*, 1995). The supertypes identified in the paper are: A2 (A*0201-06, A*6802, A*6901), A3 (A*0301, A*1101, A*3101, A*3301, A*6801), B7 (B*0702-5, B*3501-3, B*5101-5, B*5301, B*5401, B*5501-2, B*5601, B*6701and B*7801) and B44 (B37, B41, B44, B45, B47, B49, B50, B60, B61). The same group later published review papers in which the four supertypes were revised. A*0207 was added to the A2 supertype and B*1508 and B*5602 were added to the B7 supertype (Sette and Sidney, 1998; Sette and Sidney, 1999). Sette and Sidney carried out further analysis in 1999 and defined a total of nine supertypes including the previously defined supertypes (Sette and Sidney, 1999). The nine supertypes were estimated to cover 99% of the world population (Sette *et al.*, 2001).

The supertype definition can be applied in epitope based vaccine research. Epitopes taken from hepatitis B virus infected patients have been shown to cross react with alleles in the A2, A3 and B7 superfamilies (Bertoni *et al.*, 1997). Epitopes isolated from Epstein-Barr virus reacted with several alleles of the

B*44 family (Khanna *et al.*, 1997). Epitopes have been identified to cross-react with the A24 family (Burrows *et al.*, 2003). Many viral and tumour antigen derived vaccine candidates have also been shown to be able to bind multiple alleles (Bertoletti *et al.*, 1997; Fleischhauer *et al.*, 1996; Kawashima *et al.*, 1998; Wang *et al.*, 1998). Sette et al. predicted 223 potential cancer peptides of CEA, Her-2/neu, P53 and MAGE antigens using T cell epitope prediction algorithm, among which 115 were cross-reactive peptides of the A2 supertype. 43 peptides were tested for immunogenecity and 73% were positive (Sette *et al.*, 2002). Recently a protein sequence scan has been carried out to search T cell epitopes within the SARS virus based on the nine HLA supertypes in Sette's analysis (Sylvester-Hvid *et al.*, 2004). 15 predicted epitopes for each supertype were identified and tested experimentally. 75% of the predicted epitopes were found to be high affinity peptides (IC_{50} < 500nM) and about 112 vaccine candidates were obtained from the experiments. Table 1.1 lists the supertypes and alleles within each supertype.

Based on Sette's study, Lund et al (Lund *et al.*, 2004) classified HLA-A and B molecules using specificity matrix. The nonamer ligands of all HLA-A and B molecules were collected from SYFPEITHI and MHCPEP and aligned. The frequencies of each amino acid at each position were summarised into matrix. The matrix was used as the input for a clustering analysis and the HLA superfamilies were organised into a consensus tree. In their results, the A26 alleles were separated from the A1 cluster in Sette's results, and a new B8 superfamily was defined. The other superfamilies were the same as Sette's.

Supertype	MHC alleles
A1	0101 2501 2601 02 3201
A2	0201-07 6802 6901
A24	2301 2402-04 3001-03
A3	0301 1101 3101 3301 6801
B44	37 40012 4006 41 44 45 47 49 50
B27	1401 – 02 1503, 09, 10, 18 2701 – 08 3801. 02 3901 – 04 4801, 02 7301
B7	07 35 51 53 54 55 56 67 78
B58	1516, 17 5701, 02 58
B62	1301 – 02 1501, 02, 06, 12, 13, 14, 19, 21 4601 52

Table 1.1. Nine supertypes defined by Sette and Sidney (Sette and Sidney, 1999).

It should be noted that class II HLA molecules have also been classified by sequence approach. Chelvanayagam defined the HLA-DR roadmap by allele binding specificities and the polymorphic residues inside the binding site that contact peptides. The important residues were identified by studying the crystal structures of known HLA-DR-peptide complexes (Chelvanayagam, 1997). HLA-DP (Castelli *et al.*, 2002) and DQ (Baas *et al.*, 1999) supertypes have been defined based on binding studies to define the motifs and structural modelling of the peptide-MHC complexes. Reche and Reinherz used multiple sequence alignment to find important residues in 774 class I and 485 II HLA molecules. Consensus sequence patterns were obtained for the binding sites of HLA-A, B, C, DP, DR and DQ groups (Reche and Reinherz, 2003).

1.6 HLA and disease

The ultimate goal for T cell epitope research is to develop immuno-therapy and design vaccines that provide protection against pathogens or tumours. HLA can be either associated or linked to diseases (Thomsen *et al.*, 1979). An association is when affected individuals in a population are unrelated and have the same HLA allele. A linkage is when members of the same family have the disease and also have same HLA allele. MHC was first linked to disease by Lilly in 1964, when he observed that the H-2^K mice were susceptible to the Gross leukaemia virus, while the H-2^b mice were resistant (Lilly *et al.*, 1964). Later the disease was confirmed to be linked to the Rgv-1 gene encoded within the H-2 gene region (Lilly, 1971). Since then, the relationship between HLA and more than a hundred different diseases have been studied.

HLA is associated with many autoimmune diseases that affect 4% of the population (Merriman and Todd, 1995), many of which are long-term diseases and difficult to treat. Common ones are rheumatoid arthritis and HLA-DR4, insulin independent diabetes mellitus (IDDM) and HLA-DR/DQ alleles (Rani *et al.*, 1999), muscular sclerosis (Mehta *et al.*, 1986) and systemic lupus erythematosus and HLA-DR alleles (Gladman *et al.*, 1979; Jazwinska *et al.*, 1989; Kampf *et al.*, 1979; Marchini *et al.*, 2003; Stephansson *et al.*, 1993; Yao *et al.*, 1993; Yao *et al.*, 1994). Autoimmune diseases are complicated, there are 12 loci that are suspected to be linked to IDDM, in particular DR and DQ alleles (Choudhuri and Vergani, 1998; Fernandez-Vina *et al.*, 1993; Kelly *et al.*, 1985; Kockum *et al.*, 1994; Maruyama *et al.*, 1994; Matsumoto and Awata, 1994;

Pituch-Noworolska *et al.*, 1991). Class I HLA alleles have also been detected in some IDDM patients (Anal *et al.*, 1997; Faustman, 1995; Ono *et al.*, 1988).

Apart from autoimmune diseases, HLA alleles have also been linked to many infectious diseases. HLA-DR2 was found to be associated with leprosy and tuberculosis in Asian populations (Brahmajothi *et al.*, 1991; Singh *et al.*, 1983; van Eden *et al.*, 1980). B*5301 and DRB1*1302 alleles were associated with decreased risk of malaria in children in a study conducted in Gambia (Hill *et al.*, 1991). HLA-DRB1*1302 was also associated with clearance of hepatitis B infection (Hill, 1998). DRB1*1101 was reported to be associated with clearance of HCV infection in the UK (Tibbs *et al.*, 1996). HLA-B35 has been associated with more severe HIV progression while B27 was found to be associated with slower disease progression (Kaslow *et al.*, 1996; McNeil *et al.*, 1996).

1.7 HLA and vaccine design

Vaccination is a cost-effective strategy for disease prevention (Hellstrom and Hellstrom, 2003). The goal of vaccination is to give recipients limited exposure to a pathogen, which generates host memory cells that can elicit a strong immune response when the pathogen invades the body later (Payette and Davis, 2001). Vaccine research started in the 18th century. The first vaccine was the smallpox vaccine produced by Edward Jenner in 1796, when he infected his gardener's son with pus taken from a patient who had cowpox (cowpox is equivalent to smallpox in humans). The boy contracted mild cowpox but recovered and became immune to smallpox (Mayr, 1999). Smallpox vaccination was later used

world-wide and by 1980 WHO reported that smallpox had been eradicated from the world.

Vaccines often contain live attenuated whole viruses, nucleic acids or fragments of viruses such as peptides or subunits of proteins (Moingeon, 2001). Live attenuated viruses were commonly used in vaccines since the development of the smallpox vaccine (Smith, 1999). Viruses are either killed or mutated to stop their replication thereby reducing virulence. Many well known vaccines use live attenuated viruses, examples of this type of vaccine are the influenza, measles, mumps, rubella, polio, yellow fever and hepatitis A vaccine (Payette and Davis, 2001). This form of vaccine uses whole pathogens in their native form and induce good immunity in the host, but they cannot be applied to viruses with high mutation rates or more lethal viruses like HIV for safety reasons (Newman *et al.*, 2002). Purified protein sub-units have been tested for immunogenicity (Paschen *et al.*, 2004). Often the polysaccharide coat of the pathogen is used as it is the part that is in direct contact with the immune system. DNA encoding the antigen is introduced into non-pathogenic bacteria or yeast and the antigen is produced in large amounts to be used as vaccines (Payette and Davis, 2001). One of the most successful subunit vaccines is the hepatitis B vaccine, in which purified hepatitis B surface antigen (HBsAg) was used (Szmuness *et al.*, 1981a; Szmuness *et al.*, 1981b). Alternatively, DNA can be inserted into attenuated bacteria or viruses and injected into the recipient, so that the bacteria or virus can replicate within the host and produce antigenic proteins. The vaccinia virus vaccine uses this technique (Smith, 1999).

Introducing plasmid DNA directly into the recipient can stimulate a good immune response (Ciernik and Carbone, 1995). DNA encoded peptides are expressed and induce both B and T cell responses. DNA based HIV epitope vaccines have been developed by Bazhan et al., in which more than thirty T cell epitopes restricted by 10 different class I HLA alleles were inserted into plasmid vectors (Bazhan *et al.*, 2004). Additional mouse MHC class I epitopes were inserted into the vector to test the immunogenicity. The vaccine induced specific CTL response in the immunised animals. DNA vaccine is safer as no bacteria or virus is required. However, the risk of DNA vaccine is that the immune system may target the DNA itself which can lead to auto-immunity.

Another rapidly developing area is the use of epitopes. Advances in cloning and sequencing technology make it possible to detect the antigenic region of the protein and synthesise corresponding peptide fragments (Moingeon, 2001). The first peptide immunisation experiment was carried out in 1985, when an epitope inserted in *E. Coli* induced immunity in mice against cholera toxin and *E. Coli* heat liable toxin (Jacob *et al.*, 1985). Epitope based vaccines can be designed using either T cell epitopes or B cell epitopes (Dermime *et al.*, 2004; Meloen *et al.*, 2001). T cell epitopes are protein fragments that initiate cellular immune response while B cell epitopes can be from proteins, lipids, nucleic acids and carbohydrates and are recognised by antibodies and facilitate the humoral immune response (Dermime *et al.*, 2004). B cell epitopes are divided into continuous and discontinuous epitopes (Sundaram *et al.*, 2004). Continuous epitopes are linear protein fragments and discontinuous epitopes are formed from a surface area containing segments of different regions of proteins (Mahler *et al.*,

2003). Antibody affinities against epitope fragments are generally lower than towards proteins in their native conformation (Meloen *et al.*, 2001). The three dimensional structures of the B cell epitopes are often required for function, sometimes a whole molecule containing the epitope is required to activate an immune response (Lehner *et al.*, 1990). Because of the structural requirement, it is difficult to model and predict B cell epitopes. T cell epitopes are linear peptides, therefore they are easy to synthesise and to be inserted into plasmid vectors (van Endert, 2001). A major disadvantage of epitope vaccines is the difficulty in finding immunodominant epitopes and T cell epitopes are MHC restricted, therefore the effectiveness of the vaccine is dependent on the population having the required MHC haplotype.

Currently, several T cell epitope based vaccines are under development. Synthetic HIV peptide vaccine has been tested in mice and induced CTL responses against HIV (Belyakov *et al.*, 1998a; Belyakov *et al.*, 1998b). HIV Env and Gag epitopes have generated CTL responses in the recipients (Ferrari *et al.*, 2001). A 40mer synthetic peptide has stimulated immune responses in intestinal nematode infections (Robinson *et al.*, 1995). A phase I clinical trial has been carried out to test the effect of an Epstein Barr virus epitope vaccine (Bharadwaj and Moss, 2002). In another experiment, multiple tumour epitopes induced protective immunity in mice (Toes *et al.*, 1997). In the following section cancer vaccine development is explained as an example of epitope based vaccine research. Vaccines can be either prophylactic or therapeutic. Vaccines using tumour antigens have been considered as a possible cancer therapy. Tumour antigens have been observed in animal models to induce immune responses and

prevent cancer progression. In a murine model, vaccination with a peptide derived from the HPV16 E7 oncoprotein in IFA prevented the growth of an HPV16 induced tumour in mice (Feltkamp *et al.*, 1993). Vaccination with p53 protein prevented chemically induced skin cancer in mice (Ben-Hur *et al.*, 1998), and the injection of purified tumour antigen in mice could suppress chemically induced tumours (Ben-Hur *et al.*, 2000). Introducing mouse breast cancer antigen MUC1 into C3H/HeOuj mice reduced the incidence of breast cancer (Xing *et al.*, 2001).

Tumour associated antigens can be discovered by peptide elution from the MHC-peptide complexes or by screening recombinant DNA libraries (van der Bruggen *et al.*, 1991). The first human tumour antigen was discovered in 1991 (van der Bruggen *et al.*, 1991), when a tumour related gene MAGE was identified by transfection of a genomic DNA library into a melanoma antigen loss variant (Robbins and Kawakami, 1996). Subsequent experiments found that the gene belongs to one family and other members of the family have been discovered. An important finding is that proteins encoded by the gene are present in many forms of tumour, such as melanoma, breast carcinoma and sarcoma, suggesting that it may be possible to design a single vaccine that will be effective against several forms of tumour. Since then, several other tumour antigens have been discovered and tested as vaccine candidates, such as NY-ESO-1 (Chen *et al.*, 2004; Korangy *et al.*, 2004; Sugita *et al.*, 2004), tyrosin kinase (Hung *et al.*, 2001; Topalian *et al.*, 1997), MUC-1 (Moore *et al.*, 2004; Pantuck *et al.*, 2004; Tsang *et al.*, 2004), Her-2/neu (Disis *et al.*, 1999), gp100 (Bakker *et al.*, 1994) and p53 (Vierboom *et al.*, 1997). To improve immunogenicity, structurally altered peptides such as

cyclic and chimaeric HSV epitopes have been produced (Hudecz, 2001). Recently, an A*0201 tumour epitope has been shown to induce both CD8 and CD4 T cell responses in the recipients (Harada *et al.*, 2004). The exact mechanism of the epitope induced response is not clear.

Clinical trials of potential vaccine candidates have been carried out (Jager *et al.*, 2002). HLA-A1 restricted MAGE-3 epitopes have been shown to induce tumour regression in patients with melanoma (Marchand *et al.*, 1999). In another experiment, CTL were induced by peptides taken from an allogenic HLA-matched melanoma and these killed the autologous tumour *in vivo* (Imro *et al.*, 1999). Later on, the MAGE-3 gene was linked to influenza protein and applied to 35 patients; although no immune response was developed in the patients, two patients did show clinical remission (Marchand *et al.*, 2003). HLA-A2 restricted NY-ESO-1 peptides were injected into 12 patients, with four patients developed CD8+ T cell response and regression (Jager *et al.*, 2000). The results show that vaccination with tumour associated MHC binding peptides can be effective in inducing a tumour specific CTL response. One problem with tumour antigen based vaccination is to find a suitable antigen that can be presented by MHC molecules and is able to induce a CTL response (Bodey *et al.*, 2000). There is evidence of impaired antigen processing and decreased antigen expression by tumour cells which is a main limitation in cancer vaccine development (Restifo *et al.*, 1993a; Restifo *et al.*, 1993b).

In recent years, multi-epitope vaccines have become more popular compared with single epitope vaccines. As the immune response generated by vaccines

using a single epitope is usually less than vaccines using whole viruses, multi-epitope vaccines will be more broadly used to enhance the immune response (Arnon *et al.*, 2001). Also peptides that are restricted by different HLA alleles can be inserted into a carrying vector to increase population coverage. The simplest way of constructing a multi-epiope vaccine is to link all the epitopes together and insert the fragment into the vector (An and Whitton, 1997; Hanke *et al.*, 1998; Ishioka *et al.*, 1999). The strategy has been applied to cancer and HIV vaccines. Multi-epitope HIV and malaria vaccines have been constructed. The HIV epitope discovered from antigenic proteins like Env, Gag, Pol and Nef with murine epitopes for testing in mice were constructed. The malaria vaccine used epitopes from *Plasmodium falciparum* restricted by human and murine MHC molecules. The DNA HIV/malaria vaccine was injected intramuscularly and shown to induce a CTL response to the two murine epitopes in mice after a single vaccination (Bazhan *et al.*, 2004; Hanke *et al.*, 1998).

1.8 QSAR

Quantitative structure - activity relationships (QSAR) are a group of quantitative methods used to relate the biological activity of small molecules to their structures. QSAR techniques have been applied in many areas including chemistry, biology, drug discovery and environmental toxicology. QSAR is considered a valuable tool for predicting the biological activities of untested molecules.

Research in QSAR can be traced to the beginning of the last century, when Meyer and Overton suggested that the toxicity of organic compounds could be

related to their lipophilicity (oil-water partition coefficients) (Selassie, 2003). Two major experiments calculating molecular properties were carried out by Hammett and Taft. Hammett measured the electronic effect on benzoic acids and calculated the σ constant, and Taft was the first to introduce the steric parameter *Es*. Based on their research, in 1963 Hansch and Fujita studied the relationship between the chemical compositions of plant growth regulators and their reactivity using the octanol-water partition system (Hansch and Fujita, 1963; Hansch and Fujita, 1964). A new molecular descriptor, the octanol-water coefficient, was calculated in their study. Later Hansch developed an equation which related biological activity to the electronic and hydrophobic properties of the compound structures (Eqn. 1.1).

$$\log(\frac{1}{c}) = a\log P - b(\log P)^2 + c\sigma + k \qquad \text{Eqn. 1.1}$$

where c is the minimum concentration for the compounds to be effective, P is the octanol-water partition coefficient, σ is the electronic properties derived by Hammett, and a, b, c and k are constants.

Apart from Hansch's equation, other methods are also used in QSAR modelling. One widely used method was developed by Free and Wilson; their main theory is that each substituent of a molecule makes an additive and independent contribution to the biological activity (Free and Wilson, 1964).

$$BA = \sum k_i x_i + u \qquad \text{Eqn. 1.2}$$

Equation 1.2 describes the Free and Wilson concept. BA is the biological activity, k is the contribution of each subtituent, x is the indicator variable which represents the presence or absence of a substituent (x=0 is absent and x=1 is present) and u is the contribution of the parent molecule. Fujita and Ban later modified the equation by using the logarithm of biological activity: LogBA (Fujita and Ban, 1971).

Biological data is the basis of QSAR modelling. The training set is very important as the quality of the model is dependent on the training data. Biological data should be a measurable biological or physiological function, which can be enzyme reactions, ligand-receptor interactions, etc. Often the logarithm values of the biological activities are used. Ideally molecules included in the study are congeneric and both active and less active molecules are included to generate high quality QSAR models. Data are usually derived from the same experimental protocol to ensure data consistency. The number of compounds for a QSAR model training set should be 20 or more (Perkins *et al.*, 2003). In QSAR models, there are often a large number of descriptors in the model with a relatively small number of compounds, therefore multivariate methods are used to reduce the number of descriptors and group the most important ones into a few, uncorrelated variables called principal components. The most common methods used in multivariate analysis are multivariate linear regression (MLR) (Mielke and Berry, 2002; Saxena and Prathipati, 2003) and partial least squares (PLS) (Xing *et al.*, 2003).

Historically, inear regression analysis is the most often used technique in QSAR studies. Equation 1.3 represents the multiple linear regression (MLR) model.

$$y = b + \sum c_i x_i \qquad \text{Eqn. 1.3}$$

where b and c are constants and x_1 to x_i are the calculated properties. The goodness of fit of the model is measured by the correlation coefficient r^2, which is an indication of how much variance in the data is explained by the model.

The partial least squares (PLS) method is developed from linear regression and is good at analysing multivariate data, especially when the number of variables are greater than the number of molecules. PLS simplifies the data by grouping variables that explain similar properties and replacing the variables with a few new, uncorrelated variables called latent variables (LV). The latent variables are used to explain the biological activity as in equation 1.4, where y is the independent variable, or biological activity.

$$y = a_1 LV_1 + a_2 LV_2 + a_3 LV_3 + a_i LV_i \qquad \text{Eqn. 1.4}$$

Apart from MLR and PLS, other techniques are also used in QSAR analysis such as principal component analysis (PCA). MLR and PLS are useful in correlating biological activities to structures, while PCA is a classification method used to find similarities and difference in the molecules. The biological activities are not required in the classification. Similar to PLS, PCA produces several uncorrelated

principal components which explain the maximum variance of the data. PLS and PCA will be further discussed in section 2.2.3.

Nearly all QSAR models have outliers, that is, molecules that are badly fitted by the model. These outliers may be the result of experimental errors, alternatively, the abnormal behaviour of the molecules may be the result of their chemical composition. The latter can provide valuable information and increase the explanative power of the model, although inclusion of these outliers can reduce the predictivity of the model. Also, removing too many outliers increases the predictivity but reduces the chemical diversity of the data set.

QSAR methods can be either 2D or 3D. 2D QSAR methods use descriptors to study molecular properties. Statistical methods are used to study the relationship between the structure and activities of the molecules. The 2D QSAR methods have been widely applied since the pioneering work by Hansch in 1969, when he found that the octanol-water partition coefficient logP value can be used to describe the hydrophobicity of compounds (Hansch, 1969). Since then, hundreds of descriptors have been measured or calculated, such as the molecular surface area, molecular connectivity, molecular density and so on. Because of the potentially large number of variables, variable selection programs such as a genetic algorithm and GOLPE are required.

3D QSAR methods are similar to 2D methods, however they correlate spatial structural properties, within a group of compounds, with activity. In 3D-QSAR, the alignment of the structures is important as in many programs, the descriptors

are location dependent. Commonly used 3D QSAR techniques are CoMFA (comparative molecular field analysis) (Cramer *et al.*, 1989), CoMSIA (comparative molecular similarity index analysis) (Klebe and Abraham, 1999) and GRID/GOLPE (Cruciani and Watson, 1994). 3D QSAR methods can take more calculation time but can offer a more specific analysis about interactions in 3D space.

3D-QSAR is used widely in modelling ligand-receptor interactions, as it can provide good visualisation of potential energy surrounding the molecules. Comparative molecular field analysis (CoMFA) is one of the major tools in 3D-QSAR. CoMFA studies differences in target properties that relate to changes in activity. In CoMFA, molecules are aligned by their shared molecular features. The aligned molecules are fixed and placed in a 3D grid. A probe atom is placed at each point of the grid in turn and the interaction energy between the atoms of the molecules and the probe atom is calculated. The inter-molecular energy of the molecules is ignored. The two forms of energy studied by CoMFA are steric and electrostatic. Steric bulk is calculated by a Lennard-Jones potential, which describes the forces between two atoms that are dependent on the distance between their centres. Electrostatic energy represents the attraction of opposite charges and is calculated using Coulomb potentials. CoMFA is also used as a standard technique in modelling different ligand-receptor interactions (Barreca *et al.*, 1999; Grunewald *et al.*, 1999; Li *et al.*, 1999; Newman *et al.*, 1999; Xing *et al.*, 1999), such as ion channel inhibitors and enzyme activity (Bhongade and Gadad, 2004; Buolamwini and Assefa, 2002; Ducrot *et al.*, 2001; Khandelwal *et*

al., 2003; Kuo *et al.*, 2004; Murthy and Kulkarni, 2002; Purushottamachar and Kulkarni, 2003; Raichurkar and Kulkarni, 2003)

Comparative molecular similarity index analysis (CoMSIA) (Klebe, 1998; Klebe *et al.*, 1994) is a recently developed 3D-QSAR method and can be considered as an extension of CoMFA. CoMFA calculates energy fields of the molecules, while CoMSIA compares the similarities between the aligned molecules. Apart from steric and electrostatic, CoMSIA describes other molecular interactions: hydrophobic, hydrogen bond donor and acceptor. The hydrogen donor potential studies the ability of the molecule to form hydrogen bonds by donating a hydrogen atom, and the hydrogen acceptor describes the ability of the molecule to form hydrogen bonds by accepting a hydrogen atom. Similarities of the molecules in the data set are obtained by comparing the similarities between the molecules and a pre-defined probe atom. A disadvantage of CoMFA is that it uses a Lennard-Jones potential to calculate steric forces, which is distance dependent and can cause rapid changes in energy near the surface of the molecule. Therefore a cut off of 0.5 to 1Å is set in the program. Unlike CoMFA, CoMSIA uses a Gaussian function instead of a Lennard-Jones potential and does not need a cut off (Klebe and Abraham, 1999).

Since its development, CoMSIA has been used in drug design to study the ligand-receptor interactions and has proved to be of good predictivity. Examples of CoMSIA studies include enzyme inhibitors such as cyclooxygenase inhibitors (Lee *et al.*, 2004), kinase inhibitors (Sperandio Da Silva *et al.*, 2004), HIV integrase (Buolamwini and Assefa, 2002) and urokinase inhibitors (Bhongade

and Gadad, 2004). CoMSIA has also been used to study ion channel inhibitors (Doddareddy *et al.*, 2004; Ducrot *et al.*, 2001; Pearlstein *et al.*, 2003) and various antagonists (Choo *et al.*, 2003; Dixit *et al.*, 2004; Islam *et al.*, 2003; Khandelwal *et al.*, 2003; Kunick *et al.*, 2004; Kuo *et al.*, 2004; Murthy and Kulkarni, 2002; Purushottamachar and Kulkarni, 2003; Raichurkar and Kulkarni, 2003).

QSAR techniques have been applied to peptide-MHC interactions. Mallios calculated amino acid frequencies in the training data set and used discriminant analysis to build models for mouse alleles IAd and IEd (Mallios, 1993). Later, Mallios used Sette's database of synthetic peptides and used a multiple regression method to re-calculate the mouse MHC models (Mallios, 1994). Mallios also developed an iterative stepwise discriminate analysis to align known class II MHC peptides and generated a quantitative matrix based on the sequences and binding motifs (Mallios, 1997; Mallios, 1998; Mallios, 1999; Mallios, 2001). The performance of the iterative stepwise discriminant analysis was compared with two other class II MHC prediction algorithms SYFPEITHI and ProPred. Four data sets were applied and it was found that the algorithms had different predictivity with different data sets and no one algorithm was better in all tests (Mallios, 2003). Bologa et al. applied QSAR techniques to A*0201 peptides, the main properties studied were steric and side chain hydrophobicity (Bologa *et al.*, 1995). Lipophilicity was found to be favoured by the anchor residues and amino acids at position 1, 3 and 6. Similar studies have been carried out by Chersi et al. (Chersi *et al.*, 2000).

Other 2D and 3D QSAR methods have been applied in MHC-peptide interaction studies and epitope predictions of the HLA-A2 alleles (Doytchinova and Flower, 2003a; Doytchinova *et al.*, 2002; Doytchinova and Flower, 2001; Doytchinova and Flower, 2002; Doytchinova *et al.*, 2004). Initially 3D QSAR methods CoMFA and CoMSIA were applied to A*0201 binding peptides. The training set used peptides reported by previous publications and binding affinities measured by IC_{50} assays in the previous reports were used as experimental values. Partial least squares (PLS) was used to build both CoMFA and CoMSIA models. The quality of the models was determined by the predictivity q^2 and the explained variance r^2, i.e, the percentage of the properties in the training set explained by the model. In the experiment, the predictivity of the CoMSIA model was 0.542 with r^2 value of 0.679, while both values of the CoMFA model were below 0.5. The CoMSIA contour maps highlighted hydrophobic regions that were the most important in peptide binding to A*0201 allele. Subsequently CoMSIA was applied to other HLA and mice alleles.

A 2D QSAR technique, the additive method, was developed to study peptide-MHC binding (Doytchinova and Flower, 2001). The additive method is based on the Free-Wilson concept (Free and Wilson, 1964; Kubinyi and Kehrhahn, 1976). Additional terms were added to the basic QSAR model to account for the adjacent and every second side-chain interactions. For a nonamer peptide the model could be presented by equation 1.5:

$$pIC_{50} = const + \sum_{i=1}^{9} P_i + \sum_{i=1}^{8} P_i P_{i+1} + \sum_{i=1}^{7} P_i P_{i+2} \qquad \text{Eqn. 1.5}$$

where pIC_{50} is the binding affinity expressed in p-units (negative decimal logarithm of IC_{50} values), the *const* accounts for the peptide backbone contribution, $\sum_{i=1}^{9} P_i$ is the sum of amino acid contributions at each position, $\sum_{i=1}^{8} P_i P_{i+1}$ is the sum of adjacent peptide side-chain interactions, $\sum_{i=1}^{7} P_i P_{i+2}$ is the sum of every second side-chain interactions.

The additive method was first applied to HLA-A*0201, using the same training set as the CoMSIA study. The additive model generated a coefficient contribution for each of the amino acids in each position. An A*0201 nonamer model was derived from the regression equation, containing the favoured and disfavoured amino acids at each position. Using this model, affinities of other peptides can be predicted. The additive method has also been used to generate a binding motif for the A2 supertype (A*0201, A*0202, A*0203, A*0206 and A*6802) (Doytchinova and Flower, 2003a). The additive method has also been applied to some of the HLA class II DRB1 alleles (Doytchinova and Flower, 2003b) and mouse alleles (Hattotuwagama *et al.*, 2004).

GRID/CPCA has been used in studying structure – activity relationships. GRID was developed by Goodford (Cruciani and Watson, 1994; Goodford, 1985). It is a computational program used to determine energetically favourable interactions between binding sites and ligands using pre-defined chemical probes (Kastenholz *et al.*, 2000). Outputs from GRID can be used in other programs like GOLPE and SIMCA for further analysis. CPCA is used to define groups of similar molecules and outliers using interaction energy values calculated by GRID. GRID/CPCA is

often used in enzyme – substrate research and has been applied to differentiate specificities of matrix metalloproteinases MMP-3 and MMP-8 inhibitors (Matter and Schwab, 1999) and to test the substrate selectivities of ten enzymes in the matrix metalloproteinases family (Terp *et al.*, 2002). Other examples of applications include a analysis of glycogen phosphorylase b inhibitors (Cruciani and Watson, 1994), a comparison of bacterial and human dihydrofolate reductase receptor selectivity (Pastor and Cruciani, 1995), study of the chyotrypsin family (Kastenholz *et al.*, 2000) and the kinase family (Naumann and Matter, 2002).

1.9 Aims

Although QSAR methods have been widely used in drug discovery and development, they have only recently been applied in immunology. The aim of this thesis was to apply 2D and 3D QSAR techniques to analyse the interactions between peptides and HLA molecules and to design new high affinity binding peptides. Various amino acid descriptors were applied to define a binding motif for the HLA-A*0201 allele. The descriptors used included the descriptors taken from the AAindex database, the three z and the five z descriptors. Variable selection techniques SIMCA, GOLPE and GA were used to remove irrelevant and redundant descriptors. The additive method and CoMSIA were used to define a binding motif for the HLA-A3 superfamily. Results of the models generated by the additive method were incorporated into a web server, MHCPred, to facilitate online T cell epitope prediction. The predictivity of the additive method was evaluated and compared with some of the other T cell epitope prediction servers. Finally, all class I HLA alleles were classified into supertypes

using a combined GRID/CPCA approach, and the results were compared with HLA supertype definition by hierarchical clustering analysis.

Chapter 2
Material and Methods

Both laboratory and *in silico* experiments were conducted during my research. A variety of *in silico* techniques were used for defining HLA binding motifs and class I HLA superfamilies. A 3D QSAR technique Comparative Molecular Similarity Indices Analysis (CoMSIA) and a 2D QSAR technique the additive method were used to define binding motifs for HLA-A*0201 and the HLA-A3 supertype. GRID/CPCA was used to identify class I HLA superfamilies. High affinity peptides were predicted and synthesised, their binding affinities tested experimentally using a T2 stabilisation assay.

2.1 Experimental Material

Sections 2.1.1 to 2.1.5 describe material used in the T2 stabilisation assays. The remaining sections relate to *in silico* experiments.

2.1.1 Plastic Ware

Product	*Supplier*	*Catalogue Number*
50ml Falcon Tubes	Becton Dickinson Labwear, NJ, USA	352070
96 Flat Bottom Plates	Corning Costar, High Wycombe, UK	3595
96 U Bottom Plates	Corning Costar, High Wycombe, UK	3799
Cluster Tubes (FACS)	Abgene, Surrey, UK	AB-0672
FACS Tubes (5ml Round Bottom Polystyrene)	Becton Dickinson Labwear, NJ, USA	352054
Pipettes 5ml	Bibby Sterilin, Stone, UK	40105
10ml		47110
25ml		18327
Reagent Reservoirs (100ml)	Corning Inc, NY, USA	4873
Tips 20ul	Rainin Instruments Co Ltd,	GPS25

200ul 1000ul	Woburn, USA	GPS250 GPS1000

2.1.2 Tissue Culture Reagents

Reagents	*Supplier*	*Catalogue Number*
AIMV medium	Invitrogen Life Technologies, Paisley, UK	12055-091
Foetal Bovine Serum (GFCS)	Harlan Sear-Lab Ltd, Loughborough, UK	Batch No. 0010502
Geneticin (G418)	Sigma Aldrich Company Ltd, Poole, UK	A-1720
Human beta 2-microglobulin	SCIPac, Kent, UK	P122-1
Penicillin & Streptomycin	IAH Media Supplies, Compton, UK	

2.1.3 Peptides

peptides	*Supplier*
A2 and A3 peptides	Mimotopes, Cheshire, UK Dr Lawrence Hunt, IAH, Compton, UK

2.1.4 Cell Lines

Cell Line	*Supplier*
TAP-deficient cell line T2	Dr P. Borrow, Compton, UK
T2 cells transfected with an A3 plasmid	Prof Peter Cresswell, Yale University, USA.

2.1.5 Antibodies

Antibody	*Supplier*	*Catalogue*	*Clone*	*Isotype*
Anti-human HLA-A2 FITC conjugated Ab	Pharmingen, BD Biosciences, Oxford, UK	551285	BB7.2	IgG2b
FITC conjugated mouse Ab	Pharmingen, UK	555742	27-35	IgG2b
FITC conjugated Affinpure F(ab)' fragment goat anti-mouse Ab	Startech Scientific Ltd, UK	115-096-068	-	IgG & IgM

GAP A3 Ab	LGC Promochem, UK	HB-122 (ATCC number)	-	IgG2a

2.1.6 3D structural data of the HLA molecules

All class I HLA alleles (excluding those with silent mutations) in the IMGT/HLA database (Robinson *et al.*, 2003) were included in the GRID/CPCA (Cruciani and Watson, 1994) study. A total of 229 HLA-A, 447 HLA-B and 107 HLA-C molecules were selected. As only a few HLA molecules have been crystallised, the 3D structures of most HLA molecules were obtained by homology modelling using existing structures as templates. The protein backbones of the crystal structures HLA-A*0201 – 1I4F (Hillig *et al.*, 2001), B*0801 – 1AGD (Reid *et al.*, 1996) and Cw*0401 – 1IM9 (Fan *et al.*, 2001) were taken from the RCSB protein databank (Westbrook *et al.*, 2002) and were used to build the 3D structures of HLA-A, B and C, respectively. The β2-microglobulin and the α3 domain were deleted from the static template structures, as they are not involved in peptide-MHC interactions. Water molecules, co-factors and ligands were also deleted before modelling. Side chains were added to the built HLA structures using the program SCRWL (Side-Chain Placement with a Rotamer Library) version 2.8 (Bower *et al.*, 1997), which used rotamer libraries and protein main-chain coordinates to predict side-chain conformations.

2.1.7 The A3 peptides

Nonamer peptides binding to the HLA-A*0301, HLA-A*3101, HLA-A*1101 and HLA-A*6801 alleles were used to build the QSAR models. The additive and

the CoMSIA models were generated using the same training data set. Information on peptide sequences and their binding affinities was obtained from the AntiJen database. AntiJen, originally named JenPep, is an immunological database maintained in house (http://www.jenner.ac.uk/Antijen) (Blythe *et al.*, 2002; McSparron *et al.*, 2003). Only nonamers were included in the study. The HLA-A*0301 allele set included 72 peptides, the set for A*1101 included 62, A*6801 included 38 and A*3101 included 31 (appendix 2). Among the selected peptides, some bound to more than one allele. IC_{50} measurements were used in the original experiments to quantify the interactions between the peptide and the MHC molecule (Chang *et al.*, 1999; Kast *et al.*, 1994; Kawashima *et al.*, 1999; Scognamiglio *et al.*, 1999; Threlkeld *et al.*, 1997; van der Burg *et al.*, 1995; Wang *et al.*, 1998). The IC_{50} values were measured by a competition assay based on the inhibition of the binding of a radiolabeled standard peptide to detergent solubilised MHC molecules (Sidney *et al.*, 1996). In the assay, purified MHC molecules were incubated with radiolabeled probe peptide, human β_2m and protease inhibitors. After incubation, the HLA-peptide complexes were separated from free peptides by gel filtration. The percentage of bound peptides was calculated as the ratio of peptide left in the solution to the total peptide recovered (Sette *et al.*, 1989b).

2.1.8 The A2 peptides

Two sets of A*0201 peptides were used in the project (appendix 1). 266 nonamer peptides were used as a training set in the binding motif analysis, all of which were from the AntiJen database. As with the A3 peptides, IC_{50} measurements were used in the original experiments to quantify peptide affinities (del Guercio

et al., 1995; Kast *et al.*, 1994; Parkhurst *et al.*, 1998; Parkhurst *et al.*, 1996; Rivoltini *et al.*, 1995; Rongcun *et al.*, 1999; Sette *et al.*, 1994; Tsai *et al.*, 1997; Vitiello *et al.*, 1997).

A separate A*0201 data set was used as a test set to assess the predictivity of the additive method. The set was a gift from Dr. Vladimir Brusic, which included 181 T cell epitopes, 44 poly-alanine derived peptides, 56 naturally processed peptides and 245 non-binding peptides.

2.1.9 The epitopes

Epitopes used to compare the online T cell epitope prediction algorithms with the additive method were published within the last three years. The full list of epitopes and corresponding references can be found in appendix 4. The protein sequences that contained the epitopes were retrieved from either SWISS-PROT (Boeckmann *et al.*, 2003) or Genbank (Benson *et al.*, 2004).

2.1.10 Amino acid descriptors

Three sets of amino acid descriptors were used in A*0201 binding motif analysis: 93 descriptors selected from the AAindex database (Kawashima and Kanehisa, 2000), three z descriptors (Hellberg *et al.*, 1987) and five z descriptors (Hellberg *et al.*, 1987; Sandberg *et al.*, 1998).

2.1.10.1 The AAindex descriptors

The descriptors used in the first section of the A*0201 analysis were taken from the amino acid descriptor database AAindex (Kawashima and Kanehisa, 2000).

The database can be accessed from the following URL http://www.genome.ad.jp/dbget/aaindex.html. The AAindex contains descriptors explaining physico-chemical and biochemical properties of both single amino acids and interactions between amino acids. The properties are represented as numerical values. They can be properties like hydrophobicity, pKa, solubility, steric bulk, and surface area. Descriptors are either global descriptors, when describing the whole molecule, or local descriptors, when they describe single residues. The database is composed of two sections: AAindex 1 for amino acid indices (437 descriptors at the time of study) and AAindex 2 for the amino acid mutation indices (71 amino acid mutation matrices at the time of study). The descriptors used in the study were taken from index 1. An initial QSAR analysis was carried out using all the descriptors and those with correlation coefficients greater than 0.3 were selected and used in the final QSAR analysis. A total of 93 descriptors were selected, covering four major areas: hydrophobicity, flexibility, steric bulk and electrostatic properties. A list of the chosen descriptors is included in appendix 3.

2.1.10.2 The z descriptors

The z descriptors were originally defined in a peptide QSAR study by Wold and colleagues (Hellberg *et al.*, 1987), in which 29 physico-chemical variables were used to describe the 20 natural amino acids. These variables were converted into three scales z1, z2 and z3 using principal component analysis (PCA). The z1 scale is the hydrophobicity scale where negative values indicate hydrophobicity, and positive values indicate hydrophilicity. The z2 scale is useful in describing steric properties of the residue. A negative value in z2 corresponds to small

amino acids with low molecular weight and small surface area, while a positive value corresponds to large, bulky amino acids with large surface area. The z3 scale describes electronic properties. Amino acids with negative z3 values are polar and those with positive z3 values are non-polar.

In 1998, Sandberg re-examined Wold's three z descriptors and added two other properties z4 and z5 to explain molecular properties for both natural and synthetic amino acids (Sandberg *et al.*, 1998). The new scales were developed using partial least squares (PLS) and PCA. The z1 - z5 scales can be used to describe the following properties of the peptides: hydrophobicity, size/polarisability and electronic properties.

Both the three z and five z descriptors were used in the present study. Table 2.1 includes the three z descriptors reported by Wold and the five z descriptors from Sandberg for the 20 naturally occurring amino acids.

2.1.11 Epitope prediction servers

A total of nine epitope prediction servers were used in the evaluation of T cell epitope prediction algorithms, including BIMAS, ComPred, netMHC, PREDEP, ProPred, RANKPEP, SMM, SVMHC and SYFPEITHI. The algorithms tested include motif based patterning searching, matrix based predictions, machine learning methods and peptide-MHC interaction energy estimation. In the first evaluation test, the predictivities of the algorithms were tested using a ROC analysis of A*0201 binding peptides. In the second test, the ability of the algorithms to identify T cell epitopes within protein sequences was tested. The

results were compared with that of the additive method (MHCPred). A summary of the methods implemented by the servers is shown in table 2.2.

	Z3 descriptors			*Z5 descriptors*				
	z1	*z2*	*z3*	*z1*	*z2*	*z3*	*z4*	*z5*
A	0.07	-1.73	0.09	0.24	-2.32	0.60	-0.14	1.30
C	0.71	-0.97	4.13	0.84	-1.67	3.71	0.18	-2.65
D	3.64	1.13	2.36	3.98	0.93	1.93	-2.46	0.75
E	3.08	0.39	-0.07	3.11	0.26	-0.11	-3.04	-0.25
F	-4.92	1.30	0.45	-4.22	1.94	1.06	0.54	-0.62
G	2.23	-5.36	0.30	2.05	-4.06	0.36	-0.82	-0.38
H	2.41	1.74	1.11	2.47	1.95	0.26	3.90	0.09
I	-4.44	-1.68	-1.03	-3.89	-1.73	1.71	-0.84	0.26
K	2.84	1.41	-3.14	2.29	0.89	-2.49	1.49	0.31
L	-4.19	-1.03	-0.98	-4.28	-1.30	-1.49	-0.72	0.84
M	-2.49	-0.27	-0.41	-2.85	-0.22	0.47	1.94	-0.98
N	3.22	1.45	0.84	3.05	1.62	1.04	-1.15	1.61
P	-1.22	0.88	2.23	-1.66	0.27	1.84	0.70	2.00
Q	2.18	0.53	-1.14	1.75	0.50	-1.44	-1.34	0.66
R	2.88	2.52	-3.44	3.52	2.50	-3.50	1.99	-0.17
S	1.96	-1.63	0.57	2.39	-1.07	1.15	-1.39	0.67
T	0.92	-2.09	-1.40	0.75	-2.18	-1.12	-1.46	-0.40
W	-4.75	3.65	0.85	-4.36	3.94	0.59	3.44	-1.59

Table 2.1. The z descriptors. The three z descriptors developed by Wold, and the five z descriptors developed by Sandberg.

	Prediction server	*URL*	*Algorithm used by the server*	*MHC alleles predicted by the server*
Motif searching	SYFPEITHI (Rammensee *et al.*, 1999)	http://syfpeithi.bmi-heidelberg.com/Scripts/MHCServer.dll/EpPredict.htm	Motif based pattering searching	Class I and II
Matrix based algorithms	BIMAS (Parker *et al.*, 1992b)	http://bimas.dcrt.nih.gov/molbio/hla_bind/	Amino acid matrix, evaluates the binding affinity of peptides to MHC alleles by their half-time disassociation rates.	Class I
	RANKPEP (Reche *et al.*, 2002)	http://mif.dfci.harvard.edu/Tools/rankpep.html	Position specific scoring matrix (PSSM), produced by ungapped block alignment of known peptides.	Class I and II
	SMM (Peters *et al.*, 2003)	http://zlab.bu.edu/SMM/	Matrix based prediction. Considers both amino acids and their interactions.	A2
	ProPred (Singh and Raghava, 2001)	http://www.imtech.res.in/raghava/propred/	Matrix based prediction	Class II
Machine learning	netMHC (Buus *et al.*, 2003)	http://www.cbs.dtu.dk/services/NetMHC/	ANN	HLA-A2 and H-Kk

methods	SVMHC (Donnes and Elofsson, 2002)	http://www.sbc.su.se/svmhc/new.cgi	Prediction using support vector machines	
Structural based algorithm	PREDEP (Altuvia *et al.*, 1995)	http://bioinfo.md.huji.ac.il/marg/Teppred/mhc-bind/	Structural based approach using binding energy estimation.	
Combined methods	ComPred (Bhasin and Raghava, 2004)	http://www.imtech.res.in/raghava/nhlapred/comp.html	Combines artificial neural network (ANN) and matrix prediction algorithm	Class I and II

Table 2.2. T cell epitope prediction servers used in the evaluation study.

2.2 Methods

2.2.1 The T2 stabilisation assay

The binding of peptides to HLA alleles was measured by a quantitative T2 cell surface stabilisation assay (MyIntyre *et al.*, 1996; Salter and Cresswell, 1986): Aliquots of 2 x 10^5 cells/well were incubated in 96-well flat bottom microtiter plates with 100µl of test and control peptide (0.04 ~ 200µM) in the presence of AIMV and 100nM β2-microglobulin. The plates were stored at 37°C with 5% CO_2 overnight. The HBV specific peptide FLPSDFFPSV was used as a positive control for A*0201. HIV specific peptide HMYISKKAK was the positive control for A*0301. The HIV-nef peptide KAAVDLSHF was used as a negative control in both A*0201 and A*0301 experiments. A non-specific background control was used, i.e, wells with the same reagents as the others but with no peptides. This control was used to measure the level of background binding of the antibodies. After incubation, the cells were washed twice and re-suspended in HB2 buffer. FITC conjugated mouse anti-human HLA-A2 monoclonal antibody was added to the test peptides, the positive control and negative control peptides and half of the background control peptides at 1µl/10^6 cells concentration. The $mIgG_{2b}$-FITC isotype control antibody was added to the rest of the background control peptides at 10µl/10^6 cells concentration. The cells were incubated at 4°C in the dark for one hour and were fixed with 4% paraformaldehyde. The MHC-bound fluorescence level was measured by facscalibur analysis (FACS) and the results were analysed with the program Cellquest. The same procedure was used for the A*0301 binding experiments. Gap.A3 antibody (hybridoma) (ATCC) was

used as the first antibody, and FITC-conjugated $F(ab)_2$ fragment goat anti-mouse IgG + IgM was used as the secondary antibody at 1:100 concentration.

The fluorescence level of the peptides bound to HLA-A2 molecules was converted to fluorescence index (FI) values using the following equation (Yoon *et al.*, 1998):

$$FI = \frac{F_S - F_B}{F_{T2} - F_B} \times 100.00 \qquad \text{Eqn. 2.1}$$

where F_S is the mean fluorescence index (MFI) of the test peptides, F_B is the no peptide isotype antibody stained control MFI and F_{T2} is the no peptide HLA-A2 antibody stained control MFI.

2.2.2 BL_{50} calculation

The binding affinities of the test peptides to HLA-A2 and A3 alleles were obtained by converting their FI values to the half-maximal binding level (BL_{50}), which was the peptide concentration yielding the half-maximal FI value. The program used for the BL_{50} calculation was *ED50 plus 1.0*, an Excel macro written by MH Yargas for conversion of FI values to BL_{50} values, and is freely available on the Internet (URL:http://www.winsite.com/bin/Info?5387). Peptides were classified into three groups according to the BL_{50} values: high binders ($BL_{50} \leq 10^{-5}$M), medium binders (10^{-5}M < $BL_{50} \leq 10^{-4}$M) and non-binders ($BL_{50} > 10^{-4}$M).

2.2.3 Statistics

2.2.3.1 Principal component analysis (PCA)

PCA is commonly used in multivariate data analysis to reduce the number of variables. Data used in PCA are stored in a data matrix X (fig 2.1). There are N observations and K variables in the matrix. Each observation occupies one row, the variables are measurements of the observation and are stored in the columns.

PCA decomposes the matrix X into two smaller matrices: the scores matrix T and the loading matrix P', which explain the overall variance of the X matrix. The scores matrix contains a few variables M (fig 2.1), that is, the principal components (PC), which can be used to describe the observations. The loading matrix reveals the relationship between the variables in the original matrix and the principal components. Plots of the observations in the multidimensional space are called the scores plot, which identifies similarities and differences within the observations and groups them accordingly, while the loading plot relates the original variables with the PCs and identifies variables that are important in distinguishing groups of observations.

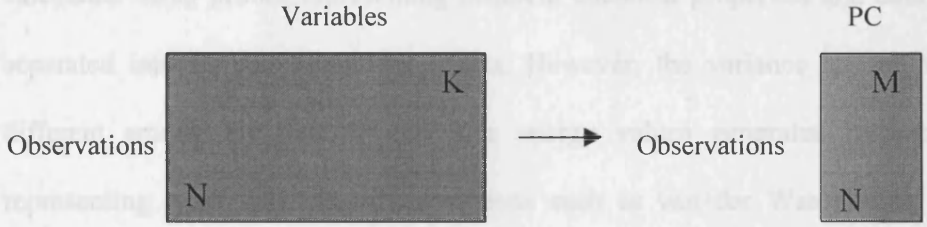
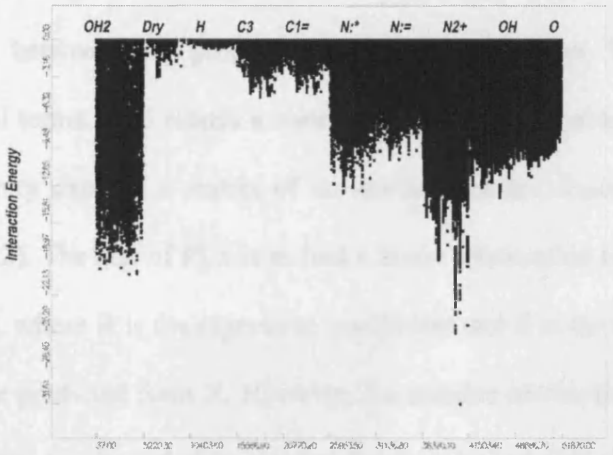


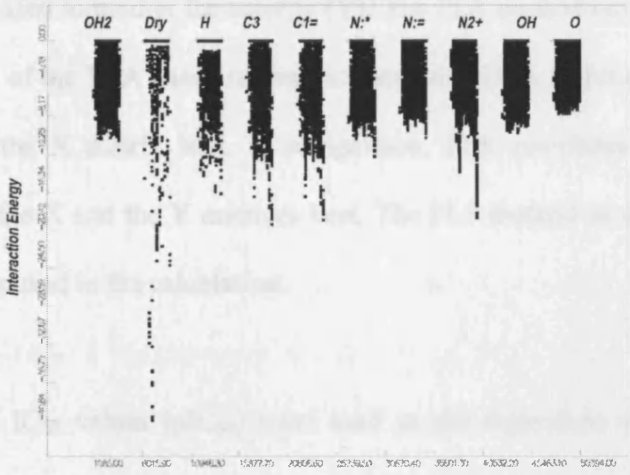
Figure 2.1. The data in PCA analysis are stored in a matrix, with N observations and K variables. The analysis builds a new model containing all the observations and the variables in the original data set are replaced by a few new uncorrelated variables M, called principal components (PC). By reducing the number of variables, the PCA model shows relationships between observations and variables and among observations themselves.

Consensus PCA (CPCA)

Some multivariate data are organised in blocks, each block describes one molecular force. For example, in GRID, the interaction energy values are calculated using probes representing different chemical properties and data are separated into the corresponding blocks. However, the variance can be very different among the data blocks. The energy values generated by probes representing weak non-bonded interactions such as van der Waals force and hydrophobic attractions will be masked by those generated by stronger interactions like hydrogen bonds. Since the weak forces are equally important in molecular interaction, it is necessary that their effects are considered in the CPCA model. To overcome this problem, a scaling process is applied to the data to normalise their importance in the model. The scaling method used in GOLPE is named block unscaled weights (BUW) scaling, in which data generated by each probe are organised into one block and weighting coefficients are calculated for each block. The probes are scaled according to the weighting coefficients, which gives each probe the same importance in the model while the relative scales of variables within the block do not change. Figure 2.2 illustrates the BUW scaling. Figure 2.2a shows the initial variable distribution in each probe, and figure 2.2b shows the normalised variable distribution after the scaling.



a



b

Figure 2.2. The distribution of the variables for each probe. a. before the block unscaled weights (BUW) scaling, and b. after the block unscaled weights (BUW) scaling.

2.2.3.2 Partial least squares (PLS)

The partial least squares (PLS) method is an effective technique for finding the relationship between the properties of a molecule and its structure. In mathematical terms, PLS relates a matrix of dependent variables Y, in this case binding affinity data, to a matrix of molecular structure descriptors X (Wold, 1995) (fig 2.2). The aim of PLS is to find a linear relationship between X and Y: $Y = XB + E$, where B is the regression coefficient and E is the residuals (noise), and Y can be predicted from X. However, the number of descriptors (X) is often greater than the number of objects (compounds, proteins) (Y) and a linear model cannot be built directly. PLS decomposes the matrix X into several latent variables that correlate best with the activity of the compounds. The latent variables are used to predict the activity (Y). The PLS method can be considered as a variation of the PCA based regression methods. PCA searches for variables that explain the X matrix best, in comparison, PLS calculates variables that explain both the X and the Y matrices best. The PLS method as implemented in Sybyl 6.7 was used in the calculation.

Experimental IC_{50} values (pIC_{50}) were used as the dependent variables in the study. Both the column filtering and the scaling were turned off. The optimal number of components was found by running cross-validation using SAMPLS (Bush and Nachbar, 1993).

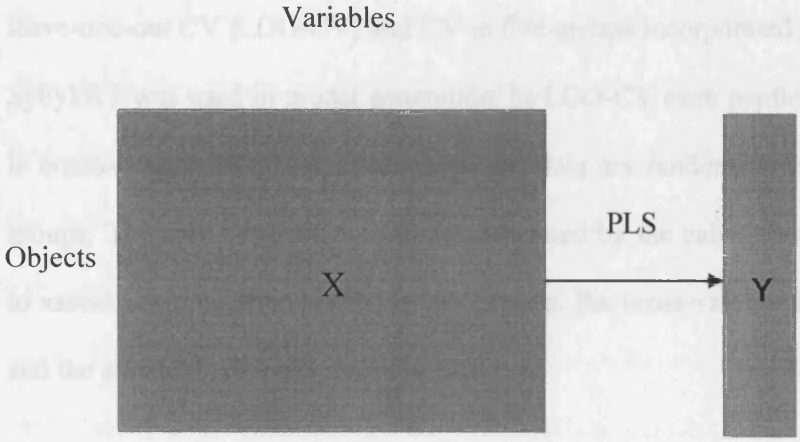


Figure 2.3. The PLS method finds the maximum covariance between the observation matrix X and the biological activity Y. PLS is often used when the number of variables (descriptors) in X is greater than that of the objects (Y). PLS decomposes the X matrix into several latent variables that describe the variance in both X and Y matrix. The latent variables can be used to predict the biological activity of the objects.

2.2.3.3 Cross-validation (CV)

Models produced by PLS are validated using cross-validation. Cross-validation (CV) estimates the predictivity of the model (Wold, 1995). In cross-validation the data are randomly divided into groups, and the activities of the compounds in one group are predicted using the model generated by the rest of the data. The leave-one-out CV (LOO-CV) and CV in five groups incorporated in the program Sybyl 6.7 was used in model generation. In LOO-CV each peptide in the model is omitted once. In CV in five groups the data are randomly divided into five groups. The following parameters are generated by the calculation and are used to assess the predictive ability of the models: the cross-validated coefficient q^2 and the standard error of prediction *SEP*:

$$q^2 = 1 - \frac{\sum_{i=1}^{n} (pIC_{50\,\mathrm{exp}} - pIC_{50\,pred})^2}{\sum_{i=1}^{n} (pIC_{50\,\mathrm{exp}} - pIC_{50\,mean})^2} \qquad \text{Eqn. 2.2}$$

$$SEP = \sqrt{\frac{\sum_{i=1}^{n} (pIC_{50\,\mathrm{exp}} - pIC_{50\,pred})^2}{n}} \qquad \text{Eqn. 2.3}$$

where n represents the number of the peptides included in the model (for LOO-CV, n equals the number of peptides-1), pIC_{50pred} and pIC_{50exp} are the values predicted by LOO-CV for the binding affinity and from the binding experiments, respectively. The q^2 represents the predictivity of the model, its value is between 0 and 1. The *SEP* is the error in prediction and is usually lower than 1. A q^2 value

above 0.5 with small *SEP* indicates good predictivity of the model. However, the q^2value is often lower than 0.5 and a q^2 value of 0.3 or better is generally accepted as good predictivity.

After cross-validation, a non-cross-validated model is generated by PLS using the number of principal components (PC) derived in CV. Three values are obtained in the calculation: the variance explained by the model (r^2), the standard error of estimate (*SEE*), and the *F* ratio. An r^2 value close to 0 shows that none of the biological activity is explained by the model, and $r^2$of 1 indicates 100% explanation. An r^2 value of 0.7 or greater could be considered as a good fit to the model. The *SEE* is how confident the r^2 is and usually is smaller than 1. The F, or Fisher ratio, is the ratio of r^2 to 1- r^2 (explained to unexplained). It estimates how significant the regression equation is. A higher F ratio means more biological properties are explained by the model.

2.2.3.4 ROC analysis

Receiver operating characteristic (ROC) curve is a standard method used to analyse scientific and clinical data. The ROC curve gives a graphical representation of the level of the true positive rate (sensitivity) and the false positive rate (specificity) of the data at different levels of cut off. The sensitivity and the specificity can be calculated using the following equations:

$$Sensitivity = \frac{TP}{TP + FN} \qquad \text{Eqn. 2.4}$$

$$Specificity = \frac{TN}{TN + FP} \qquad \text{Eqn. 2.5}$$

where *TP* is the number of true positives, *FN* is the number of false negatives, *TN* represents the number of true negatives and *FP* is the number of false positives. An example of the ROC curve is in figure 2.10.

Usually a ROC curve takes the shape of a plateau curve (fig 2.4). The area under the curve (Aroc) indicates the quality of the data. The larger the Aroc, the better the data. Usually, Aroc values between 0.6 and 0.8 are good and values higher than 0.8 are excellent.

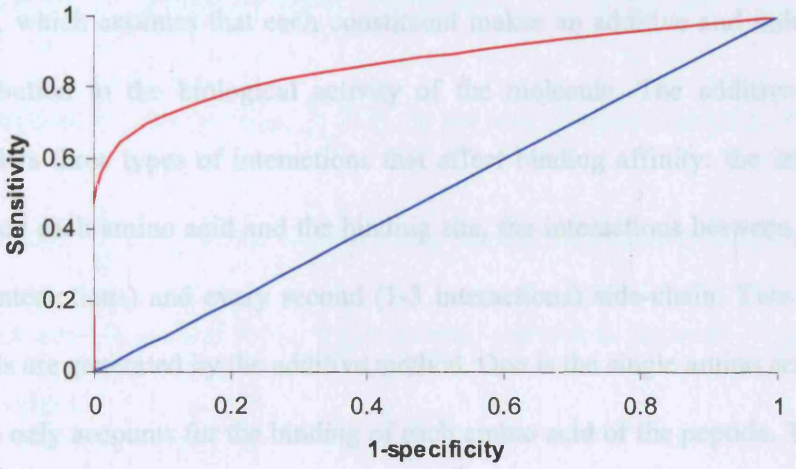


Figure 2.4. An example of an ROC curve. The red curve is a ROC curve. The blue line is when the data is random. The area under the curve indicates how good the data is. Aroc values between 0.6 and 0.8 are good and values higher than 0.8 are excellent.

2.2.4 Modelling

2.2.4.1 The additive method

The additive method is based on the Free-Wilson approach (Free and Wilson, 1964), which assumes that each constituent makes an additive and independent contribution to the biological activity of the molecule. The additive method considers three types of interactions that affect binding affinity: the interaction between each amino acid and the binding site, the interactions between adjacent (1-2 interactions) and every second (1-3 interactions) side-chain. Two types of models are generated by the additive method. One is the single amino acid model, which only accounts for the binding of each amino acid of the peptide. The other is the amino acid and interactions model, which considers both the contribution of individual amino acid and interactions between adjacent and every second amino acids. However, it should be noted that because the peptide is linear, when considering 1-3 interactions, the interactions of residues at the two ends of the peptide are considered only once (eg. P1 only interacts with P3), while residues at other positions are considered twice (eg. P3 interacts with P1 and P5). An illustration of the amino acids interactions is in figure 2.5.

For a nonamer peptide, the single amino acid model and the amino acid and interactions model are given by equation 2.6 and 2.7, respectively:

$$pIC_{50} = const + \sum_{i=1}^{9} P_i \qquad \text{Eqn.2.6}$$

$$pIC_{50} = const + \sum_{i=1}^{9} P_i + \sum_{i=1}^{8} P_i P_{i+1} + \sum_{i=1}^{7} P_i P_{i+2} \qquad \text{Eqn.2.7}$$

where pIC_{50} is the binding affinity expressed in p-units (-logIC_{50}), the *const* accounts for the peptide backbone contribution, $\sum_{i=1}^{9} P_i$ is the sum of amino acid contributions at each position, $\sum_{i=1}^{8} P_i P_{i+1}$ is the sum of adjacent peptide side-chain interactions, $\sum_{i=1}^{7} P_i P_{i+2}$ is the sum of every second side-chain interactions. Note that the term P_iP_{i+1} and P_iP_{i+2} represent the interactions between neighbouring amino acids.

A flowchart for generating additive models is given in figure 2.6. The first step is to build a data matrix. The computer-generated matrix consisted of 6181 columns. The number of rows was equal to the number of peptides in the study, data from one peptide were stored in one row. The first column of the matrix represented the dependent variable pIC_{50}. The next 180 columns represented the single amino acid contributions, the following 3200 (20x20x8) columns represented the contributions of adjacent amino acid interactions, and the last 2800 (20x20x7) columns were for the side chain interactions of amino acids at every second position. The existence of each amino acid and each interaction were recorded in the matrix. If present, the matrix recorded 1 in the corresponding column, otherwise the element would be 0. Any column containing only zero values was deleted to reduce the time used in calculation. After the matrix was constructed, the additive model was generated and validated using PLS.

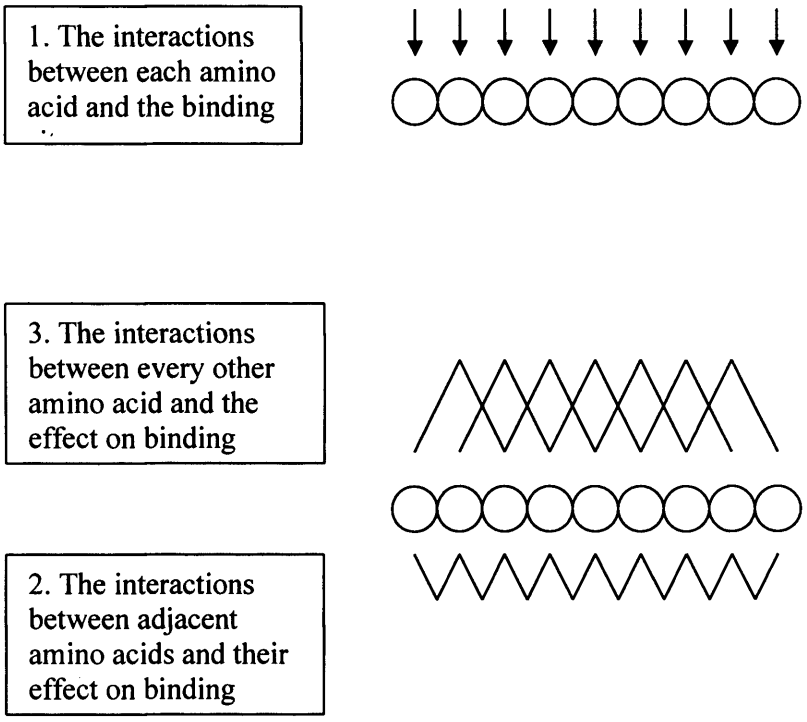


Figure 2.5. An illustration of the interactions considered by the additive model. Each circle represents one amino acid of the peptide. Three types of interactions are taken into account by the additive model: 1. the interactions between each amino acid and the binding site, 2. the interactions between adjacent amino acids and their effect on peptide binding to the MHC binding site, 3. the interactions between every other amino acid and the effect on peptide binding.

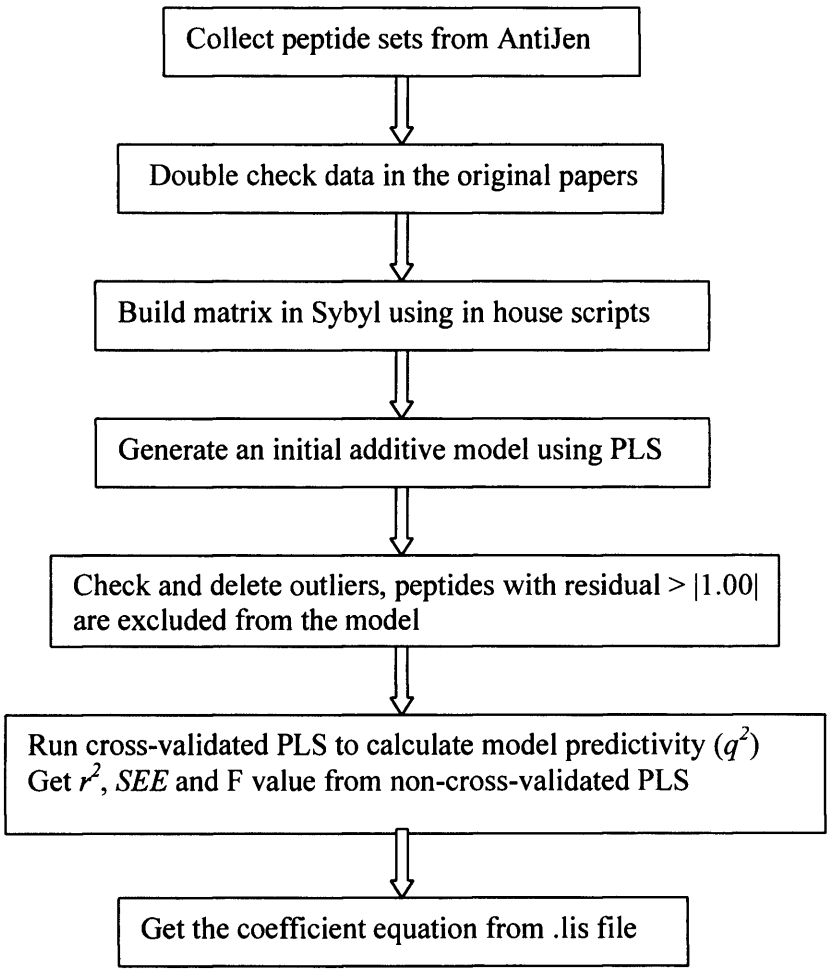


Figure 2.6. Steps in generating the additive models.

2.2.4.2 Molecular modelling and CoMSIA

No X-ray data were available for a peptide binding to the HLA-A3 alleles. As the A2 supertype is the closest to the A3 supertype, a crystallographic structure of the peptide TLTSCNTSV binding to the HLA-A*0201 allele was chosen as a starting conformation (Madden *et al.*, 1993). All molecular modelling calculations were performed on a Silicon Graphics workstation using Sybyl 6.7 as previously described (Doytchinova and Flower, 2001). A flow-chart of the CoMSIA calculation process is in figure 2.7. Peptide sequences and their binding affinities were collected from the AntiJen database. The X-ray structure of the peptide TLTSCNTSV backbone was used as a template to build all the peptides in the data set and CoMSIA models are generated using PLS.

The peptides were evaluated using the five CoMSIA physicochemical properties included in the QSAR module of Sybyl 6.7: steric, electrostatic, hydrophobic, hydrogen donor and hydrogen bond acceptor properties. The properties were evaluated using a probe atom placed at regular intervals within the grid. The probe had a radius of 1Å, charge, hydrophobicity, hydrogen bond donor and acceptor properties all equal to +1. Similarity indices were calculated using Gaussian-type distance dependence between the probe and the atoms of the peptides tested.

The CoMSIA models were built using PLS. The initial models were calibrated with respect to the grid spacing, attenuation factor and column filtering. The grid was extended 2.0Å beyond the aligned molecules. Different values were tested for grid spacing: 1.0 to 2.5Å in steps of 0.5Å. Values for the attenuation factor

varied from 0.3 to 0.7Å in steps of 0.1Å. Column filtering from 0.5 to 1.5Å in steps of 0.5Å. The cross-validated models were assessed by q^2, standard error of prediction (SEP) and the mean value of the residuals between experimental affinities and those predicted by leave-one-out cross-validation (LOO-CV), presented as negative logarithms of IC_{50}. The non-cross-validated models used the optimal number of components found by LOO-CV and were assessed by the non-cross-validated r^2, standard error of estimate (SEE) and F-ratio. The ratio of the standard errors to the affinity range was used as a more effective measure of model predictivity and goodness of fit.

The results of the non-cross-validated CoMSIA models were displayed using contour maps. The contour maps highlighted whether changes in the peptide structure favour or disfavour binding to MHC molecule. The maps from the present study were generated using the StDev*Coeff option, using the actual values. Five maps for each of the four alleles were generated for each of the five physicochemical properties.

2.2.4.3 SIMCA

SIMCA (Soft Independent Modelling of Class Analogy) is useful in solving pattern recognition and classification problems. It is commonly used in multivariate data analysis because of its ability to group variables with similar properties into smaller groups and reduce descriptor redundancy before generating the QSAR models. The statistics package used to perform the calculations in the project was SIMCA-P 8.0. SIMCA uses PLS to build QSAR models.

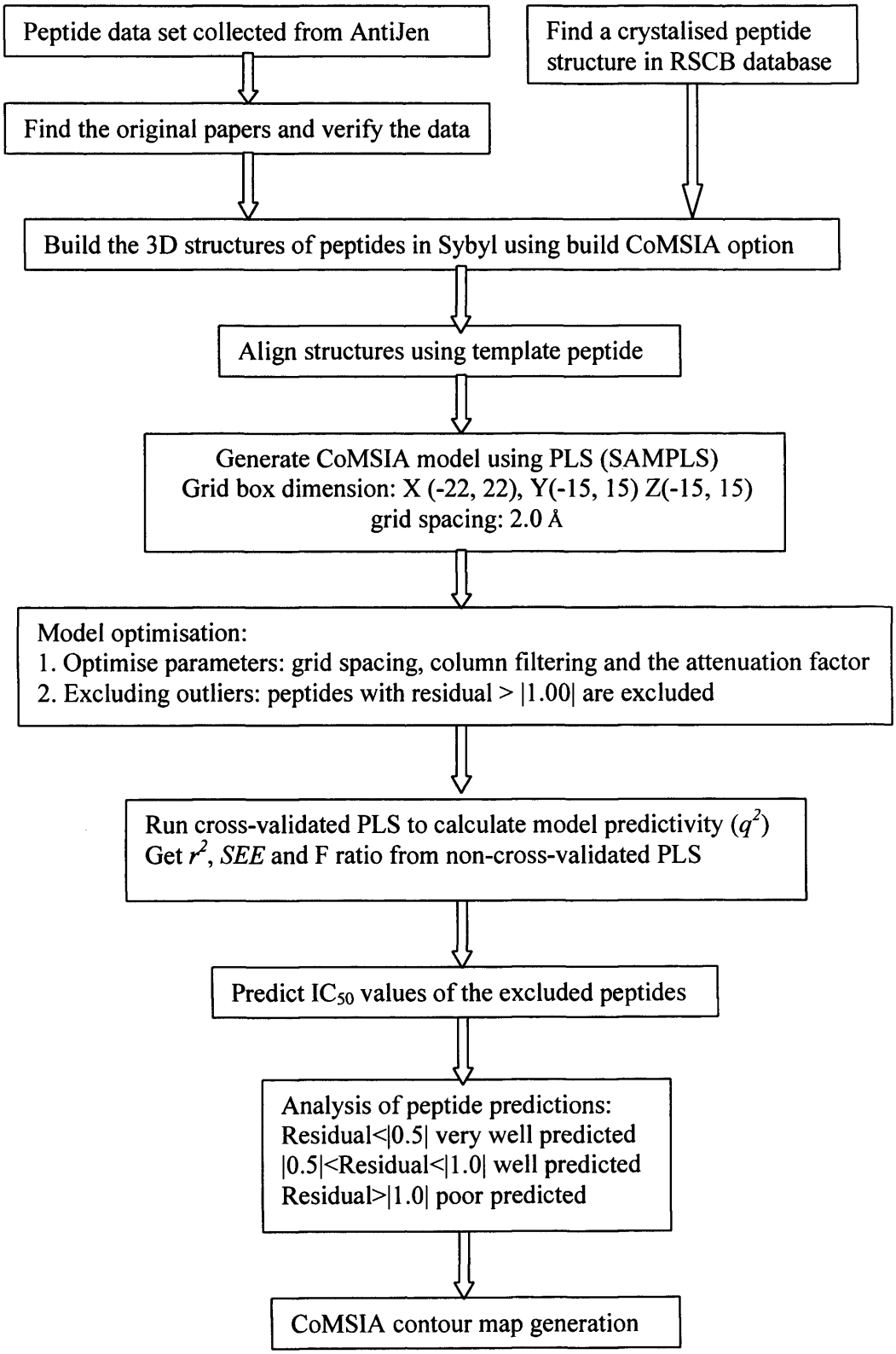


Figure 2.7. A flowchart of the CoMSIA model generation process.

All the binding data and selected descriptors (see section 2.1.10) were organised into a matrix before calculation. The first column of the matrix contained the peptide sequences. The experimental value (pIC_{50}) was in the second column. The rest of the columns contained the descriptor variables. The matrix was stored in the Excel format and was imported into SIMCA.

The models generated in the project were fitted using the 'Autofit' option, which calculated the cross-validated coefficient q^2 and the explained variance r^2. The correlation between the individual variable and the data matrix was observed in the variable importance in the projection (VIP). VIP is the sum of the variable influence over all model dimensions. Equation 2.8 was used to calculate VIP values.

$$VIP_k = \sum_a (VIN)_k^2 \qquad \text{Eqn. 2.8}$$

where a was a given PLS dimension, k was the number of variables and VIN was the individual variable influence. Higher VIP values (VIP > 0.7) indicated good correlation between the variable and the data. To improve the model quality, variables with low VIP values were excluded from the model in a stepwise manner.

2.2.4.4 The genetic algorithm (GA)

The genetic algorithm (GA) was developed by Holland (Holland, 1975). The GA is inspired by Darwin's theory of evolution, which states that those individuals whose genetic composition fit best with their environment will survive and have

a greater chance of reproduction. Also, the offspring produced from the more 'fit' individuals contain the combination of their genomes and may have a higher chance of survival.

The procedure of the GA is summarised in figure 2.8. Briefly, the variables are randomly grouped into binary strings. Each string is called a chromosome and each variable is called a gene. Initially a random group of chromosomes is chosen as the starting population (Devillers, 1996). The fitness of the chromosomes is calculated by a fitness function. Different fitness functions can be used according to the nature of the problem. The population is then improved by introducing new variables, which are produced by chromosome crossover and gene mutation. In crossover, two chromosomes are randomly selected as the parents, and new chromosomes are produced by crossing over between the parent chromosomes (figure 2.9). The crossover point is selected randomly and the genes to the right of the crossover point are exchanged between the parents. The probability of crossover is often set to a high value, for example, 0.9. In point mutation, a random gene is selected. If the gene encoded 1 originally, then it will be changed to 0 and vice versa. The number of genes mutated can be set up at the beginning to control how similar the child is to the parent. If the number is big, many genes will be changed and the child will be very different from the parent. This is not good for population convergence and finding the optimal solution may be slowed down. Therefore the percentage of chromosomes that undergo mutation is usually set to a small value, such as 0.1.

The quality of the new population is assessed by repeating the evaluation step, and the result is compared with that of the original population. If the new population is better than the original, then the original population is discarded, otherwise the original population will be kept. The evaluation, crossover and mutation steps are repeated until the quality of the population cannot be improved.

GA has been applied to generate QSAR models (Bangalore *et al.*, 1996; Hasegawa and Funatsu, 2000; Shaffer *et al.*, 1996; Yamashita *et al.*, 2002). As PLS is a standard statistical method used in QSAR modelling, most GA-QSAR studies combine GA and PLS to derive a new algorithm that is termed genetic partial least squares (GA-PLS). The fitness is usually measured by the predictive ability of the model, that is, q^2. The GA-PLS used in this project was designed by Shun Jin Chou at the Laboratory for Molecular Modelling, School of Pharmacy, UNC (http://mmlin1.pha.unc.edu/~jin/QSAR/analyze.html).

$$1-\frac{(n-1)(1-q^2)}{n-c} \qquad \text{Eqn. 2.9}$$

Equation 2.9 was used as the fitness function in the web implemented GA-PLS calculation. q^2 is the predictivity of the model, n is the number of compounds in the data set and c is the optimal number of components.

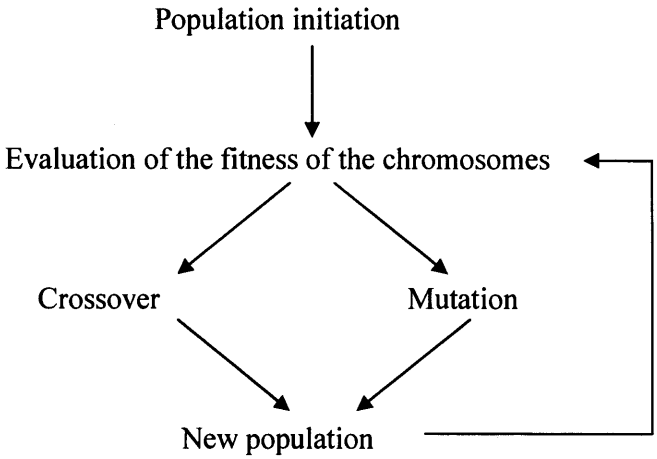


Figure 2.8. The flowchart of the genetic algorithm. A random chromosome population is defined at the start of the calculation. The fitness of the chromosomes is evaluated by the fitness function, after which new members of the population are generated by point mutation and crossover between two chromosomes. The fitness of the new chromosomes are then evaluated using the fitness function, if the new chromosomes have a higher level of fitness, then the parent chromosomes are discarded, otherwise the parent chromosomes are kept in the population. The process continues until no new chromosomes with a higher level of fitness can be generated.

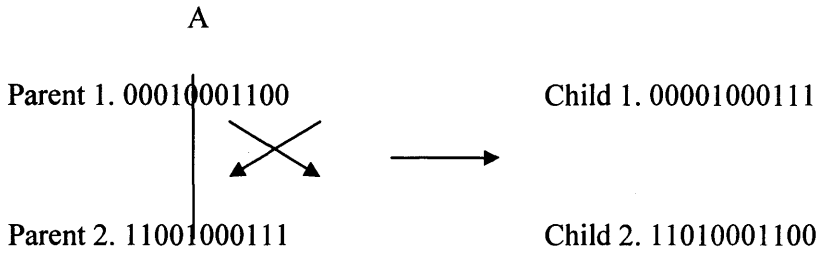


Figure 2.9. Example of crossover between parents. Crossover point A is selected randomly. Values to the right of the crossover point are exchanged between the parents.

2.2.4.5 GRID

The GRID program (version 21) finds the energetically favoured or disfavoured regions on molecules with known three-dimensional structures. Many molecules can be included in one calculation (Cruciani and Watson, 1994). A selection of chemical probes is included in the program, which represent atoms or functional groups with different properties. GRID calculates the interaction energy between selected chemical probes and each of the molecules.

A GRID box was defined to only include the peptide binding site in the calculation (figure 2.10). The dimensions of the GRID boxes used for each of the HLA classes are in table 2.3. The grid spacing was set to 2Å.

	X		*Y*		*Z*	
HLA-A	-9.44	13.43	-17.43	21.58	-19.16	15.21
HLA-B	-26.82	11.54	-17.47	8.37	-26.97	-3.36
HLA-C	-19.73	22.96	-13.13	11.23	-17.47	9.17

Table 2.3. The dimensions (Å) of the GRID box for different HLA class molecules.

GRID uses different probes placed at a regular interval throughout the grid box to calculate the interaction energy between the molecule and the probes. A total of 13 probes were used in the calculation, which covered the chemical functionality of the 20 naturally occurring amino acids (table 2.4). For other

parameters, the MOVE was set to 1 to make protein side chains flexible. The LIST value was set to –2, facilitating importing the results into GOLPE.

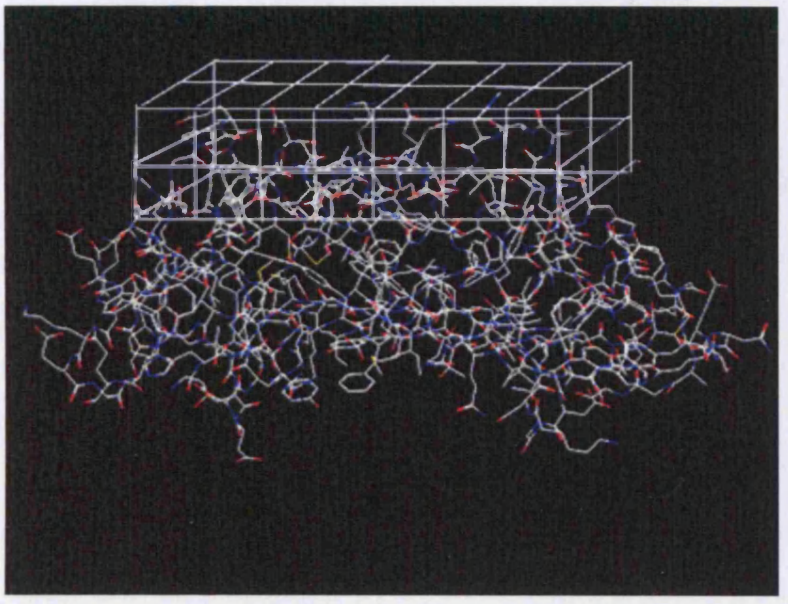


Figure 2.10. An illustration of the GRID box defined for the experiment. The molecule displayed in the graph is the α1 and α2 domain of HLA-A*0201. The grid box (shown in white) is defined to include only the peptide binding site of the MHC molecule.

Probe	*Chemical group*	*Represented amino acids*
OH2	Water	Hydrophilic amino acids
Dry	The hydrophobic probe	Hydrophobic amino acids
H	Hydrogen	Hydrogen bond donor/accepter
C3	Methyl CH3 group	Aliphatic amino acids
C1=	sp2 CH aromatic or vinyl	Phe, Tyr, Trp, His
N:*	sp N with a lone pair	His
N:=	sp2 N with a lone pair	Asn Gln
N1	Neutral flat NH eg. Amide	Any amino acids
N2+	sp3 amine NH2 cation	Arg Lys
O1	Alkyl hydroxy OH group	Ser Thr
OH	Phenol or carboxy OH	Tyr Asp Glu
O	sp2 carbonyl oxygen	Asp Asn Glu Gln
S1	Neutral SH	Cys Met

Table 2.4. List of GRID probes used in the study. A total of 13 probes are selected from probes offered in GRID. These probes are chosen to represent different characteristics of the twenty amino acids.

2.2.4.6 GOLPE

Generating Optimal Linear Partial least square Estimations (GOLPE) (Cruciani and Watson, 1994) improves the predictivity of the model by comparing the contributions of each variable, and excluding those that make very small or no contributions. In this way, the model generated by GOLPE has a higher level of predictivity than the one generated by PLS alone. The major steps in GOLPE calculation are summarised in figure 2.11 (Cruciani and Watson, 1994).

GOLPE also has one module for PCA calculation. The principal components are obtained by maximising the variance of linear functions of the matrix. The results of the GRID fields calculations were stored in files with .kont extension and were imported into GOLPE. The data were pre-treated before calculation, all the data with absolute values smaller than 0.01 or with standard deviation less than 0.01 were deleted. Positive interaction energy represented unfavourable steric repulsion between the probe and the molecule, therefore it was removed by setting the maximum cut off to 0 kcal/mol.

Additionally, when the scores plots failed to give well-defined clusters, a cut off option in GOLPE was used to reduce the number of non-significant interactions and improve the signal/noise ratio. In the GRID/CPCA study a cut off region of 4Å within the binding site was applied for HLA-B molecules.

After calculating GRID energy fields using each probe, the probes that gave the highest explained variance by the first three PCs were selected and a GRID

calculation was run using all these probes. The results were used to build a consensus PCA (CPCA) model.

When more than one probe is used in the GRID calculation, the data generated by different probes are grouped into blocks, and they are often analysed by hierarchical PCA methods such as CPCA. The advantage of such methods over PCA is that they compare the relative importance of each block in the calculation and make a 'consensus' clustering of the objects. CPCA uses the same principle as PCA: a CPCA model tries to explain the overall variance of the original data matrix. The algorithm used in CPCA is an adaptation of the NIPALS algorithm used in PCA (Wold *et al.*, 1987). Like PCA, CPCA calculates the principal components and gives the scores and loading matrix. In addition, CPCA also calculates the importance of each data block. It calculates the scores and the loading matrix for each probe used, and gives the weight matrix that illustrates the contribution of each probe in the overall scores.

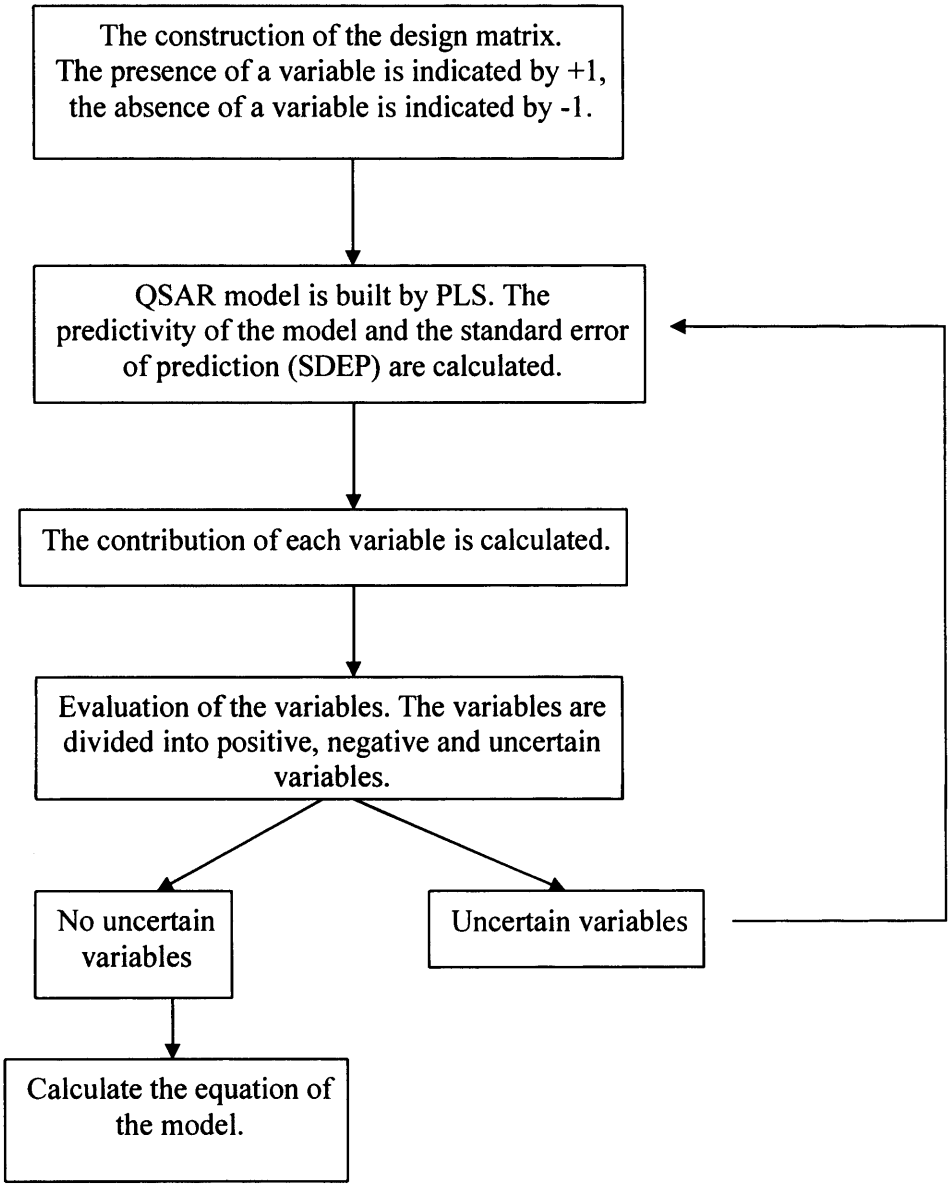


Figure 2.11. The flowchart of the GOLPE process.

Chapter 3
HLA-A2 and A3 motif definition using 2D-QSAR methods

3.1 Introduction

Quantitative structure-activity relationships (QSAR) relate the biological activity of a molecule to its structure. The hypothesis behind QSAR is that the biological property of a molecule is determined by its chemical structure: changes in the structure lead to changes in the property (Sneath, 1966). QSAR techniques can be 2D or 3D, the former uses 2D or physico-chemical descriptors to study the molecules while the latter also takes the spatial features of the molecules into account (Collantes and Dunn, 1995; Fauchere *et al.*, 1988; Norinder, 1991). Statistical methods, such as partial least square (PLS) analysis, are used in QSAR methods to produce models that explain properties of the protein that lead to changes in the activity (Felipe-Sotelo *et al.*, 2003; Hasegawa and Funatsu, 2000).

Amino acid descriptors are often used in peptide QSAR studies (Cui *et al.*, 2002; Eriksson *et al.*, 1990; Gupta *et al.*, 2002; Nadasdi and Medzihradszky, 1981). They describe the physico-chemical properties of the peptides quantitatively. Many of the properties are measured experimentally, such as pKa, hydrophobicity and logP, using methods such as TLC, HPLC, and spectroscopy(Sun, 2004; Xing and Glen, 2002; Zhao *et al.*, 2002). Other properties cannot be measured but can be calculated, such as the molecular surface area and atomic charges. The quality of the amino acid descriptors is an important factor in producing models with a good level of predictivity (Hunt, 1999).

The first study of amino acid descriptors was undertaken by Sneath (Sneath, 1966), who used physico-chemical semi-quantative data to derive descriptors for the 20 naturally occurring amino acids. Since then, a number of studies have generated many new descriptors. Kidera et al. statistically analysed 188 amino acid indices and divided them into groups according to the properties they represent (Kidera *et al.*, 1985). On the basis of Kidera's data and some new additions, Nakai et al. carried out another cluster analysis of 222 amino acid indices, dividing them into four major groups: α and β turn propensities, β propensity, hydrophobicity and physico-chemical properties (Nakai *et al.*, 1988). In the same year, Fauchere et al. chose a group of 15 physico-chemical descriptors and applied them to 20 natural and 26 synthetic amino acids (Fauchere *et al.*, 1988). To facilitate public access to these descriptors, Kawashima collected most published descriptors and established a database named AAindex (Kawashima and Kanehisa, 2000).

One set of the amino acid descriptors commonly used in peptide QSAR studies is the z descriptors, which is obtained by applying principal component analysis to groups of physico-chemical variables (Hellberg *et al.*, 1987). z descriptors, which mainly explain hydrophilicity, size and polarity of the amino acids, have been used to characterise amino acids and synthetic peptides (Eriksson *et al.*, 1989). Later Sandberg used the z descriptors to classify 89 synthetic elastase substrate and 29 neurotensin peptide analogues (Sandberg *et al.*, 1998). The quantitative structure-activity modelling (QSAM) generated models with high predictivity and a high level of explained variance (q^2 = 0.77 r^2 = 0.83 for electase substrates and q^2 = 0.78 r^2 = 0.93 for neorotensin analogues. Later, the z descriptors were

used to model the relationship between the functions of the peptides and their amino acid composition (Siebert, 2001).

The additive method is a 2D-QSAR method (Doytchinova *et al.*, 2002). Previously the additive method was applied to build an A*0201 peptide-binding model and an A2 supermotif (Doytchinova and Flower, 2003). In the present study the method was applied to alleles of the A3 serotype to define an A3 supermotif (Guan *et al.*, 2003).

The additive method is based on the Free-Wilson concept (Craig, 1974; Kubinyi and Kehrhahn, 1976), whereby each constituent makes an additive and independent contribution to the biological activity (Free and Wilson, 1964). Additional terms were added to the basic QSAR model to account for the adjacent and every second side-chain interactions. For a nonamer peptide the model could be presented by equation 3.1 (Doytchinova *et al.*, 2002):

$$pIC_{50} = const + \sum_{i=1}^{9} P_i + \sum_{i=1}^{8} P_i P_{i+1} + \sum_{i=1}^{7} P_i P_{i+2} \qquad \text{Eqn. 3.1}$$

where pIC_{50} is the binding affinity expressed in p-units (negative decimal logarithm of IC_{50} values), the *const* accounts for the peptide backbone contribution, $\sum_{i=1}^{9} P_i$ is the sum of amino acid contributions at each position, $\sum_{i=1}^{8} P_i P_{i+1}$ is the sum of adjacent peptide side-chain interactions, and $\sum_{i=1}^{7} P_i P_{i+2}$ is the sum of every second side-chain interaction. Two types of models were created: one based solely on the amino-acid contributions (single amino acid

model) and another based on both amino-acid contributions and the contributions of amino-acid interactions (amino acid and interaction model).

3.2 Results

A full protein sequence alignment of some HLA A, B and C alleles is shown in figure 3.1 to 3.3. The multiple sequence alignment was performed using the online protein sequence analysis server clustalw (Combet *et al.*, 2000), URL: http://npsa-pbil.ibcp.fr/cgi-bin/npsa_automat.pl?page=/NPSA/npsa_clustalw.html. The sequence alignment showed that most of the polymorphic residues were within the α1 and α2 domains. The crystal structure of the A*0201 allele revealed that the binding site was within the α1/α2 domain (Saper *et al.*, 1991). A total of 58 positions were polymorphic in the α1/α2 domain alignment, among which 16 positions were inside the binding site: position 9, 24, 45, 63, 66, 70, 73, 77, 80 of α1 domain, and position 5, 7, 9, 24, 26, 62, 66 of α2 domain.

```
                    10        20        30        40        50        60
                    |         |         |         |         |         |
A0201x1     GSHSMRYFFTSVSRPGRGEPRFIAVGYVDDTQFVRFDSDAASQRMEPRAPWIEQEGPEYW
A0202x2     ..........................................R.................
A3101x5     ........T.....................................................R....
A0101x0     ..............................................K.................
A0301x3     ..................................................................
A1101x4     ........Y.........................................................
A6801x6     ........Y.........................................................
B2704x7     ........H.............T.......L..........P.E.....................
B070201     ........Y.............S..................P.E.....................
Cw010201    C....K.................S..................P.G.....V.........
Cw04010101  ........S....W...........................P.G...E..V.........
             ****:** **** *********:******* ********** : *** **:*** ****
Prim.cons.  GSHSMRYFFTSVSRPGRGEPRFIAVGYVDDTQFVRFDSDAASQRMEPRAPWIEQEGPEYW

                    70        80        90
                    |         |         |
A0201x1     DGETRKVKAHSQTHRVDLGTLRGYYNQSEA
A0202x2     ..............................
A3101x5     .Q...N......ID................
A0101x0     .Q...NM.....D.AN............D
A0301x3     .Q...N...Q...D................
A1101x4     .Q...N...Q...D...............D
A6801x6     .RN..N...Q...D................
B2704x7     .R..QIC..KA..D.ES.R..LR.......
B070201     .RN.QIY..QA..D.ES.RN..........
Cw010201    .R..Q.Y.RQA..D..S.RN..........
Cw04010101  .R..Q.Y.RQA.AD..N.RK.........D
            * :*:  * ::* .* .* .*  ******
Prim.cons.  DRETRNVKAQSQTDRVDLGTLRGYYNQSEA
```

Figure 3.1. Multiple sequence alignment of α1 domain using the clustalw program. All parameters are set to default. Sequences of the HLA-A, B and C molecules were taken from the IMGT/HLA database. Identical residues are in red. Non-identical residues are in black. The polymorphic residues are in blue or green.

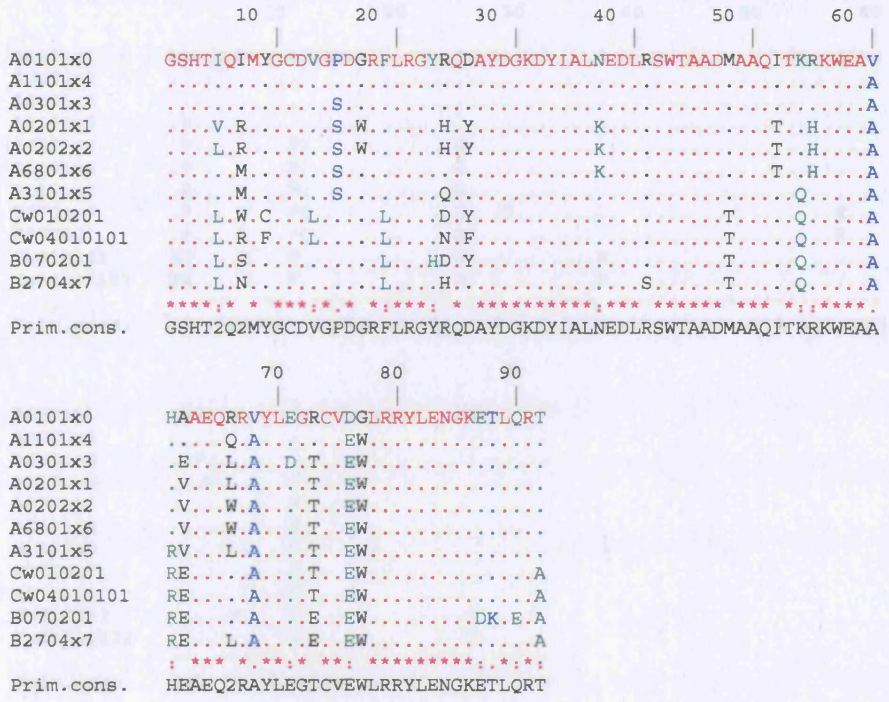


Figure 3.2. Multiple sequence alignment of α2 domain using clustalw program. Identical residues are in red. Non-identical residues are in black. The polymorphic residues are in blue or green.

Figure 3.3. Multiple sequence alignment of $\alpha 3$ domain using clustalw program. Identical residues are in red. Non-identical residues are in black. The polymorphic residues are in blue or green.

3.2.1 The additive HLA-A3 supermotif study

3.2.1.1 The additive models

The additive models were generated for 4 of the A3 supertype alleles: A*1101, A*0301, A*3101 and A*6801. The statistical parameters of the models are given in table 3.1. The number of peptides included in the amino acid only and amino acid and interactions models was different for the A*0301 and A*6801 alleles, because some of the poorly predicted peptides by LOO-CV (peptides with residual value over |1.5|) were excluded. The peptides were excluded in a stepwise fashion and q^2 was re-calculated after each exclusion. The process was repeated until q^2 reached the highest value and started to decrease. Most of the excluded peptides had low experimental affinity due to the absence of anchor or secondary anchor residues.

In general, q^2 of the single amino acid models were higher than that of the amino acid and interaction models. The difference ranged from about 3% for A*1101 and A*3101 to 13% for A*0301 and 16% for A*6801. This was because some amino acid interactions occur only once in the data set. Such interactions created a column in the matrix with only one value and many zeros. They appeared as missing terms in the cross-validated equation used to predict the binding affinity of a peptide with such interactions. The prediction error was proportional to the number of missing terms. The number of missing terms in the single amino acid models was lower therefore their predictivity was higher. As more experimental data becomes available, more peptides will be included in the set and the number of unique amino acids and amino-acid interactions will be reduced. The percentage of well predicted peptides in the training set (with the absolute difference between predicted and

experimental binding affinities less than 0.5) was more than 50% in all models, and the percentage of poorly predicted peptides (with the absolute difference between predicted and experimental binding affinities (pIC_{50}) more than 1) was between 6 and 23% for most of the models.

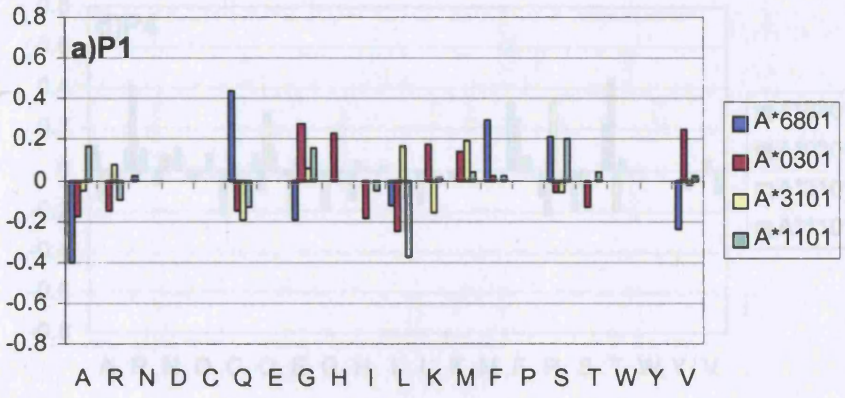
In contrast, the r^2 values were slightly lower for the single amino acid models than for the amino acid and interaction models. This showed that the amino acid side-chain interactions were important in explaining the variance of the peptides and should be included in the models. The $r^2_{bootstrap}$ values are calculated by randomly choosing the sample rows repeatly and calculating the mean r^2 values. The present calculation is the mean r^2 of 20 runs, and it was found that the values were slightly higher than the r^2. If the number of runs were increased to 1000, then there is a slight drop in the mean r^2.

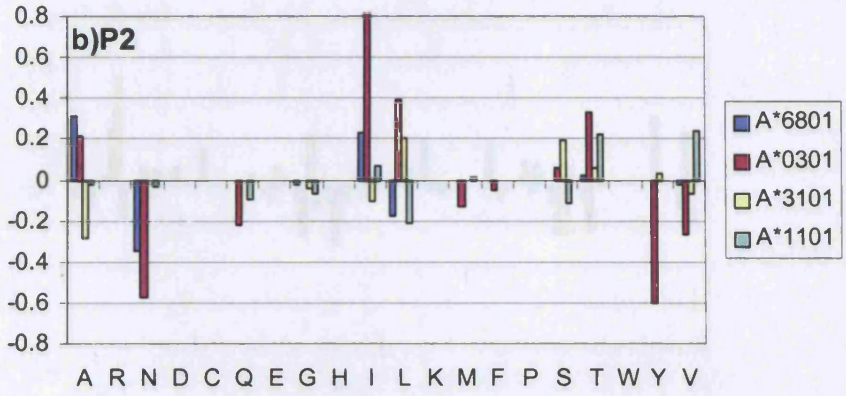
The amino acid and interactions models were used to draw bar-charts for the amino acid contributions at each position of the peptide (figure 3.4). Amino acids with contributions greater than 0.2 were considered as preferred at the specific position and those with contributions less than –0.2 were considered as deleterious. Residues identified as preferred for two or more A3 alleles without being deleterious for any other alleles were considered as preferred for the A3 supermotif. Residues identified as deleterious for two or more alleles were considered as deleterious for the motif.

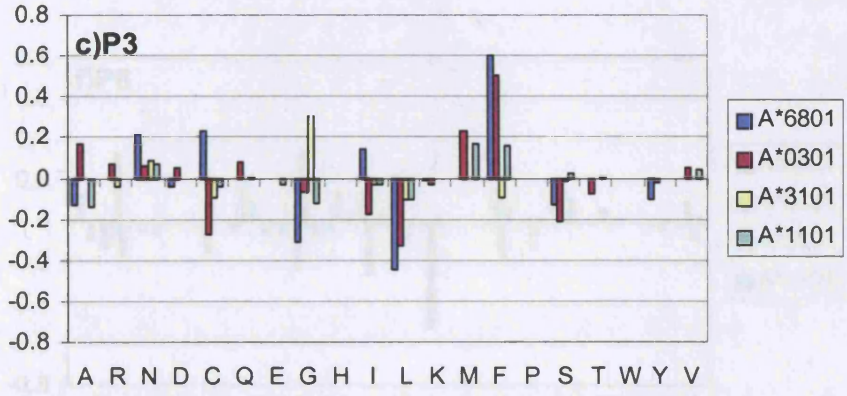
	*A*1101*		*A*0301*		*A*3101*		*A*6801*	
Model	S	I	S	I	S	I	S	I
N[a]	62	62	72	70	30	31	38	37
q^2	0.458	0.428	0.436	0.305	0.482	0.453	0.531	0.370
NC[b]	2	3	6	4	3	6	4	4
SEP[c]	0.572	0.593	0.680	0.699	0.710	0.727	0.594	0.664
q^2_{cv5}[d]	0.433	0.397	0.360	0.294	0.453	0.401	0.510	0.326
r^2	0.829	0.977	0.959	0.972	0.892	0.990	0.959	0.974
SEE[e]	0.321	0.119	0.181	0.136	0.325	0.098	0.175	0.136
$r^2_{bootstrap}$	0.988	0.997	0.976	0.975	0.986	0.994	0.987	0.993
F ratio	143.005	821.098	246.895	557.374	71.356	399.955	194.845	297.481
|res.|≤0.5	36 58.10%	39 62.90%	43 59.71%	44 62.88%	15 50%	16 51.61%	26 68.42%	21 56.76%
0.5 < |res.| ≤ 1.0	16 25.80%	19 30.65%	20 27.70%	15 21.42%	8 26.6%	10 32.26%	8 21.05%	12 32.43%
|res.| > 1.0	10 16.10%	4 6.45%	9 12.50%	11 15.71%	7 23.3%	5 16.13%	4 10.53%	4 10.81%
Mean |residual|	0.507	0.467	0.504	0.527	0.602	0.502	0.418	0.485
Standard deviation	0.423	0.354	0.407	0.420	0.400	0.402	0.363	0.373

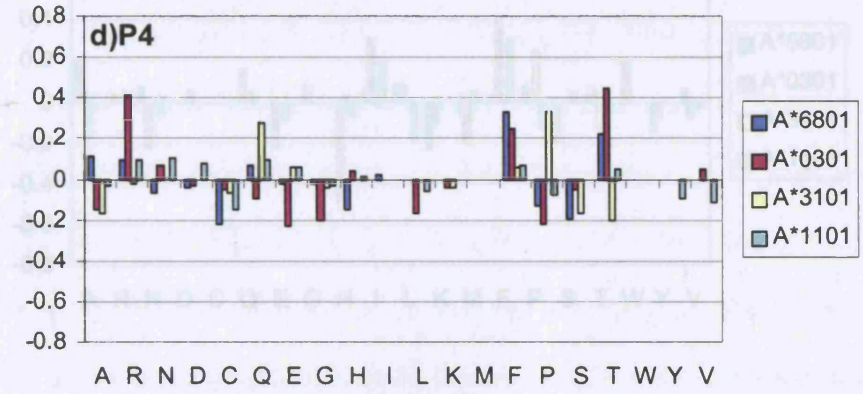
[a]number of peptides. [b]optimal number of components. [c]standard error of prediction. [d]q^2 obtained by cross-validation in five groups [e]standard error of estimate

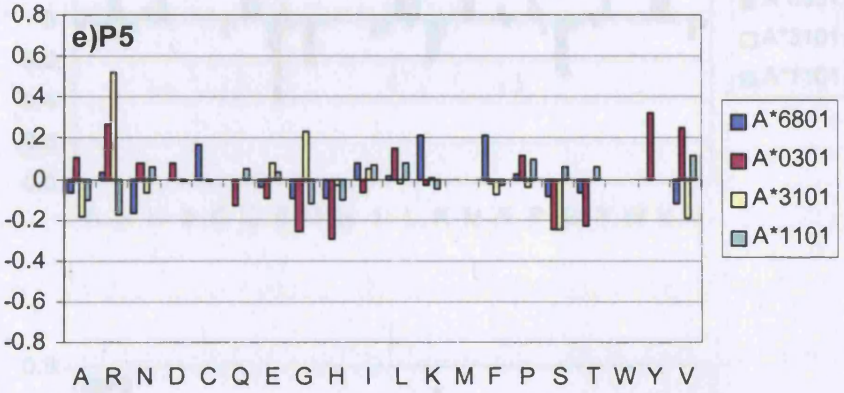
Table 3.1. The A3 models generated by the additive method. A single amino acid model (S) and an amino acid with interaction model (I) were generated for each of the HLA-A3 alleles. The model predictivities were measured by the cross-validated (leave-one-out and cross-validation in 5 groups) q^2 and the standard error of prediction (*SEP*). The r^2 and the standard error of estimate (*SEE*) indicated how much variance in the training set was explained by the model. The F ratio was the ratio of r^2 to 1- r^2, and indicated how much biological activity was explained by the models. The peptide residual values (errors in the predictions) were divided into 3 groups, very well predicted (|res.|≤ 0.5), intermediate (0.5 < |res.| ≤ 1.0) and badly predicted (|res.| > 1.0). The mean residual values and the standard deviation of the residual values were also calculated.

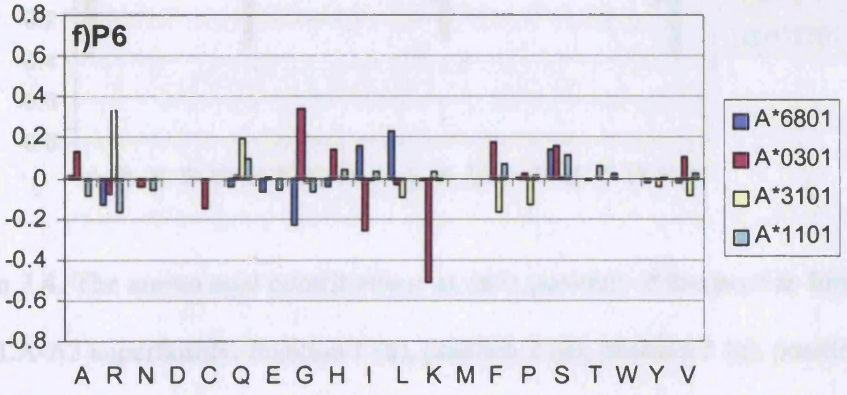

a)P1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
A R N D C Q E G H I L K M F P S T W Y V
A*6801
A*0301
A*3101
A*1101


b)P2
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
A R N D C Q E G H I L K M F P S T W Y V
A*6801
A*0301
A*3101
A*1101


c)P3
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
A R N D C Q E G H I L K M F P S T W Y V
A*6801
A*0301
A*3101
A*1101


d)P4
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
A R N D C Q E G H I L K M F P S T W Y V
A*6801
A*0301
A*3101
A*1101


e)P5
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
A R N D C Q E G H I L K M F P S T W Y V
A*6801
A*0301
A*3101
A*1101


f)P6
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
A R N D C Q E G H I L K M F P S T W Y V
A*6801
A*0301
A*3101
A*1101

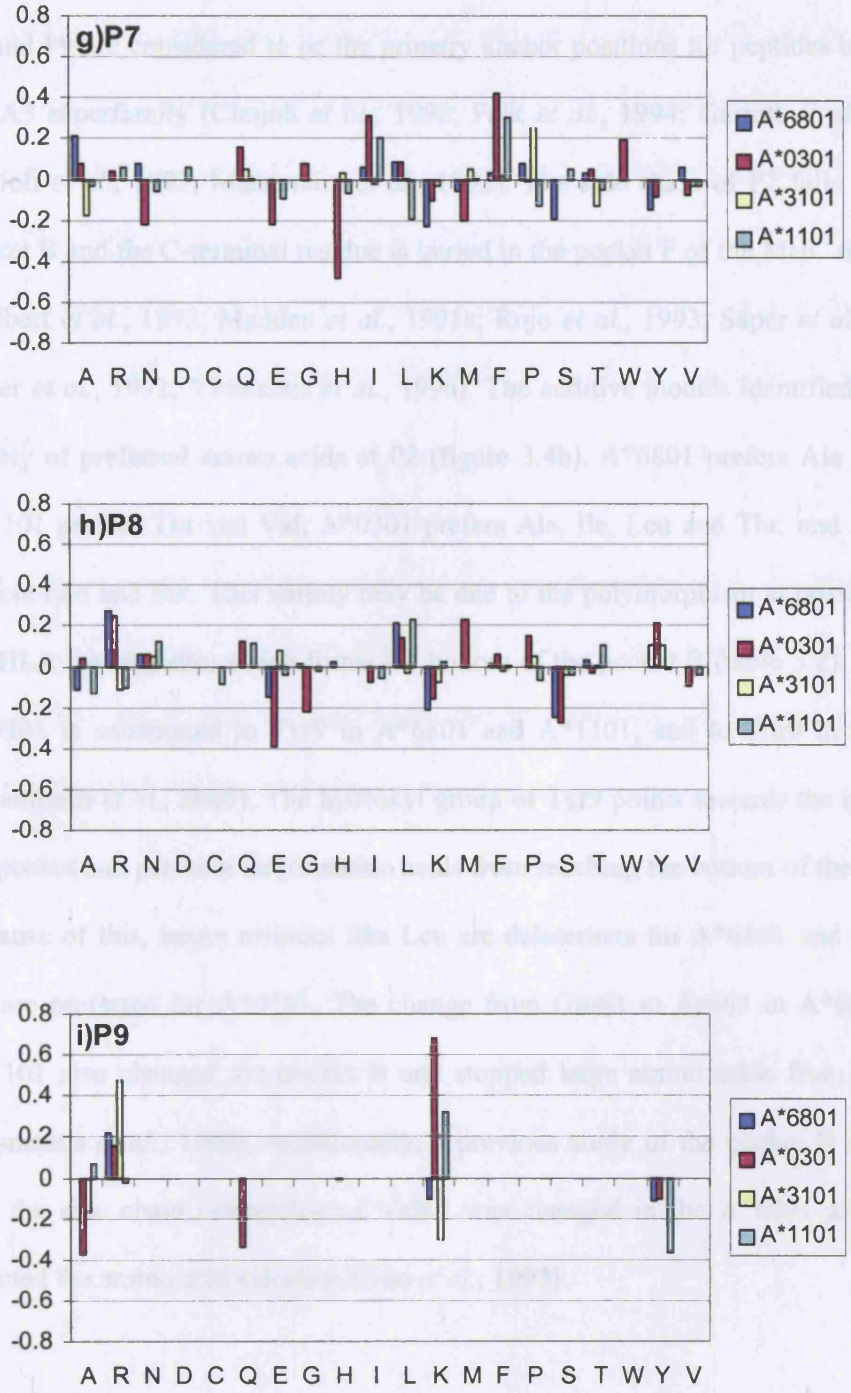


Figure 3.4. The amino acid contributions at each position of the peptide binding to the HLA-A3 superfamily. Position 1 (a), position 2 (b), position 3 (c), position 4 (d), position 5 (e), position 6 (f), position 7 (g), position 8 (h), and position 9 (i).

3.2.1.2 Primary anchor positions

P2 and P9 are considered to be the primary anchor positions for peptides bound to the A3 superfamily (Chujoh *et al.*, 1998; Falk *et al.*, 1994; Garrett *et al.*, 1989; Gavioli *et al.*, 1993; Matsumura *et al.*, 1992). The side chain of P2 falls into the pocket B and the C-terminal residue is buried in the pocket F of the MHC molecule (Colbert *et al.*, 1993; Madden *et al.*, 1991a; Rojo *et al.*, 1993; Saper *et al.*, 1991; Silver *et al.*, 1992; Vasmatzis *et al.*, 1996). The additive models identified a great variety of preferred amino acids at P2 (figure 3.4b). A*6801 prefers Ala and Ile, A*1101 prefers Thr and Val, A*0301 prefers Ala, Ile, Leu and Thr, and A*3101 prefers Leu and Ser. This variety may be due to the polymorphism at position 9 in the HLA binding site, which forms the bottom of the pocket B (table 3.2). Phe9 in A*0301 is substituted to Tyr9 in A*6801 and A*1101, and to Thr9 in A*3101 (Schonbach *et al.*, 2000). The hydroxyl group of Tyr9 points towards the inside of the pocket and prevents larger amino acids from reaching the bottom of the pocket. Because of this, larger residues like Leu are deleterious for A*6801 and A*1101 but are preferred for A*0301. The change from Glu63 to Asn63 in A*6801 and A*1101 also changed the pocket B and stopped large amino acids from binding (Vasmatzis *et al.*, 1996). Additionally, a previous study of the pocket B revealed that the side chain orientation of Val67 was changed in the A*6801 allele and affected the amino acid selection (Guo *et al.*, 1993).

The side chain of P9 extended into pocket F. Most of the residues lining the pocket F were conserved among the HLA-A3 alleles (table 3.2). Positively charged amino acids like Arg and Lys were preferred at this position (figure 3.4i). Two negatively charged Aspartic acid residues, Asp77 and Asp116, were inside the pocket F (table

3.2). These residues provided a negatively charged environment that could make a favourable interaction with positively charged amino acids. Falk and Rotzschke postulated that Asp amino acids at position 77, 74 and 116 were absolutely required in all class I alleles that had an acidic C terminus in their binding motif (Falk and Rotzschke, 1993). This ability could be abolished by the absence of any of the Asps at these positions. The additive models showed that Tyr was deleterious at P9, which may be because of the bulky aromatic ring that could not fit into the narrow pocket.

There was a slight difference in the side chain preference at P9 among the A3 alleles. A*6801 and A*3101 preferred Arg, A*1101 favoured relatively smaller residue Lys and A*0301 accepted both. The difference seemed to be important in determining the binding affinity of the peptide.

	Pocket A									
	5	7	59	63	66	99	159	163	167	171
A*1101	M	Y	Y	E	N	Y	Y	R	W	Y
A*0301	-	-	-	-	-	-	-	T	-	-
A*3101	-	-	-	-	-	-	-	T	-	-
A*6801	-	-	-	N	-	-	-	T	-	-

	Pocket B									
	7	9	24	34	45	63	66	67	70	99
A*1101	Y	Y	A	V	M	E	N	V	Q	Y
A*0301	-	F	-	-	-	-	-	-	-	-
A*3101	-	T	-	-	-	-	-	-	H	-
A*6801	-	Y	-	-	-	N	-	-	-	-

	Pocket C								
	9	22	70	73	74	97	99	114	116
A*1101	Y	F	Q	T	D	I	Y	R	D
A*0301	F	-	-	-	-	-	-	-	-
A*3101	T	-	H	I	-	M	-	Q	-
A*6801	Y	-	-	-	-	M	-	-	-

	Pocket D					
	99	114	155	156	159	160
A*1101	Y	R	Q	Q	Y	L
A*0301	-	-	-	L	-	-
A*3101	-	Q	-	L	-	-
A*6801	-	-	-	W	-	-

	Pocket E						
	97	114	116	147	152	155	156
A*1101	I	R	D	W	A	Q	Q
A*0301	-	-	-	-	E	-	L
A*3101	M	Q	-	-	V	-	L
A*6801	M	-	-	-	V	-	W

	Pocket F										
	73	77	80	81	84	95	116	118	123	124	143
A*1101	T	D	T	L	Y	I	D	Y	Y	I	T
A*0301	-	-	-	-	-	-	-	-	-	-	-
A*3101	I	-	-	-	-	-	-	-	-	-	-
A*6801	-	-	-	-	-	-	-	-	-	-	-

Table 3.2. The alignment of the residues in the HLA-A3 binding pockets.

3.2.1.3 Secondary anchor positions

The presence of the primary anchor residues alone does not guarantee high affinity peptide binding, several secondary anchor positions are also crucial (Zhang *et al.*, 1993). The presence of P3, together with the anchor residues facilitated high affinity binding in previous experiments (DiBrino *et al.*, 1993). In the additive models, P3 preferred aromatic residue Phe (figure 3.4c). The side chain of Phe extended into pocket D (Matsumura *et al.*, 1992), and contacted the side chains of the two conserved Tyr residues at position 99 and 159. Previous peptide binding experiments by Sidney et al. gave similar results (Sidney *et al.*, 1996).

Another secondary anchor position was P7 (Rammensee *et al.*, 1995). Hydrophobic residues were preferred at this position. Phe and Ile were strongly preferred by A*0301 and A*1101. Peptide binding studies showed either P3 or P7, together with residues at P2 and P9, induced stable binding of the peptide (Sidney *et al.*, 1996). Part of the side chain of P7 could contact with pocket E, which accepted residues with a variety of side chains (Madden, 1995).

3.2.1.4 Other positions

The study of the crystal structure of Aw68 (Silver *et al.*, 1992) suggested that P1, P4 and P8 pointed away from the peptide-MHC complex and towards the T cell receptor. In the additive models, Ser and Met were preferred at P1, and Phe, Arg and Tyr were favoured at P4. Arg, Tyr and Leu were slightly favoured at P8, while Ser, Lys and Gln were deleterious. The variance of amino acids accepted at these positions showed that it was less likely that these positions contributed significantly

to the binding to the MHC molecules, and that they may be more important in antigen recognition by T cells.

In the structure of Aw68 (A*6801), residues at P5 and P6 were found lying across the top of the binding groove and have contact with the T cell receptor. In the present study no amino acids were preferred at P5 and Ser was favoured at P6. Similarly to the discussion above, these positions were not particularly important in the binding of the peptide and they might participate in reactions with T cells. The summarised HLA-A3 supermotif is shown in table 3.3.

Preferred	SM	IT	F	FRQ	-	S	FI	RLY	R
position	1	2	3	4	5	6	7	8	9
deleterious	ALQ	N	L	S	GHS	-	-	KSE	Y

Table 3.3. HLA-A3 superfamily binding motif defined by the additive models.

3.2.1.5 Discussion

The present study defined an epitope-binding motif for the A3 superfamily using the additive method. The superfamily classification was based on the peptide binding specificities of the alleles. Class I HLA alleles A*1101, A*0301, A*3101 and A*6801 bound to similar peptides. Six pockets were present in the binding site of the HLA alleles, which interacted with the side chains of the peptides and determined the binding specificity. Sequence analysis showed that only 11 of the residues inside the binding pockets were polymorphic (table 3.2). A good, if incomplete, consensus was found for the preferences at the primary anchor positions P2 and P9. Thr and short hydrophobic residues like Ala and Ile were

favoured at P2, and nearly all the peptides bound to A3 alleles had positively charged residues Arg or Lys at the C terminus.

The amino acids involved in peptide binding were similar among the HLA-A2 and A3 alleles. Pocket B interacted with the side chain of P2, which was one of the anchor positions in nearly all the class I MHC alleles. Most of the amino acids in the pocket B were conserved in HLA-A2 and A3 families, and both families accept hydrophobic residues at P2. The amino acid at sequence position 9 of the HLA protein is important in peptide binding in the two families. Alleles with small to medium sized residues at position 9, such as Phe9 or Thr9, were able to accept peptides with long side chains at P2. Examples of such alleles were A*3101, A*0301 and A*0201. On the other hand, only small residues, like Ala and Val, could bind to A* 6801, A*1101 and A*0206, all of which had the larger residue Tyr9 in the pocket B.

The five residues in pocket F that directly interacted with P9 are identical in both the A3 family and HLA-B27 (Leu81, Asp116, Tyr123, Thr143 and Trp147). Positively charged residues bound in pocket F interact with negatively charged residues Asp116 or Asp77 in the A3 family and HLA-B27. B27 had been shown to accept hydrophobic residues like Leu, Ala and Tyr because they can interact with Leu81, Tyr123, Thr143 and Trp147 in binding pocket F (Jardetzky *et al.*, 1991). In the present study, the specificity at P9 was restricted to Arg and Lys; both Ala and Tyr had deleterious effects on peptide binding. This suggested a possible difference in the conformation of the binding pockets in different alleles in spite of their sequence similarity.

A peptide binding motif for the HLA-A3 superfamily was previously defined by Sidney et al. (Sidney *et al.*, 1996) and Rammensee et al. (Rammensee *et al.*, 1995). Some similarities can be found by comparing the present motif with the ones defined by those two groups. The amino acid preferences for the primary anchor residues were similar. All the motifs had Arg and Lys at P9 and various hydrophobic residues at P2, such as Ile and Thr. The preferences for secondary anchors P3 and P7 in the three motifs were for hydrophobic amino acids. The motif defined by the additive model, while in good agreement with previous motifs, is more extensive, covering all the 9 positions of the peptide.

To conclude, the supermotif of the HLA-A3 familiy can be found in table 3.3. Good binders of the A3-superfamily have a small to medium sized residue at P2, such as Ile or Thr, and a positively charged residue Arg at P9. Phe at P3 and P7 is also required for stable binding. Ser is well accepted at P1 and P6. Although P4 and P8 are more solvent-exposed than MHC-bound, they also had some well-defined preferences. P4 favours Phe, Arg and Gln, and P8 favours Arg, Leu and Tyr.

3.2.2 HLA-A*0201 study using amino acid descriptors

The aim of this study is to use amino acid descriptors and 2D-QSAR techniques to describe the binding motif for the A*0201 allele. The class I allele A*0201 was chosen as it was the best studied HLA allele and had the most binding data available (266 nonamer peptides in AntiJen at the time of the study). Two sets of descriptors were used: the AAindex descriptors and the z descriptors. The question was how to pick up only those that were relevant to the problem from a large

selection of descriptors? To solve this problem, variable selection techniques were used. Three variable selection techniques were applied to the A*0201 peptides: SIMCA, genetic algorithm (GA) and GOLPE.

3.2.2.1 A*0201 models with AAindex descriptors

There were 437 amino acid descriptors in the AAindex database at the time of the study, many of the descriptors described whole protein properties such as helix and β-sheet conformations. As the present study was focused on short peptides, therefore, descriptors that were used for amino acids and small peptides were collected from the database manually. A total of 93 descriptors was selected and were used to build the A*0201 model.

QSAR models were generated for the A*0201 data set using the SIMCA package. The training set includes 266 nonamer peptides, logarithm of the peptide IC_{50} values range from 4.3 to 9. The 93 descriptors were applied to each position of the nonamer peptide, generating a total of (93 x 9) 837 columns in the matrix. Initially, the q^2 value was low for both leave-one-out cross-validation (q^2 = 0.259), and cross-validation in seven groups (q^2 = 0.268). The *VIP* value of each variable was calculated and q^2 was improved by excluding variables with *VIP* values lower than 0.7. A list of q^2 values after variable exclusion was listed in table 3.4.

Model	*Variable Number*	q^2	*Number of components*	r^2
1	837	0.268	1	0.458
2	440	0.305	1	0.358
3	337	0.317	1	0.361
4	276	0.323	1	0.363
5	260	0.324	1	0.364
6	237	0.327	1	0.365
7	229	0.329	1	0.368
8	223	0.332	1	0.370
9	216	0.329	1	0.366
10	206	0.324	1	0.361

Table 3.4. Changes in q^2 of A*0201 models after variable selection in SIMCA.

Table 3.4 showed that there was a 4% improvement in q^2 after excluding nearly half of the variables (model 2, $q^2 = 0.305$), there was no significant changes in q^2 in the subsequent models. There was another slight increase in q^2 when two-thirds of the variables were excluded (model 8, $q^2 = 0.332$), after which q^2 started to decrease (model 9 and 10). In contrast, r^2 was highest when all variables were included, and was decreased as the number of variables dropped (model 2, $r^2 = 0.358$). The r^2 values were generally low for all the models, which indicated that variables used in the study were not informative in describing properties of the peptides. The best q^2 value was 0.332 from model 8, which indicated that the model has some predictivity but not high. Similar results were obtained when applying GOLPE and GA to the data set ($q^2 = 0.298$, $r^2 = 0.445$ for GOLPE and $q^2 = 0.260$ and $r^2 = 0.410$

for GA). The low q^2 and r^2 values of the models suggested that the models generated using AAindex descriptors were not predictive, and were not appropriate for the analysis of peptide-MHC interactions.

3.2.2.2 A*0201 models with the z descriptors

One possible reason for the poor performance of the AAindex descriptors was the quality of descriptors. There were a total of 93 descriptors found in the database that described amino acids and short peptides. It is possible that some descriptors were redundant, therefore the signal to noise ratio was small and useful information was masked by noise. Hence in the second part of the experiment the three z and the five z descriptors were applied to peptides binding to HLA-A*0201. The z descriptors are a class of properties obtained by PCA analysis and are representative of large numbers of redundant, degenerate descriptors. The three z descriptors were used first. A total of 27 (3 x 9) descriptors were applied to the training set. QSAR models were built using SIMCA, GOLPE and GA. Results of the QSAR models are listed in table 3.5. q^2 of the SIMCA model was the lowest among the three (q^2 = 0.29). The predictivities of the GA and the GOLPE models were 0.396 and 0.424, respectively. The relative coefficients of the variables in each model were plotted in figure 3.5.

QSAR models using z1-z3 descriptors					
	q^2	*SEP*	*NC*	r^2	*SEE*
GOLPE	0.424	0.517	4	0.510	0.477
GA	0.396	0.534	3	0.528	0.472
SIMCA	0.292	-	2	0.383	-

Table 3.5. Results of the three z descriptors models calculated by three methods: GOLPE, GA and SIMCA (SIMCA does not report SEP or SEE values).

The coefficients reflected the contributions of each variable (descriptor) at each position. A property was considered to be favoured if it had positive contributions, or coefficients, from all models, and was disfavoured if it had negative coefficients from all three models.

Considering the individual properties, hydrophobic amino acids were favoured at P2, P3, P6 and P7 (with positive z1 coefficients), and disfavoured at P4 and P8 (with negative z1 coefficients) (figure. 3.2). Large amino acids were favoured at P2, P3, P4 and P6 (with positive z2 coefficients), disfavoured at P5, P7 and P8 (negative z2 coefficients). Polar residues were preferred at P4, P8 (with negative z3 coefficients), and disfavoured at P2, P3, P6 and P7 (positive z3 coefficients).

The five z descriptors were then used to describe the A*0201 training set. The meanings of the first three z descriptors were similar to the three z descriptors. The

z1 descriptor represented the hydrophobicity scale. Large negative values indicated hydrophobic amino acid and positive values hydrophilic amino acid. The z2 descriptor represented the steric bulk property. Amino acids with negative z2 value had small molecular weight and surface area. The z3 descriptor represented polarity. Negative z3 values indicated the ability to accept electrons, while positive z3 values described the ability to attract electrons. The property assigned to the last two z properties, z4 and z5, were more complicated, as the two properties took into account a mixture of polar and other chemical properties such as heat of formation (the heat absorbed during the formation of one mole of the substance from its component elements).

The five descriptors were applied to each of the nine positions of the peptide, giving a total of 45 variables. Results of the QSAR models were listed in table 3.6. Relative coefficients of each position were shown in figure 3.6.

QSAR models using z1-z5 descriptors					
	q^2	*SEP*	*NC*	r^2	*SEE*
GOLPE	0.619	0.452	4	0.684	0.412
GA	0.606	0.464	4	0.732	0.383
SIMCA	0.702	-	2	0.897	-

Table 3.6. Results of the five z descriptors models calculated by three methods: GOLPE, GA and SIMCA (SIMCA does not include SEP or SEE values in the result).

q^2 values of the five z models ranged from 0.6 to 0.7, which indicated good predictivity of the models. The model built by SIMCA had the highest $q^2$of 0.702 using the first two components, the r^2 value was 0.897, which was also the highest. The q^2 values of the GOLPE and the GA model were slightly lower: 0.619 and 0.606, respectively. The r^2 values of the two models were 0.684 and 0.732. The five z descriptors QSAR model gave the best results among all descriptors. The relative coefficients of the z5 descriptors revealed that hydrophilic amino acids are favoured at P4 and P8 (with positive z1 coefficients), but disfavoured at P1, P2, P3, P5, P6 and P7 (with negative z1 values) (figure 3.6). Large bulky amino acids were preferred at P1, P2, P3, P5, P6 and P8 (with positive z2 coefficients), while small amino acids were preferred at P4 and P9 (with negative z2 coefficients). Polar amino acids are likely to appear at P1 and P4 (with positive z3 and z4 coefficients), but not at P2, P3, P5 and P7 (with negative z3 coefficients).

The QSAR models generated by the three z and five z descriptors gave similar results. Both the three and the five z descriptors model showed that P2 favoured bulky, non-polar amino acids. P9 preferred small amino acids in the five z descriptors model. No consensus was found at P9 in the three z descriptors model. Secondary anchor positions P3 and P7 favoured non-polar amino acids in both models. Bulky hydrophobic amino acids were identified at P1 in the five z descriptors model. P4 and P8 accept hydrophilic amino acids. P6 favoured bulky amino acids in both models.

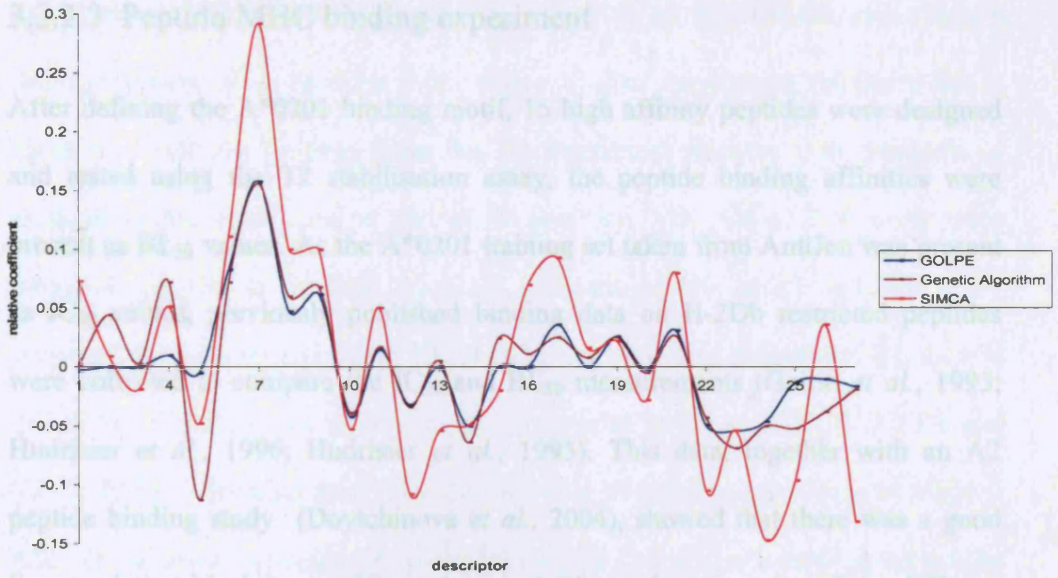


Figure 3.5. The relative coefficients of the QSAR models built by GOLPE, GA and SIMCA using the three z descriptors

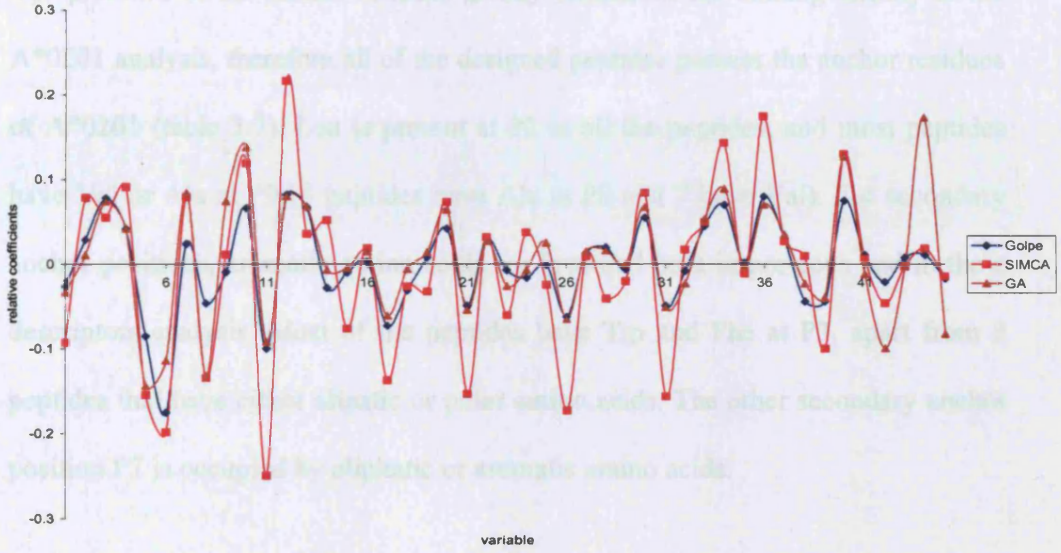


Figure 3.6. The relative coefficients of the QSAR models built by GOLPE, GA and SIMCA using the five z descriptors.

3.2.2.3 Peptide-MHC binding experiment

After defining the A*0201 binding motif, 15 high affinity peptides were designed and tested using the T2 stabilisation assay, the peptide binding affinities were present as BL_{50} values. As the A*0201 training set taken from AntiJen was present as IC_{50} values, previously published binding data on H-2Db restricted peptides were collected to compare the IC_{50} and BL_{50} measurements (Gairin *et al.*, 1995; Hudrisier *et al.*, 1996; Hudrisier *et al.*, 1995). This data, together with an A2 peptide binding study (Doytchinova *et al.*, 2004), showed that there was a good linear relationship between IC_{50} values and BL_{50} values in spite of the different techniques used (figure 3.7).

The presence of the anchor residues greatly influenced the binding affinity in the A*0201 analysis, therefore all of the designed peptides possess the anchor residues of A*0201 (table 3.7). Leu is present at P2 in all the peptides, and most peptides have Val or Ala at P9 (8 peptides have Ala at P9 and 7 have Val). For secondary anchor positions, aromatic amino acids are favoured both in previous and in the z descriptors analysis. Most of the peptides have Trp and Phe at P3, apart from 3 peptides that have either alipatic or polar amino acids. The other secondary anchor position P7 is occupied by aliphatic or aromatic amino acids.

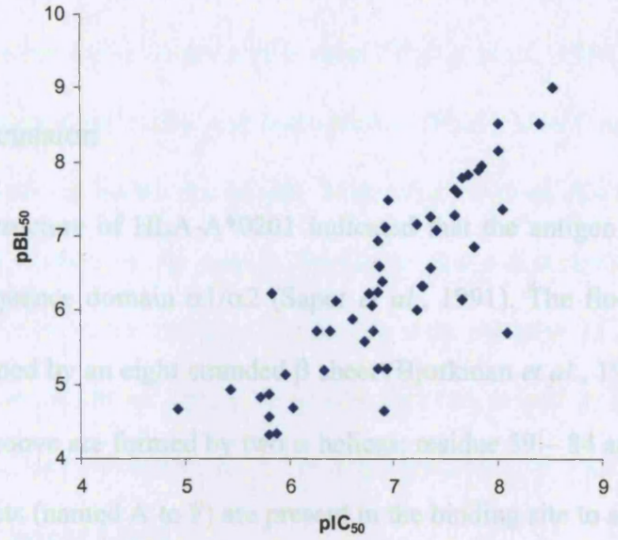
The BL_{50} values of the test peptides were plotted in figure 3.8. The binding affinities for all the peptides in the set were within the range of $4 \sim 6$.

For A*0201 peptides, those with pBL_{50} values above 4 were considered as between intermediate and good binders, and those with pBL_{50} values above 6 were very

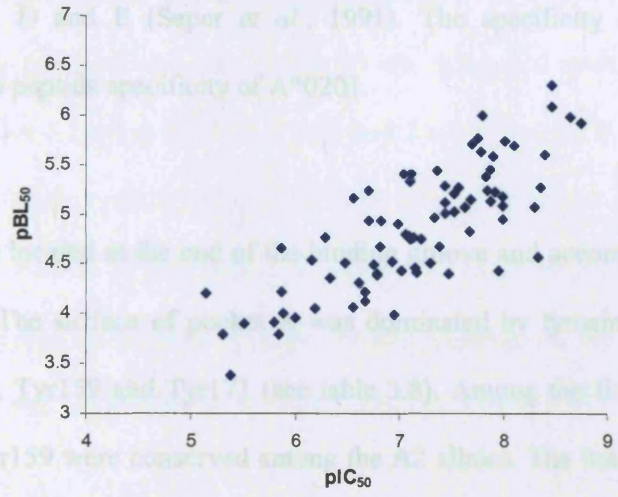
good binders. In this experiment, the BL_{50} values for all the peptides were above 4 and the values of 4 peptides were above 6. The experiment results were in agreement with the findings from the 2D descriptors analysis. The presence of aliphatic amino acids Leu or Met at P2 and Gly, Val, Ala or Tyr at P9 were important in peptide binding. Also, the presence of Phe at P3 and Val at P7 increased the binding affinity of the peptide. Tyr was well accepted at P1, as was Pro at P4. There are two poorly predicted peptides in the set, YLCPGPVTA and VLFNGPVTV, showing that Val at P1 or Cys at P3 result in a decrease of affinity. Also the presence of negatively charged residues Lys or Arg nearly abolishes the ability of the peptide to bind to A*0201, although the peptide has the preferred anchor residues, indicating that MHC-peptide binding is much more complex than the simple motif requirement.

Peptide	*-logIC_{50}*	*Experimental pBL_{50}*
KLPQLCTEL	6.716	4.49
YMLDLQPET	7.59	5.54
RLWPFYHNV	8.206	4.27
FLWPIYHNV	8.298	4.94
YLFPGPMTA	8.169	5.43
YLFDGPVTA	7.838	4.5
YLFPGPFTA	8.276	4.72
YLFPPPVTA	7.863	5.25
YLCPGPVTA	7.143	5.84
YLFPGVVTA	8.147	5.66
YLFPCPVTA	7.902	5.81
YLFPDPVTA	7.82	6.58
YLFPGPVTG	8.035	5.5
YLFPGPMTV	8.525	6.09
YLFDGPVTV	8.194	5.38
YLFPGPFTV	8.632	5.98
YLFPPPVTV	8.224	6.34
VLFNGPVTV	4.662	6.06

Table 3.7. The A*0201 test peptides, their predicted and experimental binding affinities. The first three peptides are reference peptides.



a



b

Figure 3.7. Comparison of IC_{50} and BL_{50} measurements. Peptides that had been measured by both experiments were used in the graph. Previously published H-2Db data (a), and A2 binding study (b) showed that there was a linear relationship

between IC_{50} and BL_{50} values with correlation coefficient of 0.837 and 0.796, respectively.

3.2.2.4 Discussion

The crystal structure of HLA-A*0201 indicated that the antigen binding site was located in sequence domain α1/α2 (Saper *et al.*, 1991). The floor of the binding groove is formed by an eight stranded β sheet (Bjorkman *et al.*, 1987b) and the two sides of the groove are formed by two α helices: residue 59 – 84 and 143 – 171. Six binding pockets (named A to F) are present in the binding site to accommodate side chains of the antigenic peptide. In A*0201, all the pockets were either between the helices like pocket A and F, or at the junction of the β sheet and the helix, like pocket B, C, D and E (Saper *et al.*, 1991). The specificity of these pockets influenced the peptide specificity of A*0201.

Pocket A was located at the end of the binding groove and accommodated the side chain of P1. The surface of pocket A was dominated by tyrosine residues: Tyr7, Tyr59, Tyr99, Tyr159 and Tyr171 (see table 3.8). Among the five residues, Tyr7, Tyr59 and Tyr159 were conserved among the A2 alleles. The bottom of the pocket was occupied by Tyr7. The composition of the surface suggested a preference for aromatic residues in pocket A. The hydroxyl group on the side chain of tyrosine was a potential hydrogen bond acceptor, and could interact with amino acids with potential hydrogen donor side chains. The results of the present study confirmed that aromatic residues were favoured at P1. Met was also accepted at P1, the side chains of which could form hydrogen bonds with Tyr7.

The side chain of P2 interacted with pocket B, which is positioned on one side of pocket A, between the α1 helix and β sheet (Young *et al.*, 1994). Residues lining inside the pocket were bulky and hydrophobic (Phe9, Met45 and Val67), which reduced the volume inside the pocket. Non-polar residues Ala24 and Val34 are located at the bottom of the pocket. Results of the z descriptors model showed medium size hydrophobic residues Leu and Ile were preferred at P2. Side chains of these two amino acids are long and narrow and can extend to the bottom of the pocket B. This was confirmed in the binding experiment, in which all high binders of A*0201 possessed Leu at P2.

Aromatic residues such as Tyr and Trp were favoured at P3 and P7. The side chain of the residue extended into the pocket D and interacted with the hydrophobic residues inside the pocket, Leu156, Tyr99 and Tyr159. Pocket E accepted the side chain of the amino acid at P7, which may interact with the aromatic residues Trp133 at the bottom of the pocket. Hydrophobic residues such as Val and Met were also accepted at P7. In the peptide binding experiment, the four high binders had Phe at P3 and Val or Met at P7 (YLFPDPVTA, YLFPGPMTV, VLFNGPVTV, YLFPPPVTV).

The side chain of P6 interacted with the pocket C, which was shallow with polar residues lining the inside (His70, Thr73, His74 and Arg97). The bottom of the pocket was defined by aromatic residue Phe9. In the present study, P6 accommodated a variety of amino acids: aromatic residue Tyr, Trp and Phe,

medium size hydrophobic residue Leu, Ile and Pro were all accepted. Tyr and Trp were potential hydrogen bond donors that could interact with the polar residues inside the pocket, while Phe can reach the bottom of the pocket and stabilise the binding.

P9 was known to be an important anchor residue in HLA-A*0201. The side chain of P9 reached pocket F, which was relatively deep with side chains of Leu81, Tyr123 and Tyr116 at the bottom. In the present study, P9 favoured medium size, non-polar residues such as Leu, Ile and Met. Two small non-polar amino acids Ala and Val were also well accepted, as demonstrated in the peptide binding experiment. Thr was the only polar residue favoured at this position, which may form hydrogen bonds with Thr and Tyr residues in pocket F (Thr80, Tyr84, Thr143).

Side chains of the other positions (4, 5, and 8) did not bind to the inside of the binding groove, they were orientated towards the outside of the groove and possibly interacted with the T cell receptor. The amino acids at these positions were more diverse. P4 preferred small amino acids and P5 preferred hydrophobic residues. In the peptide binding experiment, peptides with Pro at P4 and Gly at P5 were well accepted. P8 favoured more hydrophilic residues.

Previously, A*0201 motifs have been defined by quantitative binding assays (Ruppert *et al.*, 1993) and by the grouping of naturally occurring epitopes (Rammensee *et al.*, 1995). For anchor residues, the five z descriptors study

identified Leu, Ile, Val, Ala and Met at P2 and P9, which were also the most preferred amino acid from other studies (Falk and Rotzschke, 1993). For non-anchor residues, some of the results overlapped. The present study identified aromatic residues at P1, small hydrophilic residues at P4 and hydrophobic amino acids at P5, while the binding study by Drijfhout suggested Lys, Tyr, Thr at P1, and that P4 and P5 accept both polar and non-polar residues (Drijfhout *et al.*, 1995). As previous studies indicate (Falk *et al.*, 1991) these positions are more involved with TCR interaction and the amino acids that occupy these positions may vary greatly between epitopes from different organisms. Furthermore, the differences at the positions may also be due to the peptide data set used. Results of the peptide binding experiments showed that peptides with the favoured residues identified from the 2D-QSAR study bound to A*0201 with high affinity. Overall, the five z amino acid descriptors seemed to be a promising tool in studying peptide-MHC interaction, and may be used in combination with other QSAR methods such as CoMSIA.

	Pocket					*Residues*				
A	5	7	59	63	66	99	159	163	167	171
	M	Y	Y	E	K	Y	Y	T	W	Y
B	7	9	24	34	45	63	66	67	70	99
	Y	F	A	V	M	E	K	V	H	Y
C	9	22	70	73	74	97	99	114	116	
	F	F	H	T	H	R	Y	H	Y	
D	99	114	155	156	159	160				
	Y	H	Q	L	Y	L				
E	97	114	116	147	152	155	156			
	R	H	Y	W	V	Q	L			
F	73	77	80	81	84	95	116	118	123	124	143
	T	D	T	L	Y	V	Y	Y	Y	I	T

Table 3.8. The residues that form the peptide binding pockets of the HLA-A*0201 molecule.

Chapter 4
On-line application of the additive method – MHCPred

4.1 Introduction

In the previous chapter, the theory of the additive method and its application to the generation of MHC-peptide interaction models were explained. Apart from the additive method, many other algorithms have been developed to predict T cell epitopes: motif search methods (Rammensee *et al.*, 1999), quantitative matrices (Reche *et al.*, 2002), structure-based approaches (Altuvia *et al.*, 1997) and artificial neural networks (Del Carpio *et al.*, 2002), etc. Many algorithms have been implemented as internet-based servers where users can predict T cell epitopes in protein sequences. The internet-based epitope prediction program is an effective way of applying Bioinformatics data on a wide scale, as it also helps laboratory-based scientists world-wide to use these methods in their work. For this purpose, an Internet application of the additive method, called MHCPred, was produced. MHCPred includes models for all human and mouse MHC alleles generated in the Bioinformatics lab so far, and users can predict potential T cell epitopes restricted to these alleles (Guan *et al.*, 2003a; Guan *et al.*, 2003d; Hattotuwagama *et al.*, 2004).

In the first part of this chapter, the web interface, the underlying perl program, and the output of MHCPred are explained. In the second part of the chapter, the predictivity of MHCPred is evaluated using peptide data not used to construct its models. Two sets of peptide data are used: A*0201 binding peptides from Dr. Vladimir Brusic that includes peptides taken from the MHCPEP database (Brusic *et al.*, 1998) together with some unpublished data, and the second data set contains recently published epitopes from the literature. The predictivities of other online T

cell epitope prediction algorithms are also tested and compared with the additive method.

4.2 The MHCPred server

4.2.1 The MHCPred web interface

The MHCPred interface is shown in figure 4.1. Currently the server holds models for a total of 23 human and mouse MHC alleles (table 4.1). As the research continues, models for other alleles will be added. A summary of the model statistics is in table 4.1.

Two versions of MHCPred have been developed. The first version is available on-line and the user can access it through the following URL: http://www.jenner.ac.uk/MHCPred/ (Guan *et al.*, 2003d). An improved version has been made and is currently on the intranet only. In this chapter the term 'MHCPred server' implies the second version, as it will soon become available online and replace the first one.

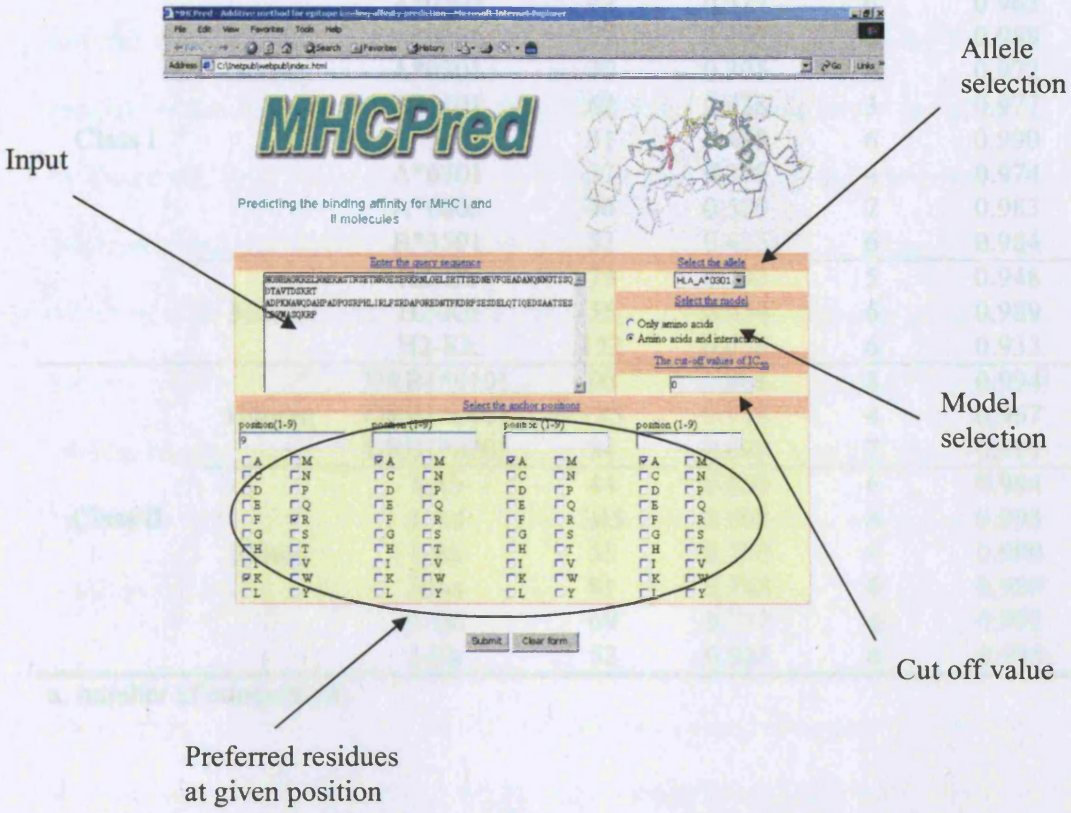


Figure 4.1. Graphical user interface (GUI) of MHCPred, from which the user can enter the query protein sequence, choose the allele and model used in the prediction. The IC_{50} threshold (nM) is used to restrict the output. If a value is entered, peptides with IC_{50} values higher than the value will not be listed in the output. Also the user can choose to have output with only peptides that have preferred residues at given positions. A total of four such positions can be selected.

			No. of peptides	q^2	NC[a]	r^2
Class I	Human	A*0101	95	0.420	4	0.997
		A*0201	335	0.377	6	0.731
		A*0202	69	0.317	9	0.943
		A*0203	62	0.327	6	0.963
		A*0206	57	0.475	6	0.989
		A*0301	70	0.305	4	0.972
		A*1101	62	0.428	3	0.977
		A*3101	31	0.453	6	0.990
		A*6801	37	0.370	4	0.974
		A*6802	46	0.500	7	0.983
		B*3501	52	0.435	6	0.984
	Mouse	H2-Db	73	0.493	5	0.948
		H2-Kb	55	0.454	6	0.989
		H2-Kk	152	0.456	6	0.933
Class II	Human	DRB1*0101	90	0.808	8	0.994
		DRB1*0401	185	0.716	4	0.967
		DRB1*0701	84	0.649	7	0.999
	Mouse	I-Ab	44	0.850	6	0.994
		I-Ad	145	0.898	6	0.993
		I-Ak	55	0.790	6	0.990
		I-As	81	0.783	6	0.980
		I-Ed	69	0.732	6	0.992
		I-Ek	52	0.925	6	0.995

a. number of components

Table 4.1. Alleles included in the MHCPred server. The table lists the human and mouse models used in the server and statistics obtained for each model. q^2 values indicated how good the ability of the models is to predict epitopes from sequences, and r^2 values show the proportion of variance explained by the model, or, how good the models fitted the training data. NC is the number of components that gave the optimal q^2 values. Note that the q^2 values for class II models are higher than those of class I. This is because of the training set for class II models are generated by the iterative technique that involves pre-selection of data. A detailed expla nation is given in Dr. Doytchinova's paper (Doytchinova and Flower, 2003b).

4.2.2 The input

To keep the calculation time relatively short, the input protein sequence length is limited to 500 amino acids, this restriction will be relaxed in the future when MHCPred will be put on a faster server. The server interface is written in HTML, and the CGI program is written in Perl. The calculation procedure of the CGI program is summarised in figure 4.2. Only sequences in flat text format are accepted by the server. Both upper and lower cases are accepted and all the non-amino acid characters will be deleted during calculation. A pull-down box is used for the selection of the alleles, where the user can choose any of the alleles in table 4.1. In the first version of MHCPred only one allele can be chosen at a time. In the second version, multiple alleles can be selected. However it is not recommended to select more than three alleles at once, as it slows down the calculation and increases the result page loading time.

Two types of models are generated by the additive method. The user can choose to use the single amino acid model, which only considers the interaction between the peptide and the binding site. Alternatively, if the user wants to take into account the interactions between adjacent amino acids of the peptide, the amino acid with interactions model should be used. Later in this chapter the predictivity of the two models will be compared.

The query sequence is chopped into nonamer peptides except for the H-2Kd and H-2Kk model, which requires octamers. If the user enters preferred residues and positions, the program uses another subroutine to check the peptide to see whether the output peptide has these residues at these positions. If it does, the program starts

the next step, if it does not have the residue at the given position (figure. 4.2), then the subroutine runs again to get the next peptide and the process continues until a peptide with the preferred residue is found at the given position. After processing the input sequence, the program opens the file containing the coefficients and reads the file into a two-dimensional matrix. For each amino acid of the peptide, the program reads the corresponding value in the matrix and adds it to the constant value to give the final result. If amino acid and interactions model is selected, the adjacent and 1-3 amino acid interactions are taken into account and their contributions are added to the result.

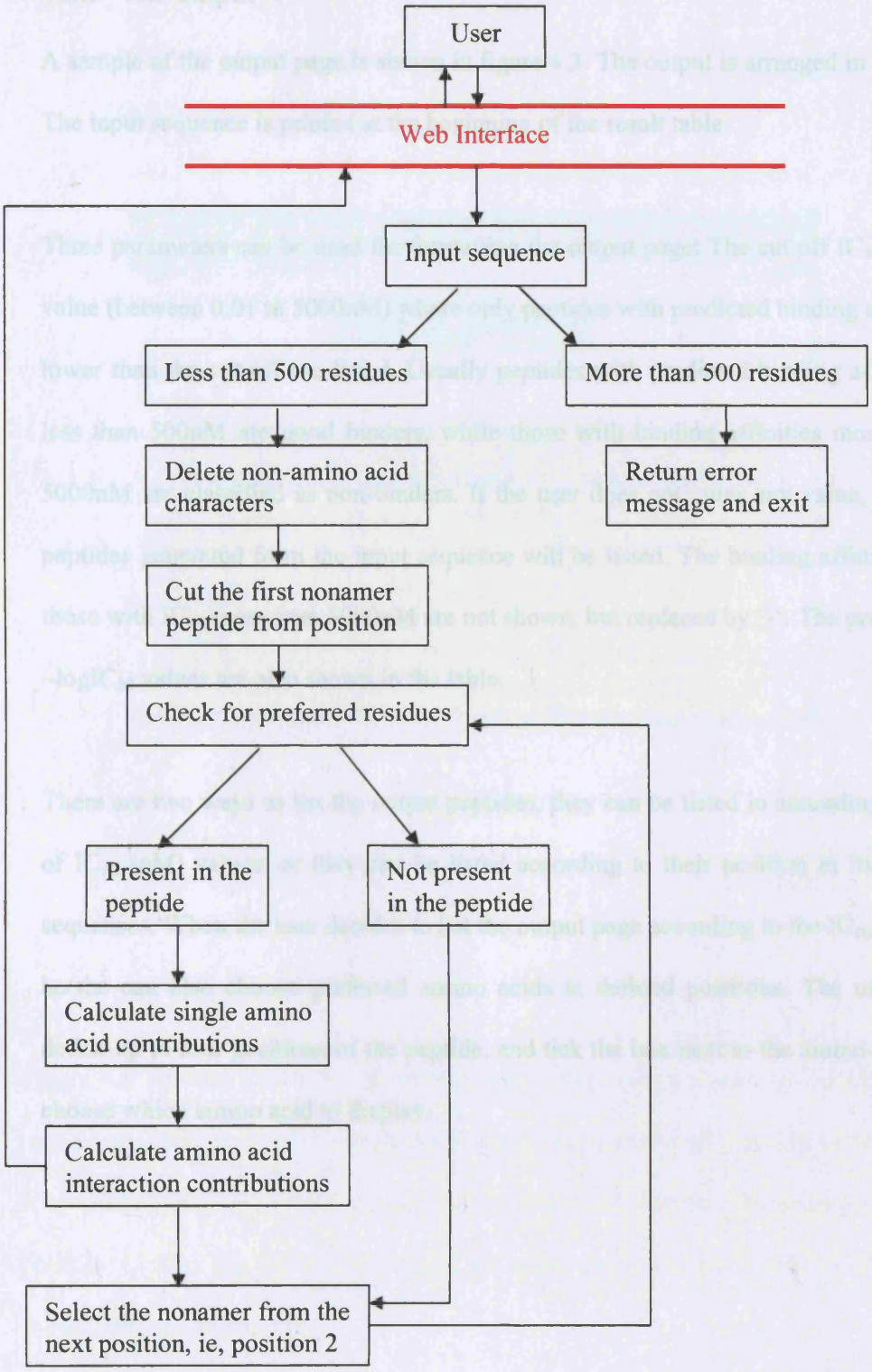


Figure 4.2. A flow chart of how the CGI program works in MHCPred.

4.2.3 The output

A sample of the output page is shown in figure 4.3. The output is arranged in a table. The input sequence is printed at the beginning of the result table.

Three parameters can be used for formatting the output page: The cut off IC_{50} (nM) value (between 0.01 to 5000nM) where only peptides with predicted binding affinity lower than the cut off are listed. Usually peptides with predicted binding affinities less than 500nM are good binders, while those with binding affinities more than 5000nM are classified as non-binders. If the user does not enter any value, all the peptides generated from the input sequence will be listed. The binding affinities of those with IC_{50} more than 5000nM are not shown, but replaced by '-'. The predicted $-\log IC_{50}$ values are also shown in the table.

There are two ways to list the output peptides, they can be listed in ascending order of IC_{50} (nM) values, or they can be listed according to their position in the input sequences. When the user decides to list the output page according to the IC_{50} values, he/she can also choose preferred amino acids at defined positions. The user can define up to four positions of the peptide, and tick the box next to the amino acid to choose which amino acid to display.

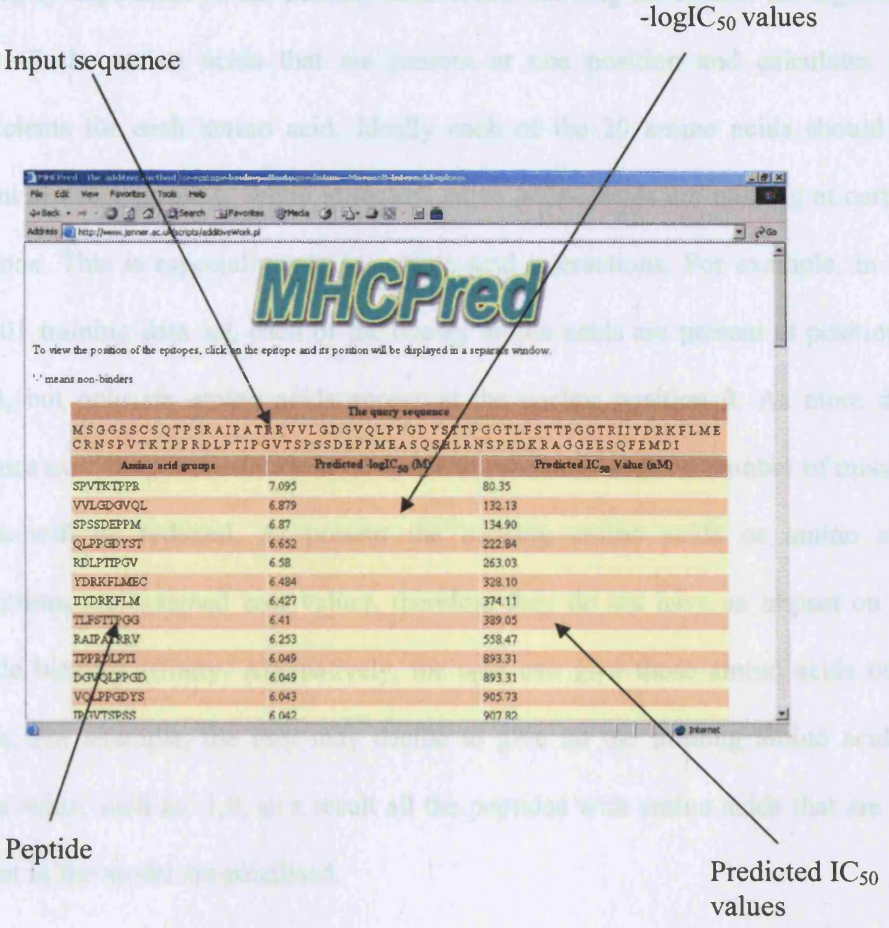


Figure 4.3. An example of the MHCPred output. Peptides generated by the additive method are listed in the first column of the table, their predicted $-\log IC_{50}$ values and IC_{50} (nM) values are listed in the second and third column. Peptides that are potential binders are listed at the top of the table, weak and non-binders are listed towards the end of the page.

The predictivities of the single amino acid and amino acid with interactions models are largely dependent on the training data. When building the model, the algorithm scans all the amino acids that are present at one position and calculates the coefficients for each amino acid. Ideally each of the 20 amino acids should be present at every position, while in reality, some amino acids are missing at certain positions. This is especially true for amino acid interactions. For example, in the A*0201 training data set, each of the twenty amino acids are present at position 3 and 4, but only six amino acids appear at the anchor position 9. As more data becomes available, more data will be added to the model and the number of missing values will be reduced. At present the missing amino acids or amino acid interactions are assigned zero values, therefore they do not have an impact on the peptide binding affinity. Alternatively, the user can give those amino acids other values. For example, the user may decide to give all the missing amino acids a minus value, such as -1.0, as a result all the peptides with amino acids that are not present in the model are penalised.

4.2.4 The peptide library

An added option in the second version of MHCPred is to calculate the effect on affinity of mono-amino acid mutations of the peptide. The program takes a single nonamer peptide as the input, substitutes the amino acid at a certain position with each of the twenty amino acids and calculates the binding affinities of the new peptides. This option is useful in comparing the binding affinities of heteroclitic analogues of the test peptide, and to test the effect of different amino acids at specific positions of the peptide. The interface is shown in figure 4.4.

Like the MHCPred webpage, the user can also choose to use either the single amino acid model or the amino acid with interactions model. The user can change up to two positions within the peptide. If the user decides to choose one position, then each of the nine positions of the peptide are substituted with each of the 20 amino acids in turn, generating a total of 180 peptides. If the user wants to change two positions, then two random positions are selected by the server. Each of the 20 amino acids will appear at each of the positions, and the total number of different peptides generated is 13680 (9x(9-1)/2!x20x19). The output cut off has three options: the input peptide affinity, peptides with affinities 5% lower than the input peptide, or the user can select all and see all the output. A sample output is shown in figure 4.5.

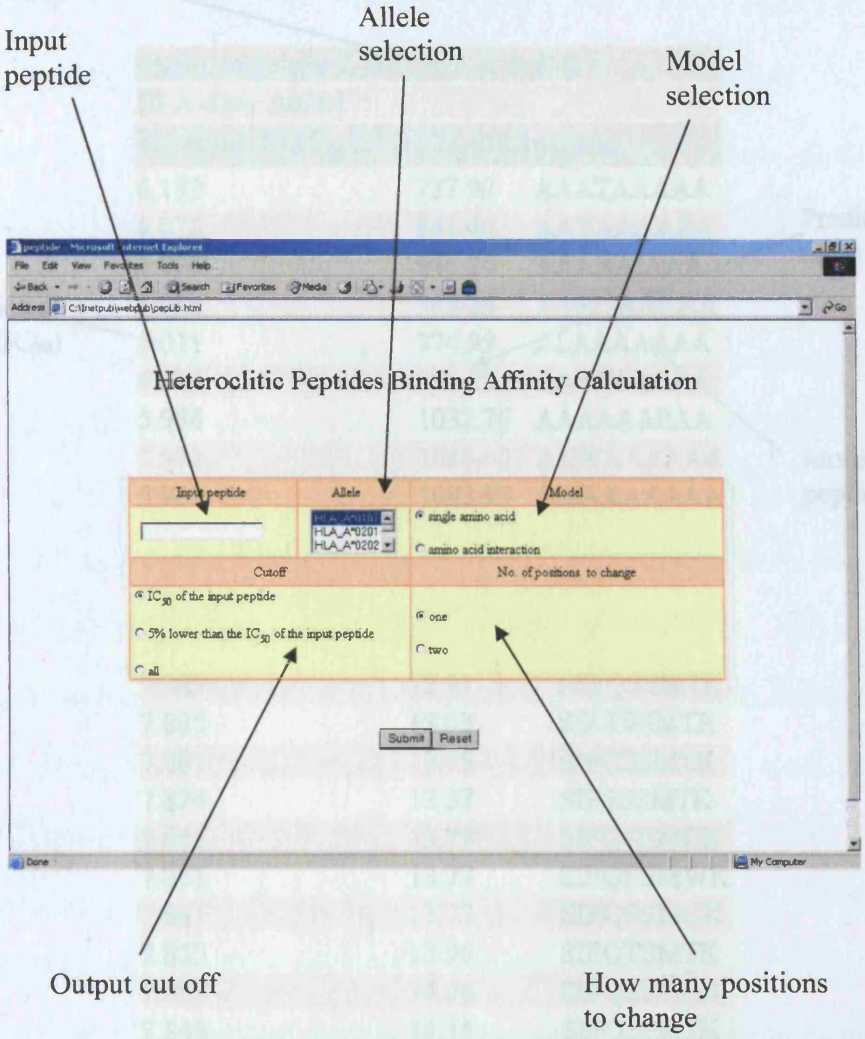


Figure 4.4 The graphical user interface of the peptide library, where the user can enter the query sequence, choose which allele the peptide is restricted to and the additive model to use. For the output, the user can choose to modify one or two positions of the peptide and select the cut off so that peptides below the cut off will not be displayed.

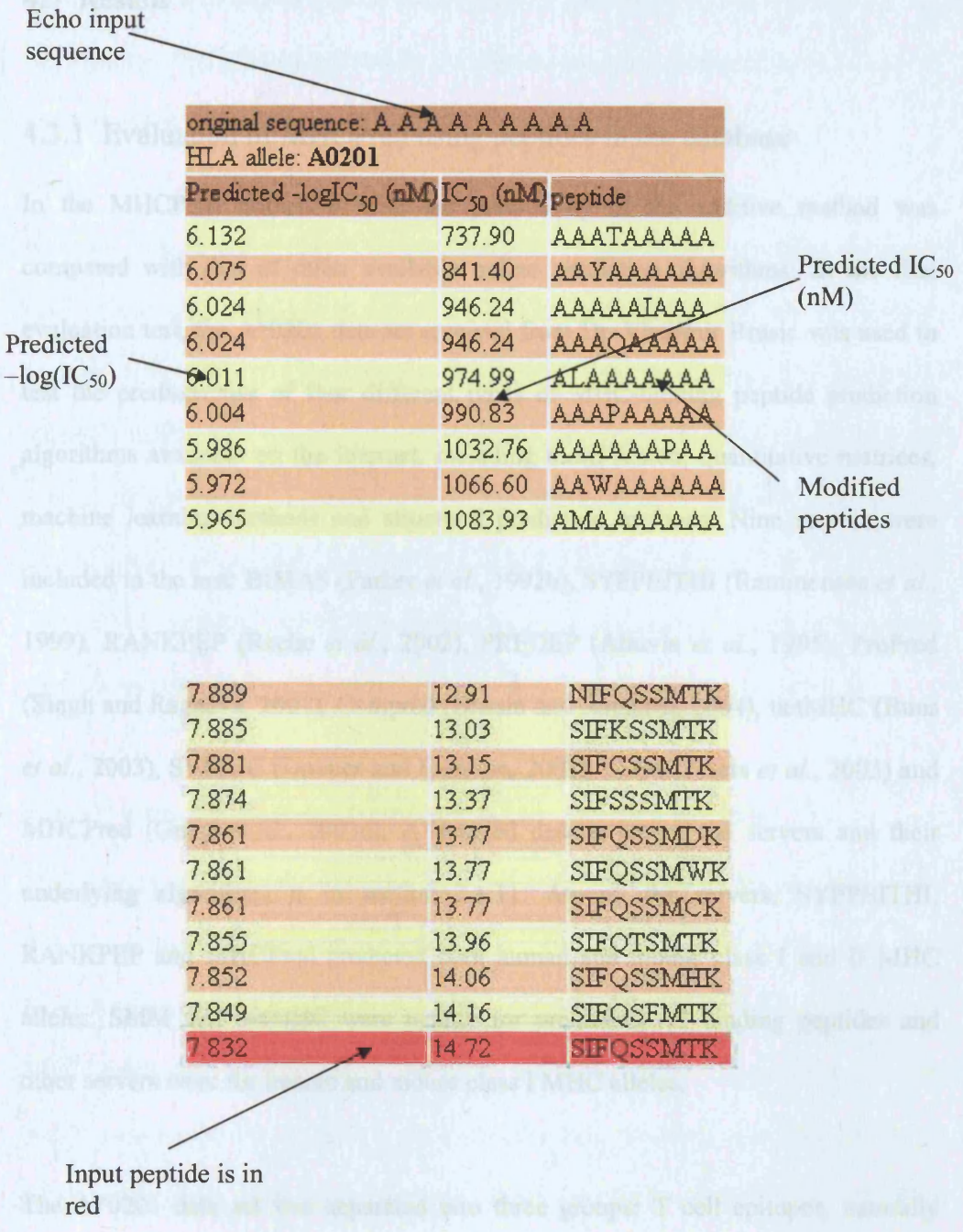


Figure 4.5. Part of the output page of the peptide library. Mutated peptides are listed in ascending order of their binding affinities. The input peptide is coloured in red.

4.3 Results

4.3.1 Evaluation of MHCPred using peptides in the database

In the MHCPred evaluation test, the predictivity of the additive method was compared with that of other available online prediction algorithms. In the first evaluation test, the A*0201 data set received from Dr. Vladimir Brusic was used to test the predictivities of four different types of MHC-binding peptide prediction algorithms available on the Internet, including motif search, quantitative matrices, machine learning methods and structural prediction methods. Nine servers were included in the test: BIMAS (Parker *et al.*, 1992b), SYFPEITHI (Rammensee *et al.*, 1999), RANKPEP (Reche *et al.*, 2002), PREDEP (Altuvia *et al.*, 1995), ProPred (Singh and Raghava, 2001), Compred (Bhasin and Raghava, 2004), netMHC (Buus *et al.*, 2003), SVMHC (Donnes and Elofsson, 2002), SMM (Peters *et al.*, 2003) and MHCPred (Guan *et al.*, 2003d). A detailed description of the servers and their underlying algorithms is in section 2.1.11. Among the servers, SYFPEITHI, RANKPEP and MHCPred predicted both human and mouse class I and II MHC alleles. SMM and netMHC were mainly for predicting A2 binding peptides and other servers were for human and mouse class I MHC alleles.

The A*0201 data set was separated into three groups: T cell epitopes, naturally processed peptides and poly-alanine peptides. T cell epitopes were peptides that had been proved to induce T cell responses. Naturally processed peptides were fragments eluted from cell surface MHC molecules. Poly-alanine peptides were synthetic peptides with mainly alanines and one or two other amino acids. Poly-alanine peptides are commonly used in MHC-peptide interaction research to define

the binding ability of non-alanine amino acids at specific positions. Affinities of the three groups of peptides predicted by the nine servers were presented as ROC curves and compared by the area under the ROC curves (Aroc).

Average Aroc values of the test set were plotted in figure 4.6 and the ROC curves of the different test sets were in figure 4.7, 4.8 and 4.9. RANKPEP had the highest average Aroc value (0.955), and had exceptional good predictions with the poly-alanine peptide set (Aroc=0.999). The additive single amino acid model had the second highest predictivity in the test and with the Aroc value of 0.947. SYFPEITHI and BIMAS also had high scores of 0.937 and 0.935, respectively. The Aroc values of PREDEP, netMHC, SVMHC and the additive amino acid and interactions model were above 0.88. The Aroc values of COMPRED and SMM were about 0.85.

4.3.1.1 Comparing the predictivity of two additive models

The additive method generates two types of models: the single amino acid model and the amino acid plus interactions model. The first model considers only the interaction between individual amino acids of the peptide and the binding site of the MHC, while the second model also takes into account the interactions between nearby residues of the peptide. In the evaluation test, the Aroc values of the single amino acid model were on average 5% higher than those of the amino acid plus interactions model. The single amino acid model had much higher Aroc values in predicting poly-alanine derivatives and naturally processed peptides. However, the difference in the prediction of T cell epitopes using the two models was small, in which the Aroc value of MHCPred with single amino acid model (Aroc = 0.929) is about 3% higher than the model with amino acid and interactions (Aroc = 0.901).

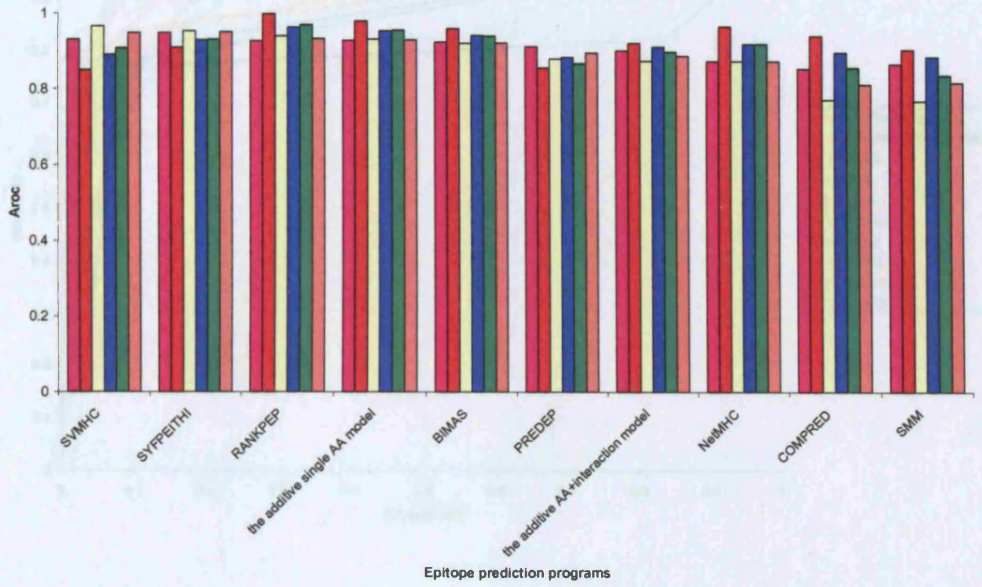


Figure 4.6. The overall performance of the T cell epitope prediction servers. The Aroc values for the T cell epitopes (T), poly-alanine peptides (A) and naturally processed peptides (N) were in pink, red and cream, respectively. The average Aroc values of T+A were in blue, the average Aroc values of A+N were in green, and the average Aroc values of T+N were in orange.

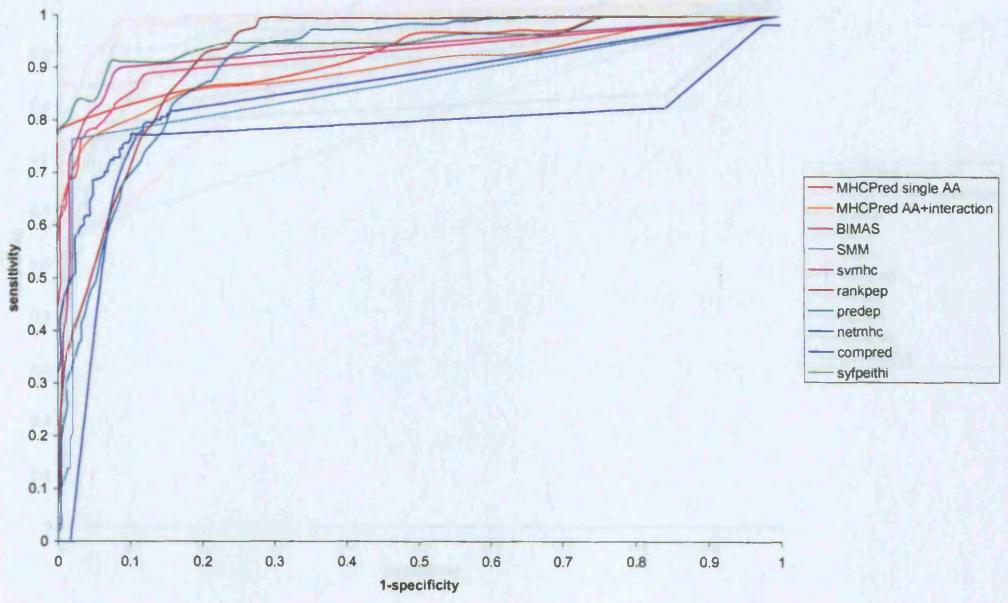


Figure 4.7. ROC curve of the T cell prediction. T cell epitopes data from Dr. Brusic was used as the input for the different algorithms. The ROC values of the prediction servers are (in descending order): SYFPEITHI (0.949), SVMHC (0.931), MHCPred single amino acid model (0.929), RANKPEP (0.927), BIMAS (0.924), PREDEP (0.912), MHCPred amino acid and interactions model (0.901), netMHC (0.873), SMM (0.866), and compred (0.853).

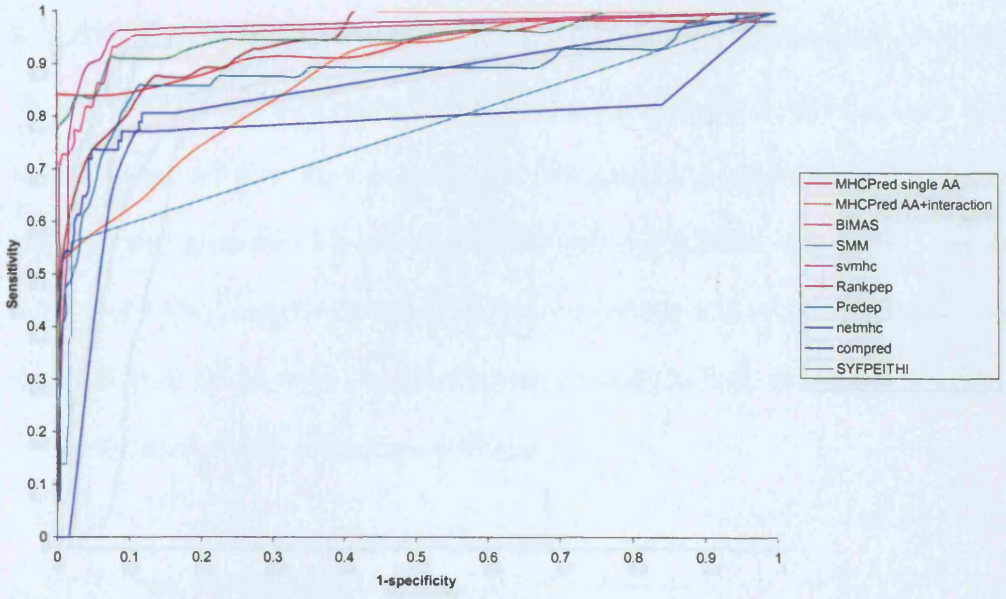


Figure 4.8. The ROC curve of naturally processed peptide prediction. The ROC values of the prediction servers are (in descending order): SVMHC (0.966), SYFPEITHI (0.953), RANKPEP (0.94), MHCPred single amino acid model (0.932), BIMAS (0.921), PREDEP (0.88), MHCPred amino acid and interactions model (0.874), netMHC (0.873), compred (0.772) and SMM (0.768).

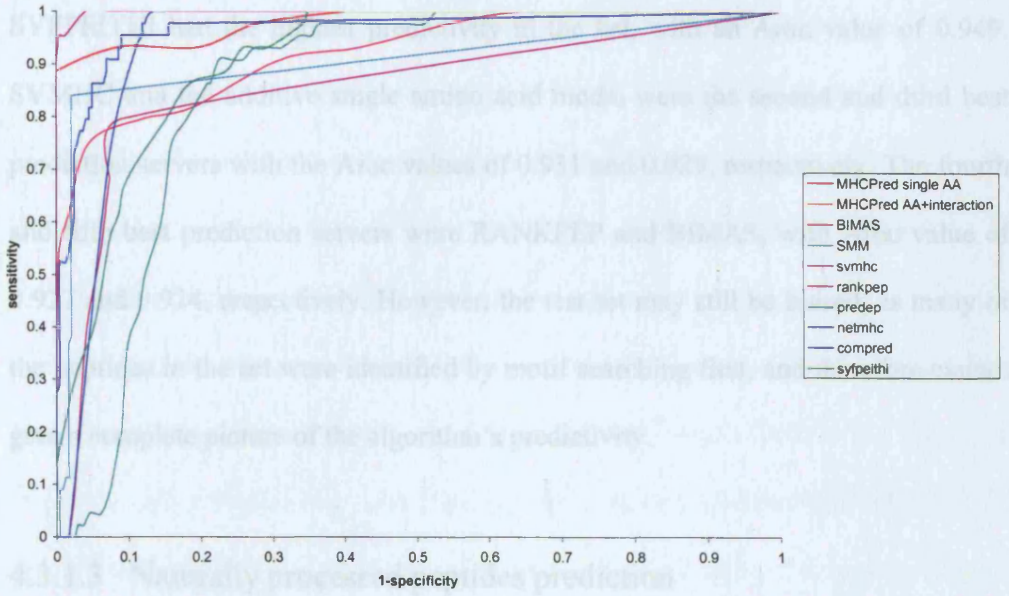


Figure 4.9. ROC curve of the poly-alanine peptide prediction. The ROC values of the prediction servers are (in descending order): RANKPEP (0.999), MHCPred single amino acid model (0.98), netMHC (0.965), BIMAS (0.96), compred (0.94), MHCPred amino acid and interactions model (0.921), SYFPEITHI (0.91), SMM (0.904), PREDEP (0.856) and SVMHC (0.85).

4.3.1.2 T cell epitope prediction

The ability to correctly identify T cell epitopes within a protein sequence is the ultimate goal of any MHC-peptide binding affinity prediction server. In the T cell epitope predictions (figure 4.7), matrix-based algorithms had good predictivity. SYFPEITHI had the highest predictivity in the list, with an Aroc value of 0.949. SVMHC and the additive single amino acid model were the second and third best prediction servers with the Aroc values of 0.931 and 0.929, respectively. The fourth and fifth best prediction servers were RANKPEP and BIMAS, with Aroc value of 0.927 and 0.924, respectively. However, the test set may still be biased, as many of the peptides in the set were identified by motif searching first, and therefore cannot give a complete picture of the algorithm's predictivity.

4.3.1.3 Naturally processed peptides prediction

The support vector machines based server SVMHC was the best server for the prediction of naturally processed peptides in the evaluation test (Aroc=0.966) (figure 4.8). Matrix based algorithms also had high levels of predictivity. SYFPEITHI was the second best in this category, with the Aroc value of 0.953. RANKPEP and the additive single amino acid model have similar Aroc values (RANKPEP Aroc = 0.94, the additive model Aroc = 0.932)

4.3.1.4 Poly-alanine peptides prediction

The Aroc values were relatively high (Aroc > 0.85 for all servers) in the prediction of poly-alanine peptides for all algorithms (figure 4.9). RANKPEP had the highest Aroc value of 0.999 in predicting poly-alanine derivatives, and was closely followed by the additive single amino acid model, which had the Aroc value of 0.98. netMHC

is the third best server in predicting poly-alanine peptides with a Aroc value of 0.965.

4.3.2 Evaluation using recently published epitopes

The evaluation test using Dr. Vladimir Brusic's data examined the ability of the prediction servers to distinguish established T cell epitopes and ligands from non-binding peptides. All prediction algorithms showed good predictivity in the test with average Aroc values of 75% or more. To test the predictivity of the algorithms in real world situations, the second part of the evaluation used recently published T cell epitopes. To avoid replicating data from existing databases, only epitopes that have been published within the last few years were used (2001-2004). However there was still a chance that a small part of epitopes may overlap with ones already present in the training set of the tested algorithms, yet as most of the epitopes were new and the set was less biased than using extant data.

In the first evaluation test, RANKPEP, BIMAS, SYFPEITHI, SVMHC and the additive amino acid only model were the top five servers, therefore they were chosen to be used in the second evaluation test. Both human and mouse epitopes were included in the data set. As BIMAS did not include any Class II human alleles, the Class II MHC prediction server, ProPred, was used for the prediction of Class II epitopes. As SVMHC did not have any mouse alleles, only RANKPEP, BIMAS, SYFPEITHI and the additive single amino acid model were used to predict mouse epitopes.

A total of 83 epitopes from 60 protein sequences were collected from the literature, including 36 human class I HLA, 27 class II HLA and 20 mouse class I MHC epitopes. The class I epitopes mainly bind to A1, A3 and A2 alleles. Class II alleles tested were restricted to DRB*0101, *0401 and *0701 alleles, and mouse alleles tested were restricted to H-2Kb and Db. The protein sequences were retrieved from either SWISS-PROT (Bairoch and Boeckmann, 1991; Boeckmann *et al.*, 2003) or Genbank (Benson *et al.*, 2004; Cassatt and Peterson, 1987), and were used as the input sequences for the prediction servers. The full list of epitopes and corresponding protein sequences are given in appendix 4.

To test whether a server can predict the epitopes, a cut off threshold was set. If the predicted affinity or score (some servers did not directly predict the binding affinity, but gave scores for each peptide) of the epitope was above the cut off, then the server was able to predict the epitope. The cut off points were different for each server. SVMHC highlighted predicted epitopes. For RANKPEP and ProPred server, default thresholds were used, which were the top 2% and 3% of generated peptides, respectively. For SYFPEITHI and MHCPred, it was set to the top twenty peptides. For BIMAS, a peptide-MHC dissociation half-life of 5 minutes was used. The results of the predictions are summarised in figure 4.10, 4.11 and 4.12.

The overall predictivities of the servers for epitopes from real proteins were not as high as using peptides stored in the database. It should be noted that most prediction servers had models for more alleles than the ones tested. Due to restrictions on the available data, many alleles were not tested. Therefore the predictivity of other alleles may differ from the present results.

The average predictivity of human class I epitopes was the highest for all the servers, followed by class II predictions. The mouse epitope predictions were the lowest in the test. The additive model was the best algorithm in the test with about 85% accuracy in all tests. For class I epitopes, RANKPEP, with a predictivity of 79%, was the second best server. BIMAS was also good at predicting class I HLA epitopes (66%), and SYFPEITHI and SVMHC had similar performance, with 74% and 64%, respectively. For Class II epitope predictions, MHCPred was the best among all the servers, with a predictivity of 74% followed by ProPred with a predictivity of 67% (figure 4.11). RANKPEP and SYFPEITHI had a similar level of predictivity, both were about 40% accurate. In mouse class I MHC epitope prediction (figure 4.12), MHCPred again was the best prediction server with a predictivity of 90%, followed by SYFPEITHI with 65% predictivity. BIMAS and RANKPEP had the lowest levels of predictivity at 0.3.

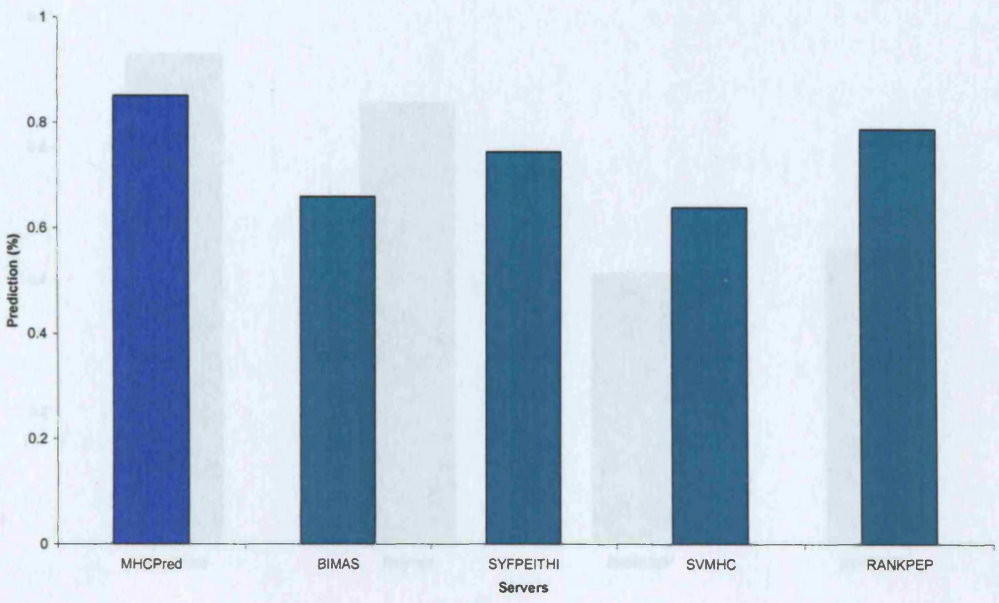


Figure 4.10. Percentage of correct predictions for the five epitope prediction servers using Class I HLA epitopes. The MHCPred server had the best prediction rate of 85%, followed by RANKPEP 79%, BIMAS 66%, SYFPEITHI 74% and SVMHC 64%.

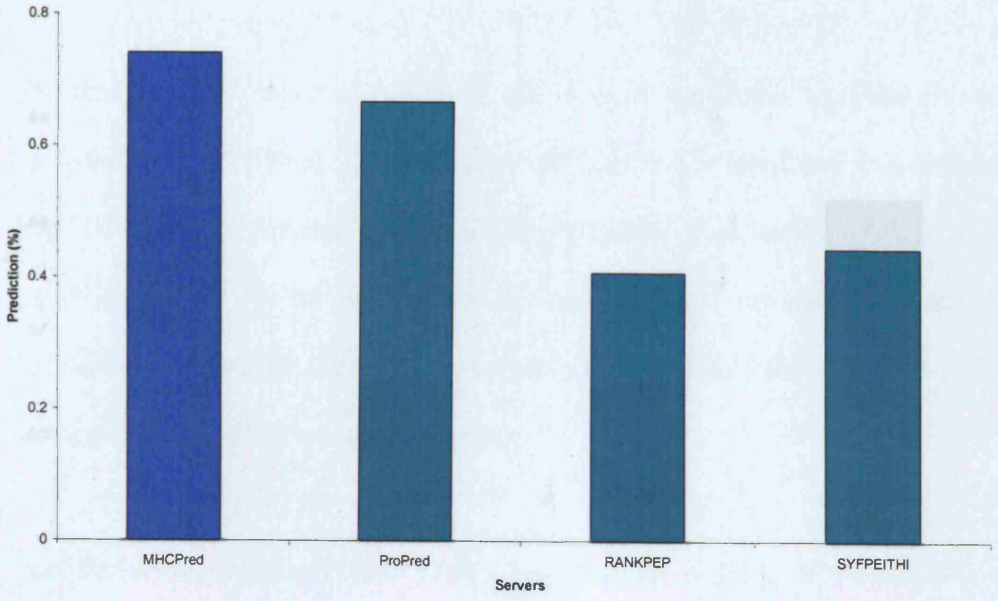


Figure 4.11. The percentage of correct Class II epitope predictions for four on-line prediction servers. MHCPred had the highest predictivity of 74%, followed by ProPred 67%, SYFPEITHI 44% and RANKPEP 40%.

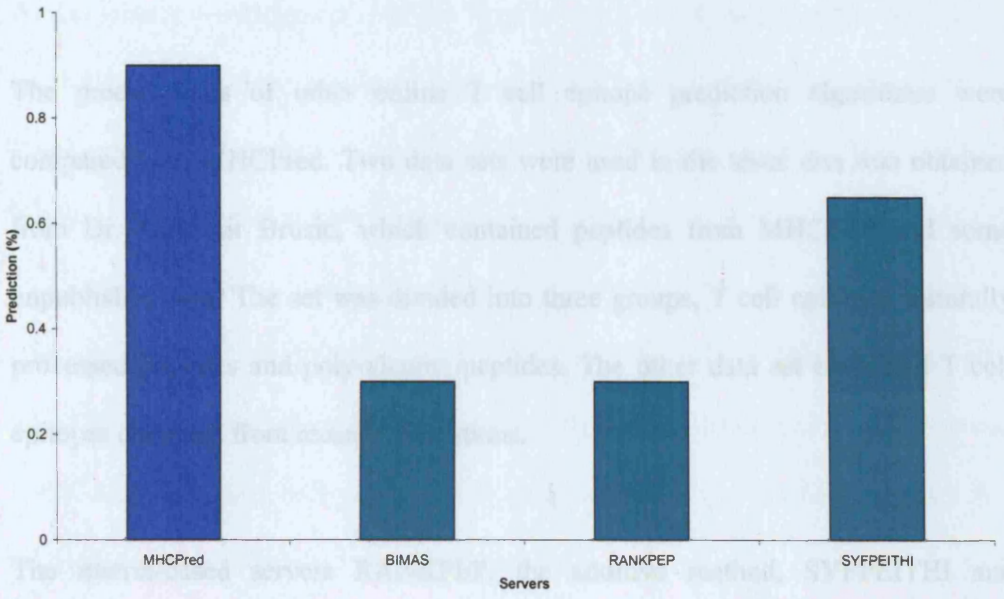


Figure 4.12. The percentage of correction mouse class I MHC epitope predictions for four servers. The correct prediction rate of MHCPred was 90%, followed by SYFPEITHI 65%. BIMAS and SYFPEITHI had the same prediction rate of 30%.

4.4 Discussion

The online application of the additive method used allele-specific additive models to predict potential epitopes for MHC alleles within a given protein sequence. Both human and mouse MHC class I and II models were included in MHCPred.

The predictivities of other online T cell epitope prediction algorithms were compared with MHCPred. Two data sets were used in the tests: one was obtained from Dr. Vladimir Brusic, which contained peptides from MHCPEP and some unpublished data. The set was divided into three groups, T cell epitopes, naturally processed peptides and poly-alanine peptides. The other data set contained T cell epitopes collected from recent publications.

The matrix-based servers RANKPEP, the additive method, SYFPEITHI and BIMAS were the four best servers in the first evaluation test (section 4.3.1). The result suggested that, in spite of the disadvantage of having to produce specific models for each allele, the matrix-based methods were useful in predicting MHC ligands and epitopes. However, the training set of the RANKPEP models was taken from the MHCPEP database, which might be the explanation of the exceptionally high accuracy in predicting poly-alanine peptides (Aroc=0.999). The predictivity of the structure-based method was lower in this analysis, which may be due to the difficulties in predicting the structures of peptide-MHC complexes and the limitations in the availability and quality of the X-ray data.

The results of the present test also suggested that no algorithm was good in all the predictions. SYFPEITHI, SVMHC and the additive method were good for

predicting T cell epitopes and naturally processed peptides, RANKPEP and the additive method was better at predicting the poly-alanine peptides. Since Dr. Brusic's data set overlapped with the RANKPEP training data, the ability of RANKPEP in predicting peptides derived from other sources remains to be tested. As poly-alanine peptides are synthetic peptides used to aid MHC-peptide research, it has little value in the prediction of natural T cell epitopes. Therefore the ability of the algorithms to correctly predict naturally processed peptides and T cell epitopes is more important for vaccine research.

The performance of different algorithms was affected by their training sets. SYFPEITHI used only T cell epitopes as the training set, which explains why it was better at predicting T cell epitopes and naturally processed peptides but was less predictive when poly-alanine peptides were used.

The best performing algorithms from the first test were used in the second test, where their ability to find T cell epitopes in protein sequences was examined. The predictivity of the algorithms was about 10% lower than that of the first test. The additive method still maintained a high level of predictivity, with a correct prediction rate of over 70% in all three test sets. This slight decrease of performance was due to the test set used. In the first test, the binding data were from the database and non-binding data were laboratory tested non-binders, which meant that there was a clear difference between the two groups. However, in the second test, the difference in affinity between epitopes and other peptides may be less clear. There was more chance that some of the protein fragments did possess one or more of the ancillary anchor residues, or may even have one anchor residue. In this case, it

required the server to have high sensitivity and specificity to be able to pick the epitope from the sequence. This is also the present bottleneck for not only MHC-binding prediction, but all proteomic prediction servers.

Previously, Kun Yu carried out a study to compare the predictivity of A*0201 models of different algorithms (Yu *et al.*, 2002). The A*0201 data set used in the first part of the present study was also used as part of his test set. In his work, the predictivity was compared between matrix based predictions (BIMAS, SYFPEITHI and their in house models) and machine learning algorithms ANN and hidden Markov models. There was a good correlation between the two experiments. In both experiments, SYFPEITHI had a high Aroc value for naturally processed peptides and T cell epitopes, and the predictivity of BIMAS was similar for both T cell epitopes and naturally processed peptides. In the two experiments Aroc values of poly-alanine peptides were both the highest among his predictions, with the average Aroc value above 0.9.

Among all the MHC alleles, A*0201 is the best studied and there are more than two hundred ligands in the AntiJen database alone, while some of the alleles have less than 30 binding peptides available. The large group of available data makes it easier to produce a good quality model for the A*0201 allele. This is reflected in the results of the second test, where the predictivity of A*0201 epitopes was significantly higher than that of class II and mice MHCs.

The cross-comparison of all the additive models (table 4.1) showed most models had q^2 value higher than 0.3 and r^2 values were higher than 0.9, which indicated

good level of predictivity in QSAR studies. The level of predictivity was proved in the two validation tests. In the first test where database peptides were used, a prediction rate of more than 90% was observed in all peptide groups. In the second test, the additive method maintained a high level of accuracy of 70% or more, while the performance of other algorithms dropped. For epitope prediction servers, to be able to correctly predict epitopes is more important than predicting good binders. A T cell epitope can both bind to MHC and be recognised by the T cell receptor, whereas a good binder may not be recognised by the T cell receptor and the immune system will not be activated. In laboratories, a good binder is usually identified first, which is then tested for immunogenicity. The additive method has demonstrated the ability to not only predict high binders, but epitopes, therefore greatly shortening the epitope discovery process.

Epitopes used in the present study were mainly restricted to HLA-A2, A3, DRB1, H-2Kb and Db alleles, which was because they were the common alleles studied and most papers found in the literature focused on these alleles. In the future, more data will be collected and the predictivity of other models can be tested.

One of the options in the MHCPred interface is to choose a cut off threshold for the output. The problem remains as to what level of cut off the user should use? At present the default threshold for MHCPred is 5000nM (IC_{50}), as peptides with affinities lower than 5000nM are considered as non-binding peptides. A more specific cut off is required to help the user to identify epitopes. Sette et al. has defined a cut off of 1000nM for epitope prediction (Sette *et al.*, 1989a), later the cut off was reduced to 500nM (Sette *et al.*, 1994). In the ROC plots, the plateaus start

after pIC_{50} of 6.5 (equivalent to IC_{50} of 300nM), that is, the performance of the algorithm reaches the peak at 6.5. According to the analysis, a threshold of 300nM would be sufficient to identify the best MHC binders.

In all the predictions, the predictivity of the additive amino acid only model was higher than that of the amino acid and interactions model. Similar results have been observed in the evaluation of T cell epitope prediction algorithms by Peters et al. (Peters *et al.*, 2003). This is because of the limitations in the size of the training data set. The additive method considers two types of interactions: the interactions between adjacent amino acid and the interactions between one amino acid and every second amino acid. The total number of possible amino acid pair combinations is 6,000 (20x20x8+20x20x7). The biggest data set used for additive model generation was 335 (A*0201), which was not able to include all the amino acid combinations. Therefore the reason of the relatively lower predictivity of the amino acid and interaction model was due to lack of data. This finding was in agreement with the comparative study by Peters et al. (Peters *et al.*, 2003). This can be improved as more peptide binding experiments become available in the literature.

To conclude, the evaluation tests showed that the additive method is a reliable algorithm for studying MHC-peptide interactions. There are a number of human and mouse class I and II models available and as the research carries on more models will be produced and put on-line.

Chapter 5

Definition of an HLA-A3 supermotif using CoMSIA

5.1 Introduction

In chapter three, a 2D QSAR technique - the additive method - was applied to the study of peptides binding to HLA-A3 alleles and the definition of a HLA-A3 peptide binding motif. The additive method compared amino acids present at each position of the peptide and their effect on binding affinity, and derived a regression equation that can be used to predict the affinity of as yet untested peptides. In the present chapter, a 3D QSAR method, CoMSIA, was used to define the HLA-A3 supermotif. Previously, CoMSIA was applied to peptides binding to the HLA-A*0201 allele and a good description of the peptide - binding site interaction was obtained (Doytchinova and Flower, 2001). More recently, the technique was applied to peptides binding to the HLA-A2 supertype and defined an A2 supermotif (Doytchinova and Flower, 2002). In this project, the CoMSIA technique is used to define the amino acid preferences of peptides binding to the HLA-A3 family A*0301, A*1101, A*3101 and A*6801 (Guan *et al.*, 2003a).

5.2 Results

5.2.1 The CoMSIA models

The peptide training set for each allele was collected from the AntiJen database (McSparron *et al.*, 2003). Only nonamers were included in the set since they are the most common peptides bound to HLA class I molecules. The data set used in

CoMSIA was identical to the set used in the additive method (see section 2.1.7). Some peptides in the training set bound to more than one allele. The correlation coefficients between the affinity data for the common peptides ranged from 0.168 for A*3101/A*6801 (n = 22) to 0.661 for A*0301/A*1101 (n = 50). The pIC_{50} ranges were from 3.3 to 3.5 log units. As with the additive method, peptides with |residual values| larger than 1.5 were stepwisely excluded during QSAR model generation. The process was repeated until q^2 started to drop. Because of the different techniques used in the experiments, the number of outliers was different between the additive method and CoMSIA. However, some of the outliers were identical in both calculations, which often did not possess preferred anchor residues and had low experimental activity.

The all-fields models for each of the four HLA-A3 alleles are presented in table 5.1. The model of A*3101 had the highest predictivity (q^2 = 0.700). The predictivity of the models for A*6801, A*1101 and A*0301 was 0.570, 0.496 and 0.486, respectively. The models produced 56-90% of their affinity predictions with residuals less than 0.5 log units and the percentage of poorly predicted peptides (residuals > 1.0) was between 0 and 18%.

The values of r^2 were greater than 0.9 for the four models, indicating a good correlation between peptide structures and binding affinities. The non-cross-validated analyses showed that the local hydrophobicity and hydrogen bond donor ability had the highest fractional values, followed by electrostatic, hydrogen-bond acceptor and steric properties. All the models provided a high level of peptide prediction, ranging from 50 to 90%, and the percentage of poorly predicted peptides

was between 0 and 10%. As the affinity range for each allele was slightly different, the ratio of the SEP to affinity range and SEE to affinity range were used to assess the fitness and predictivity of the models. This ratio should generally be <10% for good QSAR models and as a rule the ratio SEP/affinity range is higher than the ratio SEE/affinity range. The present models had ratios from 16.5 to 18.6% for SEP/affinity range and from 2.8 to 8.4% for SEE/affinity range.

Five contour maps were generated for each allele, representing the five physicochemical properties: steric bulk, electrostatic potential, local hydrophobicity, and hydrogen donor and acceptor abilities. The maps were produced using non-cross-validated PLS (figure 5.1 to figure 5.5). The favoured and disfavoured areas for each property were highlighted in different colours in the map. A property that was favoured by two or more alleles without being disfavoured by any of them was considered a favoured property for the supertype. A property disfavoured for two or more alleles was defined as being disfavoured for the supertype.

Comparing the fractional values of the different properties in table 5.1, steric complementary was not the major property involved in peptide-MHC binding. Figure 5.1 showed that the contribution of steric bulk was very different among the alleles. Steric bulk was favoured at P5 for two of the alleles (A*1101 and A*3101) without being deleterious for the other alleles. Electron density was favoured at P3, P4, P5, P8 and P9 (figure 5.2). Hydrophobicity was favoured at P7 and disfavoured at P6 (figure 5.3). Hydrogen-bond donor groups were favoured at P1, P4 and P6 and disfavoured at P5 (figure 5.4). As was evident from figure 5.5 hydrogen acceptors were well accepted at P6, but they were disfavoured at pP5.

		*A*1101*		*A*0301*		*A*3101*		*A*6801*	
Number of peptides		59		69		30		39	
Grid spacing (Å)		2		2		1.5		2	
α[a]		0.6		0.6		0.6		0.5	
Column filtering (kcal/mol)		0.5		0.5		0.5		0.5	
q^2_{loo}[b]		0.496		0.486		0.700		0.570	
Number of components		8		6		4		10	
SEP[c]		0.588		0.629		0.551		0.655	
SEP/affinity range (%)		17.3		18.1		16.5		19.1	
q^2_{cv5}[d]		0.416		0.424		0.640		0.349	
r^2		0.972		0.959		0.921		0.950	
SEE[e]		0.141		0.177		0.282		0.119	
SEE/affinity range (%)		4.1		5.1		8.4		3.4	
$r^2_{bootstrap}$		0.989		0.971		0.950		0.961	
F ratio		167.666		241.818		73.177		126.217	
Fraction	Steric	0.114		0.104		0.071		0.126	
	Electrostatic	0.234		0.277		0.254		0.251	
	Hydrophobic	0.250		0.260		0.225		0.257	
	H donor	0.261		0.226		0.364		0.241	
	H acceptor	0.141		0.133		0.087		0.125	
$\|res.\| \leq 0.5$		40	67.80%	42	60.87%	27	90%	22	56.41%
$0.5 < \|res.\| \leq 1.0$		15	25.42%	20	27.03%	3	10%	10	25.64%
$\|res.\| > 1.0$		4	6.78%	7	9.46%	0	0	7	17.95%
Mean |residual|		0.443		0.585		0.179		0.516	
Standard deviation		0.343		0.500		0.188		0.377	

[a]Attenuation factor.
[b]q^2 factor obtained after leave-one-out crossvalidation.
[c]Standard error of prediction.
[d]q^2 obtained by crossvalidation in five groups.
[e]Standard error of estimate

Table 5.1. CoMSIA models for A3 alleles A*1101, A*3101, A*0301 and A*6801.

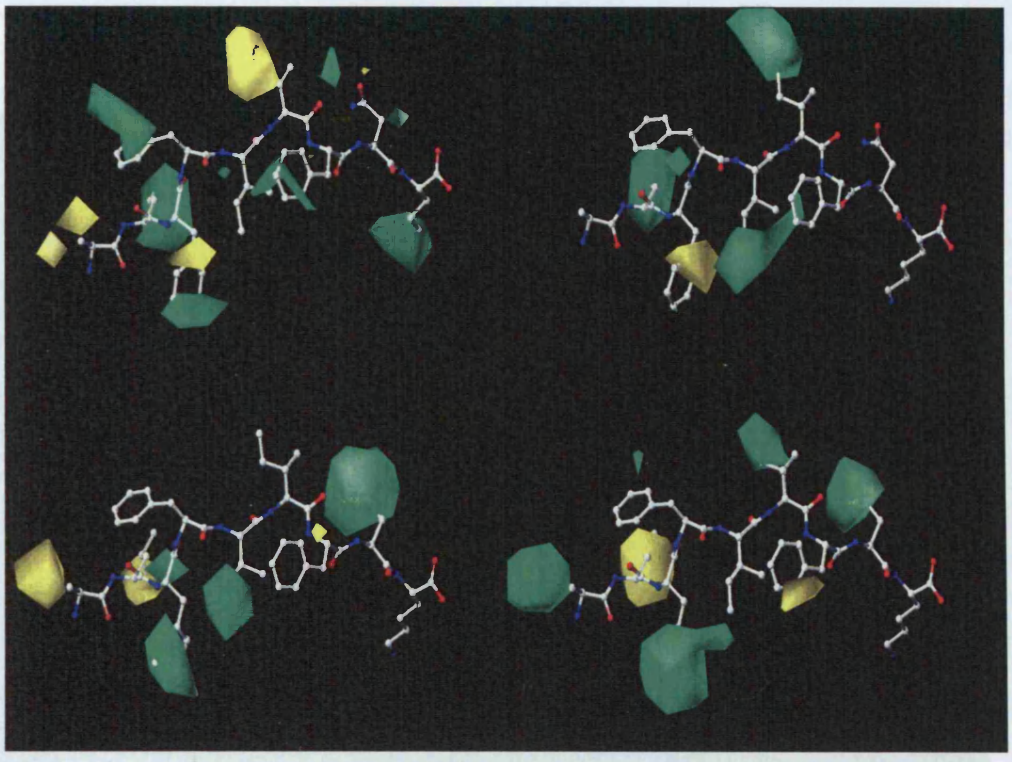


Figure 5.1. CoMSIA stdev*coeff contour maps displaying the steric bulk property surrounding the peptide. Peptide ALFFIIFNK is shown inside the fields. The peptide is positioned with the N-terminus and position 1 to the left. Green and yellow areas indicate where steric bulk will increase or decrease the affinity, respectivity. Upper left: A*0301. Upper right: A*3101. Lower left: A*1101. Lower right: A*6801.

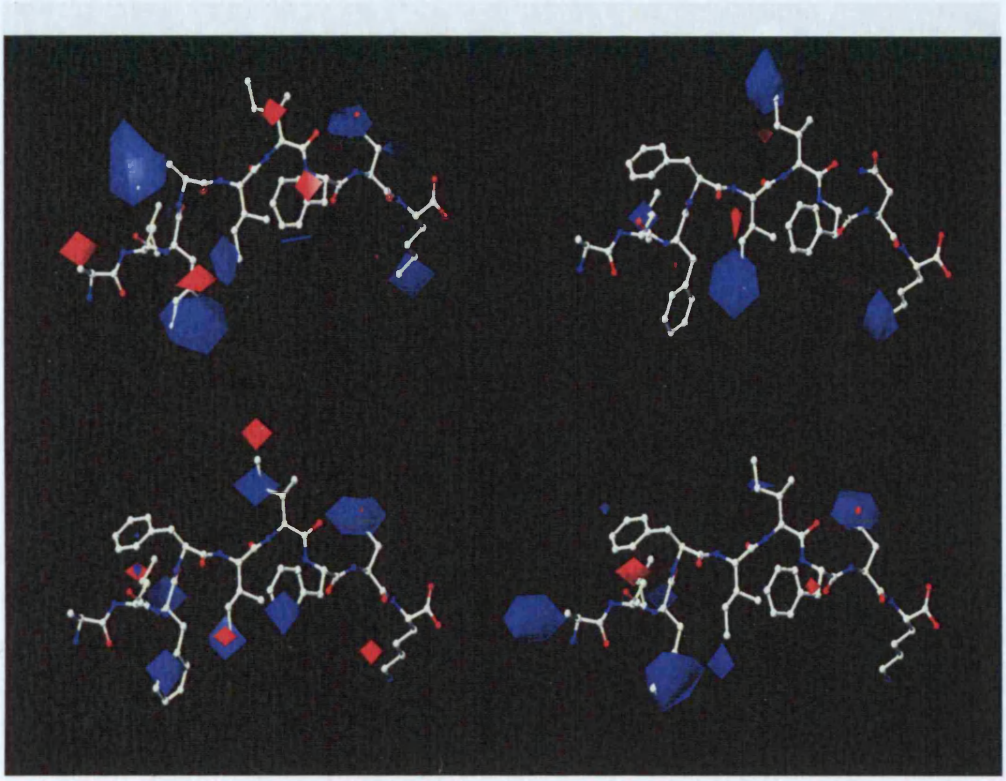


Figure 5.2. Electrostatic potential. Peptide ALFFIIFNK is shown inside the fields. The peptide is positioned with the N-terminus and position 1 to the left. Blue and red areas indicate where negative electrostatic potential will increase or decrease the affinity, respectively. Upper left: A*0301. Upper right: A*3101. Lower left: A*1101. Lower right: A*6801.

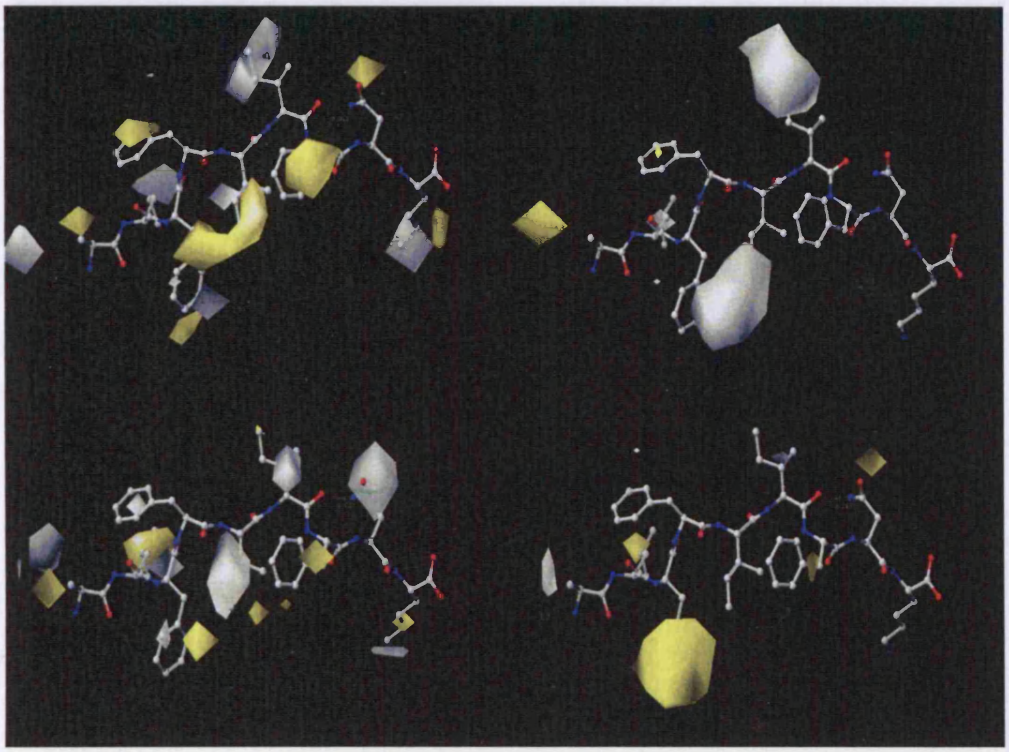


Figure 5.3. Local hydrophobicity contour map. Peptide ALFFIIFNK is shown inside the fields. The peptide is positioned with the N-terminus and position 1 to the left. Yellow and white areas indicate where hydrophobic amino acid residues will increase or decrease the affinity, respectively. Upper left: A*0301. Upper right: A*3101. Lower left: A*1101. Lower right: A*6801.

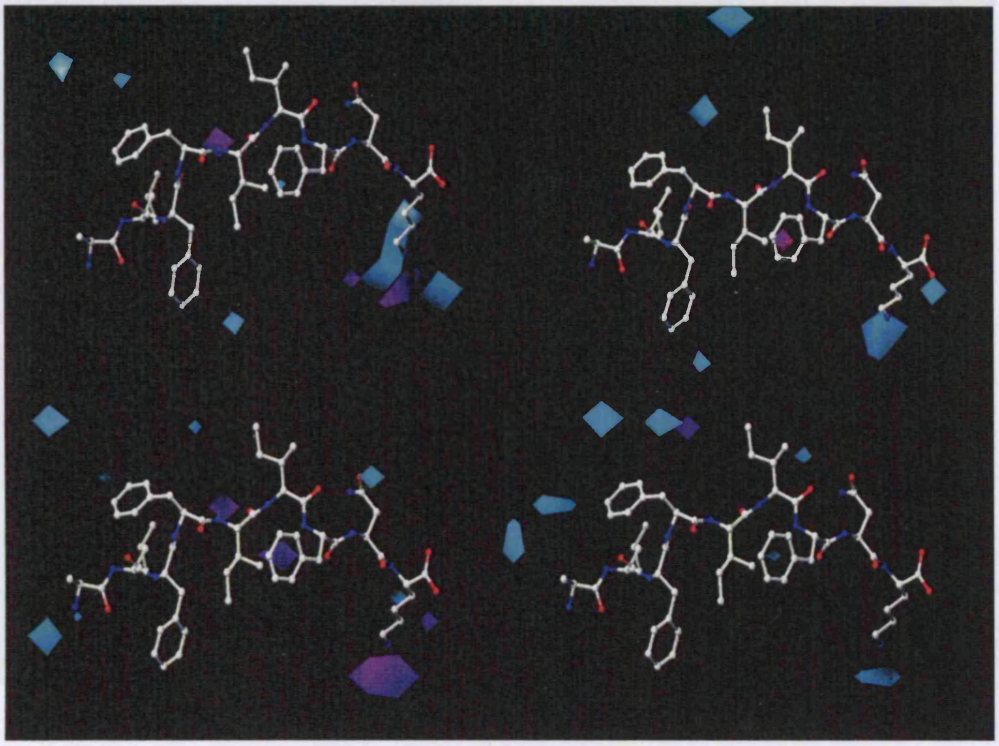


Figure 5.4. Hydrogen donor contour map. Peptide ALFFIIFNK is shown inside the fields. The peptide is positioned with the N-terminus and position 1 to the left. Cyan and purple areas indicate where hydrogen-bond donor group on the ligand will increase of decrease the affinity, resepectively. Upper left: A*0301. Upper right: A*3101. Lower left: A*1101. Lower right: A*6801.

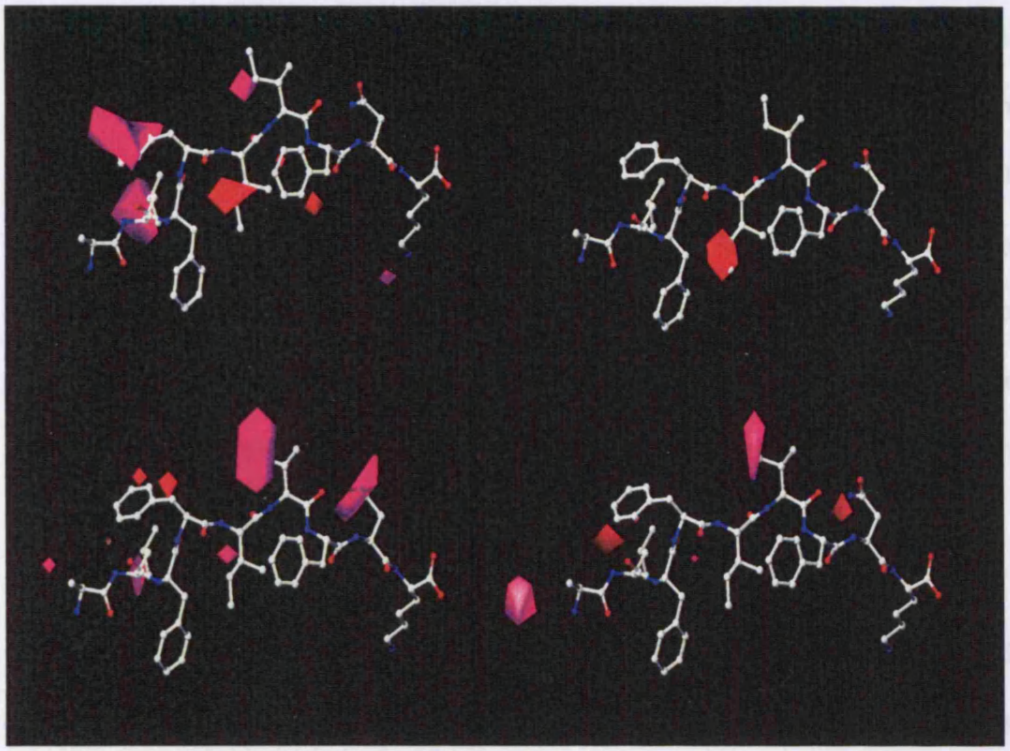


Figure 5.5. Hydrogen bond acceptor abilities. Peptide ALFFIIFNK is shown inside the fields. The peptide is positioned with the N-terminus and position 1 to the left. Magenta and red areas indicate where hydrogen-bond acceptor groups on the ligand will increase or decrease the affinity, respectively. Upper left: A*0301. Upper right: A*3101. Lower left: A*1101. Lower right: A*6801.

5.2.2 The peptide binding experiment

Using the results from the HLA-A2 additive model and CoMSIA contour maps, some high binders for A*0301 allele were designed and their binding affinities were tested experimentally. The experimental protocol is the same as for A*0201 binding peptides, a T2 stabilisation assay. The binding affinities of eleven already defined peptides from the literature were measured to ensure that the test was reliable and that there was a linear relationship between IC_{50} and BL_{50} measurements (correlation coefficient of 0.778). The peptides selected include high, medium and low binders. Sequences of the reference peptides, their measured IC_{50} values in the original papers, and BL_{50} values measured in the present experiments were listed in table 5.2.

The experimental binding affinities of the reference peptides and their predicted binding affinities were plotted in figure 5.6. The measured BL_{50} values (presented as $-\log BL_{50} = pBL_{50}$) were plotted against the negative logarithm of the literature values (pIC_{50}). For A*0301 peptides used in the experiment, peptides with pBL_{50} values below 4 ($BL_{50} > 10^{-4}$) are medium or low binders and those above 4 ($BL_{50} < 10^{-4}$) are high binders. A total of nine test peptides were designed and tested. Among the designed peptides, amino acids that occupy anchor position 2 and 9 were the preferred amino acids from the model, Ile at position 2 and Lys/Arg at position 9. Similarly, the secondary anchor positions 3 and 7 also had the preferred amino acid Phe. Other positions were more flexible and contained different amino acids. The BL_{50} values of all test peptides were above 4, indicating good binding to the MHC allele. The experimental and predicted binding affinities of the peptides were listed in table 5.3.

Peptide	*Experimental IC_{50}*	*BL_{50}*
RINEEKHEK	6.105	3.621
LLIFHINGK	6.321	3.550
GTGSGVSSK	6.9	4.234
VLSHNSYEK	7.102	4.592
AIFQSSMTR	7.301	4.612
LVKSPNHVK	7.64	5.028
SIFQSSMTK	7.921	5.580
ALNFPGSQK	8.071	5.383
GTMTTSLYK	8.469	5.322
HLFGYSWYK	8.658	5.420
ALFQRSMTR	7.432	4.166

Table 5.2. The reference peptides, their experimental binding affinities (IC_{50}) recorded in the literature, and measured BL_{50} values in the present experiments.

Peptides	*Predicted binding affinity - log IC_{50} (M) by additive models*	*Experimental binding affinity (BL_{50})*
GIFTYGFRK	8.53	5.030
GIFTYGFMK	8.44	4.600
VIFTYGFRK	8.129	5.054
GIFTYGFYK	8.241	4.514
GIFRYGFRK	8.327	5.078
VIFTYGFMK	8.084	4.274
HIFTYGFRK	8.135	5.010

Table 5.3. The designed peptides, their predicted binding affinities (pIC_{50}) using the additive model and their measured binding affinities.

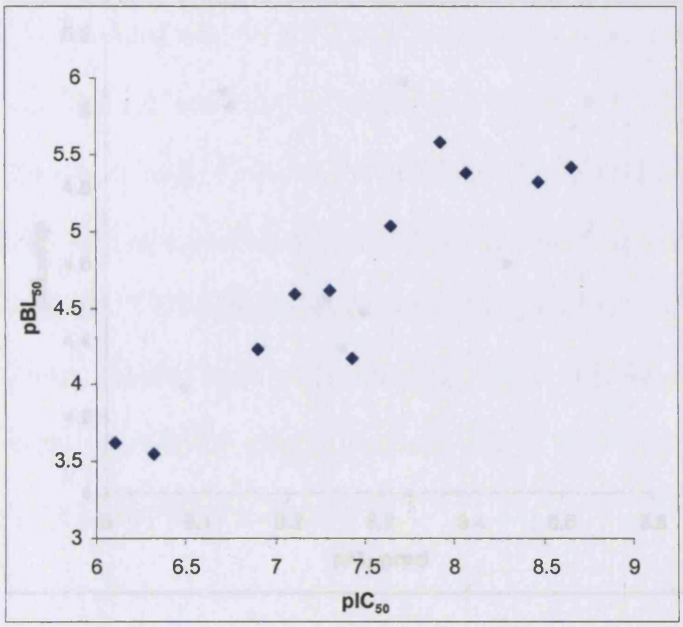


Figure 5.6 The correlation between literature values and experimental values of the peptides. The IC_{50} values (the logarithm values of the experimental IC_{50} value) were on the X axis and the BL_{50} values (converted from experimentally measured fluorescence intensity values) were on the Y axis. The two values form a near-linear relationship with a correlation coefficient of 0.778.

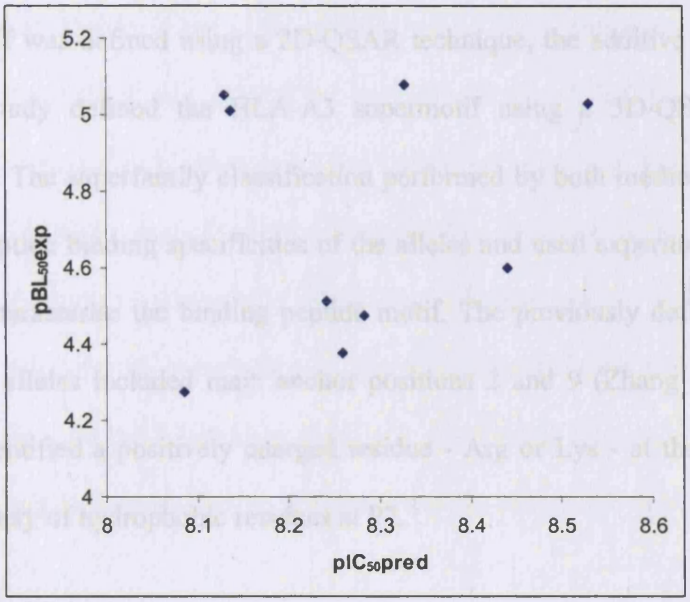


Figure 5.7. The predicted binding affinities of the designed peptides and the measured binding affinity BL_{50} values.

5.3 Discussion

HLA-A3 alleles A*1101, A*0301, A*3101 and A*6801 bound peptides with similar anchor residues. Sequence analysis showed that only 11 of the residues inside the binding pockets were polymorphic. In chapter three, a HLA-A3 supermotif was defined using a 2D-QSAR technique, the additive method. The present study defined the HLA-A3 supermotif using a 3D-QSAR method, CoMSIA. The superfamily classification performed by both methods was based on the peptide binding specificities of the alleles and used experimental binding data to characterise the binding peptide motif. The previously defined motif of HLA-A3 alleles included main anchor positions 2 and 9 (Zhang *et al.*, 1993), which identified a positively charged residue - Arg or Lys - at the C terminus, and a variety of hydrophobic residues at P2.

Some differences in side chain preferences by the A3 alleles were found at P2. In the additive models, it was found that small residues were preferred by A*6801 and A*1101. A*0301 and A*3101 preferred medium or large hydrophobic residues. Similar results were found in the CoMSIA contour maps. Steric bulk was favoured at P2 for A*3101 and A*0301 but disfavoured in A*6801 and A*1101 models. The study of crystal structures of MHC molecules showed that the side chain of the P2 residue bound in pocket B (Madden *et al.*, 1991a). There are different residues lining pocket B in the different HLA-A3 molecules, Tyr9 in A*1101 and A*6801, Phe9 in A*0301 and Thr9 in A*3101. The presence of Tyr9 in A*1101 and A*6801 allows pocket B to accommodate lager side chains like Leu. Electrostatic potential, hydrophobicity and hydrogen bond acceptance maps varied at this position, which was in agreement with the broad spectrum of

well-accommodated amino acids found here, from the bulky, hydrophobic Leu to the small polar amino acid Thr.

The most important property for the amino acid at position 9 is hydrogen-bond donor ability. It was favoured in A*6801 and A*3101, and was disfavoured in A*1101. The side chains of A*0301 both favoured and disfavoured hydrogen bond donor potential. In some cases, the change of Lys to the larger residue Arg could affect the expression of the molecule (Sidney *et al.*, 1996). Results from the present study suggested the interaction between residue 9 and the MHC molecule may play an important role. The side chain of larger basic residue Arg could extend to the bottom of pocket F in A*6801 and A*3101, forming hydrogen bonds with residues at the bottom of the pocket and thus stabilising the complex.

The five residues that directly interact with the peptide in the F pocket were identical in both the A3 family and HLA-B27 (Leu81, Asp116, Tyr123, Thr143 and Trp147). Arg and Lys bound to pocket F by extending to the bottom of the pocket and interacting with negatively charged residues Asp116 or Asp77 in both the A3 family and HLA-B27. B27 had been shown to accept hydrophobic residues like Leu, Ala and Tyr because of their interaction with Leu81, Tyr123, Thr143 and Trp147 (Jardetzky *et al.*, 1991). In the present study, the specificity of residues at position 9 was restricted to Arg and Lys only; both Ala and Tyr were deleterious in the additive method. This suggests a possible difference in conformation of the binding pocket in spite of sequence similarity. Also, this

could be the result of a change of conformation after the binding of other amino acids in the peptide.

Secondary anchor positions 1, 3, 5, 6 and 7 were also of great importance. The common favoured property for position 1 was hydrogen-bond donor ability. The electron density at position 3 was preferred for three of the alleles. Sidney and co-workers found that peptides with aromatic residue at P3, like Tyr, Phe and Trp had a 31 fold increase in binding affinity to A*0301 (Sidney *et al.*, 1996). Bulky side chains with high electron density were preferred at position 5. Hydrogen-bond donor and acceptors were disfavoured here. Hydrophilic amino acids capable of forming hydrogen bonds were well accommodated at position 6. Hydrophobic residues at P7 were preferred by both additive and CoMSIA models.

Positions 4 and 8 faced away towards the T-cell receptor (Silver *et al.*, 1992), but still could contribute to affinity. Electron density was favoured at both positions. Additionally, hydrogen-bond ability was important for position 4 and steric bulk was disfavoured at position 8.

To conclude, in order to bind to members of the HLA-A3 superfamily, a peptide requires a small to medium sized residue at position 2, such as Ile or Threonine, and a positively charged residue Arg at position 9. Phe at either position 3 or 7 was required for stable binding (Guan *et al.*, 2003a).

According to the supermotif derived from chapter 3, and the present CoMSIA results, nine high binders of the A*0301 allele were designed and their binding affinity tested. All of the peptides had anchor amino acid Lys at position 9 and Ile at position 2, which bind to pocket F and B, respectively. These are the most important amino acids of the peptide and are required for high level binding. Secondary anchor positions 3 and 7 of the peptides were occupied by Phe, which was identified as the preferred amino acid at these two positions by CoMSIA. All peptides displayed good binding to the A*0301 allele, the BL_{50} values of all peptides were above 4, the average binding affinity is 4.7. Three peptides were found to be the best binders, with BL_{50} values above 5: VIFTYGFRK, GIFRYGFRK and HIFTYGFRK, none of them have been recorded previously in the AntiJen database.

In the present study, CoMSIA was applied to HLA-A3 alleles. Five contour maps were generated, describing the steric, electrostatic, hydrophobic, hydrogen bond donor and acceptor forces that were favoured or disfavoured by the A3 peptides. Besides the detailed explanatory ability, the results can also be used to design high affinity peptides of the A3 alleles. CoMSIA is an effective method for describing ligand-receptor interactions in drug design. The present study demonstrated that it could also be used in immunology to characterise binding motifs for MHC molecules. In the future, CoMSIA can be applied to other HLA alleles and can also be used in other immunology problems such as antibody-antigen reaction and modelling B cell epitopes.

Chapter 6
Class I HLA supertype classification by GRID/CPCA

6.1 Introduction

Sette et al. was the first to group class I HLA alleles into superfamilies according to their binding motifs (Sidney *et al.*, 1996a). Several HLA supertypes were described - A2 (del Guercio *et al.*, 1995; Sidney *et al.*, 1996a), A3 (Sidney *et al.*, 1996b) and B44 (Sidney *et al.*, 2003). Later the number of defined supertypes was extended to nine (Sette and Sidney, 1999), which were A1 (A*0101, A*2501, A*2601, A*2601, A*3201), A2 (A*0201-07 A*6802 A*6901), A24 (A*2301 A*2402-04 A*3001-03), A3 (A*0301 A*1101 A*3101 A*3301 A*6801), B7 (B*07 B*35 B*51 B*53 B*54 B*55 B*56 B*67 B*78), B27 (B*1401 – 02 B*1503 B*1509 B*1510 B*1518 B*2701 – 08 B*3801 B*3802 B*3901 – 04 B*4801 B*4802 B*7301), B44 (B*37 B*4001 B*4002 B*4006 B*41 B*44 B*45 B*47 B*49 B*50), B58 (B*1516 B*1517 B*5701 B*5702 B*58) and B62 (B*1301 – 02 B*1501 B*1502 B*1506 B*1512 B*1513 B*1514 B*1519 B*1521 B*4601 B*52). Sette's classification was a motif-based approach and required binding motifs for each allele. However, most of the 783 known class I HLA alleles have not been studied experimentally. To characterise all HLA alleles using experimental binding assays is both expensive and time consuming.

In this chapter, I describe a chemometric strategy for classifying class I HLA molecules into supertypes, using information drawn solely from the protein sequences. The techniques used were GRID (Cruciani and Watson, 1994) and principal component analysis (PCA) (Inoue and Kajiya, 1976; van der Voet and

Franke, 1985), in which the molecular interaction fields (MIFs) between the chemical probes and the HLA molecules were calculated in GRID and the MIFs were then used to build PCA/CPCA models. Results of the GRID/CPCA analysis were compared with the classification using hierarchical clustering analysis on CoMSIA fields; together the results were used to classify HLA molecules and generate 'supertype fingerprints', that is, the sequence features for supertype classification (Doytchinova *et al.*, 2004a).

In chemical or pharmacological analysis, often many drug targets are studied in one experiment and little information can be extracted from the data directly (Pate *et al.*, 2004). PCA simplifies the data by replacing the large number of variables in the original data set with a few new, uncorrelated variables called principal components (PC) (Inoue and Kajiya, 1976). The principal components are calculated in the order of importance, and most of the variance in the data can be explained by the first few components. A variation of the PCA, consensus PCA (CPCA) is also commonly used for calculations with multiple probes (Kastenholz *et al.*, 2000). CPCA divides values generated by each probe into blocks and it is easier to see which property is the most important in the model (Myshkin and Wang, 2003; Terp *et al.*, 2002).

Hierarchical clustering analysis is a statistical technique used in classifying large numbers of objects to reveal how closely the objects are related (Johnson, 1967). A common form of hierarchical clustering is the agglomerative algorithm, in which the calculation of hierarchical clusters starts by separating each object into a separate cluster (Guess and Wilson, 2002). The distance between two clusters

is dependent on the similarities between the two objects. The clustering is then improved by merging clusters that have the shortest distance (Guess and Wilson, 2002). The distance between the new clusters is recalculated. The steps are repeated until all clusters are clustered into a single cluster (Glazko and Mushegian, 2004). The result of the clustering is a binary tree with a root and many leaves, each leaf represents one object (Levenstien *et al.*, 2003). The order of the leaves is arbitrary. An HLA classification was carried out by Dr. Irini Doytchinova using hierarchical clustering based on CoMSIA fields, in which the alleles were clustered by comparing the generated CoMSIA fields of each molecule (Doytchinova *et al.*, 2004a).

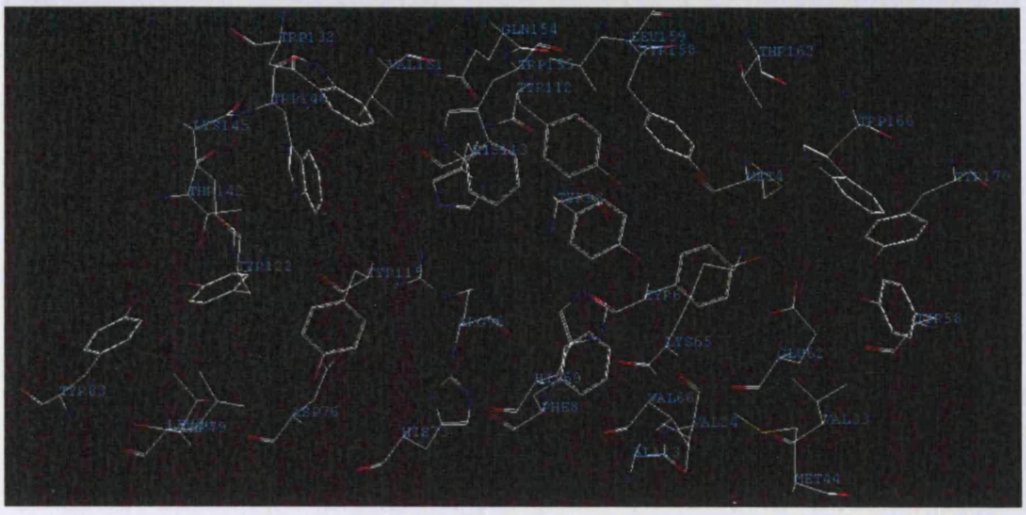
6.2 Results

6.2.1 Peptide binding site

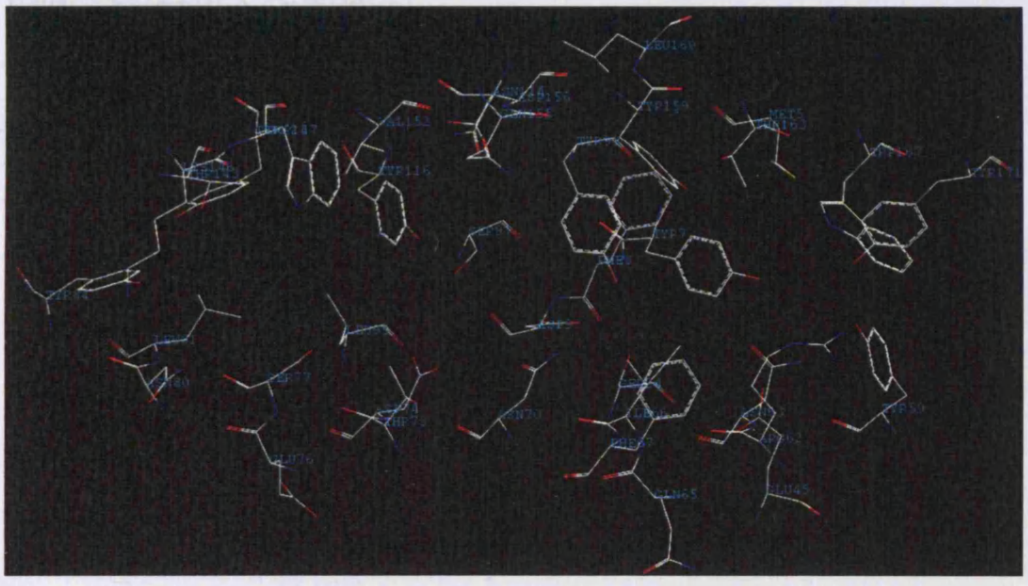
The structures of the peptide binding sites of the HLA-A, B and C molecules were constructed using the homology modelling program SCRWL 2.8 (Bower *et al.*, 1997). The binding sites of the HLA-A*0201 (Ding *et al.*, 1998), B*0801 (Reid *et al.*, 1996) and Cw*0401 (Fan *et al.*, 2001) molecules were used as the templates to define the GRID box (fig. 6.1). The dimensions of HLA-A*0201 and B*0801 binding sites were similar, the HLA-A*0201 binding site consisted of 35 residues and the HLA-B*0801 binding site had 37 residues. The binding site of the HLA-Cw*0401 molecules was smaller with 32 residues. Table 6.1 listed the residues that formed the binding site of HLA-A*0201, B*0801 and Cw*0401.

	Positions inside the binding site
HLA-A	5, 7, 9, 24, 25, 34, 45, 59, 63, 66, 67, 70, 74, 77, 80, 81, 84, 97, 99, 113, 114, 116, 123, 133, 143, 146, 147, 152, 155, 156, 159, 160, 163, 167, 171
HLA-B	5, 7, 8, 9, 24, 45, 59, 62, 63, 65, 66, 67, 70, 73, 74, 76, 77, 80, 81, 84, 95, 97, 99, 114, 116, 123, 143, 146, 147, 152, 155, 156, 159, 160, 163, 167, 171
HLA-C	5, 7, 9, 22, 59, 62, 63, 66, 67, 69, 70, 73, 74, 77, 80, 81, 84, 95, 97, 99, 116, 123, 124, 143, 146, 147, 156, 159, 163, 164, 167, 171

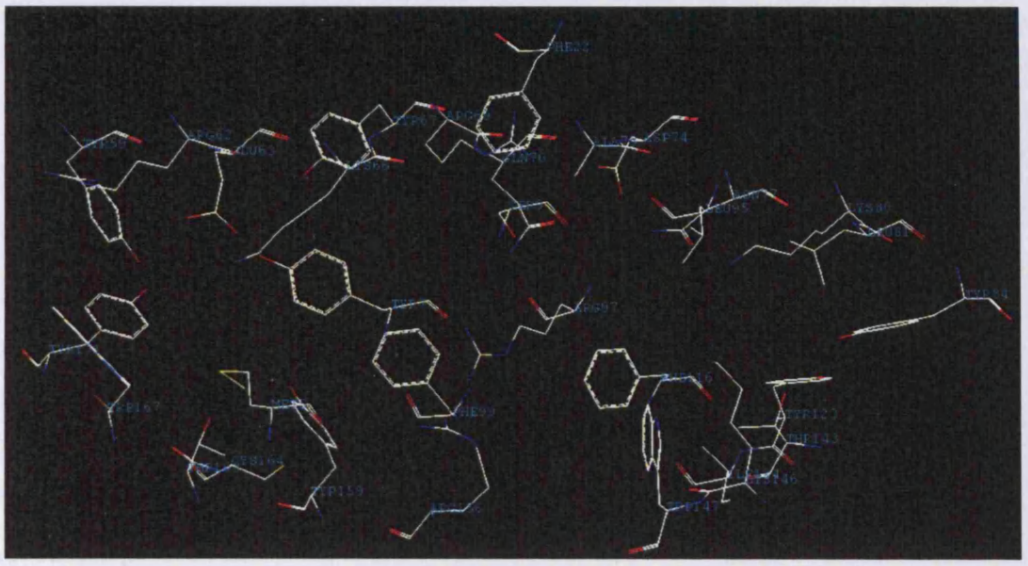
Table 6.1. List of residues that formed the peptide binding site of HLA--A*0201, B*0801 and Cw*0401.



a.



b.



c.

Figure 6.1. The 3D structures of the binding site: HLA-A*0201 (a), HLA-B*0801(b) and HLA-Cw*0401(c).

6.2.2 The HLA-A classification

PCA models were built using the program GOLPE (Cruciani and Watson, 1994). A PCA model was built for each of the 13 probes listed in table 6.2. The structures of 229 HLA-A molecules were built by homology modelling using the program SCRWL. A GRID interaction box was defined to only include the peptide binding site in the calculation. The GRID program placed the probe at each point of the grid box, and calculated the interaction energy between the molecule and the probe (see section 2.2.4.5). The energy values were used to build the PCA models in GOLPE. The accumulated explained variance of the first three components (PC1, PC2 and PC3) of the PCA models was used as a criterion for comparing the models, as most of the variance in the PCA models was explained by the first three PCs. PC4 and PC5 explained less than 10% of the total variance and were not used. Ten probes that produced the best PCA models, that is, the models with the highest explained variance by the first three PCs, were used to build a final CPCA model. The probes were OH_2, Dry, H, C3, C1=, N:#, N:=, N2+, OH and O.

In the CPCA model, PC1 explained 25% of the total variance and PC2 added a further 17%. The explained variance of the CPCA model was slightly less than that of the PCA models, indicating that the molecular forces represented by the probes are complicated and do not follow a simple additive pattern. However, the CPCA model is more important because it includes different interactions like electrostatic, hydrogen bonding, etc. The 3D scores plot of the CPCA model was shown in figure 6.2, in which the X, Y and Z axes represented PC1, PC2 and PC3, respectively.

	Probes	*Explained variance by the first three components of the PCA model (%)*		
		PC1	*PC2*	*PC3*
Single probe PCA model	OH_2	23.02	41.19	52.82
	Dry	34.96	55.09	67.55
	C3	24.90	44.17	54.88
	N:#	26.37	42.28	52.01
	H	29.25	47.50	56.51
	C1=	25.42	44.66	55.18
	N:=	25.42	42.21	54.01
	N1	25.59	41.06	51.18
	OH	23.71	42.11	53.37
	S1	22.48	41.66	52.32
	O1	22.66	38.81	50.70
	N2+	31.80	46.92	55.85
	O	26.85	42.82	54.05
10 probes model CPCA model	OH2, dry, H, C3, C1=, N:*, N:=, N2+, OH, O	24.98	41.56	51.60

Table 6.2. The chemical probes used in the GRID calculation. The cumulative explained variance of the first three principal components (PC1, PC2 and PC3) by the corresponding PCA model was listed. The cumulative explained variance of the final CPCA model using 10 probes is in the last row.

In the scores plot, each dot represented one HLA-A molecule, and each ellipse represented one cluster (figure 6.2). The first component of the CPCA model separated A23 and most of the A24 molecules on the left, with negative PC1 scores, from the rest of the HLA-A molecules. The second principal component separated the HLA-A*1, A*11, A*25, A*26, A*29, A*03, A*31, A*32, A*33, A*34, A*36, A*66, A*68 and A*74 families with positive PC2 scores from the others. The CPCA analysis showed that the HLA-A molecules were grouped into

three clusters as demonstrated in the 3D scores plot: the A3 cluster on the top right of figure 6.2, including the alleles A*01, A*03, A*11, A*25, A*26, A*29, A*30, A*31, A*32, A*33, A*34, A*36, A*4301, A*66, A*74 and A*8001. Most of the A*68 alleles (except A*6802 and A*6815, which were in the A2 cluster) were also included in the A3 family. The A24 cluster is on the top left of the figure including the A*23 and A*24 alleles. The A2 cluster is at the bottom of the figure, with most of the A*02 alleles. Other alleles in the A2 cluster were A*57, A*6802, A*6815, A*6823 and A*6901.

Figure 6.3 was the result of the hierarchical clustering analysis using CoMSIA fields (Doytchinova *et al.*, 2004a). Three clusters were also defined using the hierarchical clustering method. The cluster on the left includes HLA alleles A*02, A*25, A*26, A*3401, A*3405, A*4301, A*66, A*6802, A*6815, A*6823 and A*6901. This cluster was called the A2 cluster. The A24 cluster was well distinguished and included A*23 and A*24. Finally, the A3 cluster included A*01, A*03, A*11, A*29, A*30, A*31, A*32, A*33, A*36. Some A*34 and A*68 alleles, A*74 and A*8001 were also in this cluster.

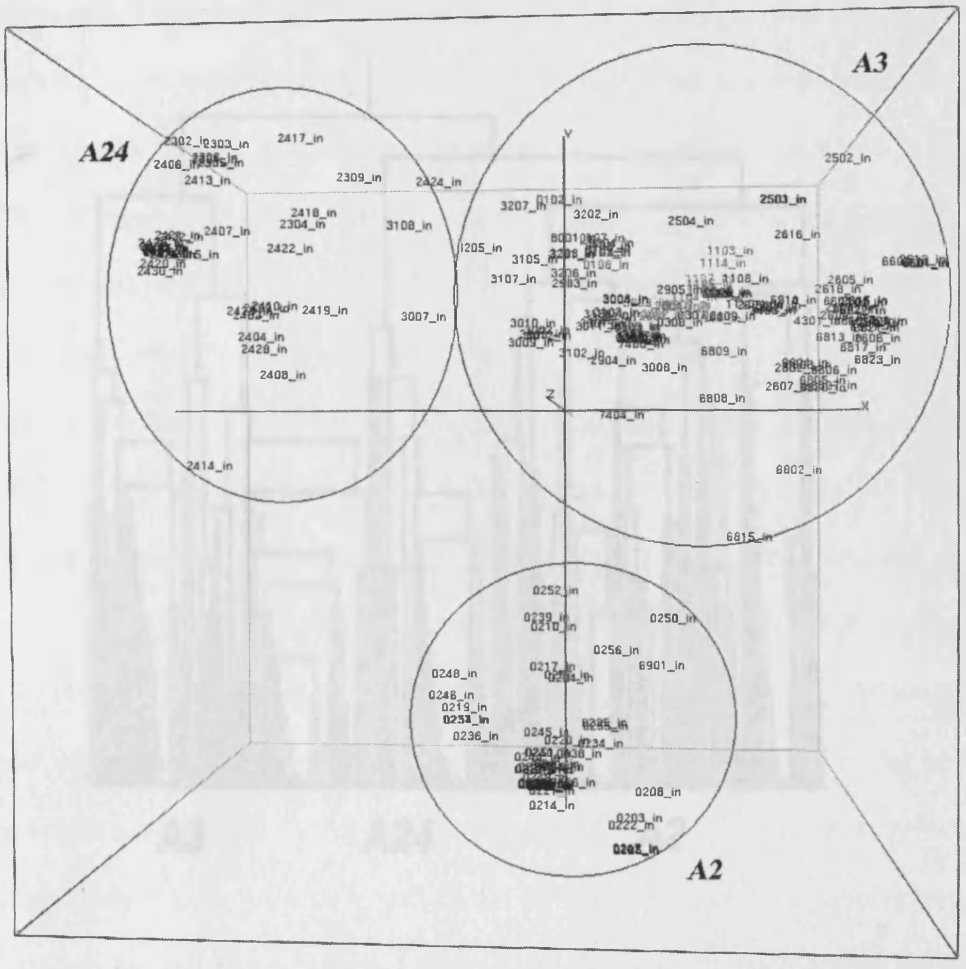


Figure 6.2. The 3D scores plot of the CPCA analysis for HLA-A molecules. The A24 cluster is on the top left of the plot, the A3 cluster is on the top right of the plot and the A2 cluster is below the X axis.

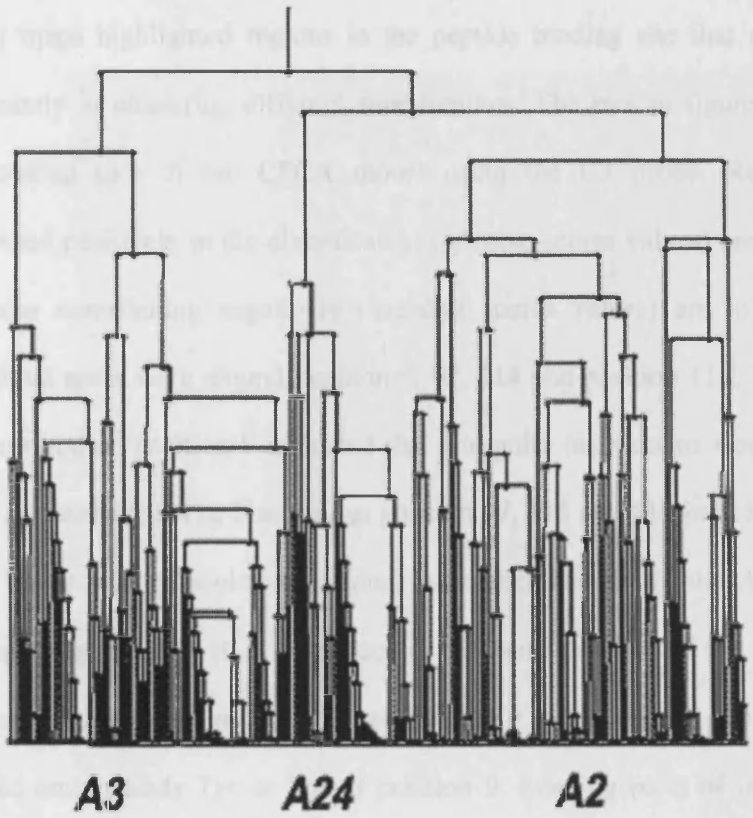
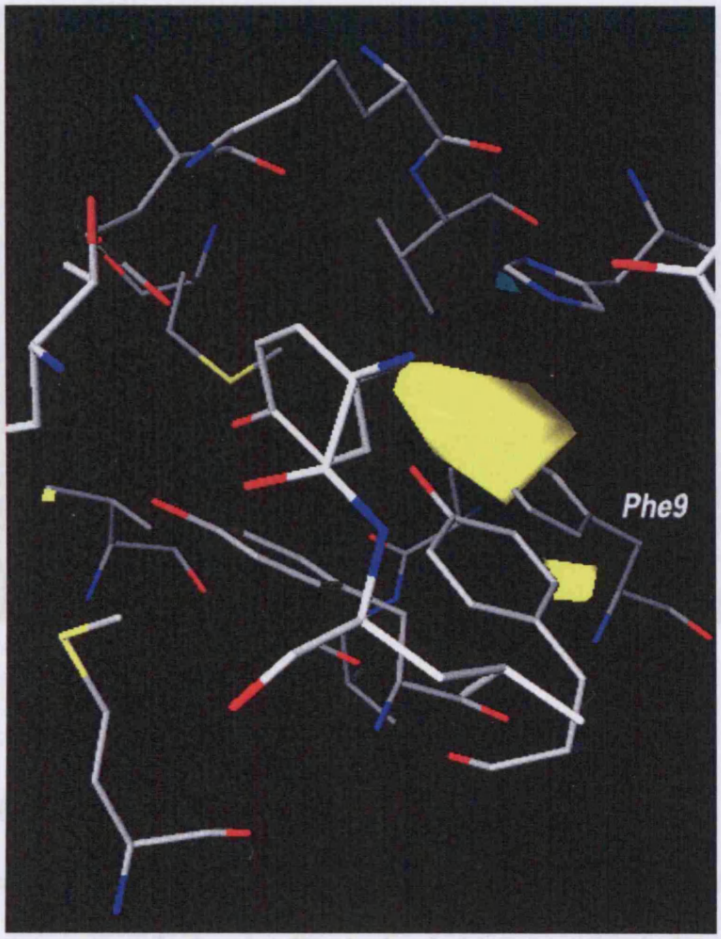
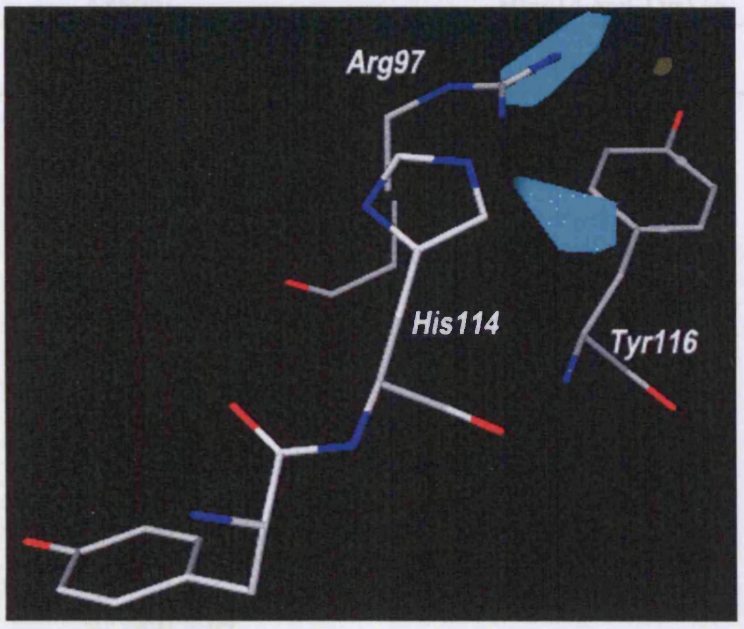


Figure 6.3. The HLA-A classification defined by hierarchical clustering (Doytchinova *et al.*, 2004a). A hierarchical tree was built for the 229 HLA-A alleles. Each leaf represented one allele. The results of the clustering were similar to that of the GRID/CPCA analysis, the three clusters were defined in both experiments: A2, A3 and A24.

In addition to the scores plot, the loading plot of the CPCA model was also generated. The scores plot showed the clustering of the HLA alleles, whereas the loading maps highlighted regions in the peptide binding site that contributed significantly in clustering different superfamilies. The plot in figure 6.4 is the PC1 loading plot of the CPCA model using the C3 probe. Regions that contributed positively in the classification (positive scores values) are in yellow, and those contributing negatively (negative scores values) are in blue. The highlighted areas were around position 9, 97, 114 and position 116. The yellow area surrounding position 9 indicated that non-polar interactions were favoured by the A3 supertype. The blue area at position 97, 114 and 116 indicated regions where bulky and hydrophobic residues were disfavoured by the A24 family. Sequence alignment of HLA-A molecules showed that most of the A24 alleles had dominant polar amino acid Ser at position 9, while the A3 molecules had aromatic amino acids Tyr or Phe at position 9. Loading plots of other probes highlighted the same residues, and are not shown. A summary of the molecules included in each cluster and the important amino acids for each cluster are in table 6.3.



(a)



(b)

Figure 6.4. The loading plot of the HLA-A CPCA model. The binding site of A*0201 is used in the plot to display the positions of the amino acids. There were two important interactions in the plot. The hydrophobic interaction is favoured at position 9 (a), and disfavoured around position 97, 114 and 116 (b).

Supertype	*Consensus PCA*	*Supertype fingerprint*
A2	A*0201 – 60 without 04, 17, 57	Tyr9/Phe9
	A*6802, 15, 23	Arg97
	A*6901	His114 and Tyr116
A24	A*2301 – 09	Ser9
	A*2402 – 38	Met97
A3	A*0101 - 09	Tyr9/Phe9/Ser9
	A*0301 – 10	Ile97/Met97
	A*1101 – 14	Glu114 and Asp116
	A*2501 – 04	
	A*2601 – 18	
	A*2901 – 07	
	A*3001 – 12	
	A*3101 – 09	
	A*3201 - 07	
	A*3301 – 06	
	A*3401 – 05	
	A*3601 – 04	
	A*4301	
	A*6601 - 04	
	A*6801 – 23 without 02, 15	
	A*7401 – 09	
	A*8001	

Table 6.3. A list of HLA alleles included in each cluster in the scores plot. For simplicity only the beginning and the end of the alleles were listed. For example, A*0201 – 60 meant that all sixty alleles from A*0201, A*0202, A*0203 ... to A*0260 were included in the cluster, etc. The amino acids used to define each cluster are shown in the last column.

6.2.3 The HLA-B classification

The structures of the HLA-B molecules were modelled using SCRWL, with the B*0801 structure as a template. After comparing the PCA models generated using single probes in table 6.3, the 10 best probes were selected for the CPCA model, which were OH_2, dry, C3, N:#, N1, H, N:=, OH, N2+ and O. Initially, when interactions within the whole binding site were considered in the CPCA model, no consensus pattern was found in the scores plot, which may be due to the slightly larger size of the HLA-B binding site and to the presence of irrelevant amino acids in the GRID interaction box. To increase the signal to noise ratio, a region of 4Å was applied so as to only include interactions within 4Å of the peptide binding site. Comparing the PCA models, the H probe model had the highest explained variance of nearly 100%, showing that hydrogen bonding is an important force in the HLA-B - peptide interaction. However, as the H probe only considers hydrogen bond donor and acceptor and cannot explain all molecular interactions of the HLA-peptide interaction, the CPCA model using the 10 best probes was used to classify HLA-B alleles. PC1 of the final CPCA model explained 18.40% of the total variance, and the second component explained a further 18.13% of the variance.

The scores plot of the first three components (figure 6.5) reveals that the HLA-B molecules are divided into three clusters: B7 (B*07, B*08, B*14, some B*15, B*18, B*35, B*3705, B*3904, B*41, B*42, B*45, B*48, B*50, B*55, B*56, B*6701, B*6702, B*7301, B*78, B*81, B*82 and B*83), which is on the left of the Y axis, B27 (B*27, B*37, B*38, B*4013, B*4019 and B*4028) in the top right corner of the plot, and B44 (B13, B44, B*47, B*49, B*51, B*52, B*53,

B*5607, B*57, B*58 and B*5901), lower right of the plot. The complete list of all the molecules included in each cluster is in the table 6.5. Similar clusters were found using hierarchical clustering method (figure 6.6), in which three clusters (B7, B27 and B44) were identified.

	Probes	*Explained variance by the first three components of the PCA model (%)*		
		PC1	*PC2*	*PC3*
Single probe PCA model	OH_2	35.14	51.28	66.96
	Dry	42.02	65.31	76.70
	C3	67.98	80.46	86.22
	N:#	66.56	79.49	85.46
	H	99.86	99.99	100.00
	C1=	23.09	39.29	52.12
	N:=	34.71	54.93	70.26
	N1	36.31	55.50	68.18
	OH	25.01	44.64	61.89
	S1	32.16	53.70	66.90
	O1	26.85	48.19	61.64
	N2+	47.07	62.21	74.79
	O	28.19	48.47	65.53
10 probes model CPCA model	OH2, dry, C3, N:*, N:=, N2+, OH, O, N1 and H	18.40	36.53	50.50

Table 6.4. The chemical probes used in HLA-B GRID/CPCA analysis.

The PC1 loading plot (figure 6.6) showed that two areas were important in the classification. Position 63 and 66 (figure 6.6a) were inside pocket A and B. Position 66 was conserved while position 63 was polymorphic with two amino acid variations Glu and Asn. The other important area in the loading plot was around position 77 and 81 in the pocket F. Asn, Ser and Asp were found at position 77, and Leu and Ala at position 81.

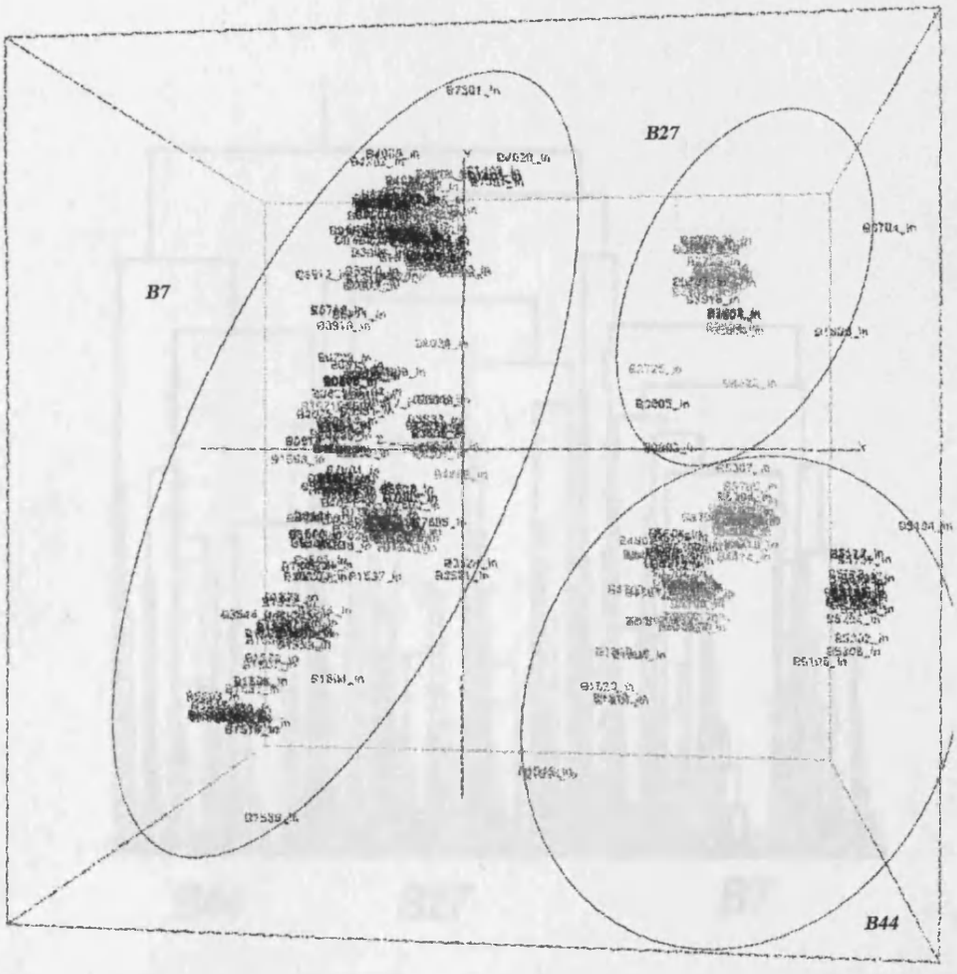


Figure 6.5. The 3D scores plot of the CPCA analysis for HLA-B molecules. Three clusters were identified in the plot: B7, B27 and B44.

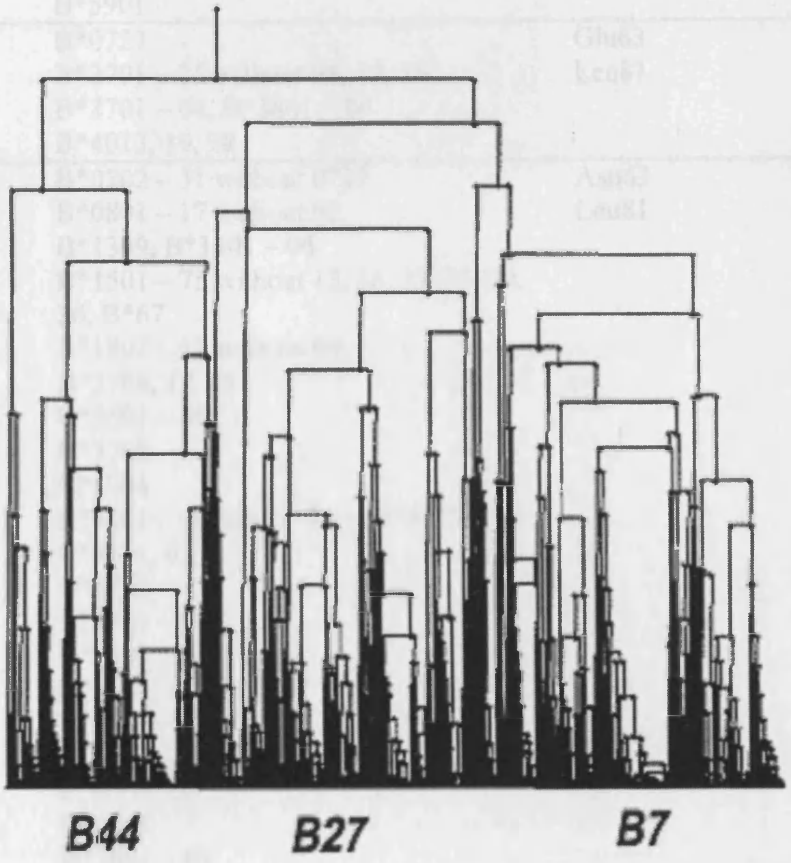
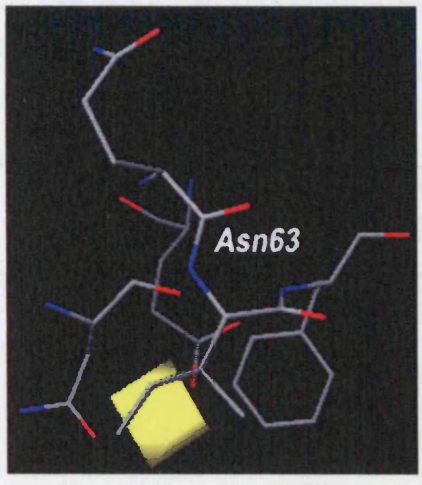


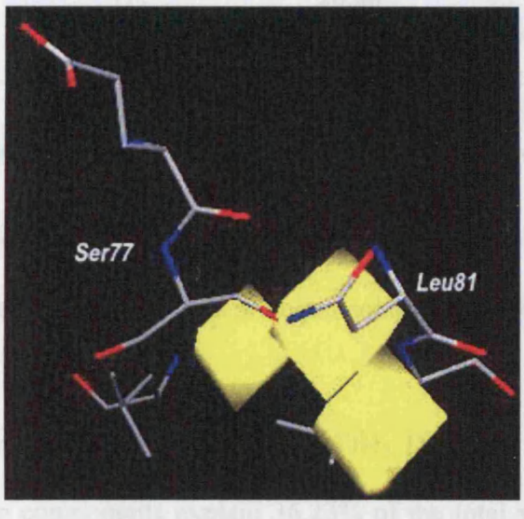
Figure 6.6. HLA clusters produced using hierarchical clustering (Doytchinova *et al.*, 2004a). A hierarchical tree was produced for the 447 HLA-B alleles. Each leaf represents one allele.

Supertype	*Consensus PCA*	*Supertype fingerprint*
B44	B*0802, B*1301 – 1311 without 09 B*1513, 16, 17, 23, 24, 36, 67 B*1809, B*3805 B*4402 – 33 without 09 B*4701 – 04 without 02 B*4901 – 03, B*5101 – 34, B*5201 – 05 B*5301 – 09, B*5607 B*5701 – 09, B*5801 – 07 B*5901	Ala81
B27	B*0727 B*2701 – 25 without 08, 12, 18 B*3701 – 04, B*3801 – 09 B*4013, 19, 28	Glu63 Leu81
B7	B*0702 – 31 without 0727 B*0801 – 17 without 02 B*1309, B*1401 – 06 B*1501 – 75 without 13, 16, 17, 23, 24, 36, B*67 B*1801 – 18 without 09 B*2708, 12, 18 B*3501 – 45 B*3705 B*3904 B*4101 – 06 4201 – 04 4409 4501 – 06 B*4601, 02 B*4702 B*4801 – 07 B*5001 – 04 B*5401, 02 B*5501 – 10 B*5601 – 11 without 5607 B*6701, 02 B*7301 B*7801 – 05 B*8101 8201, 02 8301	Asn63 Leu81

Table 6.5. A list of the HLA-B molecules in each dimension of the scores plot. For simplicity only the beginning and the end of the alleles were listed. For example, B*4901 – 03 meant that alleles B*4901, B*4902 and B*4903 were included in the cluster, etc. The amino acid used to define each cluster is in the last column.



a



b

Figure 6.7. Loading plot of the CPCA model for the HLA-B superfamilies classification. Part of the B*0801 binding site is shown in the plot. The hydrophobic interaction is found around position 63 and 66 (a), 77 and 81 (b)

6.2.4 The HLA-C classification

Like HLA-A and HLA-B, PCA models for HLA-C, using single probes, was generated. The results are in table 6.6.

	Probes	*Explained variance by the first three components of the PCA model (%)*		
		PC1	*PC2*	*PC3*
Single probe PCA model	OH_2	20.91	38.85	50.45
	Dry	29.91	45.06	54.15
	C3	23.27	36.02	47.20
	N:#	20.78	39.07	49.92
	H	22.14	34.60	45.65
	C1=	19.57	37.04	47.33
	N:=	22.75	34.98	46.01
	N1	20.55	40.61	52.06
	OH	22.90	35.72	49.69
	S1	21.77	35.82	48.78
	O1	26.06	43.52	54.62
	N2+	17.06	32.42	45.04
	O	30.99	47.34	58.08
8 probes model CPCA model	OH_2, Dry, N:#, N1, OH, S1, O1 and O	20.96	36.23	47.42

Table 6.6. Probes used in the PCA and CPCA models and the cumulative variance explained by the first three principal components (PC1, PC2 and PC3).

Eight probes were used in the CPCA model (OH_2, Dry, N:#, N1, OH, S1, O1 and O). The first two components explain 36.23% of the total variance. The scores plot of the CPCA model is in figure 6.8, in which HLA-C molecules were divided into two clusters. Cw*01, 03, 07, 08, 12 and 16 are grouped above the X axis, and Cw*02, 03, 04, 05, 06, 15, 17 and 18 are clustered below the X axis. Some of the 03, 07 and 12 are also grouped into the second cluster. The first cluster is named C1 and the second cluster is named C4. The result from hierarchical clustering gave nearly identical groups (figure 6.9), with only eight

amino acids mis-placed *0308, *0310, *0701, *0706, *0716, *0718, *1208 and *1404. The list of molecules in the C1 and C4 clusters is in table 6.7.

It should be noted that the classification is based on the scores and the loading map of CPCA models and the hierarchical clustering trees, i.e, the clusters defined must be present in both analyses. The scores map of the HLA-C CPCA model showed that there might be smaller clusters within the C1 and C4 family, such as Cw*1502, *03, *05, *06, *08, *10, *11 and *0206 on the lower left of the C4 cluster. Some of the Cw*05 and 06 molecules are clustered on the right of the C4 cluster and many Cw*08 molecules are separated from the others in the C1 cluster. These small clusters could be small families and further analysis is required to define them.

The PC2 loading plots showed that positions 70, 74, 77 and 81 of the HLA-C molecules were involved in the classification (figure 6.10). Among the HLA-C molecules, only position 77 was polymorphic. The amino acids presented at this position were Ser and Asn. The molecules in the C4 class all have Asn at position 77. The ones in the C1 cluster, on the other hand, all have serine at this position. As Asn is more polar than serine, they are more favoured for interaction with polar probes and hydrogen-bond formation.

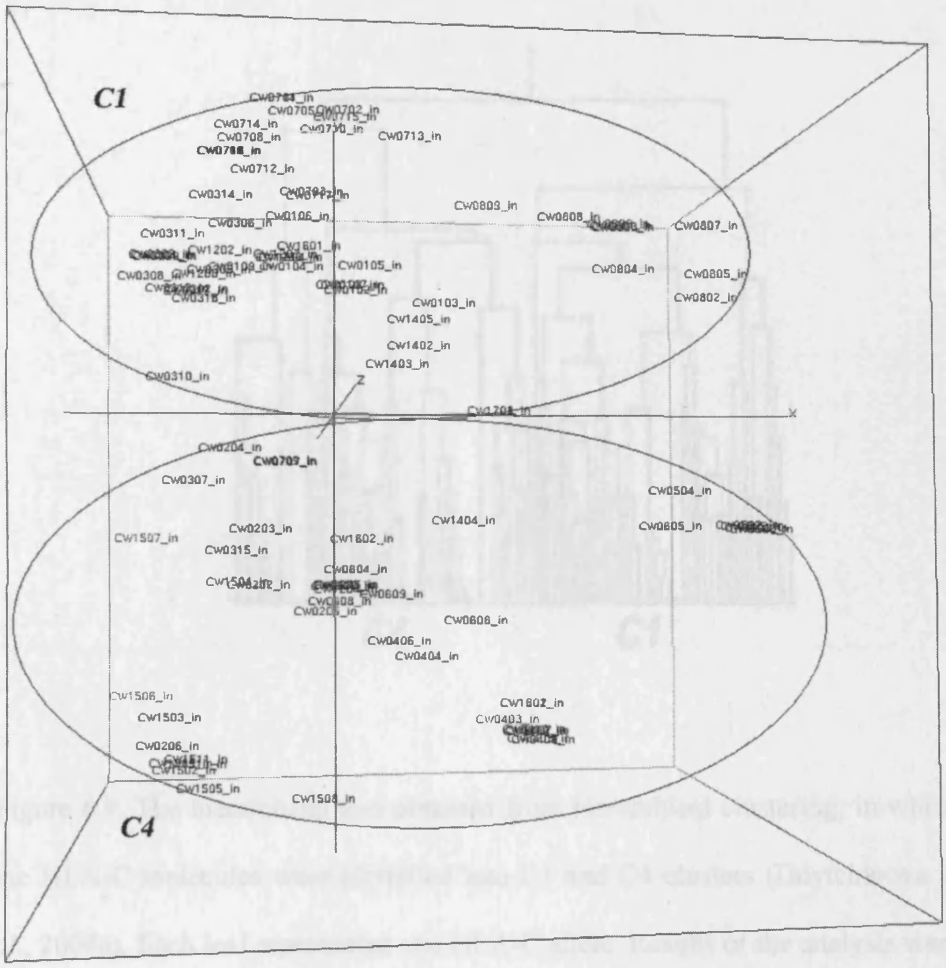


Figure 6.8. The 3D scores plot of the HLA-C CPCA analysis. Two clusters were displayed in the plot. The main cluster above the X axis had many C1 molecules and was named the C1 cluster. The cluster below the X axis had lots of C4 molecules and was named the C4 cluster.

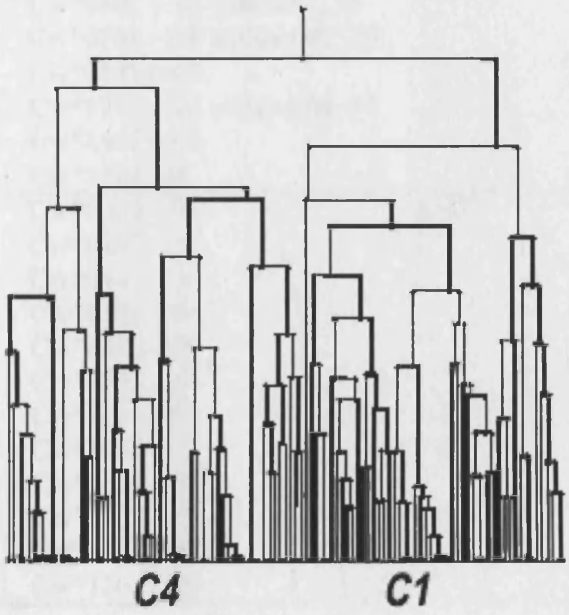


Figure 6.9. The hierarchical tree obtained from hierarchical clustering, in which the HLA-C molecules were classified into C1 and C4 clusters (Doytchinova *et al.*, 2004a). Each leaf represented one HLA-C allele. Results of the analysis were in accordance with the GRID/CPCA classification.

Supertype	*Consensus PCA*	*Supertype fingerprint*
C1	Cw*0102 – 09 Cw*0302 – 16 without 7, 15 Cw*0701 – 18 without 07, 09 Cw*0801 – 09 Cw*1202 – 08 without 04, 05 Cw*1402 – 05 Cw*1601, 04	Ser77/Gly77
C4	Cw*0202 – 06 Cw*0307, 15 Cw*0401 – 10 Cw*0501 – 06 Cw*0602 – 09 Cw*0707, 09 Cw*1204, 05 Cw*1404 Cw*1502 – 11 Cw*1602 Cw*1701 – 03 Cw*1801, 02	Asn77

Table 6.7. A list of the HLA-C molecules in each cluster. The important residues in defining the clusters were listed in the last column.

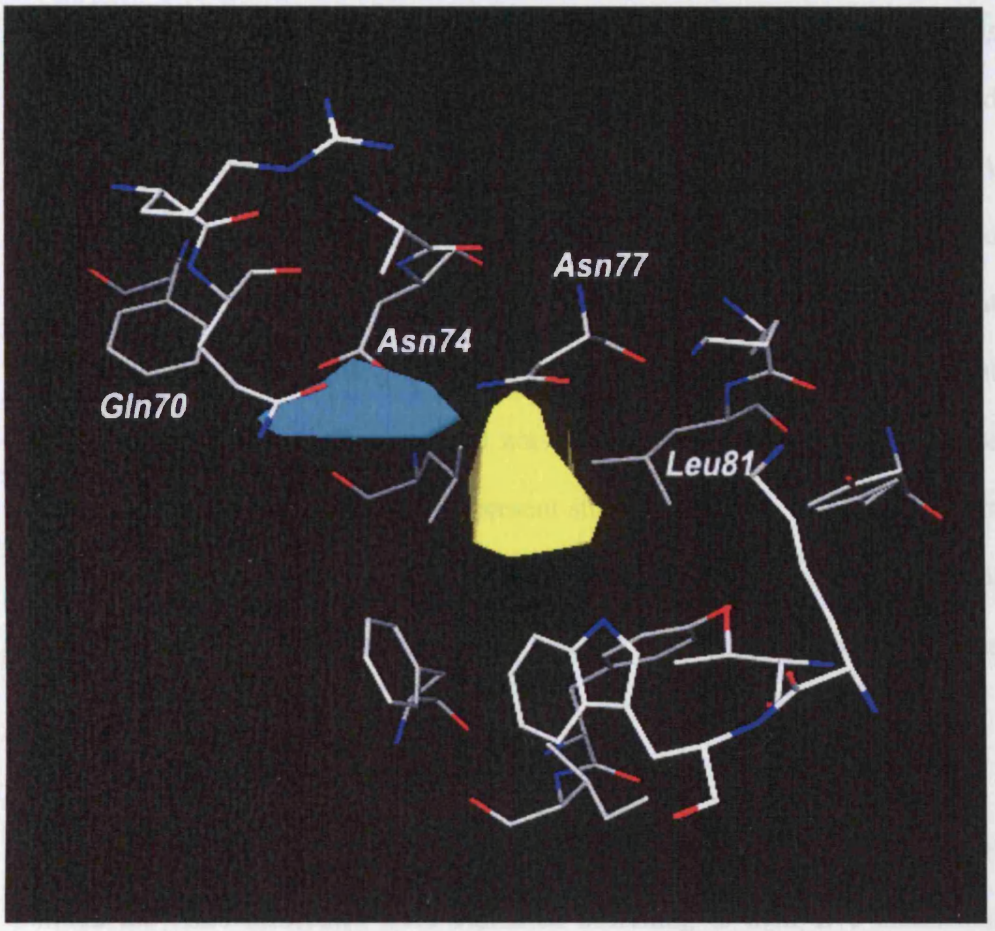


Figure 6.10. The loading plot of the HLA-C CPCA model for the water probe. The binding site of Cw*0401 is shown in the plot. The highlighted area is around position 70, 74 and 81.

6.3 Discussion

In the present project, the HLA-A, B and C molecules were classified into superfamilies using the GRID/CPCA approach: molecular interaction fields (MIFs) between different chemical probes and the HLA proteins were calculated in GRID, and these were used to build PCA and CPCA models in GOLPE. A total of 783 HLA sequences were found in the HLA database and were included in the study. The sequences were selected on the basis of the differences at protein sequence level. Many other sequences in the HLA database have silent mutations, that is, mutations that did not change the protein sequences. Those sequences were not included in the present study. The scores and loading plots were generated by the CPCA models. The scores plot was a graphical presentation of the HLA classifications, and the loading plot highlighted areas upon which the classifications were based.

The analysis was compared with hierarchical clustering using CoMSIA fields, in which the HLA molecules were classified according to their five interaction potentials: steric bulk, electrostatic potential, hydrophobicity and hydrogen donor and acceptor abilities. Although based on different molecular interactions, the two analyses gave a 77% consensus. HLA-A classification by both methods was 88% identical, some A*02, A*25, A*26, A*34, A*66 and A*68 alleles were classified into the A2 cluster by hierarchical clustering, but were in the A3 cluster in the CPCA plot. Molecules in the A24 cluster were the same in both classifications. HLA-B classification by the two methods gave a slightly lower consensus (68%), which may be because the group had the largest number of molecules among the three (447 HLA-B alleles) and the binding site consisted of

more amino acids. The classification of the cluster B27 was debatable, as most of the molecules in the B27 cluster, as defined by hierarchical clustering, were in the B7 cluster in the CPCA model. The HLA-C classification gave the best agreement using the two methods (93% consensus). Only 8 molecules were classified into different subtypes by the two methods. Molecules that have been classified into different clusters by the two methods were considered as outliers as it was not possible to classify them into clusters according the present results. They needed to be re-classified in the future using other techniques. A closer look at the protein sequence level showed that these outliers do not have significant resemblance to the classified alleles. For example, A*2501 - A*2503 alleles had Tyr at 9 and Asp at 116, which were identical as A*11 alleles, but they also had Glu at position 114 like the A*31 and A*32 alleles.

Also, the scores plot showed there may be smaller clusters within C1 and C4. For example, some of the Cw*15 molecules and Cw*0206 formed a small cluster at the left bottom of the plot. Some Cw*05 and Cw*06 molecules were separated from other C4 molecules and formed a tight cluster near the X axis. Since the HLA-C group is the least studied HLA locus at present, and no other classification has been made, it was not possible to compare the results with other studies. Therefore these small clusters remained to be confirmed by other techniques.

Based on the CPCA model, HLA-A molecules were divided into three clusters: A2, A3 and A24. The loading plot showed that the classification of HLA-A was focused on residue differences at four positions of the HLA molecule: 9, 97, 114

and 116. Molecules with small polar amino acid Ser at position 9 were clustered as the A24 superfamily and are separated from the other molecules with Tyr, Phe or Thr at this position (mainly A2 and A3 molecules). Position 9 is situated at the bottom of the pocket B, which accepts the anchor amino acid at P2. A2 and A3 peptides have aliphatic amino acids Ala, Val, Leu, Ile and Met at P2, while A24 peptides have Tyr (figure 6.11). The A2 and A3 superfamily members are further separated by the differences at position 97, which is in pocket C and E and contacts P6 and P7 of the peptide, respectively. Most of the A3 molecules accept non-polar Ile or Met at 97 (figure 6.12b), while the A2 family has Met and also the polar amino acid Arg (figure 6.12a). Some A3 molecules also have Arg97, which overlaps with the A2 family. These are separated further by differences at position 114 in the pockets D and E, and position 116 in the pocket F. A2 molecules have basic residue His114 and aromatic amino acid Tyr at position 116 (figure 6.13), while A3 molecules have acidic amino acids at both positions (Glu114 and Asp116). Tyr is a relatively large amino acid and restricts the size of the pocket, so that the pocket can only accept small hydrophobic amino acids at position 9. For A3 molecules, Asp and Glu are small therefore the pocket can hold large charged amino acids, often Arg, at P9.

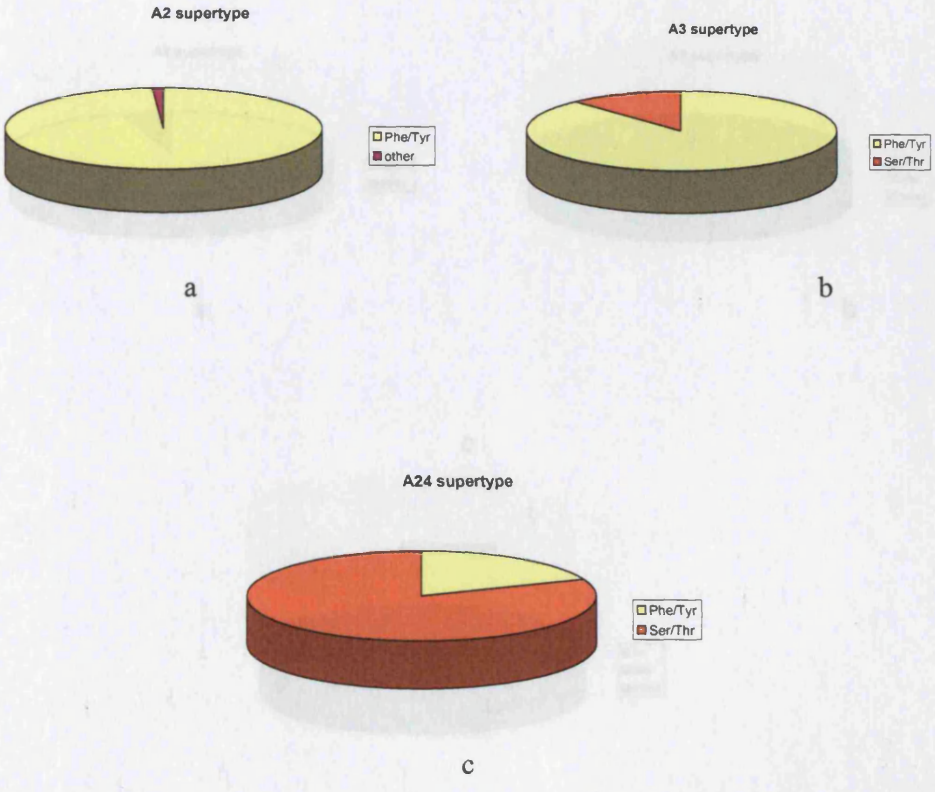


Figure 6.11. Percentage of different amino acids occupying position 9 in a) A2, b) A3 and c) A24 clusters.

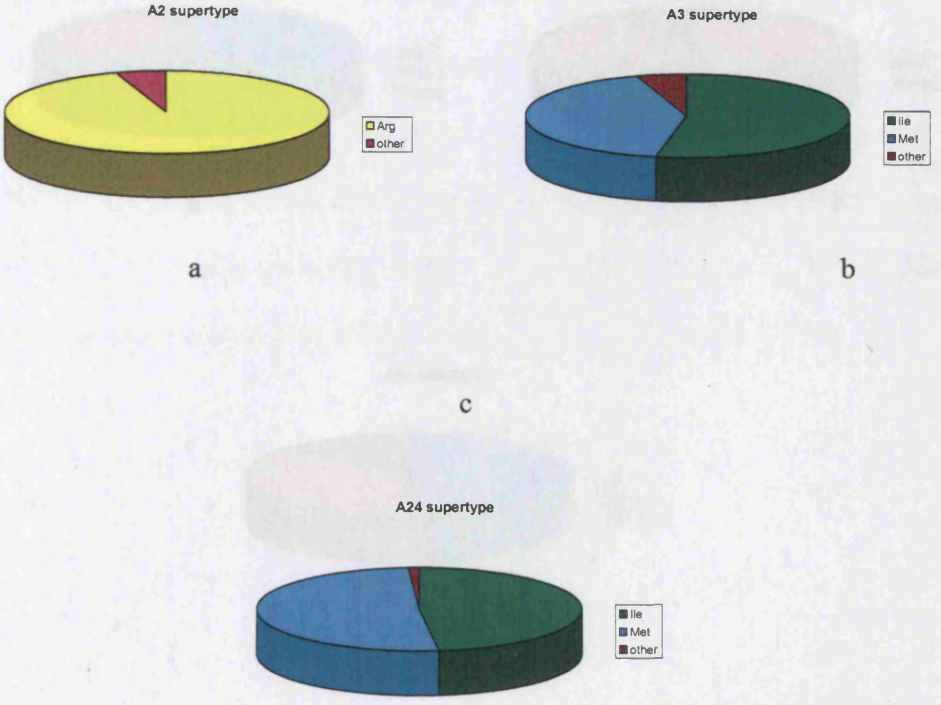


Figure 6.12. Percentage of different amino acids occupying position 97 in a) A2, b) A3 and c) A24 clusters.

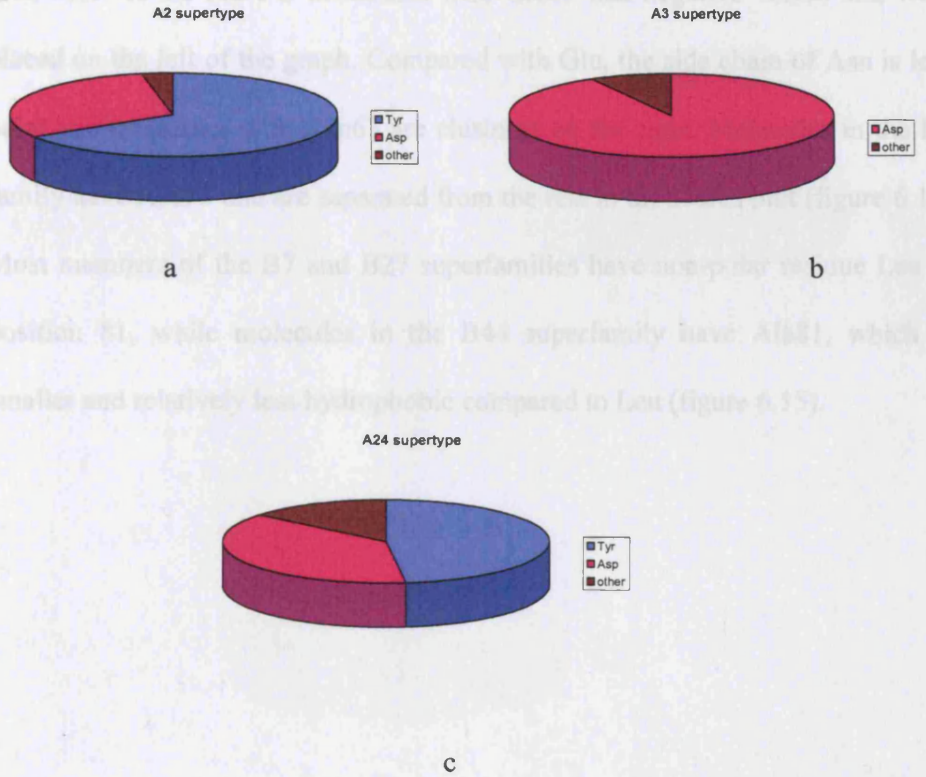


Figure 6.13. Percentage of different amino acids occupying position 116 in a) A2, b) A3 and c) A24 clusters.

HLA-B superfamilies were defined in the CPCA model scores plot: B7, B27 and B44. There were two highlighted positions - 63 and 81 - in the HLA-B loading plot, which are inside the binding pockets A and F, respectively. In the scores plot, most of the HLA-B molecules with Glu63 had negative values and were placed on the left of the graph. Compared with Glu, the side chain of Asn is less polar and molecules with Asn63 are clustered on the right. Molecules in the B7 family have Asn63 and are separated from the rest in the scores plot (figure 6.14). Most members of the B7 and B27 superfamilies have non-polar residue Leu at position 81, while molecules in the B44 superfamily have Ala81, which is smaller and relatively less hydrophobic compared to Leu (figure 6.15).

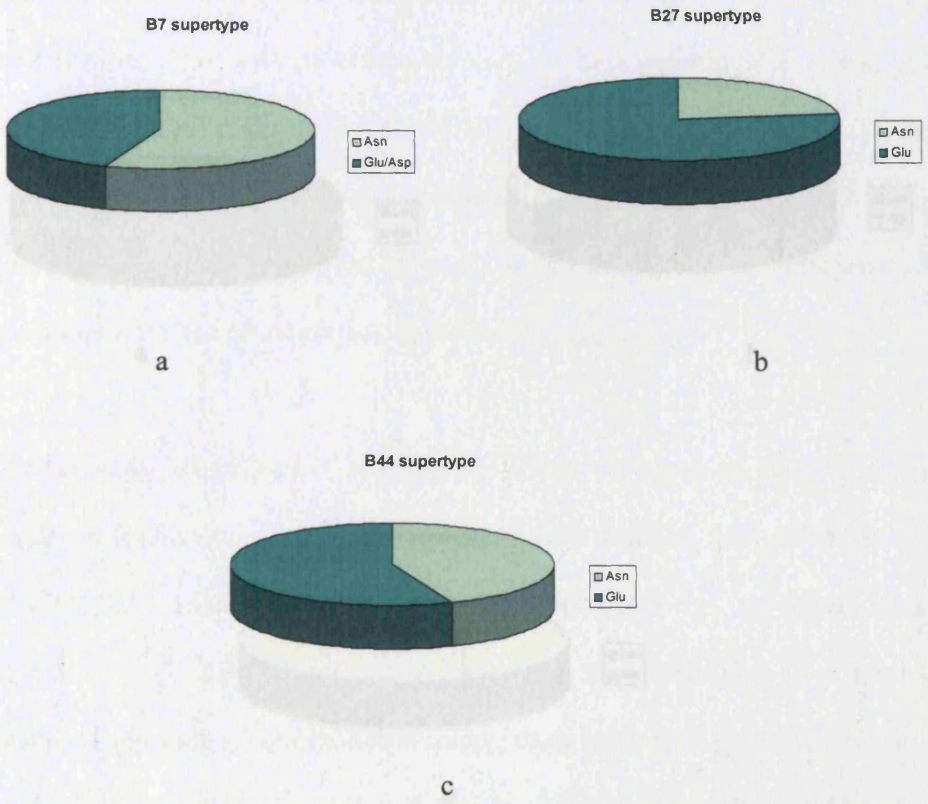


Figure 6.14. Percentage of different amino acids occupying position 63 in a) B7, b) B27 and c) B44 clusters.

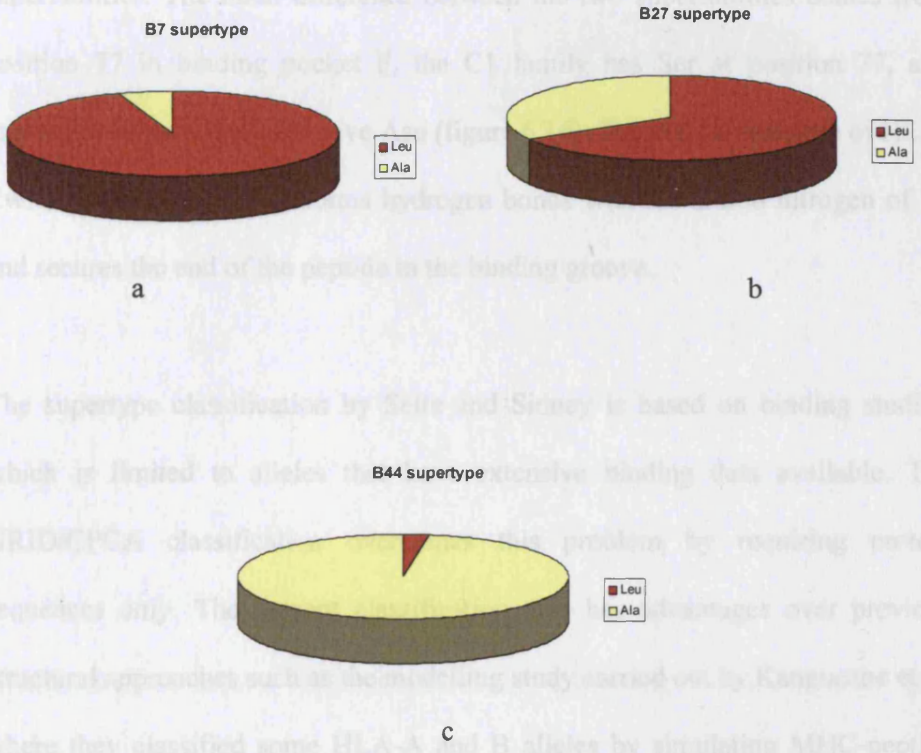
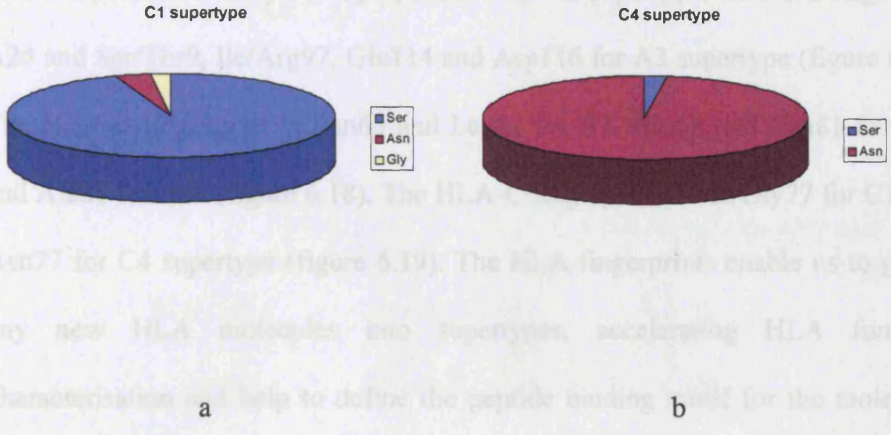


Figure 6.15. Percentage of different amino acids occupying position 81 in a) B7, b) B27 and c) B44 clusters.

The HLA-C molecules have not been as extensively studied. Only two structures of HLA-C have been published: Cw3 and Cw4 molecules co-crystallised with a human natural killer cell inhibitory receptor (Boyington *et al.*, 2000; Fan *et al.*, 2001). Compared with HLA-A and B, HLA-C has fewer molecules and a much smaller peptide binding site. In this study, the HLA-C class is separated into two superfamilies. The main difference between the two superfamilies comes from position 77 in binding pocket F, the C1 family has Ser at position 77, and molecules in the C4 cluster have Asn (figure 6.16). The crystal structure of HLA-Cw3 indicates that Ser77 forms hydrogen bonds with the amino nitrogen of P9 and secures the end of the peptide in the binding groove.

The supertype classification by Sette and Sidney is based on binding studies, which is limited to alleles that have extensive binding data available. The GRID/CPCA classification overcomes this problem by requiring protein sequences only. The present classification also has advantages over previous structural approaches such as the modelling study carried out by Kangueane et al., where they classified some HLA-A and B alleles by simulating MHC-peptide complex structures (Kangueane *et al.*, 2000). Like Sette's classification, the study by Kangueane et al. also required binding data for each allele tested, which restricted a wide application of the classification The GRID/CPCA classification, on the other hand, only requires protein sequences of the alleles and can be applied to all available alleles. The other advantage of the present classification is that the interactions between peptides and the whole binding site are considered, while some other classifications only consider interactions between motif residues and the binding site.



a b

Figure 6.16. Percentage of different amino acids occupying position 77 in a) C1, b) C4 clusters.

To conclude, the HLA-A, B and C molecules can be classified into supertypes using only their sequence information. The present classification identifies crucial, cluster determining differences at several important positions in the binding site. These positions are the HLA 'fingerprints'. The HLA-A fingerprint includes position Phe/Tyr9, Arg97, His114 for A2 supertype, Ser9 and Arg97 for A24 and Ser/Thr9, Ile/Arg97, Glu114 and Asp116 for A3 supertype (figure 6.17). The HLA-B fingerprint is Asn63 and Leu81 for B7, Glu63 and Leu81 for B27 and Ala81 for B44 (figure 6.18). The HLA-C fingerprint is Ser/Gly77 for C1 and Asn77 for C4 supertype (figure 6.19). The HLA fingerprints enable us to group any new HLA molecules into supertypes, accelerating HLA function characterisation and help to define the peptide binding motif for the molecule. Also, the HLA supertype classification allows immunologists to use similarities in sequence and structure to make educated guesses about peptide binding specificity which will help in identifying good MHC binders and potential epitopes to test.

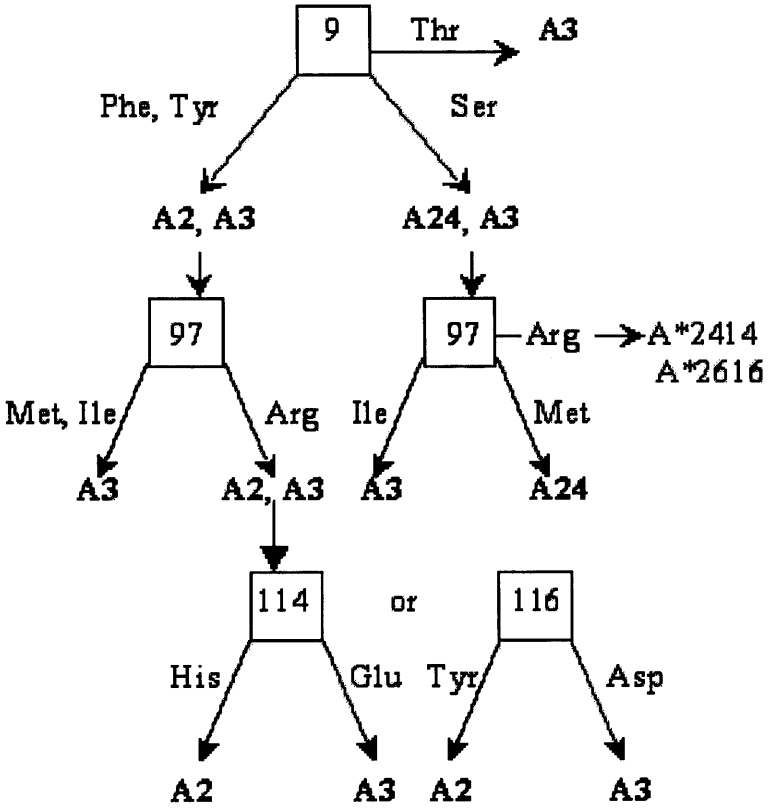


Figure 6.17. The HLA-A fingerprint.

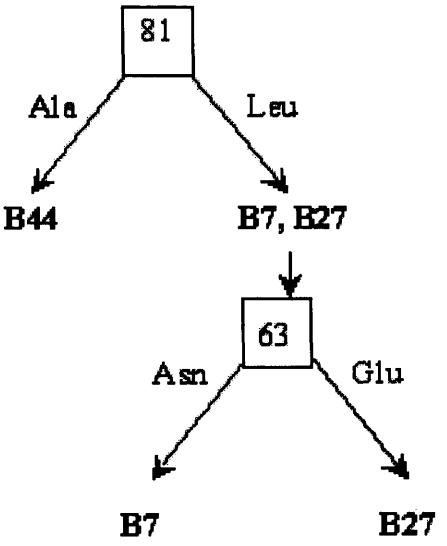


Figure 6.18. The HLA-B fingerprint.

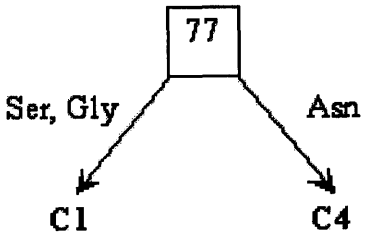


Figure 6.19. The HLA-C fingerprint.

Chapter 7
General discussion

MHC molecules are amongst the most polymorphic proteins in mammals. They recognise peptide fragments through interactions with their peptide binding site and present these fragments to T cells in order to initiate adaptive immune responses (Hauptmann and Bahram, 2004). A MHC molecule is able to recognise many different peptides. The interactions between the peptide and the binding site of the MHC molecules are not completely understood. Most research on peptide-MHC interactions focuses on peptide binding motifs obtained from mass spectrometry, pool sequencing and peptide binding assays.

Peptide binding assays have shown that most high affinity peptides for class I HLA molecules possess identical amino acids or amino acids with similar chemical properties at anchor positions 2 and 9 (Ruppert *et al.*, 1993; Sidney *et al.*, 1995; Sidney *et al.*, 1996b; Sidney *et al.*, 2001). Crystal structures of peptide-MHC complexes indicate that the amino acids at the anchor positions interact with the binding pockets inside the peptide binding site and stabilise the complex (Saper *et al.*, 1991). Apart from the anchor positions, amino acids at other positions also influence peptide binding, such as secondary anchors. These positions, together with the anchor positions, form an extended binding motif which is specific for each MHC molecule. However, the binding between peptide and MHC can not be explained by binding motifs alone and there is evidence that certain peptides may bind to MHC molecules in the absence of a binding motif (Jiang *et al.*, 2002). Therefore, peptide-MHC binding is not dependent on anchor

residues alone, but results from cumulative interactions between each amino acid and the binding site.

This thesis defined refined binding motifs for the HLA-A*0201 allele and the HLA-A3 superfamily using 2D and 3D QSAR techniques. In contrast to other motif studies, the contribution of each residue of the peptide to binding is analysed. The A*0201 binding motif was defined using different amino acid descriptors from the AAindex, and the three z and five z descriptors. Three variable selection methods SIMCA, GOLPE and GA were used to reduce the redundant variables. The HLA-A3 supermotif was defined by two QSAR techniques, the additive method (Guan *et al.*, 2003b) and CoMSIA (Guan *et al.*, 2003a). The coefficient equations generated by the additive method were used as the basis for the online T cell epitope prediction server MHCPred (Guan *et al.*, 2003c; Guan *et al.*, 2003d), and the predictivity of the additive method was compared with other T cell epitope prediction algorithms. Also, a supertype classification was carried out for all available class I HLA molecules. As the number of known HLA molecules is very large (currently 1814 alleles in the HLA/IMGT database) and increasing, it is extremely labour intensive and time consuming to characterise motifs for each allele using binding assays. The GRID/CPCA supertype classification groups all class I HLA molecules into eight supertypes based on a structural analysis of the peptide binding site (Doytchinova *et al.*, 2004a).

Previous studies observed that the A*0201 binding motif included two anchor positions P2 and P9 and two secondary anchor positions P3 and P7 (Doytchinova

and Flower, 2003; Doytchinova *et al.*, 2002; Doytchinova and Flower, 2001; Falk and Rotzschke, 1993; Falk *et al.*, 1991; Madden *et al.*, 1993; Parker *et al.*, 1992a; Ruppert *et al.*, 1993; Sidney *et al.*, 2001; Sudo *et al.*, 1995). In the 2D descriptor study (chapter 3), Leu was the most preferred residue at P2. Ile was also preferred at P2 but to a lesser degree. Similarly, hydrophobic and non-polar amino acids were favoured at P9, Leu, Ile and Ala were the most favoured residues, but Met, Val and Thr were also accepted. P3, P5 and P7 accepted aromatic residues like Phe, Trp and Tyr. In previous studies, aromatic residues were favoured at P3 and P5 and small hydrophobic residues were favoured at P7 (Doytchinova *et al.*, 2004; Falk *et al.*, 1991; Madden *et al.*, 1993; Ruppert *et al.*, 1993). The motif defined by the 2D descriptors study showed that aromatic residues could also be accepted at P7.

For the other positions, the 2D descriptor analysis identified aromatic residues, like Phe and Trp, were accepted at P1, although Cys, Gly and Met were also accepted. Serine was favoured at P4, and aliphatic residues like Ala, Leu and Ile were also accepted. Aliphatic amino acids Leu, Ile and Pro were favoured at P6. P8 accepted hydrophilic amino acids such as Ser, Lys, Asn, Glu and His.

Four alleles classified into the HLA-A3 superfamily by Sette et al. were used in the HLA-A3 supermotif analysis: A*1101, A*0301, A*3101 and A*6801 (Sette and Sidney, 1998). Two techniques were applied, the additive method (chapter 3) and CoMSIA (chapter 5). The additive method generated a matrix containing coefficients for each of the 20 amino acids at each position of the peptide, positive and negative coefficients corresponded to preferred and deleterious

amino acids, respectively. The CoMSIA models generated contour maps that displayed favoured and disfavoured molecular forces surrounding the peptide. The results of the two models correlated well with each other. In the HLA-A3 additive model, Arg was preferred by A*6801 and A*3101 at P9. In contrast, A*1101 preferred Lys and A*0301 accepted both residues. The reason for this selection may be due to the larger side chain of Arg that could extend into and form hydrogen bonds with the bottom of the pocket F. Similar results were found in the CoMSIA contour maps where hydrogen bond donor groups were favoured at P9 by A*6801 and A*3101 and disfavoured by A*1101. Both favoured and disfavoured areas have been observed in the A*0301 contour map. Steric bulk force was favoured at the other anchor position P2 in the A*0301 and A*3101 CoMSIA models, but was disfavoured in the A*1101 and A*6801 models. Similar results were found in the additive models. A*0301 and A*3101 accepted a variety of residues at P2, Ala, Ile, Leu, Thr and Ser, whereas A*6801 and A*1101 preferred residues with smaller side chains such as Ala, Val and Ile. In both CoMSIA and the additive models, aromatic residue Phe was preferred at secondary anchor position P3 and P7. In the CoMSIA model, P6 favoured hydrophilic amino acids, and in the additive models, Ser was the preferred residue at P6. For the rest of the positions, polar amino acids were favoured at P1 and P4, bulky hydrophilic amino acids were favoured at P8 by both models. The supermotif defined for the A3 superfamily is consistent with findings from other binding experiments (Falk *et al.*, 1994; Sette and Sidney, 1998; Sette and Sidney, 1999; Sidney *et al.*, 1996b; Threlkeld *et al.*, 1997).

Although based on different algorithms, both CoMSIA and the additive method were effective in characterising peptides bound to MHC molecules. The additive method was superior when predicting MHC binding peptides in a given protein sequence, as it only used protein sequences in the calculation whereas to build CoMSIA models the structures of the peptides were required therefore the calculation time was much longer. An online server - MHCPred - was able to predict T cell epitopes from protein sequences using the models generated by the additive method (Guan *et al.*, 2003c; Guan *et al.*, 2003d). A comparison of the additive method with other online T cell epitope prediction algorithms showed that the predictivity of MHCPred for human and mice epitopes was excellent, comparable with the very best of alternative servers, and that the additive method was able to correctly predict T cell epitopes within protein sequences.

One of the most obvious limitations of the present binding motif study is the availability of binding data. All the peptides and their binding affinity data used in this thesis were obtained from the AntiJen database. Although most of the original experiments followed the same protocol, factors such as different batches of antibodies and different cell lines used in the experiments and other sources of experimental errors can influence data quality and result in prediction error and inconsistency. These problems are unavoidable in most modeling exercises. Another problem is data selection bias. A large proportion of the original experiments are focused on epitope identifications, thus peptides with anchor residues and/or secondary anchor residues are much more likely to be tested. Therefore, the impact of other amino acids at these positions is not investigated. For example, in the A3 data set, only five different amino acids

appear at anchor position 2. This in particular affects the predictions by the additive method, which requires each of the 20 amino acids to be present in every position. At the moment, the influence of the 'missing' amino acids is set to zero in the additive coefficient equation. Also, most experiments tested peptides binding to several 'popular' alleles such as A*0201, A*0301 and B27, these alleles have sufficient binding data (60-200 peptides) to build reliable models while other alleles may have only 20-30 peptides or even less. The imbalance of data is reflected in the epitope prediction models. Most of the prediction algorithms produce good quality A*0201 models whereas the predictivity of other models is often much lower. An example is the additive model evaluation test (chapter 4), where the predictivity of tested algorithms was about 80% for A*0201 peptide prediction, but was much lower in predicting class II alleles. The situation may be improved in the future when more data are available.

One practical application of T cell epitope prediction is to find high affinity epitopes from viruses and bacteria and use them in vaccine development. 1814 different HLA alleles have been discovered to date (Robinson *et al.*, 2003), and HLA alleles are expressed differently in each ethnic group. Vaccines using a single epitope will not be effective in the whole population. HLA superfamilies have good phenotype frequencies across ethnic groups (table 7.1). For example, vaccines using an epitope which binds the A3 superfamily will be effective in about 50% of the global population. Vaccines using epitopes from the A2, A3 and B7 superfamily are predicted to be 90% effective in the population (Sette and Sidney, 1999).

The GRID/CPCA study grouped all class I HLA-A, B and C alleles into several supertypes. Some of the HLA-A alleles had been classified previously into supertypes. A*0201, A*0202, A*0204, A*0206 and A*0207 had been grouped into the A2 supertype by binding studies (del Guercio *et al.*, 1995; Sidney *et al.*, 1995; Sidney *et al.*, 1996a; Sidney *et al.*, 1996b; Sudo *et al.*, 1995) and motif studies (Rammensee *et al.*, 1999). All these alleles were grouped into the A2 supertype in the GRID/CPCA study with the exception of A*0204, which, like the A3 alleles, possessed Met at position 97 and was classified as belonging to the A3 family. A*0204 differed from A*0201 by having one amino acid mutation Arg -> Met at position 97. Met97 is inside pocket F. The side chain of Met97 is smaller compared with Arg, therefore increasing the volume of pocket F. However, the A*0204 binding motif (L_2L_9) was closer to A*0201, therefore it was possible that A*0204 was an outlier from the A3 superfamily. The previously classified A2 supertype also included A*6801 and A*6901, which were in the A2 superfamily in the present study.

Apart from the A2 supertype, other HLA-A supertypes are less well studied. There were three more HLA-A families in Sette's classification, the A1 superfamily (A*0101, A*2501, A*2601, A*2602 and A*3201), the A3 superfamily (A*0301, A*1101, A*3301, A*3101 and A*6801) and A24 superfamily (A*2301, A*2402, A*2403, A*2404, A*3001, A*3002, A*3003). The A1 and A3 families were grouped into the A3 superfamily in the GRID/CPCA analysis. The A*23 and A*24 alleles were in the A24 superfamily, but A*3001, A*3002 and A*3003 were placed in the A3 superfamily. The classification was also compared with the classification by Lund et al., where he

classified five HLA-A clusters (A1, A2, A3, A24, A26) using both motif information and binding site structure analysis (Lund *et al.*, 2004). The A1, A3 and A26 cluster in Lund's classification were grouped into the A3 superfamily in the present classification, and the A2 and A24 families in the two analyses were in good agreement.

HLA-B7 (B*07, B*35, B*51, B*53,B*54, B*55, B*56, B*67, B*78), B27(B*1401-02, B*1503, B*1508, B*1509, B*1510, B*1518, B*2701-08, B*3801, B*3802, B*3901-04, B*4801-02, B*7301 and B44 (B*37, B*4001-2, B*4006, B*41, B*44, B*47, B*49, B*50) families have been previously classified and tested in many binding experiments (Doolan *et al.*, 1997; Lamas *et al.*, 1998; Sidney *et al.*, 1995; Sidney *et al.*, 1996a; Sidney *et al.*, 2003). Most of the B7 alleles in Sette's classification were in the B7 cluster defined by GRID/CPCA, apart from B*51 and B*53, which were in the B44 cluster. Alleles in the B7 and B44 family of Sette's classification were scattered in B7, B27 and B44 superfamilies in the present analysis. In Sette's classification two more clusters B58 (B*1516-17, B*5701-02, B*5702, B*5708) and B62 (B*1301, B*1302, B*1501, B*1502, B*1506, B*1512-14, B*1519, B*1521, B*4601, B*4652) were defined. Molecules in the B62 cluster of Sette's classification were located in either B7 or B44 superfamilies in the GRID/CPCA analysis. The B58 cluster in Sette's classification is in the B44 cluster in the present study. Compared with Lund's classification (B7, B8, B27, B44, B58, B62), the B8 cluster was included in the B7 supertype and alleles in the in the B58 and B62 cluster were in the B7 or B27 cluster in the current analysis.

HLA supertype	*Alleles*	*Predicted phenotype frequency*				
		Asian	*Black*	*European Caucasians*	*North American Caucasians*	*Average*
A2	A*0201 A*0202 A*0203 A*0206 A*6802	42.7	40.5	50.0	51.1	46.1
A3	A*0301 A*1101 A*3101 A*3301 A*6801	56.7	51.6	48.0	47.8	51.0
B7	B*0702 B*3501 B*5101 B*5301 B*5401	43.5	55.1	51.5	52.8	50.7
A2/A3/B7		90.3	90.2	91.1	91.6	90.8
A1	A*0101 A*2601 A*2902 A*3002 A*3201	18.7	54.8	53.9	52.0	44.8
A24	A*2301 A*2402 A*2403 A*2405 A*2407	49.6	21.7	19.4	19.7	27.6
Total		100.0	100.0	100.0	100.0	100.0

Table 7.1. Phenotype frequencies of some HLA superfamilies defined by Sette, the table is adapted from Sette et al. (Sette and Sidney, 1999).

Although there is no previous HLA-C classification available for comparison, there was an interesting observation that NK cell inhibitor receptor KIR2DL was divided into two groups based on their HLA-C specificity. KIR2DL1 recognised HLA-Cw2, Cw4, Cw5 and Cw6, all of which possessed Asn77, whereas KIR2DL2 recognised HLA-Cw1, Cw3, Cw7 and Cw8, which had Ser at position 77 (Fan *et al.*, 2001). The specificity of KIR2DL was in agreement with our HLA-C classification, which suggested that position 77 was important in substrate specificity and HLA-C molecules with the same residue at position 77 tend to share the same specificity.

In the study, HLA structures were built by homology modelling using the backbone of A*0201, B*0801 and Cw*0401 as templates. Although HLA molecules are structurally similar, there may be some differences in the binding site conformation, and this can be a limitation of the technique. However, compared with HLA classifications based on peptide binding motifs, the GRID/PCA method has many advantages. GRID/CPCA is more flexible as it only requires the sequence information of molecules, therefore all the HLA molecules available, whether or not they have been studied experimentally, can be classified. In contrast, the motif-based method can only identify a small number of HLA molecules that have enough binding data. Most of the motifs include only anchor residues of the peptide, which mainly interact with binding pocket B and F, therefore only part of the peptide binding site interaction is studied in motif based classification. In contrast, the GRID/CPCA method takes the whole binding site into consideration and identifies important positions involved in the classification. Also, motif based classifications use experimental

binding data, which may be biased and contain experimental inconsistencies. On the other hand, GRID/CPCA classification uses sequence information, albeit manifest as homology modelled 3D structures, and minimise data inconsistency.

Possible future work is to apply GRID/CPCA and hierarchical clustering analysis to classify other MHC alleles such as class II HLA. Also some experimental work can be carried out to test alleles in some of the supertypes and confirm the classification. In this thesis, the binding motifs were characterised using 2D and 3D QSAR methods. In the future, QSAR methods can be used in combination with machine learning methods. Recently, support vector machines have been used together with partial least squares to generate QSAR/QSPR models (Xue *et al.*, 2004a; Xue *et al.*, 2004b), in which PLS was used to select the input variables for SVM calculation. This method can also be applied to epitope prediction. Other structural based techniques such as virtual screening and docking can also be applied. Protein degradation, transport and presentation to the T cells is a complicated process involving proteasome degradation, peptide transportation by TAP and MHC presentation. At the moment the epitope prediction is focused on peptides that bind well to the MHC molecules. In the future models of proteasome cleavage and peptides binding to the TAP molecules can be generated and these models can be used together with the additive models for T cell epitope predictions.

Chapter 8
Conclusion

To conclude, this thesis focused on the interactions between peptides and the HLA binding site. The binding motifs for the HLA-A*0201 and A3 superfamily were derived using 2D and 3D QSAR techniques. Descriptor analysis coupled with variable selection methods was applied to define the A*0201 binding motif, and two QSAR techniques, CoMSIA and the additive method, were used to define a supermotif for the HLA-A3 superfamily. The predictivity of the additive method was evaluated and the results were compared with other T cell epitope prediction algorithms. An epitope prediction server, MHCPred, was set up to facilitate online T cell epitope prediction.

The class I HLA supertype classification by GRID/CPCA defined eight supertypes, alleles within each supertypes share the same or similar binding motifs. Epitopes that are restricted to one allele can be applied to the whole family. These results can be used to identify cross-reactive epitopes that are restricted to the superfamily, which can be used in epitope based vaccine development.

MHC molecules are polymorphic proteins with most of the polymorphism in the peptide binding site. Some of the polymorphic residues are in contact with the peptide, therefore directly affect the peptide specificity of the MHC molecule, while other residues may be the result of silent mutations and are not involved with peptide binding. The GRID/CPCA classification highlighted important residues on which the classification was based. These residues are conserved

within each supertype which may come from convergent evolution. New alleles can be classified according to the differences in these residues.

Appendix

Appendix 1.

A*0201 peptides used in the 2D-QSAR analysis.

1) Peptides taken from the AntiJen database.

Peptide	pIC_{50}
AAAKAAAAV	6.398
AIAKAAAAV	6.176
AIIDPLIYA	6.623
AIYHPQQFV	6.504
ALAKAAAAA	6.947
ALAKAAAAI	6.211
ALAKAAAAL	6.511
ALAKAAAAM	7.398
ALAKAAAAV	6.597
ALCRWGLLL	7
ALIHHNTHL	6.623
ALLAGLVSL	7.117
ALLSDWLPA	7.025
ALMDKSLHV	7.767
ALMPLYACI	8
ALPYWNFAT	5.82
ALSTGLIHL	6.505
ALTVVWLLV	6.893
ALVGLFVLL	7.553
ALVLLMLPV	7.506
ALYGALLLA	8.143
AMFQDPQER	5.74
AMKADIQHV	6.777
AMLQDMAIL	7.009
AMVGAVLTA	7.122
AVAKAAAAV	6.495
AVIGALLAV	7.747
CLALSDLLV	6.447
CLTSTVQLV	6.832
DLMGYIPLV	7.097
DMWEHAFYL	6.879
DPKVKQWPL	6.176
FAFRDLCIV	6.963
FLAGALLLA	6.223
FLCWGPFFL	7.415
FLDQVPFSV	8.658
FLEPGPVTA	6.898
FLGGTPVCL	6.623
FLLLADARV	7.747
FLLPDAQSI	6.415
FLLRWEQEI	7.592
FLLSLGIHL	8.053
FLLTRILTI	8.073
FLPWHRLFL	6.95

FLWGPRALV	7.215
FLYGAALLA	8.469
FLYGALALA	8.62
FLYGALLAA	8.201
FLYGALLLA	8.585
FLYGALRLA	8.149
FLYGALVLA	7.409
FLYGGLLLA	8.959
FLYNRPLSV	7.212
FMGAGSKAV	6.2
FTDQVPFSV	7.212
FVDYNFTIV	6.62
FVNHDFTVV	6.523
FVNHRFTVV	6.523
FVTWHRYHL	5.869
FVVALIPLV	8.119
FVWLHYYSV	7.821
GIGILTVIL	6
GILTVILGV	8.342
GIRPYEILA	7.481
GLACHQLCA	6.38
GLCFFGVAL	5.38
GLFLTTEAV	7.509
GLGQVPLIV	6.301
GLIMVLSFL	7.658
GLLGNVSTV	7.62
GLLGWSPQA	8.027
GLMTAVYLV	8.051
GLQDCTMLV	7.638
GLSRYVARL	7.174
GLVDFVKHI	6.663
GLYGAQYDV	6.602
GLYLSQIAV	7.017
GLYRQWALA	6.733
GLYSSTVPV	7.577
GLYYLTTEV	7.682
GTLGIVCPI	6.666
HLAVIGALL	6.986
HLESLFTAV	5.301
HLLVGSSGL	5.792
HLYQGCQVV	6.832
HLYSHPIIL	7.131
HMWNFISGI	7.818
IAATYNFAV	7.032
IAGGVMAVV	6.708
IIDQVPFSV	7.398
IISCTCPTV	6.58
ILAGYGAGV	6.937
ILAQVPFSV	7.939
ILDEAYVMA	6.623
ILDQVPFSV	7.284
ILFTFLHLA	8.268
ILHNGAYSL	7.127
ILLLCLIFL	6.845

ILLSIARVV	6.342
ILMQVPFSV	8.125
ILSPFMPLL	7.347
ILSQVPFSV	7.699
ILSSLGLPV	7.301
ILTVILGVL	6.419
ILWQVPFSV	8.77
ILYQVPFSV	8.31
IMDQVPFSV	7.719
IMPGQEAGL	7.188
ITAQVPFSV	7.02
ITDQVPFSV	6.947
ITFQVPFSV	7.179
ITMQVPFSV	7.398
ITSQVPFSV	6.196
ITWQVPFSV	7.457
ITYQVPFSV	7.48
IVGAETFYV	8.456
IVMGNGTLV	6.001
KIFGSLAFL	7.478
KILSVFFLA	8.301
KLAGGVAVI	6.447
KLFPEVIDL	6.693
KLTPLCVTL	6.991
KTWGQYWQV	7.957
LIGNESFAL	6.38
LLACAVIHA	6.602
LLAGLVSLL	7.021
LLAQFTSAI	7.301
LLAVGATKV	6.477
LLAVLYCLL	7.478
LLCLIFLLV	6.996
LLDVPTAAV	7.77
LLFGYPVYV	7.886
LLFLGVVFL	7.301
LLFLLLADA	6.663
LLFRFMRPL	7.447
LLGCAANWI	5.301
LLGRNSFEV	6.447
LLLCLIFLL	7.585
LLLEAGALV	8.174
LLLLGLWGL	7.658
LLPLGYPFV	6.477
LLPSLFLLL	6.903
LLSCLGCKI	5.342
LLSSNLSWL	6.342
LLVFACSAV	6.342
LLVVMGTLV	5.869
LLWFHISCL	6.682
LLWQDPVPA	7.343
LLWSFQTSA	7.818
LMAVVLASL	6.954
LMIGTAAAV	7.102
LMLPGMNGI	6.623

LQTTIHDII	5.501
LTVILGVLL	5.58
LVSLLTFMI	5.716
MALLRLPLV	7.279
MLASTLTDA	6.602
MLGNAPSVV	6.644
MLGTHTMEV	7.845
MLLAVLYCL	6.478
MLQDMAILT	6.777
MMWYWGPSL	7.921
MTYAAPLFV	7.86
NLGNLNVSI	7.119
NLQSLTNLL	6
NLYVSLLLL	7.114
NMVPFFPPV	8.403
PLLPIFFCL	6.926
QLFEDNYAL	7.764
QLFHLCLII	6.886
QMTFHLFIA	5.778
QVMSLHNLV	6.025
RIWSWLLGA	7
RLLDDTPEV	7.017
RLLGSLNST	6.778
RLLQETELV	7.682
RLMIGTAAA	6.644
RLMKQDFSV	7.338
RLPLVLPAV	8.292
RLTEELNTI	6.06
RLVSGLVGA	6.818
RMFAANLGV	7.447
RMPAVTDLV	6.903
RMYGVLPWI	7.538
SAANDPIFV	5.342
SIIDPLIYA	6.342
SIISAVVGI	7.159
SLADTNSLA	6.342
SLAGFVRML	6.954
SLDDYNHLV	7.583
SLHVGTQCA	5.842
SLLEIGEGV	7.009
SLLPAIVEL	7.62
SLLTFMIAA	8.027
SLNFMGYVI	5.881
SLSRFSWGA	6.041
SLYADSPSV	7.658
SLYFGGICV	7.975
SVMDPLIYA	7.079
SVYDFFVWL	7.289
SVYVDAKLV	6.991
TLDSQVMSL	6.58
TLGIVCPIC	6.964
TLLVVMGTL	5.58
TTAEEAAGI	5.38
TVILGVLLL	6.072

TVLRFVPPL	7.114
VALVGLFVL	5.079
VCMTVDSLV	5.146
VIHAFQYVI	5.914
VILGVLLLI	6.785
VLAGLLGNV	7.721
VLAKDGTEV	7.174
VLHSFTDAI	6.17
VLIQRNPQL	7.644
VLLDYQGML	7.095
VLLLDVTPL	7.301
VLLPSLFLL	7.444
VLTALLAGL	7.086
VLVGGVLAA	6.732
VMGTLVALV	7.547
VVHFFKNIV	4.301
VVLGVVFGI	7.845
VVMGTLVAL	7.069
WILRGTSFV	6.556
WLDQVPFSV	7.939
WLEPGPVTA	6.082
WLLIDTSNA	6.447
WLSLLVPFV	8.164
WMNRLIAFA	6.914
WTDQVPFSV	6.145
YAIDLPVSV	7.801
YALTVVWLL	6.924
YLAPGPVTA	8.032
YLAPGPVTV	7.818
YLDLALMSV	8.26
YLDQVPFSV	8.638
YLEPGPVTI	7.187
YLEPGPVTL	7.058
YLEPGPVTV	7.342
YLFPGPVTA	8.495
YLFPGPVTV	8.237
YLLALRYLA	8
YLLPAIVHI	7.745
YLMPGPVTA	8.367
YLMPGPVTV	7.932
YLSEGDMAA	6.532
YLSPGPVTA	7.383
YLSPGPVTV	7.642
YLSQIAVLL	7.917
YLVAYQATV	7.304
YLVSFGVWI	8.721
YLVTRHADV	6.342
YLWPGPVTA	8.495
YLWPGPVTV	8.125
YLYPGPVTA	7.772
YLYPGPVTV	8.051
YLYVHSPAL	8.268
YMDDVVLGA	6.699
YMDDVVLGV	8.301

YMIMVKCWM	6.663
YMLDLQPET	7.373
YMNGTMSQV	7.398
YTDQVPFSV	7.066
YTYKWETFL	7.538
YVITTQHWL	6.877

2) Dr. Brusic's data set. (NPP – naturally processed peptides)

Poly-alanine	*NPP*	*T cell eiptopes*	*Non-binder*
AAAKAAAAV	ALIVGINDD	AAGIGIIQI	DSRSGSPMA
AIAKAAAAV	ALIVGLNDD	AAGIGILTV	AHKGFKGVD
ALACAAAAV	ALNELLQHV	AAPTPAAPA	AIYKQSQHM
ALADAAAAV	ALSNLEVKL	AFHHVAREL	GPGRAFVTI
ALAKAAAAA	FLLWATAEA	AIMDKNIIL	RLVTLKDIV
ALAKAAAAI	GIVPFIVSV	ALGLGLLPV	DSIGRFFGG
ALAKAAAAL	GIVPFLVSV	ALGRNSFEV	GRTQDENPV
ALAKAAAAM	GLDVLTAKV	ALMDKSLHV	PGSTAPPAH
ALAKAAAAT	GLVPFIVSV	ALMPLYACI	APRLPITGL
ALAKAAAAV	GLVPFLVSV	ALQDSGLEV	LLRRNSFEV
ALAKAAAEV	ILDQKINEV	ALSTGLIHL	RPSGPGPEL
ALAKAAAFV	ILFGHENRV	AMFQDPQER	VLASTAKAM
ALAKAAAGV	ILIDFALYL	AVGIGIAVV	NPVVHFFKN
ALAKAAALV	ILKEPVHGV	CINGVCWTV	ALAKAAAAS
ALAKAAAPV	ILMEHIHKL	CLGGLITMV	QIAKGMSYL
ALAKAAEAV	ISKKFDQSQ	CLGGLLTMV	PLERFAELV
ALAKAALAV	KINEPVIII	CLTSTVQLV	PSLKIFIAG
ALAKAANAV	KINEPVIIL	DLCGSVFLV	DIQKLVGKL
ALAKAAPAV	KINEPVILI	DLMGYIPLV	KIFIAGNSA
ALAKAAYAV	KINEPVILL	ELVSEFSRM	PGFGYGGRA
ALAKAGAAV	KINEPVLII	FAFRDLCIV	KGRGLSLSR
ALAKAIAAV	KINEPVLIL	FIDSYICQV	DYKSAHKGF
ALAKAPAAV	KINEPVLLI	FLDQVPFSV	APRLPITGI
ALAKARAAV	KINEPVLLL	FLGAAGSTM	DLAARNVLV
ALAKAYAAV	KKREEAPSL	FLGGTPVCL	RPSGPGPEI
ALAKEAAAV	KLLEPVLLL	FLKEPVHGV	FLPRHRDTG
ALAKGAAAV	KLNEPVIII	FLLLADARV	SRHKKLMFK
ALAKLAAAV	KLNEPVIIL	FLLSLGIHL	APRLPLTGL
ALAKNAAAV	KLNEPVILI	FLLTRILTI	RRIKEIVKK
ALAKYAAAV	KLNEPVILL	FLWGPRAYA	LLDDEAGPL
ALAPAAAAV	KLNEPVLII	FLYGALLLA	PPAHGVTSA
ALATAAAAV	KLNEPVLIL	GAGIGVAVL	PQKSHGRTQ
ALAVAAAAV	KLNEPVLLI	GAGIGVLTA	LLGRNSFEA
ALEKAAAAV	KLNEPVLLL	GELGFVFTL	TLQEQIGWM
ALFKAAAAV	LERPGGNEI	GIAGGLALL	ARTAHYGSL
ALKKAAAAV	LLDVPIAAV	GIGIGVLAA	EIAQRLEDV
ALMKAAAAV	LLDVPTAAV	GILGFVFTL	PLPIHTAEL
ALSKAAAAV	LLIDFALYL	GILGFVFTM	LRVEYLDDR
AMAKAAAAV	LLIENVASL	GILGFVFTV	NNMDKAVKL
ATAKAAAAV	LSKKFDQSQ	GLAPPQHLI	VFAGKNTDL
AVAKAAAAV	MVDGTLLLL	GLCTLVAML	NAPPAYEKL
FLAKAAAAV	QVCERIPTI	GLHCYEQLV	SGKDSHHPA
KLAKAAAAV	RILGAVAKV	GLIMVLSFL	TSAPDTRPA

VLAKAAAAV	SIIVRALEV	GLLGFVFTL	MVLASTAKA
	SILVRALEV	GLLGNVSTV	EGQRPGFGY
	SIPSGGIGV	GLLGWSPQA	APRLPLTGI
	SIPSGGLGV	GLQDCTMLV	KPIVQYDNF
	SLAGGIIGV	GLSRYVARL	RDTGILDSI
	SLIVRALEV	GLVPFLVSV	LRGRNSFEV
	SLLPAIVEL	GLYSSTVPV	EMFRELNEA
	SLLVRALEV	GMLGFVFTL	FGGDRGAPK
	SLPSGGIGV	GQLGFVFTL	DLMLSPDDI
	SLPSGGLGV	GTLGFVFTL	ALAAAAAAK
	STNRQSGRQ	GTLGIVCPI	KNIVTPRTP
	TLWVDPYEV	GVALQTMKQ	QGKGRGLSL
	YLLPAIVHI	GVLGFVFTL	PLIRHENRM
		HLGNVKYLV	LLGLPAAEY
		HLHQNIVDV	ENPVVHFFK
		HLLVGSSGL	PVVHFFKNI
		HLYQGCQVV	GRLTKHTKF
		HLYSHPIIL	AHGVTSAPD
		HMWNFISGI	GKGRGLSLS
		IAGIGILAI	LIKKEKVYL
		IISCTCPTV	GIGILTVIL
		IISLWDQSL	PSQGKGRGL
		ILAGYGAGV	RHGSKYLAT
		ILAKFLHWL	HGSKYLATA
		ILDSFDPLV	IFIAGNSAY
		ILFEPVHGV	NLQAYQKRM
		ILGFVFTLT	GSGKDSHHP
		ILHNGAYSL	GRERFEMFR
		ILKEYVHGV	ITFHGAKEI
		ILKSPVHGV	LLGRNSREV
		ILLLCLIFL	VLVKSPNHV
		ILSPFMPLL	APRIPITGL
		ILWEPVHGV	AYAKAAAAF
		ILYEPVHGV	KDSHHPART
		ITDQVPFSV	VHFFKNIVT
		KIFGSLAFL	ELAENREIL
		KILSVFFLA	AADKAAAAY
		KLHLYSHPI	GGDRGAPKR
		KLPQLCTEL	DLEVLMEWL
		KLTPLCVTL	SRAHSSHLK
		KLTSLCNTV	GILDSIGRF
		KTWGQYWQV	AVDLSHFLK
		LAGIGLIAA	APRIPITGI
		LIVIGILIL	NIVTPRTPP
		LLAQFTSAI	APRASRPSL
		LLARNSFEV	LAGRNSFEV
		LLCLIFLLV	AQGTLSKIF
		LLCPAGHAV	PRTPPPSQG
		LLGANSFEV	CLTFGRETV
		LLGATCMFV	ELRSRYWAI
		LLGRASFEV	HRDTGILDS
		LLGRDSFEV	APRIPLTGL
		LLGRNAFEV	YGGRASDYK
		LLGRNSAEV	LLGRNSFER

LLGRNSFAV	GSLPQKSHG
LLGRNSFEV	WGAEGQRPG
LLGRRSFEV	LIYNRMGAV
LLLCLIFLL	ASDYKSAHK
LLLLTVLTV	SLHVGTQCA
LLMDCSGSI	PAPGSTAPP
LLMGTLGIV	APRASRPSI
LLNATAIAV	DAQGTLSKI
LLNATDIAV	GDRGAPKRG
LLPENNVLS	APRIPLTGI
LLQYWSQEL	LSLSRFSWG
LLSSNLSWL	GRFFGGDRG
LLWAARPRL	YGSLPQKSH
LLWFHISCL	KRPSQRHGS
LLWTLVVLL	RTAHYGSLP
LMIIPLINV	CMGLIYNRM
LMWAKIGPV	IVTDFSVIK
LQTTIHDII	PIETVPVKL
LVVLGLLAV	LIRHENRMV
MLDLQPETT	KGVDAQGTL
MLGTHTMEV	VTPRTPPPS
MLLAVLYCL	GVTSAPDTR
MMWYWGPSL	YKSAHKGFK
NLQSLTNLL	GQRPGFGYG
NLSWLSLDV	ATDKAAAAY
NMFTPYIGV	FKGVDAQGT
PLDGEYFTL	APAAAAAAA
PLKQHFQIV	QRRRFVQNA
PLLPIFFCL	RVMAPRALL
PLSSSVPSQ	ATAKAAAAY
QAGIGILLA	ATAKAAAAK
QLFHLCLII	GLSLSRFSW
QLSLLMWIT	TVQLVTQLM
RLGRNSFEV	TQDENPVVH
RLMKQDFSV	PDTRPAPGS
RLNMFTPYI	QRHGSKYLA
RLPRIFCSC	PARTAHYGS
RLTRFLRSV	RGLSLSRFS
RVIEVLQRA	AAAAAAAAL
SLDQSVVEL	RVMAPRALI
SLFNTVATL	KLGGRDSRS
SLGGLLTMV	RPAPGSTAP
SLLLELEEV	RFSWGAEGQ
SLLMWITQC	FFGGDRGAP
SLVIVTTFV	EPRGSDIAG
SLYADSPSV	LYRKLKREI
SLYNTIAVL	SVTCTYSPA
SLYNTVATL	APPAHGVTS
SMVGNWAKV	PSQRHGSKY
STAPPAHGV	HARHGFLPR
STAPPHVNV	AEGQRPGFG
STPPPGTRV	LKAEIAQRL
SVRDRLARL	APKRGSGKD
SVYDFFVWL	APRAAAAAL

TLFIGSHVV	RHRDTGILD
TLGIVCPIC	RSGSPMARR
TLGIVVPIC	AAAAAAAAA
TLHEYMLDL	STMDHARHG
TLNAWVKVV	GSKYLATAS
VDGIGILTI	PAHGVTSAP
VIYQYMDDL	GSSEQAAEA
VLAGLLGNV	VRRCPHHER
VLFSSDFRI	ASQKRPSQR
VLHDDLLEA	RIAWARTEL
VLLDYQGML	GRGLSLSRF
VLPDVFIRC	IAGNSAYEY
VLQAGFFLL	GFGYGGRAS
VLQWASLAV	LQTTIHDII
VLSPLPSQA	IRHENRMVL
VLVKSPNHV	QKSHGRTQD
VVLGVVFGI	SQRHGSKYL
WILRGTSFV	LVMAPRTVL
WLSLLVPFV	APRAAAAAA
WLWYIKIFI	PPPSQGKGR
YIGEVLVSV	GSTAPPAHG
YLEPGPVTA	HGFLPRHRD
YLGEVIVSV	LDSIGRFFG
YLGEVLVSV	LGGRDSRSG
YLKEPVHGV	GAEGQRPGF
YLVSFGVWI	TTAEEAAGI
YMDDVVLGA	DTRPAPGST
YMDGTMSQV	EVHAADLLR
YMLDLQPET	TAHYGSLPQ
YMNGTMSQV	IFKLGGRDS
	ATASTMDHA
	ASTMDHARH
	TASTMDHAR
	SAPDTRPAP
	VAPAPAAPT
	SRSGSPMAR
	QGTLSKIFK
	AARAAAAAA
	RGAPKRGSG
	GDPNNMDKA
	RPGFGYGGR
	SQGKGRGLS
	GVDAQGTLS
	HGVTSAPDT
	TGILDSIGR
	LPRHRDTGI
	DSHHPARTA
	GTAKSVTCT
	HGRTQDENP
	RVMAPRAIL
	GRDSRSGSP
	RASDYKSAH
	QKRPSQRHG
	HYGSLPQKS

RRTEEENLR
QDENPVVHF
LATASTMDH
RVMAPRAII
SLSYSAGAL
RGSGKDSHH
PPSQGKGRG
DRGAPKRGS
HPARTAHYG
SHGRTQDEN
GAPKRGSGK
DVRLVHRDL
LPQKSHGRT
TRPAPGSTA
SWGAEGQRP
DTGILDSIG
KSHGRTQDE
GFLPRHRDT
CTTIHYNYM
RHGFLPRHR
RTPPPSQGK
GYGGRASDY
GGRASDYKS
PRHRDTGIL
PKRGSGKDS
VTSAPDTRP
HVDIRTLED
IGRFFGGDR
GGRDSRSGS
KAEIAQRLE
HKGFKGVDA
ERELVRKTR
RPSQRHGSK
TAPPAHGVT
TPPPSQGKG
DGETRKVKA
RGPGRAFVT
HHPARTAHY
LKIFIAGNS
LSRFSWGAE

Appendix 2

Peptides bound to the HLA–A3 superfamily.

*A*0301 peptide*	pIC_{50}	*A*1101 peptide*	pIC_{50}
AAFQSSMTK	7.8	AAFQSSMTK	8.292
AIAQSSMTK	7.432	AIAQSSMTK	7.921
AIFASSMTK	7.77	AIFASSMTK	8.367
AIFQASMTK	8.201	AIFQASMTK	8
AIFQCSMTK	7.854	AIFQCSMTK	8.678
AIFQRSMTR	7.432	AIFQRSMTR	8.032
AIFQSAMTK	7.886	AIFQSAMTK	8.041
AIFQSSATK	7.959	AIFQSSATK	8.367
AIFQSSMAK	7.721	AIFQSSMAK	8.167
AIFQSSMTA	5.963	AIFQSSMTA	6.417
AIFQSSMTK	8.06	AIFQSSMTK	8.174
AIFQSSMTR	7.301	AIFQSSMTR	7.796
AIFQSSMTY	7.509	AIFQSSMTY	8.301
AILQSSMTR	6.162	AILQSSMTR	7.569
ALFFIIFNK	8.036	AKFQSSMTK	5.046
ALLAVGATK	7.421	ALAETSYVK	7.785
ALNFPGSQK	8.071	ALFFIIFNK	8.658
AMFQDPQER	6.538	ALLAVGATK	7.398
AMSAARSSR	5.18	ALNFPGSQK	8.337
ANFQSSMTK	5.781	AMFQDPQER	7.921
ASFDKAKLK	7.377	ANFQSSMTK	7.658
AVDLSHFLK	6.8267	ASFDKAKLK	7.821
AYFQSSMTK	5.718	AVDLSHFLK	8.268
FIFQSSMTK	7.824	AYFQSSMTK	6.13
FLKENKLNK	6.839	FIFQSSMTK	8.387
GIFQSSMTK	8.071	FLKENKLNK	5.426
GTATLRLVK	8.161	GIFQSSMTK	8.276
GTGSGVSSK	6.9	GTGSGVSSK	7.398
GTMTTSIYK	8.469	GTMTTSIYK	8.377
GTMTTSLYK	8.469	GTMTTSLYK	8.377
GVSENIFLK	6.821	GVSENIFLK	8.301
HLDKKQRFH	7.167	IILECVYCK	6.903
HLFGYSWYK	8.658	ISEYRHYCY	6.78
IILECVYCK	5.431	IVCPICSQK	6.146
ILIKRRQQK	7.539	IVTDFSVIK	7.658
ILWKDIFHK	7.406	IVYRDGNPY	7.678
IVCPICSQK	5.954	KILSVFFLA	8.495
IVTDFSVIK	6.503	KSLYDEHIK	6.636
IVYRDGNPY	6.18	KTSERSQPR	7.028
KFYSKISEY	5.523	KVVNPLFEK	8.092
KILSVFFLA	7.745	LACAGLAYK	6.845
KIRKYTMRR	7.839	LGFGAYMSK	7.13
KLRKPKHKK	6.602	LIFCHSKKK	7.955
KTSERSQPR	7.162	LLACAGLAY	5.301
KVVNPLFEK	7.071	LLGPGRPYK	8.056
LACAGLAYK	6.374	LLGPGRPYR	6.183
LGFGAYMSK	6.867	LLIFHINGK	6.213
LIFCHSKKK	7.69	LTQDLVQEK	7.81

LIYRRRLMK	8.26	LVQEKYLEY	6.955
LLACAGLAY	6.721	MSLQRQFLR	8.081
LLGPGRPYK	8.301	QLFTFSPRR	8.097
LLGPGRPYR	6.244	QQLLRREVY	5.924
LLIFHINGK	6.321	QTNFKSLLR	7.854
LVKSPNHVK	7.64	RGDNFAVEK	6.979
MSLQRQFLR	7.268	RINEEKHEK	7.398
QLFTFSPRR	7.833	RLGVRATRK	8.194
QLVLHQILK	7.281	RTQNVLGEK	7.509
QQLLRREVY	5.301	SIFQSSMTK	8.699
QTNFKSLLR	7.301	SLFRAVITK	8.638
RINEEKHEK	6.105	SLLSTNLPY	6.452
RINGIPQQH	6.959	SLYDEHIKK	8.066
RLGVRATRK	7.932	SVLNLVIVK	8.159
RLQLSNGNR	6.39	SVMEVYDGR	8.347
SIFQSSMTK	7.921	TSYVKVLEY	6.34
SLFRAVITK	8.553	TTINFTRQR	8.229
SLLSTNLPY	5.941	TTLEQQYNK	8.222
SLYDEHIKK	7.553	VAGALVAFK	7.602
SLYGTTLEQ	5.426	VLSHNSYEK	7.721
SVLNLVIVK	6.893	VLYNTEKGR	5.501
TTLEQQYNK	6.416		
VAGALVAFK	7.339		
VLRENTSPK	7.561		
VLSHNSYEK	7.102		
VLYNTEKGR	6.55		

*A*3101 peptide*	pIC_{50}	*A*6801 peptide*	pIC_{50}
AIFASSMTK	5.313	AAFQSSMTK	7.276
AIFQASMTK	5.119	AIAQSSMTK	6.495
AIFQCSMTK	5.743	AIFASSMTK	7.06
AIFQRSMTR	7.959	AIFQASMTK	6.896
AIFQSSATK	5.407	AIFQCSMTK	7.444
AIFQSSMTK	5.038	AIFQRSMTR	7.319
AIFQSSMTR	6.539	AIFQSAMTK	6.979
AILQSSMTR	5.906	AIFQSSATK	7.367
ALFFIIFNK	6.143	AIFQSSMAK	7.018
ALLAVGATK	5.477	AIFQSSMTK	6.842
AYFQSSMTK	5.906	AIFQSSMTR	7.796
GIFQSSMTK	6.327	AILQSSMTR	5.903
GVSENIFLK	5.648	ALFFIIFNK	7.137
KILSVFFLA	5.106	ALLAVGATK	5.512
KLRKPKHKK	5.648	ANFQSSMTK	6.108
KTSERSQPR	7.176	FIFQSSMTK	8.097
LACAGLAYK	5.276	FLKENKLNK	6.353
LGFGAYMSK	5.53	GTGSGVSSK	5.459
LIFCHSKKK	5.596	GVSENIFLK	8
LLGPGRPYK	7.959	KILSVFFLA	5.051
LLGPGRPYR	7.959	KTSERSQPR	6.837
MSLQRQFLR	8.081	LACAGLAYK	7.495
QAFTSPTYK	4.745	LGFGAYMSK	6.653
QLFTFSPRR	6.207	LIFCHSKKK	6.477
QTNFKSLLR	6.745	LLACAGLAY	5.602

RINEEKHEK	6.313	LLGPGRPYK	5.804
RLGVRATRK	6.368	LLGPGRPYR	5.964
SAICSVVRR	5.921	LLIFHINGK	6.777
SLFRAVITK	5.516	MSLQRQFLR	8.071
VAGALVAFK	5.426	NVSIPWTHK	6.602
VLYNTEKGR	5.067	PVNRPIDWK	5.051
VVDFSQFSR	4.859	QAFTSPTYK	8.174
		QLFTFSPRR	8.585
		QTNFKSLLR	8.357
		RLGVRATRK	4.097
		SAICSVVRR	7.678
		SIFQSSMTK	7.62
		SLFRAVITK	7.174
		SLYDEHIKK	6.211
		VAGALVAFK	6.588
		VLSHNSYEK	5.544
		VLYNTEKGR	5.91
		VVDFSQFSR	6.616

Appendix 3

List of amino acid descriptors and references as taken from the AAindex database.

No	*Amino acid descriptors*	*Reference*
1	Hydrophobicity index	Argos, P., Rao, J.K.M. and Hargrave, P.A. Eur. J. Biochem. 128, 565-575 (1982)
2	Retention coefficient in TFA	Browne, C.A., Bennett, H.P.J., and Solomon, S. Anal. Biochem. 124, 201-208 (1982)
3	Retention coefficient in HFBA	Browne, C.A., Bennett, H.P.J., and Solomon, S. Anal. Biochem. 124, 201-208 (1982)
4	Normalized average hydrophobicity scales	Cid, H., Bunster, M., Canales, M. and Gazitua, F. Protein Engineering 5, 373-375 (1992)
5	Consensus normalized hydrophobicity scale	Eisenberg, D. Ann. Rev. Biochem. 53, 595-623 (1984)
6	Atom-based hydrophobic moment	Eisenberg, D. and McLachlan, A.D. Nature 319, 199-203 (1986)
7	Direction of hydrophobic moment	Eisenberg, D. and McLachlan, A.D. Nature 319, 199-203 (1986)
8	Hydrophobic parameter pi	Fauchere, J.L. and Pliska, V. Eur. J. Med. Chem. 18, 369-375 (1983)
9	Partition coefficient	Garel, J.P., Filliol, D., and Mandel, P. J. Chromatogr. 78, 381-391 (1973)
10	Hydrophobicity factor	Goldsack, D.E. and Chalifoux, R.C. J. Theor. Biol. 39, 645-651 (1973)
11	Hydration number	Hopfinger, A.J. "Intermolecular Interactions and Biomolecular Organizations", Wiley, New York (1977)
12	Entropy of formation	Hutchens, J.O. In "Handbook of Biochemistry", 2nd ed. (Sober, H.A., ed.), Chemical Rubber Co., Cleveland, Ohio, pp. B60-B61 (1970)
13	Hydrophobicity	Jones, D.D. J. Theor. Biol. 50, 167-183 (1975)
14	Hydropathy index	Kyte, J. and Doolittle, R.F. J. Mol. Biol. 157, 105-132 (1982)
15	Hydrophobic parameter	Levitt, M. J. Mol. Biol. 104, 59-107 (1976)
16	Average surrounding hydrophobicity	Manavalan, P. and Ponnuswamy, P.K. Nature 275, 673-674 (1978)
17	Retention coefficient in HPLC, pH7.4	Meek, J.L. Proc. Natl. Acad. Sci. USA 77, 1632-1636 (1980)
18	HPLC parameter	Parker, J.M.R., Guo, D., and Hodges, R.S. Biochemistry 25, 5425-5432 (1986)
19	Partition coefficient	Pliska, V., Schmidt, M., and Fauchere, J.L. J. Chromatogr. 216, 79-92 (1981)
20	Hydrophobicity	Prabhakaran, M. Biochem. J. 269, 691-696 (1990)
21	Side chain hydropathy,	Roseman, M.A. J. Mol. Biol. 200, 513-522 (1988)

	uncorrected for solvation	
22	Side chain hydropathy, corrected for solvation	Roseman, M.A. J. Mol. Biol. 200, 513-522 (1988)
23	Hydration potential	Wolfenden, R., Andersson, L., Cullis, P.M., and Southgate, C.C.B. Biochemistry 20, 849-855 (1981)
24	Hydrophobicity	Zimmerman, J.M., Eliezer, N., and Simha, R. J. Theor. Biol. 21, 170-201 (1968)
25	Average flexibility indices	Bhaskaran-Ponnuswamy, 1988 Int. J. Peptide Protein Res. 32, 241-255
26	Flexibility parameter for no rigid neighbors	Karplus, P.A. and Schulz, G.E. Naturwiss. 72, 212-213 (1985)
27	Flexibility parameter for one rigid neighbor	Karplus, P.A. and Schulz, G.E. Naturwiss. 72, 212-213 (1985)
28	Flexibility parameter for two rigid neighbors	Karplus, P.A. and Schulz, G.E. Naturwiss. 72, 212-213 (1985)
29	Side chain orientational preference	Rackovsky, S. and Scheraga, H.A. Proc. Natl. Acad. Sci. USA 74, 5248-5251 (1977)
30	Residue volume	Bigelow, 1967 J. Theor. Biol. 16, 187-211 (1967)
31	Apparent partial specific volume	Bull, H.B. and Breese, K. Arch. Biochem. Biophys. 161, 665-670 (1974)
32	Steric parameter	Charton, M. J. Theor. Biol. 91, 115-123 (1981)
33	The number of bonds in the longest chain	Charton, M. and Charton, B. J. Theor. Biol. 111, 447-450 (1983)
34	Partial specific volume	Cohn, E.J. and Edsall, J.T. "Protein, Amino Acid, and Peptides", Reinhold, New York (1943)
35	Size	Dawson, D.M. "The Biochemical Genetics of Man" (Brock, D.J.H. and Mayo, O., eds.), Academic Press, New York, pp.1-38 (1972)
36	Molecular weight	Fasman, G.D., ed. "Handbook of Biochemistry and Molecular Biology", 3rd ed., Proteins - Volume 1, CRC Press, Cleveland (1976)
37	Optical rotation	Fasman, G.D., ed. "Handbook of Biochemistry and Molecular Biology", 3rd ed., Proteins - Volume 1, CRC Press, Cleveland (1976)
38	Graph shape index	Fauchere, J.L., Charton, M., Kier, L.B., Verloop, A. and Pliska, V. Int. J. Peptide Protein Res. 32, 269-278 (1988)
39	Smoothed upsilon steric parameter	Fauchere, J.L., Charton, M., Kier, L.B., Verloop, A. and Pliska, V. Int. J. Peptide Protein Res. 32, 269-278 (1988)
40	Normalized van der Waals volume	Fauchere, J.L., Charton, M., Kier, L.B., Verloop, A. and Pliska, V. Int. J. Peptide Protein Res. 32, 269-278 (1988)
41	STERIMOL length of the side chain	Fauchere, J.L., Charton, M., Kier, L.B., Verloop, A. and Pliska, V. Int. J. Peptide Protein Res. 32, 269-278 (1988)
42	STERIMOL minimum width of the side chain	Fauchere, J.L., Charton, M., Kier, L.B., Verloop, A. and Pliska, V. Int. J. Peptide Protein Res. 32, 269-278 (1988)
43	STERIMOL maximum	Fauchere, J.L., Charton, M., Kier, L.B., Verloop, A.

	width of the side chain	and Pliska, V. Int. J. Peptide Protein Res. 32, 269-278 (1988)
44	Residue volume	Goldsack, D.E. and Chalifoux, R.C. J. Theor. Biol. 39, 645-651 (1973)
45	Volume	Grantham, R. Science 185, 862-864 (1974)
46	Average accessible surface area	Janin, J., Wodak, S., Levitt, M., and Maigret, B. J. Mol. Biol. 125, 357-386 (1978)
47	Side chain volume	Krigbaum, W.R. and Komoriya, A. Biochim. Biophys. Acta 576, 204-228 (1979)
48	Side chain angle theta(AAR)	Levitt, M. J. Mol. Biol. 104, 59-107 (1976)
49	Side chain torsion angle phi(AAAR)	Levitt, M. J. Mol. Biol. 104, 59-107 (1976)
50	Radius of gyration of side chain	Levitt, M. J. Mol. Biol. 104, 59-107 (1976)
51	van der Waals parameter R0	Levitt, M. J. Mol. Biol. 104, 59-107 (1976)
52	van der Waals parameter epsilon	Levitt, M. J. Mol. Biol. 104, 59-107 (1976)
53	Accessible surface area	Radzicka, A. and Wolfenden, R. Biochemistry 27, 1664-1670 (1988)
54	Bulkiness	Zimmerman, J.M., Eliezer, N., and Simha, R. J. Theor. Biol. 21, 170-201 (1968)
55	Polarizability parameter	Charton, M. and Charton, B.I. J. Theor. Biol. 99, 629-644 (1982)
56	A parameter of charge transfer capability	Charton, M. and Charton, B. J. Theor. Biol. 111, 447-450 (1983)
57	A parameter of charge transfer donor capability	Charton, M. and Charton, B. J. Theor. Biol. 111, 447-450 (1983)
58	pK-N	Fasman, G.D., ed. "Handbook of Biochemistry and Molecular Biology", 3rd ed., Proteins - Volume 1, CRC Press, Cleveland (1976)
59	pK-C	Fasman, G.D., ed. "Handbook of Biochemistry and Molecular Biology", 3rd ed., Proteins - Volume 1, CRC Press, Cleveland (1976)
60	Localized electrical effect	Fauchere, J.L., Charton, M., Kier, L.B., Verloop, A. and Pliska, V. Int. J. Peptide Protein Res. 32, 269-278 (1988)
61	Number of hydrogen bond donors	Fauchere, J.L., Charton, M., Kier, L.B., Verloop, A. and Pliska, V. Int. J. Peptide Protein Res. 32, 269-278 (1988)
62	Positive charge	Fauchere, J.L., Charton, M., Kier, L.B., Verloop, A. and Pliska, V. Int. J. Peptide Protein Res. 32, 269-278 (1988)
63	Negative charge	Fauchere, J.L., Charton, M., Kier, L.B., Verloop, A. and Pliska, V. Int. J. Peptide Protein Res. 32, 269-278 (1988)
64	pK-a(RCOOH)	Fauchere, J.L., Charton, M., Kier, L.B., Verloop, A. and Pliska, V. Int. J. Peptide Protein Res. 32, 269-278 (1988)

65	Polarity	Grantham, R. Science 185, 862-864 (1974)
66	pK (-COOH)	Jones, D.D. J. Theor. Biol. 50, 167-183 (1975)
67	Net charge	Klein, P., Kanehisa, M., and DeLisi, C. Biochim. Biophys. Acta 787, 221-226 (1984)
68	Mean polarity	Radzicka, A. and Wolfenden, R. Biochemistry 27, 1664-1670 (1988)
69	Polar requirement	Woese, C.R. Naturwiss. 60, 447-459 (1973)
70	Polarity	Zimmerman, J.M., Eliezer, N., and Simha, R. J. Theor. Biol. 21, 170-201 (1968)
71	Isoelectric point	Zimmerman, J.M., Eliezer, N., and Simha, R. J. Theor. Biol. 21, 170-201 (1968)
72	Transfer free energy to surface	Bull, H.B. and Breese, K. Arch. Biochem. Biophys. 161, 665-670 (1974)
73	Free energy of solution in water, kcal/mole	Charton, M. and Charton, B.I. J. Theor. Biol. 99, 629-644 (1982)
74	Solvation free energy	Eisenberg, D. and McLachlan, A.D. Nature 319, 199-203 (1986)
75	Melting point	Fasman, G.D., ed. "Handbook of Biochemistry and Molecular Biology", 3rd ed., Proteins - Volume 1, CRC Press, Cleveland (1976)
76	Partition energy	Guy, H.R. Biophys. J. 47, 61-70 (1985)
77	Heat capacity	Hutchens, J.O. "Handbook of Biochemistry", 2nd ed. (Sober, H.A., ed.), Chemical Rubber Co., Cleveland, Ohio, pp. B60-B61 (1970)
78	Absolute entropy	Hutchens, J.O. "Handbook of Biochemistry", 2nd ed. (Sober, H.A., ed.), Chemical Rubber Co., Cleveland, Ohio, pp. B60-B61 (1970)
79	Entropy of formation	Hutchens, J.O. "Handbook of Biochemistry", 2nd ed. (Sober, H.A., ed.), Chemical Rubber Co., Cleveland, Ohio, pp. B60-B61 (1970)
80	Transfer free energy	Janin, J. Nature 277, 491-492 (1979)
81	Transfer free energy, CHP/water	Lawson, E.Q., Sadler, A.J., Harmatz, D., Brandau, D.T., Micanovic, R. MacElroy, R.D., and Middaught, C.R. J. Biol. Chem. 259, 2910-2912 (1984)
82	Refractivity	McMeekin, T.L., Groves, M.L., and Hipp, N.J. "Amino Acids and Serum Proteins" (Stekol, J.A., ed.), American Chemical Society, Washington, D.C., p. 54 (1964)
83	Transfer energy, organic solvent/water	Nozaki, Y. and Tanford, C. J. Biol. Chem. 246, 2211-2217 (1971)
84	Transfer free energy from chx to wat	Radzicka, A. and Wolfenden, R. Biochemistry 27, 1664-1670 (1988)
85	Transfer free energy from oct to wat	Radzicka, A. and Wolfenden, R. Biochemistry 27, 1664-1670 (1988)
86	Transfer free energy	Simon, Z. Quantum Biochemistry and Specific Interactions, Abacus Press, Tunbridge Wells, Kent, England (1976)
87	Transfer free energy to lipophilic phase	von Heijne, G. and Blomberg, C. Eur. J. Biochem. 97, 175-181 (1979)

88	Principal property value z1	Wold, S., Eriksson, L., Hellberg, S., Jonsson, J., Sjostrom, M., Skagerberg, B. and Wikstrom, C. Can. J. Chem. 65, 1814-1820 (1987)
89	Principal property value z2	Wold, S., Eriksson, L., Hellberg, S., Jonsson, J., Sjostrom, M., Skagerberg, B. and Wikstrom, C. Can. J. Chem. 65, 1814-1820 (1987)
90	Principal property value z3	Wold, S., Eriksson, L., Hellberg, S., Jonsson, J., Sjostrom, M., Skagerberg, B. and Wikstrom, C. Can. J. Chem. 65, 1814-1820 (1987)
91	Unfolding Gibbs energy in water, pH7.0	Yutani, K., Ogasahara, K., Tsujita, T., and Sugino, Y. Proc. Natl. Acad. Sci. USA 84, 4441-4444 (1987)
92	Activation Gibbs energy of unfolding, pH7.0	Yutani, K., Ogasahara, K., Tsujita, T., and Sugino, Y. Proc. Natl. Acad. Sci. USA 84, 4441-4444 (1987)
93	Activation Gibbs energy of unfolding, pH9.0	Yutani, K., Ogasahara, K., Tsujita, T., and Sugino, Y. Proc. Natl. Acad. Sci. USA 84, 4441-4444 (1987)

Appendix 4

The peptide test set used in evaluation of the MHCPred server.

Protein	*Allele*	*Accession No.*	*Epitopes*	*Sequences*	*Reference*
MAGE-3	A0101	P43357	EVDPIGHLY	MPLEQRSQHCKPEEGLEARGEALGLVGAQAPATEEQEAASSSSTLV EVTLGEVPAAESPDPPQSPQGASSLPTTMNYPLWSQSYEDSSNQEEE	(Zerbini *et al.*, 2004)
	A0201		FLWGPRALV KVAELVHFL	GPSTFPDLESEFQAALSRKVAELVHFLLLKYRAREPVTKAEMLGSVV GNWQYFFPVIFSKASSSLQLVFGIELMEVDPIGHLYIFATCLGLSYDG LLGDNQIMPKAGLLIIVLAIIAREGDCAPEEKIWEELSVLEVFEGREDS ILGDPKKLLTQHFVQENYLEYRQVPGSDPACYEFLWGPRALVETSY VKVLHHMVKISGGPHISYPPLHEWVLREGEE	
CH62_MYCTU hsp65	A0201	P06806	KLAGGVAVI	AKTIAYDEEARRGLERGLNALADAVKVTLGPKGRNVVLEKKWGAPTITNDG VSIAKEIELEDPYEKIGAELVKEVAKKTDDVAGDGTTTATVLAQALVREGL RNVAAGANPLGLKRGIEKAVEKVTETLLKGAKEVETKEQIAATAAISAGDQ SIGDLIAEAMDKVGNEGVITVEESNTFGLQLELTEGMRFDKGYISGYFVTD PERQEAVLEDPYILLVSSKVSTVKDLLPLLEKVIGAGKPLLIIAEDVEGEA LSTLVVNKIRGTFKSVAVKAPGFGDRRKAMLQDMAILTGGQVISEEVGLTL ENADLSLLGKARKVVVTKDETTIVEGAGDTDAIAGRVAQIRQEIENSDSDY DREKLQERLAKLAGGVAVIKAGAATEVELKERKHRIEDAVRNAKAAVEEGI VAGGGVTLLQAAPTLDELKLEGDEATGANIVKVALEAPLKQIAFNSGLEPG VVAEKVRNLPAGHGLNAQTGVYEDLLAAGVADPVKVTRSALQNAASIAGLF LTTEAVVADKPEKEKASVPGGGDMGGMDF	(Charo *et al.*, 2001)
TRP2 Residue1-400	A0201	P40126	SVYDFFVWL	MSPLWWGFLL SCLGCKILPG AQGQFPRVCM TVDSLVNKEC CPRLGAESAN VCGSQQGRGQ CTEVRADTRP WSGPYILRNQ DDRELWPRKFFHRTCKCTGN FAGYNCGDCK FGWTGPNCER KKPPVIRQNI HSLSPQEREQ FLGALDLAKK RVHPDYVITT QHWLGLLGPN GTQPQFANCS VYDFFVWLHYYSVRDTLLGP GRPYRAIDFS HQGPAFVTWH RYHLLCLERD LQRLIGNESF ALPYWNFATG RNECDVCTDQ LFGAARPDDP TLISRNSRFS SWETVCDSLD DYNHLVTLCN GTYEGLLRRN QMGRNSMKLP TLKDIRDCLS LQKFDNPPFF QNSTFSFRNA LEGFDKADGT LDSQVMSLHN LVHSFLNGTN ALPHSAANDP IFVVLHSFTD	(Schreurs *et al.*, 2000)

Muc1_HUMAN Residue 1-420	A0201		TLAPATEPA ALGSTTPPA	MTPGTQSPFF LLLLLTVLTV VTGSGHASST PGGEKETSATQRSSVPSSTE KNAVSMTSSVLSSHSPGSGSSTTQGQDVTLAPATEPASGSAATWGQ DVTSVPVTRPALGSTTPPAHDVTSAPDNKPAPGSTAPPAHGVTS APDTRPAPGS TAPPAHGVTSAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHG VTSAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGVTSAPDT RPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGVTS APDTRPAPGS TAPPAHGVTS APDTRPAPGS TAPPAHGVTS APDTRPAPGS TAPPAHGVTSAPDTRPAPGS TAPPAHGVTS APDTRPAPGS TAPPAHGVTS APDTRPAPGS TAPPAHGVTS APDTRPAPGS APPAHGVTS	(Heukamp *et al.*, 2001)
CEA5_HUMAN	A2		ATVGIMIGV	LPVSPRLQLSNGNRTLTLFNVTRNDARAYVCGIQNSVSANRSDPVTL DVLYGPDTPIISPPDSSYLSGANLNVSCHSASNPSPQYSWRINGIPQQ HTQVLFIAKITPNNNGTYACFVSNLATGRNNSIVKSITVSASGTSPGL SAGATVGIMIGVLVGVALI	(Keogh *et al.*, 2001)
NY-ESO-1	A0201	P78358	SLLWITQC	MQAEGRGTGG STGDADGPGG PGIPDGPGGN AGGPGEAGAT GGRGPRGAGAARASGPGGGA PRGPHGGAAS GLNGCCRCGA RGPESRLLEF YLAMPFATPMEAELARRSLA QDAPPLPVPG VLLKEFTVSG NILTIRLTAA DHRQLQLSIS SCLQQLSLLM WITQCFLPVF LAQPPSGQRR	(Zeng *et al.*, 2002)
HCV NS5 Fragment Res1020-1200	A0201		CINGVCWTV	KGWRLLAPIT AYAQQTRGLL GCIITSLTGR DKNQVEGEVQ IVSTAAQTFL ATCINGVCWTVYHGAGTRTI ASPKGPVIQM YTNVDQDLVG WPAPQGSRSL TPCTCGSSDL YLVTRHADVIPVRRRGDSRG SLLSPRPISY LKGSSGGPLL CPAGHAVGIF RAAVCTRGVA KAVDFIPVEN	(Urbani *et al.*, 2001)
FETA_HUMAN Residue 1-360	A0201	P02771	PLFQVPEPV FMNKFIYEI	MKWVESIFLIFLLNFTESRTLHRNEYGIASILDSYQCTAEISLADLATI FFAQFVQEATYKEVSKMVKDALTAIEKPTGDEQSSGCLENQLPAFLE ELCHEKEILEKYGHSDCCSQSEEGRHNCFLAHKKPTPASIPLFQVPEP VTSCEAYEEDRETFMNKFIYEIARRHPFLYAPTILLWAARYDKIIPSC CKAENAVECFQTKAATVTKELRESSLLNQHACAVMKNFGTRTFQAI TVTKLSQKFTKVNFTEIQKLVLDVAHVHEHCCRGDVLDCLQDGEKI MSYICSQQDTLSNKITECCKLTTLERGQCIIHAENDEKPEGLSPNLNR FLGDRDFNQFSSGEKNIFLASFVHEYSR	(Butterfield *et al.*, 2001)
RVS	A0201	P22677	VMLRWGVLA	MALSKVKLND TFNKDQLLST SKYTIQRSTG DNIDIPNYDV	(Venter *et al.*, 2003)

NCAP_BRSV A				QKHLNKLCGM LLITEDANHKFTGLIGILYA MSRLGREDTL KILKDAGYQV RANGVDVITH RQDVNGKEMK FEVLTLVSLTSEVQGNIEIE SRKSYKKMLK EMGEVAPEYR HDSPDCGMIV LCVAALVITK LAAGDRSGLTAVIRRANNVL RNEMKRYKGL IPKDIANSFY EVIEKYPHYI DVFVHFGIAQ SSTRGGSRVEGIFAGLFMNA YGAGQVMLRWGVLAKSVKNIMLGHASVQAEMEQVVEVYEYAQK LGGEAGF YHILNNPKAS LLSLTQFPNF SSVVLGNAAG LGIMGEYRGTPRNQDLYDAA KAYAEQLKEN GVINYSVLDL TTEELEAIKN QLNPKDNDVE L	
PDC-E2	A0201	P10515	GDLLAEIETDK ATI	RYYSLPPHQK VPLPSLSPTM QAGTIARWEK KEGDKINEGD LIAEVETDKA TVGFESLEECYMAKILVAEG TRDVPIGAII CITVGKPEDI EAFKNYTLDS SAAPTPQAAP APTPAATASP PTPSAQAPGS SYPPHMQVLL PALSPTMTMG TVQRWEKKVG EKLSEGDLLA EIETDKATIG FEVQEEGYLAKILVPEGTRD VPLGTPLCII VEKEADISAF ADYRPTEVTD LKPQVPPPTP PPVAAVPPTP (Residue50-300)	(Shigematsu *et al.*, 2000)
VIE1_HCMV T RES1-360	A0201	P03169	YVLEETSVM	MESSAKRKMD PDNPDEGPSS KVPRPETPVT KATTFLQTML RKEVNSQLSL GDPLFPELAEESLKTFEQVT EDCNENPEKD VLAELVKQIK VRVDMVRHRI KEHMLKKYTQ TEEKFTGAFN MMGGCLQNAL DILDKVHEPF EEMKCIGLTM QSMYENYIVP EDKREMWMAC ELHDVSKG AANKLGGALQ AKARAKKDEL RRKMMYMCYR NIEFFTKNSA FPKTTNGCSQMAALQNLPQ CSPDEIMAYA QKIFKILDEE RDKVLTHIDH IFMDILTTCV ETMCNEYKVT SDACMMTMYGGISLLSEFCR VLCCYVLEET SVMLAKRPLI TKPEVISVMK RRIEEICMKV FAQYILGADP	(Prod'homme *et al.*, 2003)
14KD_MYCT U	A0201	P30223	LFAAFPSFA GILTVSVAV	ATTLPVQRHP RSLFPEFSEL FAAFPSFAGL RPTFDTRLMR LEDEMKEGRY EVRAELPGVDPDKDVDIMVR DGQLTIKAER TEQKDFDGRS EFAYGSFVRT VSLPVGADED DIKATYDKGI LTVSVAVSEG KPTEKHIQIR STN	(Caccamo *et al.*, 2002)
Mce2 Q7uip7	DRB1*0 101	163-175 278-290	DPIELNATLSA VA (PIELNATLS) ADLVPTATLLD TY	MPTLVTRKNR RAWLYVEGVV LLLVGALVLV LVYKQFRGE TPKTELTMVASRAGLVMEAGSKVTYNGVEI GRVGSISEIE RDGRPAAKLV LDVNPRYISL IPVNVVADIE AATLFGNKYV ALSAPKIPQQ QRISSHDVID VGSVTTEFNT LFETITSIAE KVDPIELNAT LSAVAQAPDGLGGKFGESIV NGNQILAQLN	(Panigada *et al.*, 2002)

			(DLVPTATLL)	PRLPQLGYDV RRLADLGEVY VDASPDLWSF LQNALTTART LTSQQRDLDA ALLAATGAGN TGEDVFARGG PYLARAAADL VPTATLLDTY SPELFCMIRNFHDAAPKVAD AVGGNGYSLA AAGTILGAPN PYVYPDNLPR VNAHGGPGGR PGCWQTITRE LWPAPYLVMD TGASLAPYNH VELGQPMFTE YVWGRQYGEN TINP	
Mce2 Q7uip7	DRB1*0701		EGVVLLLVGALVL (GVVLLLVGA) PRYISLIPVNVVAD (YISLIPVNV)	MPTLVTRKNR RAWLYVEGVV LLLVGALVLV LVYKQFRGE TPKTELTMVASRAGLVMEAGSKVTYNGVEI GRVGSISEIE RDGRPAAKLV LDVNPRYISL IPVNVVADIE AATLFGNKYV ALSAPKIPQQ QRISSHDVID VGSVTTEFNT LFETITSIAE KVDPIELNAT LSAVAQAPDGLGGKFGESIV NGNQILAQLN PRLPQLGYDV RRLADLGEVY VDASPDLWSF LQNALTTART LTSQQRDLDA ALLAATGAGN TGEDVFARGG PYLARAAADL VPTATLLDTY SPELFCMIRNFHDAAPKVAD AVGGNGYSLA AAGTILGAPN PYVYPDNLPR VNAHGGPGGR PGCWQTITRE LWPAPYLVMD TGASLAPYNH VELGQPMFTE YVWGRQYGEN TINP	(Panigada *et al.*, 2002)
Mage6_human	DRB1*0401	P43360	ESEFQAALSRKVAKL LLKYRAREPVTKMLGSVVGNWQ	MPLEQRSQHC KPEEGLEARG EALGLVGAQA PATEEQEAAS SSSTLVEVTL GEVPAAESPDPPQSPQGASS LPTTMNYPLW SQSYEDSSNQ EEEGPSTFPD LESEFQAALS RKVAKLVHFL LLKYRAREPV TKAEMLGSVV GNWQYFFPVI FSKASDSLQL VFGIELMEVD PIGHVYIFATCLGLSYDGLL GDNQIMPKTG FLIIILAIIA KEGDCAPEEK IWEELSVLEV FEGREDSIFG DPKKLLTQYF VQENYLEYRQ VPGSDPACYE FLWGPRALIE TSYVKVLHHM VKISGGPRISYPLLHEWALR EGEE	(Tatsumi *et al.*, 2003)
Leishmania panamensis	DRB1*0401	O43971	FKHKFAELLEQQKAAQYPSK	MATTYEEFAA KLDRLDEEFN KKMQEQNAKF FADKPDESTL SPEMKEHYEK FERMIKEHTE KFNKKMHEHS EHFKHKFAEL LEQQKAAQYP SK (Res1-200)	(Delgado *et al.*, 2003)
VS06_ROTBR	H-2Db	P04509	RLSFQLMRPPNMTP (FQLMRPPNM)	APANTQQFEH IVQLRRVLTT ATITLLPDAE RFSFPRVITS ADGATTWYFN PVILRPNNVE IEFLLNGQII NTYQARFGTI IARNFDTIRL SFQLMRPPNM TPAVAALFPN AQPFEHHATV GLTLRIESAV CESVLADASE TMLANVTSVR QEYAIPVGPV FPPGMNWTDL ITNYSPSRED NLQRVFTVAS IRSMLVK	(Choi *et al.*, 2003)
VS06_ROTBR	H-2Db	P04509	RLSFQLMRPPNMTP	APANTQQFEH IVQLRRVLTT ATITLLPDAE RFSFPRVITS ADGATTWYFN PVILRPNNVE IEFLLNGQII NTYQARFGTI	(Choi *et al.*, 2003)

			(FQLMRPPNM)	IARNFDTIRL SFQLMRPPNM TPAVAALFPN AQPFEHHATV GLTLRIESAV CESVLADASE TMLANVTSVR QEYAIPVGPV FPPGMNWTDL ITNYSPSRED NLQRVFTVAS IRSMLVK	
Ag85A	H-2Db	P17944	LTSELPGWLQA NRHVKPTGS	FSRPGLP VEYLQVPSPS MGRDIKVQFQ SGGANSPALY LLDGLRAQDD FSGWDINTPA FEWYDQSGLS VVMPVGGQSS FYSDWYQPAC GKAGCQTYKW ETFLTSELPG WLQANRHVKP TGSAVVGLSM AASSALTLAI YHPQQFVYAG AMSGLLDPSQ AMGPTLIGLAMGDAGGYKASDMWGPKEDPAWQRNDPLLNVGKLI ANNTRVWVYCGNGKPSDLGGNNLPAK FLEGFVRTSN IKFQDAYNAG GGHNGVFDFPDSGTHSWEYW GAQLNAMKPD LQRALGATPN TGPAPQGA	(D'Souza *et al.*, 2003)
Ag85B	H-2Kb	P31952	QDAYNAAGGH NAVFNFPNG (DAYNAAGG)	LTSELPQWLS ANRAVKPTGS AAIGLSMAGS SAMILAAYHP QQFIYAGSLS ALLDPSQGMG PSLIGLAMGD AGGYKAADMW GPSSDPAWER NDPTQQIPKLVANNTRLWVY CGNGTPNELG GANIPAEFLE NFVRSSNLKF QDAYNAAGGH NAVFNFPPNG THSWEYWGAQ LNAMKGDLQS SLGAG	(D'Souza *et al.*, 2003)
HCV NS3	A0201 DRB1*0 401 DRB1*0 101 (2 peptides)	P26664	KLVALGINA DECHSTDATSIL IG (STDATSILG) DLYLVTRHADVI PVR YLVTRHADV) (IIICDECHS)	KGWRLLAPITAYAQQTRGLLGCIITSLTGRDKNQVEGEVQIVSTAAQ TFLATCINGVCWTVYHGAGTRTIASPKGPVIQMYTNVDQDLVGWPA PQGSRSLTPCTCGSSDLYLVTRHADVIPVRRRGDSRGSLLSPRPISYL KGSSGGPLLCPAGHAVGIFRAAVCTRGVAKAVDFIPVENLETTMRSP VFTDNSSPPVVPQSFQVAHLHAPTGSGKSTKVPAAYAAQGYKVLVL NPSVAATLGFGAYMSKAHGIDPNIRTGVRTITTGSPITYSTYGKFLAD GGCSGGAYDIIICDECHSTDATSILGIGTVLDQAETAGARLVVLATAT PPGSVTVPHPNIEEVALSTTGEIPFYGKAIPLEVIKGGRHLIFCHSKKK CDELAAKLVALGINAVAYYRGLDVSVIPTSGDVVVVATDALMTGY TGDFDSVIDCNTCVTQ	(Wertheimer *et al.*, 2003)
Streptococcus pyogenes	H-2Db H-2Kb	Q01924	TTPQVETED SGQTTPQV	VETEDTKEP GVLMGGQSES VEFTKDTQTG MSGQTTPQVE TEDTKEPGVLMGGQSESVEFTKDTQTGMSG QTASQ	(Schulze *et al.*, 2003)
RSV G protein	H-2Db H-2Kb	Q01929	FNFVPCSICSNN PT (FVPCSICSN) PNNDFHFEVFN	MSKNKDQRTA KTLEKTWDTL NYLLFISSGL YKLNLKSIAQ ITLSILAMII STSLIITAII FIASANHKVT LTTAIIQDAT SQIKNTTPTY LTQDPQLGIS FSNLSEITSQ TTTILASTTP GVKSNLQPTT VKTKNTTTTQ TQPSKPTTKQ RQNKPPNKPN NDFHFEVFNF VPCSICSNNPTCWAICKRIP NKKPGKKTTT	(Hancock *et al.*, 2003)

			FVPC (FHFEVFNFV) (SICSNNPT)	KPTKKPTFKT TKKDLKPQTT KPKEVPTTKP TEEPTINTTK TNITTTLLTN NTTGNPKLTS QMETFHSTSS EGNLSPSQVS TTSEHPSQPS SPPNTTRQ	
Chicken OVM	H-2Kb H-2Kk	P01005	DNKTYGNKCN FCNAV D CLLCAYSIEF GTNISKEHDGE CKETVPMNCSS YANTTSEDGK VMVLCNRAFN P TDGVTYDNEC LLCAHKV	AAFGAEVDCSRFPNATDKEGKDVLVCNKDLRPICGTDGVTYTNDCL LCAY SIEFGTNISK EHDGECKETV PMNCSSYANT TSEDGKVMVL CNRAFNPVCGTDGVTYDNEC LLCAHKVEQG ASVDKRHDGG CRKELAAVSV DCSEYPKPDC TAEDRPLCGS DNKTYGNKCN FCNAVVESNG TLTLSHFGKC	(Mizumachi and Kurisaki, 2003)
SOX10 human	A0201	P56693	AWISKPPGV	PGGEAEQGGT AAIQAHYKSA HLDHRHPGEG SPMSDGNPEH PSGQSHGPPT PPTTPKTELQ SGKADPKRDG RSMGEGGKPH IDFGNVDIGE ISHEVMSNME TFDVAELDQY LPPNGHPGHV SSYSAAGYGL GSALAVASGH SAWISKPPGV ALPTVSPPGV DAKAQVKTET AGPQGPPHYT DQPSTSQIAY TSLSLPHYGS AFPSISRPQF DYSDHQPSGP YYGHSGQASG LYSAFSYMGP SQRPLYTAIS DPSPSGPQSH SPTHWEQPVY TTLSRP	(Khong and Rosenberg, 2002)
OSA1_BORBU	DRB1*0401	P14013	KSYVLEGTLTAEK (YVLEGTLTA)	MKKYLLGIGL ILALIACKQN VSSLDEKNSV SVDLPGEMNV LVSKEKNKDG KYDLIATVDKLELKGTSDKN NGSGVLEGVK ADKSKVKLTI SDDLGQTTLE VFKEDGKTLV SKKVTSKDKS STEEKFNEKG EVSEKIITRA DGTRLEYTEI KSDGSGKAKE VLKSYVLEGT LTAEKTTLVVKEGTVTLSKN ISKSGEVSVE LNDTDSSAAT KKTAAWNSGT STLTITVNSK KTKDLVFTKE NTITVQQYDS NGTKLEGSAV EITKLDEIKN ALK	(Steere *et al.*, 2003)
ALK_human Res 1200-1400	A0201	Q9UM7	GVLLWEIFSL	GGDLKSFLRE TRPRPSQPSS LAMLDLLHVA RDIACGCQYL EENHFIHRDI AARNCLLTCP GPGRVAKIGD FGMARDIYRA SYYRKGGCAM LPVKWMPPEA FMEGIFTSKT DTWSFGVLLW EIFSLGYMPY PSKSNQEVLE FVTSGGRMDP PKNCPGPVYR IMTQCWQHQP EDRPNFAIIL ERIEYCTQDP DVINTALPIE	(Passoni *et al.*, 2002)
MAG3_huma	DRB1*0	P43357	GNWQYFFPVIF	MPLEQRSQHC KPEEGLEARG EALGLVGAQA PATEEQEAAS	(Consogno *et al.*, 2003)

n	401		SKAS (QYFFPVIFS) FFPVIFSKASSS LQL (FSKASSSLQ) TSYVKVLHHM VKISG (KVLHHMVKI)	SSSTLVEVTL GEVPAAESPD PPQSPQGASS LPTTMNYPLW SQSYEDSSNQ EEEGPSTFPD LESEFQAALS RKVAELVHFL LLKYRAREPV TKAEMLGSVV GNWQYFFPVI FSKASSSLQL VFGIELMEVD PIGHLYIFAT CLGLSYDGLL GDNQIMPKAG LLIIVLAIIA REGDCAPEEK IWEELSVLEV FEGREDSILG DPKKLLTQHF VQENYLEYRQ VPGSDPACYE FLWGPRALVE TSYVKVLHHM VKISGGPHIS YPPLHEWVLR EGEE	
Tyrosinase related protein-1 Residue 1-400	DRB1*0401	P17643	ISPNSVFSQWR VVCDSLEDYD ()	MSAPKLLSLG CIFFPLLLFQ QARAQFPRQC ATVEALRSGM CCPDLSPVSG PGTDRCGSSS GRGRCEAVTA DSRPHSPQYP HDGRDDREVW PLRFFNRTCH CNGNFSGHNC GTCRPGWRGA ACDQRVLIVR RNLLDLSKEE KNHFVRALDM AKRTTHPLFV IATRRSEEIL GPDGNTPQFE NISIYNYFVW THYYSVKKTF LGVGQESFGE VDFSHEGPAF LTWHRYHLLR LEKDMQEMLQ EPSFSLPYWN FATGKNVCDI CTDDLMGSRS NFDSTLISPN SVFSQWRVVC DSLEDYDTLG TLCNSTEDGP IRRNPAGNVA RPMVQRLPEP QDVAQCLEVG LFDTPPFYSN STNSFRNTVE GYSDPTGKYD PAVRSLHNLA HLFLNGTGGQ THLSPNDPIF	(Touloukian *et al.*, 2002)
CGHB_HUMAN	A*0201 0701	P01233	VLQVGLPAL TMTRVLQGV LPQVVCNYRD VRFESI (QVVCNYRDV)	MEMFQGLLLL LLLSMGGTWA SKEPLRPRCR PINATLAVEK EGCPVCITVN TTICAGYCPT MTRVLQGVLP ALPQVVCNYR DVRFESIRLP GCPRGVNPVV SYAVALSCQC ALCRRSTTDC GGPKDHPLTC DDPRFQDSSS SKAPPPSLPS PSRLPGPSDT PILPQ	(Dangles *et al.*, 2002)
CEA5_human Residue 300-700	0401 0701	P06731	YACFVSNLATG RNNS	AHNSDTGLNR TTVTTITVYA EPPKPFITSN NSNPVEDEDA VALTCEPEIQ NTTYLWWVNN QSLPVSPRLQ LSNDNRTLTL LSVTRNDVGP YECGIQNELS VDHSDPVILN VLYGPDDPTI SPSYTYYRPG VNLSLSCHAA SNPPAQYSWL IDGNIQQHTQ ELFISNITEK NSGLYTCQAN NSASGHSRTT VKTITVSAEL PKPSISSNNS KPVEDKDAVA FTCEPEAQNT TYLWWVNGQS LPVSPRLQLS NGNRTLTLFN VTRNDARAYV CGIQNSVSAN RSDPVTLDVL YGPDTPIISP PDSSYLSGAN LNLSCHSASN PSPQYSWRIN GIPQQHTQVL FIAKITPNNN GTYACFVSNL ATGRNNSIVK SITVSASGTS PGLSAGATVG IMIGVLVGVA LI	(Kubayashi *et al.*, 2002)
MAGE_6	0101 0701	P43360	ESEFQAALSRK VAKL	MPLEQRSQHC KPEEGLEARG EALGLVGAQA PATEEQEAAS SSSTLVEVTL GEVPAAESPD PPQSPQGASS LPTTMNYPLW	(Tatsumi *et al.*, 2003)

	(the same epitopes)		YFFPVIFSKASD SLQL	SQSYEDSSNQ EEEGPSTFPD LESEFQAALS RKVAKLVHFL LLKYRAREPV TKAEMLGSVV GNWQYFFPVI FSKASDSLQL VFGIELMEVD PIGHVYIFAT CLGLSYDGLL GDNQIMPKTG FLIIILAIIA KEGDCAPEEK IWEELSVLEV FEGREDSIFG DPKKLLTQYF VQENYLEYRQ VPGSDPACYE FLWGPRALIE TSYVKVLHHM VKISGGPRISYPLLHEWALR EGEE	
MOG 35-55	H2Db	44-53	FSRVVHLYRN	MEVGWYRSPFSRVVHLYRNGK	(Sun *et al.*, 2003)
P.vivax	0401	M60807	NFVGKFLELQI PGHTDLLHL (FVGKFLELQ) LDMLKKVVLG LWKPLDNIKD (DMLKKVVL)	MKALLFFFSFIFFVTKCQCETESYKQLVANVDKLEALVVDGYEL FHKKKLGENDIKVDANANNNNNNQVSVLTSKIRNFVGKFLELQIPGHTDLL HLIRELAFEPNGIKYLVESYEEFNQLMHVINFHYDLLRANVHDMCAHDYCK IPEHLKISDKELDMLKKVVLGLWKPLDNIKDDIGKLETFITKNKETISNIN KLISDENAKRGGQSTNTTNGPGAQNNAAQGSTGNTETGTRSSASSNTLSGG DGTTVVGTSSPAPAAPSSTNEDYDEKKKIYQAMYNGIFYTSQLEEAQKLIE VLEKRVKVLKEHKGIKALLEQVEAEKKKLPKDNTTNRPLTDEQQKAAQKKI ADLESQIVANAKTVNFDLDGLFTDAEELEYYLREKAKMAGTL	(Caro-Aguilar *et al.*, 2002)
OMLK_CHLPN	H2Kb	Q9X8F4	GDYVFDRI	AGDPCDPCAT WCDAISLRAG FYGDYVFDRI LKVDAPKTFS MGAKPTGSAT ANYTTAVDRP NPAYNKHLHD AEWFTNAGFI ALNIWDRFDV FCTLGASNGY IKGNSTAFNL VGLFGVKGTS VAANELPNVS LSNGVVELYT DTSFSWSVGA RGALWECGCA TLGAEFQYAQ SKPKVEELNV ICNVAQFSVN KPKGYKGVAF PLPTDAGVAT ATGTKSATIN YHEWQVGASL SYRLNSLVPY IGVQWSRATF DADNIRIAQP KLPTAVLNLT AWNPSLLGNT TTLPTSDSFS DFMQIVSCQI NKFKSRKACG VTVGATLVDA DKWSLTAEAR LIN	(Saren *et al.*, 2002)
Vaccinia virus	A0201		KVDDTFYYV	MGIQHEFDIIINGDIALRNLQLHKGDNYGCKLKIISNDYKKLKF RFIIRPDWSEIDEVKGLTVFANNYAVKVNKVDDTFYYVIYEAVIHLY NKKTEILIYSDDENELFKHYYPYISLNMISKKYKVKEENYSSPYIEHP LIPYRDYESMD	(Terajima *et al.*, 2003)
Vaccinia virus	A0201		CLTEYILWV	MKPKVNNIGNTPLHNYVSQYDITLIPHPQPIKKWKLKPSISINGYRST FTMAFPCAQFRPCHCHATKDSLNTVADVRHCLTEYILWVSHRWTH RESAGSLYRLLISFRTDATELFGGELKDSLPWRSLNDSMKTAEELRAI IGLCTQSAIVSGRVFNDKYIDILLMLRKILNENDYLTLLDHIRTAKY	(Terajima *et al.*, 2003)
MUC1_human Residue 1-300	A0201	P15941	TLAPATEPA	MTPGTQSPFF LLLLLTVLTV VTGSGHASST PGGEKETSAT QRSSVPSSTE KNAVSMTSSV LSSHSPGSGS STTQGQDVTL APATEPASGS AATWGQDVTS VPVTRPALGS TTPPAHDVTS	(Heukamp *et al.*, 2001)

				APDNKPAPGS TAPPAHGVTS APDTRPAPGS TAPPAHGVTS APDTRPAPGS TAPPAHGVTS APDTRPAPGS TAPPAHGVTS APDTRPAPGS TAPPAHGVTS APDTRPAPGS TAPPAHGVTS APDTRPAPGS TAPPAHGVTS APDTRPAPGS TAPPAHGVTS APDTRPAPGS TAPPAHGVTS	
MUC1_human Residue 900-1100	A0201	P15941	ALGSTAPPV	APDTRPAPGS TAPPAHGVTS APDTRPAPGS TAPPAHGVTS APDNRPALGS TAPPVHNVTS ASGSASGSAS TLVHNGTSAR ATTTPASKST PFSIPSHHSD TPTTLASHST KTDASSTHHS SVPPLTSSNH STSPQLSTGV SFFFLSFHIS NLQFNSSLED PSTDYYQELQ RDISEMFLQI YKQGGFLGLS NIKFRPGSVV	(Heukamp *et al.*, 2001)
H-2Kb a1 domain	H2Db		QEGPEYWERETQK	SDAENPRYEP RARWMEQEGP EYWERETQKA KGNEQSFRVE LRTLLGYYNQSKGGSHTIQV ISGCEVGSDG RLLRG	(Honjo *et al.*, 2000)
GAG_HV1A2 Res.300-501	A0201	P03349	VLAEAMSQV EMMTACQGV	FYKTLRAEQA SQDVKNWMTE TLLVQNANPD CKTILKALGP AATLEEMMTA CQGVGGPGHK ARVLAEAMSQ VTNPANIMMQ RGNFRNQRKT VKCFNCGKEG HIAKNCRAPR KKGCWRCGRE GHQMKDCTER QANFLGKIWP SYKGRPGNFL QSRPEPTAPP EESFRFGEEK TTPSQKQEPI DKELYPLTSL RSLFGNDPSS Q	(Corbet *et al.*, 2003)
env_HV1A2 Res.500-700	A0201	P03378	YIKIFIMIV	RRVVQREKRA VGIVGAMFLG FLGAAGSTMG AVSLTLTVQA RQLLSGIVQQ QNNLLRAIEA QQHLLQLTVW GIKQLQARVL AVERYLRDQQ LLGIWGCSGK LICTTAVPWN ASWSNKSLED IWDNMTWMQW EREIDNYTNT IYTLLEESQN QQEKNEQELL ELDKWASLWN WFSITNWLWY IKIFIMIVGG LVGLRIVFAV	(Corbet *et al.*, 2003)
VIF_HV1A2	A0201	P03402	ALAALITPK	MENRWQVMIV WQVDRMRIRT WKSLVKHHMY ISKKAKGWFY RHHYESTHPR VSSEVHIPLG DAKLVITTYW GLHTGEREWH LGQGVAIEWR KKKYSTQVDP GLADQLIHLH YFDCFSESAI KNAILGYRVS PRCEYQAGHN KVGSLQYLAL AALITPKKTK PPLPSVKKLT EDRWNKPQKT KGHRGSHTMN GH	(Corbet *et al.*, 2003)
POLG_HCVBK Residue 600-800	A0201	P26663	RLWHYPCTV	PWLTPRCMVD YPYRLWHYPC TVNFTIFKVR MYVGGVEHRL NAACNWTRGE RCDLEDRDRP ELSPLLLSTT EWQVLPCSFT TLPALSTGLI HLHQNIVDVQ YLYGIGSAVV SFAIKWEYVL LLFLLLADAR VCACLWMMLL IAQAEAALEN LVVLNSASVA GAHGILSFLV FFCAAWYIKG RLVPGATYAL YGVWPLLLLL	(Sarobe *et al.*, 2001)
CEA5_HUMAN Residue 600-	A0201	P06731	YLSGANLNL VLYGPDTPI	LPVSPRLQLSNGNRTLTLFNVTRNDARAYVCGIQNSVSANRSDPVTL DVLYGPDTPIISPPDSSYLSGANLNVSCHSASNPSPQYSWRINGIPQQ HTQVLFIAKITPNNNGTYACFVSNLATGRNNSIVKSITVSASGTSPGL	(Keogh *et al.*, 2001)

700				SAGATVGIMIGVLVGVALI	
ERB2_human (Her2/neu) Res. 1-450	A0201	P04626	KIFGSLAFL	MELAALCRWG LLLALLPPGA ASTQVCTGTD MKLRLPASPE THLDMLRHLY QGCQVVQGNL ELTYLPTNAS LSFLQDIQEV QGYVLIAHNQ VRQVPLQRLR IVRGTQLFED NYALAVLDNG DPLNNTTPVT GASPGGLREL QLRSLTEILK GGVLIQRNPQ LCYQDTILWK DIFHKNNQLA LTLIDTNRSR ACHPCSPMCK GSRCWGESSE DCQSLTRTVC AGGCARCKGP LPTDCCHEQC AAGCTGPKHS DCLACLHFNH SGICELHCPA LVTYNTDTFE SMPNPEGRYT FGASCVTACP YNYLSTDVGS CTLVCPLHNQ EVTAEDGTQR CEKCSKPCAR VCYGLGMEHL REVRAVTSAN IQEFAGCKKI FGSLAFLPES FDGDPASNTA PLQPEQLQVF ETLEEITGYL YISAWPDSLP DLSVFQNLQV IRGRILHNGA YSLTLQGLGI	(Keogh *et al.*, 2001)
FGF5	A0301	AF535149	NTYASPRFK	K F R E R F Q E N S Y N T Y A S P R F K	(Hanada *et al.*, 2004)
LPPX_MYCTU	DRB1*0401	P96286	SARPATVWIAQ DGSHHLVRASI DLGSGSIQ (TVWIAQDGS)	MNDGKRAVTS AVLVVLGACL ALWLSGCSSP KPDAEEQGVP VSPTASDPAL LAEIRQSLDATKGLTSVHVA VRTTGKVDSL LGITSADVDV RANPLAAKGV CTYNDEQGVP FRVQGDNISV KLFDDWSNLG SISELSTSRV LDPAAGVTQL LSGVTNLQAQ GTEVIDGIST TKITGTIPAS SVKMLDPGAK SARPATVWIA QDGSHHLVRA SIDLGSGSIQ LTQSKWNEPV NVD	(Al-Attiyah and Mustafa, 2004)
Major pollen allergen Art v 1	0101	Q84ZX5	CDKKCIEWEK AQHGA (CIEWEKAQK)	MAKCSYVFCA VLLIFIVAIG EMEAAGSKLC EKTSKTYSGK CDNKKCDKKC IEWEKAQHGA CHKREAGKES CFCYFDCSKS PPGATPAPPG AAPPPAAGGS PSPPADGGSP PPPADGGSPP VDGGSPPPPS TH	(Jahn-Schmid *et al.*, 2002)
ALL2_DERPT	H-kb	P49278	CHGSEPCIIHRG KPF (SEPCIIHRG)	MMYKILCLSL LVAAVARDQV DVKDCANHEI KKVLVPGCHG SEPCIIHRGKPFQLEAVFEA NQNTKTAKIE IKASIDGLEV DVPGIDPNAC HYMKCPLVKGQQYDIKYTWN VPKIAPKSEN VVVTVKVMGD DGVLACAIAT HAKIRD	(Wu *et al.*, 2002a)
IAPP_HUMAN	A0201	P10997	KLQVFLIVL	MGILKLQVFLIVLSVALNHLKATPIESHQVEKRKCNTATCATQRLAN FLV HSSNNFGAILSSTNVGSNTY GKRNAVEVLKREPLNYLPL	(Panagiotopoulos *et al.*, 2003)
DMD_HUMAN (1000-1300)	A0201	P11532	WLNEVEFKL	TTVKEMSKKA PSEISRKYQSEFEEIEGRWK KLSSQLVEHC QKLEEQMNKL RKIQNHIQTL KKWMAEVDVF LKEEWPALGD SEILKKQLKQ CRLLVSDIQT IQPSLNSVNE GGQKIKNEAE PEFASRLETE LKELNTQWDHMCQQVYARKE ALKGGLEKTV SLQKDLSEMH EWMTQAEEEY LERDFEYKTP DELQKAVEEM	(Ginhoux *et al.*, 2003)

				KRAKEEAQQK EAKVKLLTES VNSVIAQAPP VAQEALKKEL ETLTTNYQWL CTRLNGKCKTLEEVWACWHE LLSYLEKANK WLNEVEFKLK TTENIPGGAE	
CTG1_HUMAN	DRB1*0401	P78358	QDAPPLPVPGVL LKEFTVSGNILT IRLTAA DHR	MQAEGRGTGG STGDADGPGG PGIPDGPGGN AGGPGEAGAT GGRGPRGAGA ARASGPGGGAPRGPHGGAAS GLNGCCRCGA RGPESRLLEF YLAMPFATPM EAELARRSLA QDAPPLPVPG VLLKEFTVSG NILTIRLTAA DHRQLQLSIS SCLQQLSLLM WITQCFLPVF LAQPPSGQRR	(Zarour *et al.*, 2002)
POLG_HCV1 1200-1500	A1	P26664	ATDALMTGY	LETTMRSPVF TDNSSPPVVP QSFQVAHLHA PTGSGKSTKV PAAYAAQGYK VLVLNPSVAA TLGFGAYMSK AHGIDPNIRT GVRTITTGSP ITYSTYGKFL ADGGCSGGAY DIIICDECHS TDATSILGIG TVLDQAETAG ARLVVLATAT PPGSVTVPHP NIEEVALSTT GEIPFYGKAI PLEVIKGGRH LIFCHSKKKC DELAAKLVAL GINAVAYYRG LDVSVIPTSG DVVVVATDAL MTGYTGDFDS VIDCNTCVTQ TVDFSLDPTF TIETITLPQD AVSRTQRRGR TGRGKPGIYR	(Lauer *et al.*, 2002)
POLG_TMEVB 200-400 VP2	H-2Db H-2Kb	P08544	Db(RVQVQCNA SQFHAGSLLVF M) (RVQVQCNA) Kb(TGYRYDSR T) (TGYRYDSR)	DKVLAAERYY TIDLASWTTS QEAFSHIRIP LPHVLAGEDG GVFGATLRRH YLCKTGWRVQ VQCNASQFHA GSLLVFMAPE FYTGKGTKTG TMEPSDPFTM DTEWRSPQGA PTGYRYDSRT GFFATNHQNQ WQWTVYPHQI LNLRTNTTVD LEVPYVNVAP SSSWTQHANW TLVVAVLSPL QYATGSSPDV QITASLQPVN	(Lyman *et al.*, 2002)
Ssx2_human	A0201	Q16385	KASEKIFYV	MNGDDAFARR PTVGAQIPEK IQKAFDDIAK YFSKEEWEKM KASEKIFYVY MKRKYEAMTK LGFKATLPPF MCNKRAEDFQ GNDLDNDPNR GNQVERPQMT FGRLQGISPK IMPKKPAEEG NDSEEVPEAS GPQNDGKELC PPGKPTTSEK IHERSGPKRG EHAWTHRLRE RKQLVIYEEI SDPEEDDE	(Ayyoub *et al.*, 2002)
Sp17_human	A0101	Q15506	ILDSSEEDK	MSIPFSNTHY RIPQGFGNLL EGLTREILRE QPDNIPAFAA AYFESLLEKR EKTNFDPAEW GSKVEDRFYN NHAFEEQEPP EKSDPKQEES QISGKEEETS VTILDSSEED KEKEEVAAVK IQAAFRGHIA REEAKKMKTN SLQNEEKEEN K	(Chiriva-Internati *et al.*, 2003)
EBNA-1 nuclear protein	DRB1*0401 DRB1*0	P03211	AEGLRALLARS HVER (EGLRALLAR)	aggagaggga gagggaggag gagagggaga gggaggagag ggaggaggag agggagagggaggagaggga ggaggagagg gagaggagga ggagaggaga gggaggagga gaggagagga	(Kruger *et al.*, 2003)

Res. 121-540	701			gaggagagga ggagaggagg agaggaggag agggaggaga gggaggagag gaggagaggaggagaggagg agagggagag gagaggggrg rggsggrgrg gsggrgrggs ggrrgrgrer arggsrerar grgrgrgekr prspssqsss sgspprrppp grrpffhpvg eadyfeyhqeggpdgepdvp pgaieqgpad dpgegpstgp rgqgdggrrk kggwfgkhrg qggsnpkfen iaeglralla rshverttde gtwvagvfvy ggsktslynl rrgtalaipq crltplsrlp	
Influenza matrix protein	A0301	Q67152	RLEDVFAGK	MSLLTEVETY VLSIVPSGPL KAEIAQRLED VFAGKNTDLE ALMEWLKTRP ILSPLTKGIL GFVFTLTVPS ERGLQRRRFV QNALNGNGDP NNMDKAVRLY RKLKREITFH GAKEVALSYS AGALASCMGL IYNRMGTVTT EVAFGLVCAT CEQIADSQHR SHRQMVTTTN PLIRHENRMV LASTTAKAME QIAGSSEQAA EAMEVASQAR QMVQAMRTIG THPSSSAGLK DDLLENLQAY QRRMGVQMQR FK	(Trojan *et al.*, 2003)
KFHU Coagulation factor IXa	H2Kb	268-282	CVETGVKITVV AGEH (KITVVAGE)	mqrvnmimae spgliticll gyllsaectv fldhenanki lnrpkrynsg kleefvqgnlerecmeekcs feearevfen terttefwkq yvdgdqcesnpclnggsckddinsyecwcpfgfegkncel dvtcniknger ceqfcknsad nkvvcscteg yrlaenqksc epavpfpcgrvsvsqtsklt raeavfpdvd yvnsteaeti ldnitqstqsfndftrvvggedakpgqfpwqvvlngkvdafcggsivnekw ivtaahcvegvkitvvagehnieetehteqkrnviriiphhnynaain kynhdialle ldeplvlnsy vtpiciadke ytniflkfgs gyvsgwgrvfhkgrsalvlq ylrvplvdra tclrstkfti ynnmfcagfh eggrdscqgd sggphvtevegtsfltgiis wgeecamkgk ygiytkvsry vnwikektkl t	(Greenwood *et al.*, 2003)
Superoxide dismutase	A0201	Q7TVI9	DMWEHAFYL	MAEYTLPDLD WDYGALEPHI SGQINELHHS KHHATYVKGA NDAVAKLEEA RAKEDHSAIL LNEKNLAFNL AGHVNHTIWW KNLSPNGGDK PTGELAAAIA DAFGSFDKFR AQFHAAATTV QGSGWAALGW DTLGNKLLIF QVYDHQTNFP LGIVPLLLLD MWEHAFYLQY KNVKVDFAKA FWNVVNWADV QSRYAAATSQ TKGLTFG	(Dong *et al.*, 2004)
L-alanine dehydrogenase	A0201	Q7TXW2	VLMGGVPGVE	MSEVAGRLAA QVGAYHLMRT QGGRGVLMGG VPGVEPADVV VIGAGTAGYN AARIANGMGA TVTVLDINID KLRQLDAEFC GRIHTRYSSA YELEGAVKRA DLVIGAVLVP GAKAPKLVSN SLVAHMKPGA VLVDIAIDQG GCFEGSRPTT YDHPTFAVHD TLFYCVANMP ASVPKTSTYA LTNATMPYVL ELADHGWRAA	(Dong *et al.*, 2004)

				CRSNPALAKG LSTHEGALLS ERVATDLGVP FTEPASVLA	
Coronavirus	A0201	AY278488	RLNEVAKNL	FIEDLLFNKVTLADAGFMKQYGECLGDINARDLICAQKFNGLTVLPPLLTD DMIAAYTAALVSGTATAGWTFGAGAALQIPFAMQMAYRFNGIGVTQNVLYE NQKQIANQFNKAISQIQESLTTTSTALGKLQDVVNQNAQALNTLVKQLSSN FGAISSVLNDILSRLDKVEAEVQIDRLITGRLQSLQTYVTQQLIRAAEIRA SANLAATKMSECVLGQSKRVDFCGKGYHLMSFPQAAPHGVVFLHVTYVPSQ ERNFTTAPAICHEGKAYFPREGVFVFNGTSWFITQRNFFSPQIITTDNTFV SGNCDVVIGIINNTVYDPLQPELDSFKEELDKYFKNHTSPDVDLGDISGIN ASVVNIQKEIDRLNEVAKNLNESLIDLQELGKYEQYIKWPWYVWLGFIAGL IAIVMVTILLCCMTSCCSCLKGACSCGSCCKFDEDDSEPVLKGVKLHYT	(Wang *et al.*, 2004)
Salmonella OmpC	H-2Kb	O52503	TRVAFAGL RNTDFFGL	MKVKVLSLLV PALLVAGAAN AAEIYNKDGN KLDLFGKVDG LHYFSDDKGS DGDQTYMRIG FKGETQVNDQ LTGYGQWEYQ IQGNQTEGSN DSWTRVAFAG LKFADAGSFD YGRNYGVTYD VTSWTDVLPE FGGDTYGADN FMQQRGNGYA TYRNTDFFGL VDGLDFALQY QGKNGSVSGE NTNGRSLLNQ NGDGYGGSLT YAIGEGFSVG GAITTSKRTA DQNNTANARL YGNGDRATVY TGGLKYDANN IYLAAQYSQT YNATRFGTSN GSNPSTSYGF ANKAQNFEVV AQYQFDFGLR PSVAYLQSKG KDISNGYGAS YGDQDIVKYV DVGATYYFNK NMSTYVDYKI NLLDKNDFTR DAGINTDDIV ALGLVYQF	(Diaz-Quinonez *et al.*, 2004)
murine encephalomyelitis virus vp2	H-2Kb	P08544	FHAGSLLVFMA PEFYTGKGT (GSLLVFMAP) NFNQYFGSLNF LFVFTGAAM	DKVLAAERYY TIDLASWTTS QEAFSHIRIP LPHVLAGEDG GVFGATLRRH YLCKTGWRVQ VQCNASQFHA GSLLVFMAPE FYTGKGTKTG TMEPSDPFTM DTEWRSPQGA PTGYRYDSRT GFFATNHQNQ WQWTVYPHQI LNLRTNTTVD LEVPYVNVAP SSSWTQHANW TLVVAVLSPL QYATGSSPDV QITASLQPVN PVFNGLRHET VIAQSPIPVT VREHKGCFYS TNPDTTVPIY GKTISTPSDY MCGEFSDLLE LCKLPTFLGN PNTNNKRYPY FSATNSVPAT SMVDYQVALS CSCMANSMLA AVARNFNQYR GSLNFLFVFT GAAMVKGKFL IAYTPPGAGK PTTRDQAMQS TYAIWDLGLN SSFNFTAPFI SPTHYRQTSY TSPTITSVDG	(Lyman *et al.*, 2002)

References

Adams, H. P. and Koziol, J. A. (1995). "Prediction of binding to MHC class I molecules." J. Immunol. Methods **185**(2): 181-90.

Adrian, P. E., Rajaseger, G., Mathura, V. S., Sakharkar, M. K. and Kangueane, P. (2002). "Types of inter-atomic interactions at the MHC-peptide interface: identifying commonality from accumulated data." BMC Struct. Biol. **2**(1): 2.

Ahn, K., Gruhler, A., Galocha, B., Jones, T. R., Wiertz, E. J., Ploegh, H. L., Peterson, P. A., Yang, Y. and Fruh, K. (1997). "The ER-luminal domain of the HCMV glycoprotein US6 inhibits peptide translocation by TAP." Immunity **6**(5): 613-21.

Al-Attiyah, R. and Mustafa, A. (2004). "Computer-assisted prediction of HLA-DR binding and experimental analysis for human promiscuous Th1-cell peptides in the 24 kDa secreted lipoprotein (LppX) of Mycobacterium tuberculosis." J. Immunol. **59**: 16-24.

Alix, A. J. (1999). "Predictive estimation of protein linear epitopes by using the program PEOPLE." Vaccine **18**(3-4): 311-4.

Alonso, A., Bottini, N., Bruckner, S., Rahmouni, S., Williams, S., Schoenberger, S. P. and Mustelin, T. (2004). "Lck dephosphorylation at Tyr-394 and inhibition of T cell antigen receptor signaling by Yersinia phosphatase YopH." J. Biol. Chem. **279**(6): 4922-8.

Altfeld, M. A., Livingston, B., Reshamwala, N., Nguyen, P. T., Addo, M. M., Shea, A., Newman, M., Fikes, J., Sidney, J., Wentworth, P., Chesnut, R., Eldridge, R. L., Rosenberg, E. S., Robbins, G. K., Brander, C., Sax, P. E., Boswell, S., Flynn, T., Buchbinder, S., Goulder, P. J., Walker, B. D., Sette, A. and Kalams, S. A. (2001). "Identification of novel HLA-A2-restricted human immunodeficiency virus type 1-specific cytotoxic T-lymphocyte epitopes predicted by the HLA-A2 supertype peptide-binding motif." J. Virol. **75**(3): 1301-11.

Altuvia, Y., Berzofsky, J. A., Rosenfeld, R. and Margalit, H. (1994). "Sequence features that correlate with MHC restriction." Mol. Immunol. **31**(1): 1-19.

Altuvia, Y., Schueler, O. and Margalit, H. (1995). "Ranking potential binding peptides to MHC molecules by a computational threading approach." J. Mol. Biol. **249**(2): 244-50.

Altuvia, Y., Sette, A., Sidney, J., Southwood, S. and Margalit, H. (1997). "A structure-based algorithm to predict potential binding peptides to MHC molecules with hydrophobic binding pockets." Hum. Immunol. **58**(1): 1-11.

Altuvia, Y. and Margalit, H. (2000). "Sequence signals for generation of antigenic peptides by the proteasome: implications for proteasomal cleavage mechanism1." J. Mol. Biol. **295**(4): 879-890.

Ambagala, A. P., Gopinath, R. S. and Srikumaran, S. (2003). "Inhibition of TAP by pseudorabies virus is independent of its vhs activity." Virus Res. **96**(1-2): 37-48.

Ambagala, A. P., Gopinath, R. S. and Srikumaran, S. (2004). "Peptide transport activity of the transporter associated with antigen processing (TAP) is inhibited by an early protein of equine herpesvirus-1." J. Gen. Virol. **85**(Pt 2): 349-53.

Amicosante, M., Gioia, C., Montesano, C., Casetti, R., Topino, S., D'Offizi, G., Cappelli, G., Ippolito, G., Colizzi, V., Poccia, F. and Pucillo, L. P. (2002). "Computer-based design of an HLA-haplotype and HIV-clade independent cytotoxic T-lymphocyte assay for monitoring HIV-specific immunity." Mol Med **8**(12): 798-807.

An, L. L. and Whitton, J. L. (1997). "A multivalent minigene vaccine, containing B-cell, cytotoxic T-lymphocyte, and Th epitopes from several microbes, induces appropriate responses in vivo and confers protection against more than one pathogen." J. Virol. **71**(3): 2292-302.

Anal, O., Akkoc, N., Sen, A., Yesil, S., Yuksel, F. and Buyukgebiz, A. (1997). "MHC class I antigen expression in patients with IDDM and their siblings." J. Pediatr. Endocrinol. Metab. **10**(4): 391-4.

Andersson, M., McMichael, A. and Peterson, P. A. (1987). "Reduced allorecognition of adenovirus-2 infected cells." J. Immunol. **138**(11): 3960-6.

Andersen, M. H., Tan, L., Sondergaard, I., Zeuthen, J., Elliott, T. and Haurum, J. S. (2000). "Poor correspondence between predicted and experimental binding of peptides to class I MHC molecules." Tissue Antigens **55**(6): 519-31.

Androulakis, I. P., Nayak, N. N., Ierapetritou, M. G., Monos, D. S. and Floudas, C. A. (1997). "A predictive method for the evaluation of peptide binding in pocket 1 of HLA-DRB1 via global minimization of energy interactions." Proteins **29**(1): 87-102.

Arden, B., Clark, S. P., Kabelitz, D. and Mak, T. W. (1995). "Human T-cell receptor variable gene segment families." Immunogenetics **42**(6): 455-500.

Arnon, R., Tarrab-Hazdai, R. and Ben-Yedidia, T. (2001). "Peptide-based synthetic recombinant vaccines with anti-viral efficacy." Biologicals **29**(3-4): 237-42.

Ayyoub, M., Stevanovic, S., Sahin, U., Guillaume, P., Servis, C., Rimoldi, D., Valmori, D., Romero, P., Cerottini, J., Rammensee, H., Pfreundschuh, M., Speiser, D. and Levy, F. (2002). "Proteasome-assissted identification of a SSX-2-derived epitope recognized by tumor-reactive CTL infiltrating metastatic melanoma." J. Immunol. **168**: 1717-1722.

Azad, R. K. and Borodovsky, M. (2004). "Probabilistic methods of identifying genes in prokaryotic genomes: connections to the HMM theory." Brief Bioinform **5**(2): 118-30.

Baas, A., Gao, X. and Chelvanayagam, G. (1999). "Peptide binding motifs and specificities for HLA-DQ molecules." Immunogenetics **50**(1-2): 8-15.

Bairoch, A. and Boeckmann, B. (1991). "The SWISS-PROT protein sequence data bank." Nucleic Acids Res. **19 Suppl**: 2247-9.

Bakker, A., Schreurs, M., DeBoer, A., Kawakami, Y., Rosenberg, S., Adema, G. and Figdor, C. (1994). "Melanocyte lineage-specific antigen gp100 is recognized by melanoma-derived tumor-infiltrating lymphocytes." J. Exp. Med. **179**: 1005-1009.

Balow, J., Weissman, J. and Kearse, K. (1995). "Unique expression of major histocompatibility complex class I proteins in the absence of glucose trimming and calnexin association." J. Biol. Chem. **270**: 29025-.

Bangalore, A. S., Shaffer, R. E., Small, G. W. and Arnold, M. A. (1996). "Genetic algorithm-based method for selecting wavelengths and model size for use with partial least-squares regression: application to near-infrared spectroscopy." Anal. Chem. **68**(23): 4200-12.

Banks, T. A., Nair, S. and Rouse, B. T. (1993). "Recognition by and in vitro induction of cytotoxic T lymphocytes against predicted epitopes of the immediate-early protein ICP27 of herpes simplex virus." J. Virol. **67**(1): 613-6.

Barreca, M. L., Carotti, A., Carrieri, A., Chimirri, A., Monforte, A. M., Calace, M. P. and Rao, A. (1999). "Comparative molecular field analysis (CoMFA) and docking studies of non-nucleoside HIV-1 RT inhibitors (NNIs)." Bioorg. Med. Chem. **7**(11): 2283-92.

Bateman, A. and Haft, D. H. (2002). "HMM-based databases in InterPro." Brief Bioinform **3**(3): 236-45.

Bauer, D. and Tampe, R. (2002). "Herpes viral proteins blocking the transporter associated with antigen processing TAP--from genes to function and structure." Curr. Top. Microbiol. Immunol. **269**: 87-99.

Baumeister, W., Walz, J., Zuhl, F., Seemuller, E., (1998). "The proteasome: paradigm of a self-compartmentalizing protease." Cell **92**: 367-380.

Bazhan, S. I., Belavin, P. A., Seregin, S. V., Danilyuk, N. K., Babkina, I. N., Karpenko, L. I., Nekrasova, N. A., Lebedev, L. R., Ignatyev, G. M., Agafonov, A. P., Poryvaeva, V. A., Aborneva, I. V. and Ilyichev, A. A. (2004). "Designing and engineering of DNA-vaccine construction encoding multiple CTL-epitopes of major HIV-1 antigens." Vaccine **22**(13-14): 1672-82.

Beale, R. and Jackson, T. (1990). Neural computing: an introduction. Bristol, Adam Hilger.

Beck, S., Kelly, A., Radley, E., Khurshid, F., Alderton, R. P. and Trowsdale, J. (1992). "DNA sequence analysis of 66 kb of the human MHC class II region encoding a cluster of genes for antigen processing." J. Mol. Biol. **228**(2): 433-41.

Beecham, E. J., Ma, Q., Ripley, R. and Junghans, R. P. (2000). "Coupling CD28 co-stimulation to immunoglobulin T-cell receptor molecules: the dynamics of T-cell proliferation and death." J. Immunother. **23**(6): 631-42.

Bellgard, M. I., Tay, G. K., Hiew, H. L., Witt, C. S., Ketheesan, N., Christiansen, F. T. and Dawkins, R. L. (1998). "MHC haplotype analysis by artificial neural networks." Hum. Immunol. **59**(1): 56-62.

Belyakov, I. M., Ahlers, J. D., Brandwein, B. Y., Earl, P., Kelsall, B. L., Moss, B., Strober, W. and Berzofsky, J. A. (1998a). "The importance of local mucosal HIV-specific CD8(+) cytotoxic T lymphocytes for resistance to mucosal viral transmission in mice and enhancement of resistance by local administration of IL-12." J Clin Invest **102**(12): 2072-81.

Belyakov, I. M., Derby, M. A., Ahlers, J. D., Kelsall, B. L., Earl, P., Moss, B., Strober, W. and Berzofsky, J. A. (1998b). "Mucosal immunization with HIV-1 peptide vaccine induces mucosal and systemic cytotoxic T lymphocytes and protective immunity in mice against intrarectal recombinant HIV-vaccinia challenge." Proc. Natl. Acad. Sci. U. S. A. **95**(4): 1709-14.

Ben-Hur, H., Ben-Meir, A., Hagay, Z., Berman, V., Schwartsburd, B., Gurevich, P., Sandler, B., Tendler, Y., Zinder, O. and Zusman, I. (1998). "Tumor-preventive effects of the soluble p53 antigen on chemically-induced skin cancer in mice." Anticancer Res. **18**(6A): 4237-41.

Ben-Hur, H., Kossoy, G., Sandler, B. and Zusman, I. (2000). "Vaccination with soluble low-molecular weight tumor-associated proteins suppresses chemically-induced mammary tumorigenesis in rats." In Vivo **14**(4): 551-4.

Beninga, J., Rock, K. L. and Goldberg, A. L. (1998). "Interferon-gamma can stimulate post-proteasomal trimming of the N terminus of an antigenic peptide by inducing leucine aminopeptidase." J. Biol. Chem. **273**(30): 18734-42.

Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J. and Wheeler, D. L. (2004). "GenBank: update." Nucleic Acids Res. **32 Database issue**: D23-6.

Bentley, G. A., Boulot, G., Karjalainen, K. and Mariuzza, R. A. (1995). "Crystal structure of the beta chain of a T cell antigen receptor." Science **267**(5206): 1984-7.

Bergmann, C. C., Tong, L., Cua, R., Sensintaffar, J. and Stohlman, S. (1994). "Differential effects of flanking residues on presentation of epitopes from chimeric peptides." J. Virol. **68**: 5306-5310.

Bergmann, C., C, Yao, Q., Ho, C. K. and Buckwold, S. L. (1996). "Flanking residues alter antigenicity and immunogenicity of multi-unit CTL epitopes." J. Immunol. **157**: 3242-3249.

Bertoletti, A., Southwood, S., Chesnut, R., Sette, A., Falco, M., Ferrara, G. B., Penna, A., Boni, C., Fiaccadori, F. and Ferrari, C. (1997). "Molecular features of the hepatitis B virus nucleocapsid T-cell epitope 18-27: interaction with HLA and T-cell receptor." Hepatology **26**(4): 1027-34.

Bertoni, R., Sidney, J., Fowler, P., Chesnut, R. W., Chisari, F. V. and Sette, A. (1997). "Human histocompatibility leukocyte antigen-binding supermotifs predict broadly cross-reactive cytotoxic T lymphocyte responses in patients with acute hepatitis." J Clin Invest **100**(3): 503-13.

Bharadwaj, M. and Moss, D. J. (2002). "Epstein-Barr virus vaccine: a cytotoxic T-cell-based approach." Expert Rev Vaccines **1**(4): 467-76.

Bhasin, M. and Raghava, G. P. (2004a). "Prediction of CTL epitopes using QM, SVM and ANN techniques." Vaccine **22**(23-24): 3195-204.

Bhasin, M. and Raghava, G. P. (2004b). "ESLpred: SVM-based method for subcellular localization of eukaryotic proteins using dipeptide composition and PSI-BLAST." Nucleic Acids Res. **32**(Web Server issue): W414-9.

Bhasin, M. and Raghava, G. P. (2004c). "GPCRpred: an SVM-based method for prediction of families and subfamilies of G-protein coupled receptors." Nucleic Acids Res. **32**(Web Server issue): W383-9.

Bhasin, M. and Raghava, G. P. (2004d). "Classification of nuclear receptors based on amino acid composition and dipeptide composition." J. Biol. Chem. **279**(22): 23262-6.

Bhasin, M. and Raghava, G. P. (2004e). "SVM based method for predicting HLA-DRB1*0401 binding peptides in an antigen sequence." Bioinformatics **20**(3): 421-3.

Bhasin, M. and Raghava, G. P. (2004f). "Analysis and prediction of affinity of TAP binding peptides using cascade SVM." Protein Sci **13**(3): 596-607.

Bhasin, M., Singh, H. and Raghava, G. P. (2003). "MHCBN: a comprehensive database of MHC binding and non-binding peptides." Bioinformatics **19**(5): 665-

Bhongade, B. A. and Gadad, A. K. (2004). "3D-QSAR CoMFA/CoMSIA studies on Urokinase plasminogen activator (uPA) inhibitors: a strategic design in novel anticancer agents." Bioorg. Med. Chem. **12**(10): 2797-805.

Binz, A. K., Rodriguez, R. C., Biddison, W. E. and Baker, B. M. (2003). "Thermodynamic and kinetic analysis of a peptide-class I MHC interaction highlights the noncovalent nature and conformational dynamics of the class I heterotrimer." Biochemistry **42**(17): 4954-61.

Bisset, L. R. and Fierz, W. (1993). "Using a neural network to identify potential HLA-DR1 binding sites within proteins." J. Mol. Recognit. **6**(1): 41-8.

Bjorkman, P. J., Saper, M. A., Samraoui, B., Bennett, W. S., Strominger, J. L. and Wiley, D. C. (1987a). "Structure of the human class I histocompatibility antigen, HLA-A2." Nature **329**(6139): 506-12.

Bjorkman, P. J., Saper, M. A., Samraoui, B., Bennett, W. S., Strominger, J. L. and Wiley, D. C. (1987b). "The foreign antigen binding site and T cell recognition regions of class I histocompatibility antigens." Nature **329**(6139): 512-8.

Bjorkman, P. J. and Parham, P. (1990). "Structure, function, and diversity of class I major histocompatibility complex molecules." Annu. Rev. Biochem. **59**: 253-88.

Blythe, M. J., Doytchinova, I. A. and Flower, D. R. (2002). "JenPep: a database of quantitative functional peptide data for immunology." Bioinformatics **18**(3): 434-9.

Bodey, B., Bodey, B., Jr., Siegel, S. E. and Kaiser, H. E. (2000). "Failure of cancer vaccines: the significant limitations of this approach to immunotherapy." Anticancer Res. **20**(4): 2665-76.

Bodmer, H. C., Bastin, J. M., Askonas, B. A. and Townsend, A. R. (1989). "Influenza-specific cytotoxic T-cell recognition is inhibited by peptides unrelated in both sequence and MHC restriction." Immunology **66**(2): 163-9.

Bodmer, J. G., Marsh, S. G. and Albert, E. (1990a). "Nomenclature for factors of the HLA system, 1989." Immunol Today **11**(1): 3-10.

Bodmer, J. G., Marsh, S. G., Parham, P., Erlich, H. A., Albert, E., Bodmer, W. F., Dupont, B., Mach, B., Mayr, W. R., Sasazuki, T. and et al. (1990b). "Nomenclature for factors of the HLA system, 1989." Tissue Antigens **35**(1): 1-8.

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**(1): 365-70.

Boehncke, W. H., Takeshita, T., Pendleton, C. D., Houghten, R. A., Sadegh-Nasseri, S., Racioppi, L., Berzofsky, J. A. and Germain, R. N. (1993). "The importance of dominant negative effects of amino acid side chain substitution in peptide-MHC molecule interactions and T cell recognition." J. Immunol. **150**(2): 331-41.

Bologa, C., Drugarin, D. and Simon, Z. (1995). "Quantitative structure-activity relations by the MTD-method for binding of nonapeptides to the HLA-A2.1 molecule." Roum Arch Microbiol Immunol **54**(1-2): 3-14.

Borrego, F., Ulbrecht, M., Weiss, E. H., Coligan, J. E. and Brooks, A. G. (1998). "Recognition of human histocompatibility leukocyte antigen (HLA)-E complexed with HLA class I signal sequence-derived peptides by CD94/NKG2 confers protection from natural killer cell-mediated lysis." J. Exp. Med. **187**(5): 813-8.

Bouvier, M. and Wiley, D. C. (1994). "Importance of peptide amino and carboxyl termini to the stability of MHC class I molecules." Science **265**(5170): 398-402.

Bouvier, M. and Wiley, D. C. (1998a). "Structural characterization of a soluble and partially folded class I major histocompatibility heavy chain/beta 2m heterodimer." Nat. Struct. Biol. **5**(5): 377-84.

Bouvier, M., Guo, H. C., Smith, K. J. and Wiley, D. C. (1998b). "Crystal structures of HLA-A*0201 complexed with antigenic peptides with either the amino- or carboxyl-terminal group substituted by a methyl group." Proteins **33**(1): 97-106.

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**(5): 1268-82.

Boyington, J. C., Motyka, S. A., Schuck, P., Brooks, A. G. and Sun, P. D. (2000). "Crystal structure of an NK cell immunoglobulin-like receptor in complex with its class I MHC ligand." Nature **405**(6786): 537-43.

Brahmajothi, V., Pitchappan, R. M., Kakkanaiah, V. N., Sashidhar, M., Rajaram, K., Ramu, S., Palanimurugan, K., Paramasivan, C. N. and Prabhakar, R. (1991). "Association of pulmonary tuberculosis and HLA in south India." Tubercle **72**(2): 123-32.

Braud, V. M., Allan, D. S., O'Callaghan, C. A., Soderstrom, K., D'Andrea, A., Ogg, G. S., Lazetic, S., Young, N. T., Bell, J. I., Phillips, J. H., Lanier, L. L. and McMichael, A. J. (1998). "HLA-E binds to natural killer cell receptors CD94/NKG2A, B and C." Nature **391**(6669): 795-9.

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.

Brusic, V., Rudy, G. and Harrison, L. C. (1998a). "MHCPEP, a database of MHC-binding peptides: update 1997." Nucleic Acids Res. **26**(1): 368-71.

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**(2): 121-30.

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**(3): 280-5.

Buolamwini, J. K. and Assefa, H. (2002). "CoMFA and CoMSIA 3D QSAR and docking studies on conformationally-restrained cinnamoyl HIV-1 integrase inhibitors: exploration of a binding mode at the active site." J. Med. Chem. **45**(4): 841-52.

Burrows, G. G., Ariail, K., Celnik, B., Gambee, J. E., Bebo, B. F., Jr., Offner, H. and Vandenbark, A. A. (1996). "Variation in H-2K(k) peptide motif revealed by sequencing naturally processed peptides from T-cell hybridoma class I molecules." J. Neurosci. Res. **45**(6): 803-11.

Burrows, S. R., Elkington, R. A., Miles, J. J., Green, K. J., Walker, S., Haryana, S. M., Moss, D. J., Dunckley, H., Burrows, J. M. and Khanna, R. (2003). "Promiscuous CTL recognition of viral epitopes on multiple human leukocyte antigens: biological validation of the proposed HLA A24 supertype." J. Immunol. **171**(3): 1407-12.

Bush, B. L. and Nachbar, R. B., Jr. (1993). "Sample-distance partial least squares: PLS optimized for many variables, with application to CoMFA." J Comput Aided Mol Des **7**(5): 587-619.

Buslepp, J., Wang, H., Biddison, W. E., Appella, E. and Collins, E. J. (2003). "A correlation between TCR Valpha docking on MHC and CD8 dependence: implications for T cell selection." Immunity **19**(4): 595-606.

Butterfield, L. H., Meng, W. S., Koh, A., Vollmer, C. M., Ribas, A., Dissette, V. B., Faull, K., Glaspy, J. A., McBride, W. H. and Economou, J. S. (2001). "T cell responses to HLA-A*0201-restricted peptides derived from human alpha fetoprotein." J. Immunol. **166**(8): 5300-8.

Buus, S., Sette, A., Colon, S. M. and Grey, H. M. (1988). "Autologous peptides constitutively occupy the antigen binding site on Ia." Science **242**(4881): 1045-7.

Buus, S., Lauemoller, S. L., Worning, P., Kesmir, C., Frimurer, T., Corbet, S., Fomsgaard, A., Hilden, J., Holm, A. and Brunak, S. (2003). "Sensitive quantitative predictions of peptide-MHC binding by a 'Query by Committee' artificial neural network approach." Tissue Antigens **62**(5): 378-84.

Caccamo, N., Milano, S., Di Sano, C., Cigna, D., Ivanyi, J., Krensky, A. M., Dieli, F. and Salerno, A. (2002). "Identification of epitopes of Mycobacterium tuberculosis 16-kDa protein recognized by human leukocyte antigen-A*0201 CD8(+) T lymphocytes." J. Infect. Dis. **186**(7): 991-8.

Cai, Y. D., Zhou, G. P., Jen, C. H., Lin, S. L. and Chou, K. C. (2004). "Identify catalytic triads of serine hydrolases by support vector machines." J. Theor. Biol. **228**(4): 551-7.

Caiozzo, V. J., Haddad, F., Baker, M., McCue, S. and Baldwin, K. M. (2000). "MHC polymorphism in rodent plantaris muscle: effects of mechanical overload and hypothyroidism." Am J Physiol Cell Physiol **278**(4): C709-17.

Cano, P., Fan, B. and Stass, S. (1998). "A geometric study of the amino acid sequence of class I HLA molecules." Immunogenetics **48**(5): 324-34.

Cano, P. and Fan, B. (2001). "A geometric and algebraic view of MHC-peptide complexes and their binding properties." BMC Struct. Biol. **1**(1): 2.

Caro-Aguilar, I., Rodriguez, A., Calvo-Calle, J., Guzman, F., De la Vega, P., Patarroyo, M., Galinski, M. and Moreno, A. (2002). "Plasmodium vivax promiscuous T-helper epitopes defined and evaluated as linear peptide chimera immunogens." Infect. Immun. **70**: 3479-3492.

Carreno, B. M., Winter, C. C., Taurog, J. D., Hansen, T. H. and Biddison, W. E. (1993). "Residues in pockets B and F of HLA-B27 are critical in the presentation of an influenza A virus nucleoprotein peptide and influence the stability of peptide - MHC complexes." Int. Immunol. **5**(4): 353-60.

Carreno, B. M., Solheim, J. C., Harris, M., Stroynowski, I., Connolly, J. M. and Hansen, T. H. (1995). "TAP associates with a unique class I conformation, whereas calnexin associates with multiple class I forms in mouse and man." J. Immunol. **155**(10): 4726-33.

Cassatt, J. C. and Peterson, J. L. (1987). "GenBank information." Science **238**(4831): 1215.

Castelli, F. A., Buhot, C., Sanson, A., Zarour, H., Pouvelle-Moratille, S., Nonn, C., Gahery-Segard, H., Guillet, J. G., Menez, A., Georges, B. and Maillere, B. (2002). "HLA-DP4, the most frequent HLA II molecule, defines a new supertype of peptide-binding specificity." J. Immunol. **169**(12): 6928-34.

Chen, Q., Jackson, H., Parente, P., Luke, T., Rizkalla, M., Tai, T. Y., Zhu, H. C., Mifsud, N. A., Dimopoulos, N., Masterman, K. A., Hopkins, W., Goldie, H., Maraskovsky, E., Green, S., Miloradovic, L., McCluskey, J., Old, L. J., Davis, I. D., Cebon, J. and Chen, W. (2004a). "Immunodominant CD4+ responses identified in a patient vaccinated with full-length NY-ESO-1 formulated with ISCOMATRIX adjuvant." Proc. Natl. Acad. Sci. U. S. A. **101**(25): 9363-8.

Chen, Y. C., Lin, Y. S., Lin, C. J. and Hwang, J. K. (2004b). "Prediction of the bonding states of cysteines using the support vector machines based on multiple feature vectors and cysteine state sequences." Proteins **55**(4): 1036-42.

Chersi, A., di Modugno, F. and Rosano, L. (2000). "Flexibility of amino acid residues at position four of nonapeptides enhances their binding to human leucocyte antigen (HLA) molecules." Z Naturforsch [C] **55**(1-2): 109-14.

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**(1): 27-47.

Chiriva-Internati, M., Wang, Z., Pochopien, S., Salati, E. and Lim, S. (2003). "Identification of a sperm protein 17 CTL epitope restricted by HLA-A1." Int. J. Cancer **107**: 863-865.

Choi, A. H., McNeal, M. M., Basu, M., Bean, J. A., VanCott, J. L., Clements, J. D. and Ward, R. L. (2003). "Functional mapping of protective epitopes within the rotavirus VP6 protein in mice belonging to different haplotypes." Vaccine **21**(7-8): 761-7.

Choo, H. Y., Choi, S., Ryu, C. K., Kim, H. J., Lee, I. Y., Paeb, A. N. and Koh, H. Y. (2003). "QSAR study of quinolinediones with inhibitory activity of endothelium-dependent vasorelaxation by CoMSIA." Bioorg. Med. Chem. **11**(9): 2019-23.

Choudhuri, K. and Vergani, D. (1998). "MHC restriction to T-cell autoaggression: an emerging understanding of IDDM pathogenesis." Diabetes Metab Rev **14**(4): 285-301.

Christinck, E. R., Luscher, M. A., Barber, B. H. and Williams, D. B. (1991). "Peptide binding to class I MHC on living cells and quantitation of complexes required for CTL lysis." Nature **352**(6330): 67-70.

Chujoh, Y., Sobao, Y., Miwa, K., Kaneko, Y. and Takiguchi, M. (1998). "The role of anchor residues in the binding of peptides to HLA-A*1101 molecules." Tissue Antigens **52**(6): 501-9.

Ciernik, I. F. and Carbone, D. P. (1995). "Tumor Suppressor Gene-Derived Peptide Antigens." Methods **8**(3): 225-233.

Clark, S. S. and Forman, J. (1984). "Functional aspects of class I MHC molecule domains." Surv Immunol Res **3**(2-3): 179-83.

Cochlovius, B., Stassar, M., Christ, O., Raddrizzani, L., Hammer, J., Mytilineos, I. and Zoller, M. (2000). "In vitro and in vivo induction of a Th cell response toward peptides of the melanoma-associated glycoprotein 100 protein selected by the TEPITOPE program." J. Immunol. **165**(8): 4731-41.

Colbert, R. A., Rowland-Jones, S. L., McMichael, A. J. and Frelinger, J. A. (1993). "Allele-specific B pocket transplant in class I major histocompatibility complex protein changes requirement for anchor residue at P2 of peptide." Proc. Natl. Acad. Sci. U. S. A. **90**(14): 6879-83.

Coligan, J. E., Gates, F. T., Kindt, T. J., Ewenstein, B. M., Martinko, J. M., Uehara, H. and Nathenson, S. G. (1981). "Primary structure of murine H-2Kb and beta 2-microglobulin." Transplant Proc. **13**: 1792-1796.

Collantes, E. R. and Dunn, W. J., 3rd (1995). "Amino acid side chain descriptors for quantitative structure-activity relationship studies of peptide analogues." J. Med. Chem. **38**(14): 2705-13.

Collins, R. W., Stephens, H. A. and Vaughan, R. W. (2003). "Molecular typing of the MHC class I chain-related gene locus." Methods Mol. Biol. **210**: 305-21.

Combet, C., Blanchet, C., Geourjon, C. and Deleage, G. (2000). "NPS@: network protein sequence analysis." Trends Biochem. Sci. **25**(3): 147-50.

Consogno, G., Manici, S., Facchinetti, V., Bachi, A., Hammer, J., Conti-Fine, B. M., Rugarli, C., Traversari, C. and Protti, M. P. (2003). "Identification of immunodominant regions among promiscuous HLA-DR-restricted CD4+ T-cell epitopes on the tumor antigen MAGE-3." Blood **101**(3): 1038-44.

Corbet, S., Nielsen, H., Vinner, L., Lauemoller, S., Therrien, D., Tang, S., Kronborg, G., Mathiesen, L., Chaplin, P., Brunak, S., Buus, S. and Fomsgaard, A. (2003). "Optimization and immune recognition of multiple novel conserved HLA-A2, human immunodeficiency virus type 1-specific CTL epitopes." J. Gen. Virol. **84**: 2409-2421.

Cossins, J., Gould, K. G., Smith, M., Driscoll, P. and Brownlee, G. G. (1993). "Precise prediction of a Kk-restricted cytotoxic T cell epitope in the NS1 protein of influenza virus using an MHC allele-specific motif." Virology **193**(1): 289-95.

Craig, P. N. (1974). "Proceedings: Comparison of Hansch and free-Wilson methods for structure-activity correlation." Cancer Chemother Rep 2 **4**(4): 39.

Cramer, R. D., 3rd, Patterson, D. E. and Bunce, J. D. (1989). "Recent advances in comparative molecular field analysis (CoMFA)." Prog. Clin. Biol. Res. **291**: 161-5.

Creusot, R. J., Mitchison, N. A. and Terazzini, N. M. (2002). "The immunological synapse." Mol. Immunol. **38**(12-13): 997-1002.

Cruciani, G. and Watson, K. A. (1994). "Comparative molecular field analysis using GRID force-field and GOLPE variable selection methods in a study of inhibitors of glycogen phosphorylase b." J. Med. Chem. **37**(16): 2589-601.

Cui, M., Huang, X., Luo, X., Briggs, J. M., Ji, R., Chen, K., Shen, J. and Jiang, H. (2002). "Molecular docking and 3D-QSAR studies on gag peptide analogue inhibitors interacting with human cyclophilin A." J. Med. Chem. **45**(24): 5249-59.

Dai, Z., Konieczny, B. T. and Lakkis, F. G. (2000). "The dual role of IL-2 in the generation and maintenance of CD8+ memory T cells." J. Immunol. **165**(6): 3031-6.

D'Amaro, J., Houbiers, J. G., Drijfhout, J. W., Brandt, R. M., Schipper, R., Bavinck, J. N., Melief, C. J. 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**(1): 13-8.

Dangles, V., Halberstam, I., Scardino, A., Choppin, J., Wertheimer, M., Richon, S., Quelvennec, E., Moirand, R., Guillet, J. G., Kosmatopoulos, K., Bellet, D. and Zeliszewski, D. (2002). "Tumor-associated antigen human chorionic gonadotropin beta contains numerous antigenic determinants recognized by in vitro-induced CD8+ and CD4+ T lymphocytes." Cancer Immunol. Immunother. **50**(12): 673-81.

Davenport, M. P., Ho Shon, I. A. and Hill, A. V. (1995). "An empirical method for the prediction of T-cell epitopes." Immunogenetics **42**(5): 392-7.

Davies, M. N., Sansom, C. E., Beazley, C. and Moss, D. S. (2003). "A novel predictive technique for the MHC class II peptide-binding interaction." Mol Med **9**(9-12): 220-5.

Davis, M. M., Boniface, J. J., Reich, Z., Lyons, D., Hampl, J., Arden, B. and Chien, Y. (1998). "Ligand recognition by alpha beta T cell receptors." Annu. Rev. Immunol. **16**: 523-44.

De Groot, A. S., Bosma, A., Chinai, N., Frost, J., Jesdale, B. M., Gonzalez, M. A., Martin, W. and Saint-Aubin, C. (2001). "From genome to vaccine: in silico predictions, ex vivo verification." Vaccine **19**(31): 4385-95.

de la Hera, A., Muller, U., Olsson, C., Isaaz, S. and Tunnacliffe, A. (1991). "Structure of the T cell antigen receptor (TCR): two CD3 epsilon subunits in a functional TCR/CD3 complex." J. Exp. Med. **173**(1): 7-17.

Del Carpio, C. A., Hennig, T., Fickel, S. and Yoshimori, A. (2002). "A combined bioinformatic approach oriented to the analysis and design of peptides with high affinity to MHC class I molecules." Immunol. Cell Biol. **80**(3): 286-99.

del Guercio, M. F., Sidney, J., Hermanson, G., Perez, C., Grey, H. M., Kubo, R. T. and Sette, A. (1995). "Binding of a peptide antigen to multiple HLA alleles allows definition of an A2-like supertype." J. Immunol. **154**(2): 685-93.

Del Val, M., Schlicht, H.-J., Ruppert, T., Reddehase, M. and Koszinowshi, U. (1991). "Efficient processing of an antigenic sequence for presentation by MHC class I molecules depends on its neighboring residues in the protein." Cell **66**: 1145-1153.

Delgado, G., Parra-Lopez, C. A., Vargas, L. E., Hoya, R., Estupinan, M., Guzman, F., Torres, A., Alonso, C., Velez, I. D., Spinel, C. and Patarroyo, M. E. (2003). "Characterizing cellular immune response to kinetoplastid membrane protein-11 (KMP-11) during Leishmania (Viannia) panamensis infection using dendritic cells (DCs) as antigen presenting cells (APCs)." Parasite Immunol **25**(4): 199-209.

Delorenzi, M. and Speed, T. (2002). "An HMM model for coiled-coil domains and a comparison with PSSM-based predictions." Bioinformatics **18**(4): 617-25.
Deres, K., Beck, W., Faath, S., Jung, G. and Rammensee, H. G. (1993). "MHC/peptide binding studies indicate hierarchy of anchor residues." Cell. Immunol. **151**(1): 158-67.

Dermime, S., Gilham, D. E., Shaw, D. M., Davidson, E. J., Meziane el, K., Armstrong, A., Hawkins, R. E. and Stern, P. L. (2004). "Vaccine and antibody-directed T cell tumour immunotherapy." Biochim. Biophys. Acta **1704**(1): 11-35.

Devillers, J. (1996). Genetic algorithms in molecular modeling. London, Academic press Ltd.

Diaz-Quinonez, A., Martin-Orozco, N., Isibasi, A. and Ortiz-Navarrete, V. (2004). "Two Salmonella OmpC K(b)-restricted epitopes for CD8+-T-cell recognition." Infect. Immun. **72**: 3059-3062.

DiBrino, M., Parker, K. C., Shiloach, J., Knierman, M., Lukszo, J., Turner, R. V., Biddison, W. E. and Coligan, J. E. (1993). "Endogenous peptides bound to HLA-A3 possess a specific combination of anchor residues that permit identification of potential antigenic peptides." Proc. Natl. Acad. Sci. U. S. A. **90**(4): 1508-12.

Dick, T. P., Stevanovic, S., Keilholz, W., Ruppert, T., Koszinowski, U., Schild, H. and Rammensee, H. G. (1998). "The making of the dominant MHC class I ligand SYFPEITHI." Eur. J. Immunol. **28**(8): 2478-86.

Ding, C. H. and Dubchak, I. (2001). "Multi-class protein fold recognition using support vector machines and neural networks." Bioinformatics **17**(4): 349-58.

Ding, Y. H., Smith, K. J., Garboczi, D. N., Utz, U., Biddison, W. E. and Wiley, D. C. (1998). "Two human T cell receptors bind in a similar diagonal mode to the HLA-A2/Tax peptide complex using different TCR amino acids." Immunity **8**(4): 403-11.

Disis, M. L., Grabstein, K. H., Sleath, P. R. and Cheever, M. A. (1999). "Generation of immunity to the HER-2/neu oncogenic protein in patients with breast and ovarian cancer using a peptide-based vaccine." Clinical Cancer Research **5**: 1289-1297.

Dittel, B. N., Stefanova, I., Germain, R. N. and Janeway, C. A., Jr. (1999). "Cross-antagonism of a T cell clone expressing two distinct T cell receptors." Immunity **11**(3): 289-98.

Dixit, A., Kashaw, S. K., Gaur, S. and Saxena, A. K. (2004). "Development of CoMFA, advance CoMFA and CoMSIA models in pyrroloquinazolines as thrombin receptor antagonist." Bioorg. Med. Chem. **12**(13): 3591-8.

Doddareddy, M. R., Jung, H. K., Cha, J. H., Cho, Y. S., Koh, H. Y., Chang, M. H. and Pae, A. N. (2004). "3D QSAR studies on T-type calcium channel blockers using CoMFA and CoMSIA." Bioorg. Med. Chem. **12**(7): 1613-21.

Dong, H. L., Sui, Y. F., Ye, J., Li, Z. S., Qu, P., Zhang, X. M., Chen, G. S. and Lu, S. Y. (2003). "[Prediction synthesis and identification of HLA-A2-restricted cytotoxic T lymphocyte epitopes of the tumor antigen MAGE-n]." Zhonghua Yi Xue Za Zhi **83**(12): 1080-3.

Dong, Y., Demaria, S., Sun, X., Santori, F. R., Jesdale, B. M., De Groot, A. S., Rom, W. N. and Bushkin, Y. (2004). "HLA-A2-restricted CD8+-cytotoxic-T-cell responses to novel epitopes in Mycobacterium tuberculosis superoxide dismutase, alanine dehydrogenase, and glutamine synthetase." Infect Immun. **72**: 2412-2415.

Donnes, P. and Elofsson, A. (2002). "Prediction of MHC class I binding peptides, using SVMHC." BMC Bioinformatics **3**(1): 25.

Doolan, D. L., Hoffman, S. L., Southwood, S., Wentworth, P. A., Sidney, J., Chesnut, R. W., Keogh, E., Appella, E., Nutman, T. B., Lal, A. A., Gordon, D. M., Oloo, A. and Sette, A. (1997). "Degenerate cytotoxic T cell epitopes from P. falciparum restricted by multiple HLA-A and HLA-B supertype alleles." Immunity 7(1): 97-112.

Doolan, D. L., Southwood, S., Chesnut, R., Appella, E., Gomez, E., Richards, A., Higashimoto, Y. I., Maewal, A., Sidney, J., Gramzinski, R. A., Mason, C., Koech, D., Hoffman, S. L. and Sette, A. (2000). "HLA-DR-promiscuous T cell epitopes from Plasmodium falciparum pre-erythrocytic-stage antigens restricted by multiple HLA class II alleles." J. Immunol. **165**(2): 1123-37.

Dooley, C. T. and Houghten, R. A. (1993). "The use of positional scanning synthetic peptide combinatorial libraries for the rapid determination of opioid receptor ligands." Life Sci. **52**(18): 1509-17.

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**(22): 3572-81.

Doytchinova, I. A. and Flower, D. R. (2002). "A comparative molecular similarity index analysis (CoMSIA) study identifies an HLA-A2 binding supermotif." J Comput Aided Mol Des **16**(8-9): 535-44.

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 I molecule HLA-A*0201." J. Proteome Res. **1**(3): 263-72.

Doytchinova, I. and Flower, D. (2003a). "The HLA-A2-supermotif: a QSAR definition." Org Biomol Chem **1**(15): 2648-54.

Doytchinova, I. A. and Flower, D. R. (2003b). "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**(17): 2263-70.

Doytchinova, I. A., Guan, P. and Flower, D. R. (2004a). "Identifiying human MHC supertypes using bioinformatic methods." J. Immunol. **172**(7): 4314-23.

Doytchinova, I. A., Walshe, V. A., Jones, N. A., Gloster, S. E., Borrow, P. and Flower, D. R. (2004). "Coupling In Silico and In Vitro Analysis of Peptide-MHC Binding: A Bioinformatic Approach Enabling Prediction of Superbinding Peptides and Anchorless Epitopes." J. Immunol. **172**: 7495-7502.

Drijfhout, J. W., Brandt, R. M., D'Amaro, J., Kast, W. M. and Melief, C. J. (1995). "Detailed motifs for peptide binding to HLA-A*0201 derived from large random sets of peptides using a cellular binding assay." Hum. Immunol. **43**(1): 1-12.

D'Souza, S., Rosseels, V., Romano, M., Tanghe, A., Denis, O., Jurion, F., Castiglione, N., Vanonckelen, A., Palfliet, K. and Huygen, K. (2003). "Mapping of murine Th1 helper T-Cell epitopes of mycolyl transferases Ag85A, Ag85B, and Ag85C from Mycobacterium tuberculosis." Infect. Immun. **71**(1): 483-93.

Ducrot, P., Andrianjara, C. R. and Wrigglesworth, R. (2001). "CoMFA and CoMSIA 3D-quantitative structure-activity relationship model on benzodiazepine derivatives, inhibitors of phosphodiesterase IV." J Comput Aided Mol Des **15**(9): 767-85.

Duran, L. W. and Pease, L. R. (1986). "Relating the structure of major transplantation antigens to immune function." Transplantation **41**(3): 279-85.

Dustin, M. L. (2002). "The immunological synapse." Arthritis Res **4 Suppl 3**: S119-25.

Dustin, M. L., Bromley, S. K., Kan, Z., Peterson, D. A. and Unanue, E. R. (1997). "Antigen receptor engagement delivers a stop signal to migrating T lymphocytes." Proc. Natl. Acad. Sci. U. S. A. **94**(8): 3909-13.

Eddy, S. R. (1998). "Profile hidden Markov models." Bioinformatics **14**(9): 755-63.

Elliott, T. (1997). "Transporter associated with antigen processing." Adv. Immunol. **65**: 47-109.

Elvin, J., Cerundolo, V., Elliott, T. and Townsend, A. (1991). "A quantitative assay of peptide-dependent class I assembly." Eur. J. Immunol. **21**(9): 2025-31.

Emmerich, N. P., Nussbaum, A. K., Stevanovic, S., Priemer, M., Toes, R. E., Rammensee, H. G. and Schild, H. (2000). "The human 26 S and 20 S proteasomes generate overlapping but different sets of peptide fragments from a model protein substrate." J. Biol. Chem. **275**(28): 21140-8.

Engel, I. and Hedrick, S. M. (1988). "Site-directed mutations in the VDJ junctional region of a T cell receptor beta chain cause changes in antigenic peptide recognition." Cell **54**(4): 473-84.

Eriksson, L., Jonsson, J., Sjostrom, M. and Wold, S. (1989). "Multivariate parametrization of coded and non-coded amino acids by thin layer chromatography." Prog. Clin. Biol. Res. **291**: 131-4.

Eriksson, L., Jonsson, J., Hellberg, S., Lindgren, F., Skagerberg, B., Sjostrom, M. and Wold, S. (1990). "Peptide QSAR on substance P analogues, enkephalins and bradykinins containing L- and D-amino acids." Acta Chem. Scand. **44**(1): 50-5.

Fahnestock, M. L., Johnson, J. L., Feldman, R. M., Tsomides, T. J., Mayer, J., Narhi, L. O. and Bjorkman, P. J. (1994). "Effects of peptide length and composition on binding to an empty class I MHC heterodimer." Biochemistry **33**(26): 8149-58.

Falk, K., Rotzschke, O. and Rammensee, H. G. (1990). "Cellular peptide composition governed by major histocompatibility complex class I molecules." Nature **348**(6298): 248-51.

Falk, K., Rotzschke, 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**(2): 425-34.

Falk, K., Rotzschke, 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**(6324): 290-6.

Falk, K. and Rotzschke, O. (1993). "Consensus motifs and peptide ligands of MHC class I molecules." Semin Immunol **5**(2): 81-94.

Falk, K., Rotzschke, O., Takiguchi, M., Grahovac, B., Gnau, V., Stevanovic, S., Jung, G. and Rammensee, H. G. (1994). "Peptide motifs of HLA-A1, -A11, -A31, and -A33 molecules." Immunogenetics **40**(3): 238-41.

Fan, Q. R., Long, E. O. and Wiley, D. C. (2001). "Crystal structure of the human natural killer cell inhibitory receptor KIR2DL1-HLA-Cw4 complex." Nat. Immunol. **2**(5): 452-60.

Fauchere, J. L., Charton, M., Kier, L. B., Verloop, A. and Pliska, V. (1988). "Amino acid side chain parameters for correlation studies in biology and pharmacology." Int J Pept Protein Res **32**(4): 269-78.

Faustman, D. L. (1995). "Altered MHC class I expression: a role for transplantation and IDDM autoimmunity." Diabetes Metab Rev **11**(1): 1-19.

Feito, M. J., Jimenez-Perianez, A., Ojeda, G., Sanchez, A., Portoles, P. and Rojo, J. M. (2002). "The TCR/CD3 complex: molecular interactions in a changing structure." Arch Immunol Ther Exp (Warsz) **50**(4): 263-72.

Felipe-Sotelo, M., Andrade, J. M., Carlosena, A. and Prada, D. (2003). "Partial least squares multivariate regression as an alternative to handle interferences of Fe on the determination of trace Cr in water by electrothermal atomic absorption spectrometry." Anal. Chem. **75**(19): 5254-61.

Feltkamp, M. C., Smits, H. L., Vierboom, M. P., Minnaar, R. P., de Jongh, B. M., Drijfhout, J. W., ter Schegget, J., Melief, C. J. and Kast, W. M. (1993). "Vaccination with cytotoxic T lymphocyte epitope-containing peptide protects against a tumor induced by human papillomavirus type 16-transformed cells." Eur. J. Immunol. **23**(9): 2242-9.

Fernandez-Vina, M., Ramirez, L. C., Raskin, P. and Stastny, P. (1993). "Genes for insulin-dependent diabetes mellitus (IDDM) in the major histocompatibility complex (MHC) of African-Americans." Tissue Antigens **41**(2): 57-64.

Ferrari, G., Neal, W., Jones, A., Olender, N., Ottinger, J., Ha, R., McElrath, M. J., Goepfert, P. and Weinhold, K. J. (2001). "CD8 CTL responses in vaccines: emerging patterns of HLA restriction and epitope recognition." Immunol. Lett. **79**(1-2): 37-45.

Festenstein, H. and Ollier, B. (1987). "Cellular typing and functional heterogeneity of MHC-encoded products." Br. Med. Bull. **43**(1): 122-55.

Fields, B. A., Ober, B., Malchiodi, E. L., Lebedeva, M. I., Braden, B. C., Ysern, X., Kim, J. K., Shao, X., Ward, E. S. and Mariuzza, R. A. (1995). "Crystal structure of the V alpha domain of a T cell antigen receptor." Science **270**(5243): 1821-4.

Flad, T., Schiestel, T., Brunner, H., Tolson, J., Ouyang, Q., Pawelec, G., Mueller, G. A., Mueller, C. A., Tovar, G. E. and Beck, H. (2003). "Development of an MHC-class I peptide selection assay combining nanoparticle technology and matrix-assisted laser desorption/ionisation mass spectrometry." J. Immunol. Methods **283**(1-2): 205-13.

Fleckenstein, B., Kalbacher, H., Muller, C. P., Stoll, D., Halder, T., Jung, G. and Wiesmuller, K. H. (1996). "New ligands binding to the human leukocyte antigen class II molecule DRB1*0101 based on the activity pattern of an undecapeptide library." Eur. J. Biochem. **240**(1): 71-7.

Fleischhauer, K., Tanzarella, S., Wallny, H. J., Bordignon, C. and Traversari, C. (1996). "Multiple HLA-A alleles can present an immunodominant peptide of the human melanoma antigen Melan-A/MART-1 to a peptide-specific HLA-A*0201+ cytotoxic T cell line." J. Immunol. **157**(2): 787-97.

Free, S. M., Jr. and Wilson, J. W. (1964). "A Mathematical Contribution to Structure-Activity Studies." J. Med. Chem. **53**: 395-9.

Fremont, D. H., Matsumura, M., Stura, E. A., Peterson, P. A. and Wilson, I. A. (1992). "Crystal structures of two viral peptides in complex with murine MHC class I H-2Kb." Science **257**(5072): 919-27.

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**(6): 1293-301.

Fujita, T. and Ban, T. (1971). "Structure-activity study of phenethylamines as substrates of biosynthetic enzymes of sympathetic transmitters." J. Med. Chem. **14**(2): 148-52.

Gairin, J. E., Mazarguil, H., Hudrisier, D. and Oldstone, M. B. (1995). "Optimal lymphocytic choriomeningitis virus sequences restricted by H-2Db major histocompatibility complex class I molecules and presented to cytotoxic T lymphocytes." J. Virol. **69**(4): 2297-305.

Gallego, R. G., Blanco, J. L., Thijssen-van Zuylen, C. W., Gotfredsen, C. H., Voshol, H., Duus, J. O., Schachner, M. and Vliegenthart, J. F. (2001). "Epitope diversity of N-glycans from bovine peripheral myelin glycoprotein P0 revealed by mass spectrometry and nano probe magic angle spinning 1H NMR spectroscopy." J. Biol. Chem. **276**(33): 30834-44.

Gao, G. F., Tormo, J., Gerth, U. C., Wyer, J. R., McMichael, A. J., Stuart, D. I., Bell, J. I., Jones, E. Y. and Jakobsen, B. K. (1997). "Crystal structure of the complex between human CD8alpha(alpha) and HLA-A2." Nature **387**(6633): 630-4.

Garboczi, D. N., Ghosh, P., Utz, U., Fan, Q. R., Biddison, W. E. and Wiley, D. C. (1996). "Structure of the complex between human T-cell receptor, viral peptide and HLA-A2." Nature **384**(6605): 134-41.

Garcia, K. C., Degano, M., Stanfield, R. L., Brunmark, A., Jackson, M. R., Peterson, P. A., Teyton, L. and Wilson, I. A. (1996a). "An alphabeta T cell receptor structure at 2.5 A and its orientation in the TCR-MHC complex." Science **274**(5285): 209-19.

Garcia, K. C., Scott, C. A., Brunmark, A., Carbone, F. R., Peterson, P. A., Wilson, I. A. and Teyton, L. (1996b). "CD8 enhances formation of stable T-cell receptor/MHC class I molecule complexes." Nature **384**(6609): 577-81.

Garrett, T. P., 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.

Gavioli, R., Kurilla, M. G., de Campos-Lima, P. O., Wallace, L. E., Dolcetti, R., Murray, R. J., Rickinson, A. B. and Masucci, M. G. (1993). "Multiple HLA A11-restricted cytotoxic T-lymphocyte epitopes of different immunogenicities in the Epstein-Barr virus-encoded nuclear antigen 4." J. Virol. **67**(3): 1572-8.

Geffrotin, C., Popescu, C. P., Cribiu, E. P., Boscher, J., Renard, C., Chardon, P. and Vaiman, M. (1984). "Assignment of MHC in swine to chromosome 7 by in situ hybridization and serological typing." Ann. Genet. **27**(4): 213-9.

Ghendler, Y., Teng, M. K., Liu, J. H., Witte, T., Liu, J., Kim, K. S., Kern, P., Chang, H. C., Wang, J. H. and Reinherz, E. L. (1998). "Differential thymic selection outcomes stimulated by focal structural alteration in peptide/major histocompatibility complex ligands." Proc Natl Acad Sci U S A. **95**: 10061-10066.

Ginhoux, F., Doucet, C., Leboeuf, M., Lemonnier, F., Danos, O., Davoust, J. and Firat, H. (2003). "Identification of an HLA-A*0201-restricted epitopic peptide from human eystrophin: application in Duchenne muscular dystrophy gene therapy." Molecular Therapy **8**: 274-283.

Gladman, D. D., Terasaki, P. I., Park, M. S., Iwaki, Y., Louie, S., Quismorio, F. P., Barnett, E. V. and Liebling, M. R. (1979). "Increased frequency of HLA-DRW2 in SLE." Lancet **2**(8148): 902.

Glazko, G. V. and Mushegian, A. R. (2004). "Detection of evolutionarily stable fragments of cellular pathways by hierarchical clustering of phyletic patterns." Genome Biol **5**(5): R32.

Goldstein, J. S., Chen, T., Gubina, E., Pastor, R. W. and Kozlowski, S. (2000). "ICAM-1 enhances MHC-peptide activation of CD8(+) T cells without an organized immunological synapse." Eur. J. Immunol. **30**(11): 3266-70.

Goodford, P. J. (1985). "A computational procedure for determining energetically favorable binding sites on biologically important macromolecules." J. Med. Chem. **28**(7): 849-57.

Gopalakrishnan, B. and Roques, B. P. (1992). "Do antigenic peptides have a unique sense of direction inside the MHC binding groove? A molecular modelling study." FEBS Lett. **303**(2-3): 224-8.

Gorer, P. A. (1936). "The detection of antigenic differences in mouse erythrocytes by the employment of immune sera." Br. J. Exp. Pathol. **17**: 42-50.

Gorer, P. A. (1937). "The genetic and antigenic basis for tumor transplantation." J. Pathol. Bacteriol. **44**: 691-697.

Gouaillard, C., Huchenq-Champagne, A., Arnaud, J., Chen Cl, C. L. and Rubin, B. (2001). "Evolution of T cell receptor (TCR) alpha beta heterodimer assembly with the CD3 complex." Eur. J. Immunol. **31**(12): 3798-805.

Goulder, P. J., Edwards, A., Phillips, R. E. and McMichael, A. J. (1997). "Identification of a novel HLA-B*3501-restricted cytotoxic T lymphocyte epitope using overlapping peptides." Aids **11**(7): 930-2.

Goust, J. M. (1993). "Major histocompatibility complex." Immunol Ser **58**: 29-48.

Graff, R. J., Mann, D. L. and Nathenson, S. G. (1970). "Immunogenic properties of papain-solubilized H-2 alloantigens." Transplantation **10**(1): 59-65.

Grakoui, A., Bromley, S. K., Sumen, C., Davis, M. M., Shaw, A. S., Allen, P. M. and Dustin, M. L. (1999). "The immunological synapse: a molecular machine controlling T cell activation." Science **285**(5425): 221-7.

Greenwood, R., Wang, B., Midkiff, K., White, G., Lin, H. and Frelinger, A. (2003). "Identification of T-cell epitopes in clotting factor IX and lack of tolerance in inbred mice." Journal of Thrombosis and Haemostasis **1**: 95-102.

Grunewald, G. L., Caldwell, T. M., Dahanukar, V. H., Jalluri, R. K. and Criscione, K. R. (1999). "Comparative molecular field analysis (CoMFA) models of phenylethanolamine N-methyltransferase (PNMT) and the alpha2-adrenoceptor: the development of new, highly selective inhibitors of PNMT." Bioorg. Med. Chem. Lett. **9**(3): 481-6.

Guan, P., Doytchinova, I. A. and Flower, D. R. (2003). "HLA-A3 supermotif defined by quantitative structure-activity relationship analysis." Protein Eng. **16**(1): 11-8.

Guan, P., Doytchinova, I. A. and Flower, D. R. (2003a). "A comparative molecular similarity indices (CoMSIA) study of peptide binding to the HLA-A3 superfamily." Bioorg. Med. Chem. **11**(10): 2307-11.

Guan, P., Doytchinova, I. A. and Flower, D. R. (2003b). "HLA-A3 supermotif defined by quantitative structure-activity relationship analysis." Protein Eng. **16**(1): 11-8.

Guan, P., Doytchinova, I. A., Zygouri, C. and Flower, D. R. (2003c). "MHCPred: A server for quantitative prediction of peptide-MHC binding." Nucleic Acids Res. **31**(13): 3621-4.

Guan, P., Doytchinova, I. A., Zygouri, C. and Flower, D. R. (2003d). "MHCPred: bringing a quantitative dimension to the online prediction of MHC binding." Appl Bioinformatics **2**(1): 63-6.

Guess, M. J. and Wilson, S. B. (2002). "Introduction to hierarchical clustering." J Clin Neurophysiol **19**(2): 144-51.

Gulukota, K., Sidney, J., Sette, A. and DeLisi, C. (1997). "Two complementary methods for predicting peptides binding major histocompatibility complex molecules." J. Mol. Biol. **267**(5): 1258-67.

Guo, H. C., Madden, D. R., Silver, M. L., Jardetzky, T. S., Gorga, J. C., Strominger, J. L. and Wiley, D. C. (1993). "Comparison of the P2 specificity pocket in three human histocompatibility antigens: HLA-A*6801, HLA-A*0201, and HLA-B*2705." Proc. Natl. Acad. Sci. U. S. A. **90**(17): 8053-7.

Gupta, M. K., Mishra, P., Prathipati, P. and Saxena, A. K. (2002). "2D-QSAR in hydroxamic acid derivatives as peptide deformylase inhibitors and antibacterial agents." Bioorg. Med. Chem. **10**(12): 3713-6.

Gussow, D., Rein, R., Ginjaar, I., Hochstenbach, F., Seemann, G., Kottman, A. and Ploegh, H. L. (1987). "The human beta 2-microglobulin gene. Primary structure and definition of the transcriptional unit." J. Immunol. **139**(9): 3132-8.

Haeney, M. (1995). "The immunological background to transplantation." J. Antimicrob. Chemother. **36 Suppl B**: 1-9.

Hakenberg, J., Nussbaum, A. K., Schild, H., Rammensee, H. G., Kuttler, C., Holzhutter, H. G., Kloetzel, P. M., Kaufmann, S. H. and Mollenkopf, H. J. (2003). "MAPPP: MHC class I antigenic peptide processing prediction." Appl Bioinformatics **2**(3): 155-8.

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**(6): 2353-8.

Hampl, J., Schild, H., Litzenberger, C., Baron, M., Crowley, M. P. and Chien, Y. H. (1999). "The specificity of a weak gamma delta TCR interaction can be modulated by the glycosylation of the ligand." J. Immunol. **163**(1): 288-94.

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

Hancock, G. E., Tebbey, P. W., Scheuer, C. A., Pryharski, K. S., Heers, K. M. and LaPierre, N. A. (2003). "Immune responses to the nonglycosylated ectodomain of respiratory syncytial virus attachment glycoprotein mediate pulmonary eosinophilia in inbred strains of mice with different MHC haplotypes." J. Med. Virol. **70**(2): 301-8.

Hanke, T., Schneider, J., Gilbert, S. C., Hill, A. V. and McMichael, A. (1998). "DNA multi-CTL epitope vaccines for HIV and Plasmodium falciparum: immunogenicity in mice." Vaccine **16**(4): 426-35.

Hansch, C. and Fujita, T. (1963). "The correlation of biological activity of plant growth regulators and chloromycetin derivatives with hammett constants and partition coefficients." J. Amer. Chem. Soc. **82**: 2817-2824.

Hansch, C. and Fujita, T. (1964). "A method for the correlation of biological activity and chemical structure." J. Amer. Chem. Sco. **86**: 1616-1626.

Hansch, C. (1969). "A quantitative approach to biological structure-activity relationships." Accounts of Chemical Research **2**: 232-239.

Hansson, L., Rabbani, H., Fagerberg, J., Osterborg, A. and Mellstedt, H. (2003). "T-cell epitopes within the complementarity-determining and framework regions of the tumor-derived immunoglobulin heavy chain in multiple myeloma." Blood **101**(12): 4930-6.

Harada, M., Gohara, R., Matsueda, S., Muto, A., Oda, T., Iwamoto, Y. and Itoh, K. (2004). "In vivo evidence that peptide vaccination can induce HLA-DR-restricted CD4+ T cells reactive to a class I tumor peptide." J. Immunol. **172**(4): 2659-67.

Hasegawa, K. and Funatsu, K. (2000). "Partial least squares modeling and genetic algorithm optimization in quantitative structure-activity relationships." SAR QSAR Environ Res **11**(3-4): 189-209.

Hattotuwagama, C. K., Guan, P., Doytchinova, I. A., Zygouri, C. and Flower, D. R. (2004). "Quantitative online prediction of peptide binding to the major histocompatibility complex." J. Mol. Graph. Model. **22**(3): 195-207.

Hauptmann, G. and Bahram, S. (2004). "Genetics of the central MHC." Curr. Opin. Immunol. **16**(5): 668-72.

Heinermeyer, W., Fischer, M., Krimmer, T., Stachon, U. and Wolf, D. (1997). "The active sites of the eukaryotic 20S proteasome and their involvement in sub-unit precursor processing." J. Biol. Chem. **272**: 25200-25209.

Hellberg, S., Sjostrom, M., Skagerberg, B. and Wold, S. (1987). "Peptide quantitative structure-activity relationships, a multivariate approach." J. Med. Chem. **30**(7): 1126-35.

Hellstrom, K. E. and Hellstrom, I. (2003). "Novel approaches to therapeutic cancer vaccines." Expert Rev Vaccines **2**(4): 517-32.

Heukamp, L. C., van der Burg, S. H., Drijfhout, J. W., Melief, C. J., Taylor-Papadimitriou, J. and Offringa, R. (2001). "Identification of three non-VNTR MUC-1derived HLA-A*0201-restricted T-cell epitopes that induce protective anti-tumor immunity in HLA-A2/Kb-transgenic mice." Int. J. Cancer **91**: 385-392.

Hewitt, E. W., Gupta, S. S. and Lehner, P. J. (2001). "The human cytomegalovirus gene product US6 inhibits ATP binding by TAP." Embo J **20**(3): 387-96.

Higgins, C. F. (1992). "ABC transporters: from microorganisms to man." Annu Rev Cell Biol **8**: 67-113.

Hill, A. V. (1998). "The immunogenetics of human infectious diseases." Annu. Rev. Immunol. **16**: 593-617.

Hill, A. V., Allsopp, C. E., Kwiatkowski, D., Anstey, N. M., Twumasi, P., Rowe, P. A., Bennett, S., Brewster, D., McMichael, A. J. and Greenwood, B. M. (1991). "Common west African HLA antigens are associated with protection from severe malaria." Nature **352**(6336): 595-600.

Hillig, R. C., Coulie, P. G., Stroobant, V., Saenger, W., Ziegler, A. and Hulsmeyer, M. (2001). "High-resolution structure of HLA-A*0201 in complex with a tumour-specific antigenic peptide encoded by the MAGE-A4 gene." J. Mol. Biol. **310**(5): 1167-76.

Holland, J. H. (1975). Adaptation in natural and artificial systems. Ann Arbor, University of Michigan Press.

Holt, A., Wieland, B. and Baker, G. B. (2004). "Allosteric modulation of semicarbazide-sensitive amine oxidase activities in vitro by imidazoline receptor ligands." Br. J. Pharmacol.

Holzhutter, H. G., Frommel, C. and Kloetzel, P. M. (1999). "A theoretical approach towards the identification of cleavage-determining amino acid motifs of the 20 S proteasome." J. Mol. Biol. **286**: 1251-1265.

Hombach, A., Sent, D., Schneider, C., Heuser, C., Koch, D., Pohl, C., Seliger, B. and Abken, H. (2001). "T-cell activation by recombinant receptors: CD28 costimulation is required for interleukin 2 secretion and receptor-mediated T-cell proliferation but does not affect receptor-mediated target cell lysis." Cancer Res. **61**(5): 1976-82.

Honeyman, M. C., Brusic, V., Stone, N. L. and Harrison, L. C. (1998). "Neural network-based prediction of candidate T-cell epitopes." Nat. Biotechnol. **16**(10): 966-9.

Honjo, K., Xu, X. and Bucy, R. (2000). "Heterogeneity of T cell clones specific for a single indirect alloantigenic epitope (I-Ab/H-2Kd 54-68) that mediate transplant rejection." Transplantation **70**: 1516-1524.

Hosoyama, H., Obata, A., Bando, N., Tsuji, H. and Ogawa, T. (1996). "Epitope analysis of soybean major allergen Gly m Bd 30K recognized by the mouse monoclonal antibody using overlapping peptides." Biosci. Biotechnol. Biochem. **60**(7): 1181-2.

Hudecz, F. (2001). "Manipulation of epitope function by modification of peptide structure: a minireview." Biologicals **29**(3-4): 197-207.

Hudrisier, D., Mazarguil, H., Laval, F., Oldstone, M. B. A. and Gairin, J. E. (1996). "Binding of viral antigens to major histocompatibility complex class I H-2Db molecules is controlled by dominant negative elements at peptide non-anchor residues." J. Biol. Chem. **271**: 17829-17836.

Hudrisier, D., Mazarguil, H., Oldstone, M. B. and Gairin, J. E. (1995). "Relative implication of peptide residues in binding to major histocompatibility complex class I H-2Db: application to the design of high-affinity, allele-specific peptides." Mol. Immunol. **32**(12): 895-907.

Hulo, N., Sigrist, C. J., Le Saux, V., Langendijk-Genevaux, P. S., Bordoli, L., Gattiker, A., De Castro, E., Bucher, P. and Bairoch, A. (2004). "Recent improvements to the PROSITE database." Nucleic Acids Res. **32 Database issue**: D134-7.

Hung, C. F., Hsu, K. F., Cheng, W. F., Chai, C. Y., He, L., Ling, M. and Wu, T. C. (2001). "Enhancement of DNA vaccine potency by linkage of antigen gene to a gene encoding the extracellular domain of Fms-like tyrosine kinase 3-ligand." Cancer Res. **61**(3): 1080-8.

Hunt, D. F., Yates, J. R., 3rd, Shabanowitz, J., Winston, S. and Hauer, C. R. (1986). "Protein sequencing by tandem mass spectrometry." Proc. Natl. Acad. Sci. U. S. A. **83**(17): 6233-7.

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**(5049): 1261-3.

Hunt, P. A. (1999). "QSAR using 2D descriptors and TRIPOS' SIMCA." J Comput Aided Mol Des **13**(5): 453-67.

Imbert, V., Farahifar, D., Auberger, P., Mary, D., Rossi, B. and Peyron, J. F. (1996). "Stimulation of the T-cell antigen receptor-CD3 complex signaling pathway by the tyrosine phosphatase inhibitor pervanadate is mediated by inhibition of CD45: evidence for two interconnected Lck/Fyn- or zap-70-dependent signaling pathways." J Inflamm **46**(2): 65-77.

Imro, M. A., Manici, S., Russo, V., Consogno, G., Bellone, M., Rugarli, C., Traversari, C. and Protti, M. P. (1999). "Major histocompatibility complex class I restricted cytotoxic T cells specific for natural melanoma peptides recognize unidentified shared melanoma antigen(s)." Cancer Res. **59**(10): 2287-91.

Inoue, M. and Kajiya, F. (1976). "[Multivariate analysis in computer diagnosis. 3. Principal component analysis]." Iyodenshi To Seitai Kogaku **14**(1): 52-7.

Ishioka, G. Y., Fikes, J., Hermanson, G., Livingston, B., Crimi, C., Qin, M., del Guercio, M. F., Oseroff, C., Dahlberg, C., Alexander, J., Chesnut, R. W. and Sette, A. (1999). "Utilization of MHC class I transgenic mice for development of minigene DNA vaccines encoding multiple HLA-restricted CTL epitopes." J. Immunol. **162**(7): 3915-25.

Islam, M. N., Song, Y. and Iskander, M. N. (2003). "Investigation of structural requirements of anticancer activity at the paclitaxel/tubulin binding site using CoMFA and CoMSIA." J. Mol. Graph. Model. **21**(4): 263-72.

Jacob, C. O., Leitner, M., Zamir, A., Salomon, D. and Arnon, R. (1985). "Priming immunization against cholera toxin and E. coli heat-labile toxin by a cholera toxin short peptide-beta-galactosidase hybrid synthesized in E. coli."

Embo J **4**(12): 3339-43.

Jager, E., Gnjatic, S., Nagata, Y., stockert, E., Jager, D., Karbach, J., Neumann, A., Rieckenberg, J., Chen, Y., Ritter, G., Hoffman, E., Arand, M., Old, L. and Knuth, A. (2000). "Induction of primary NY-ESO-1 immunity: CD8+ T lymphocyte and antibody responses in peptide-vaccinated patients with NY-ESO-1+ cancers." Proc. Natl. Acad. Sci. U. S. A. **97**: 12198-12203.

Jager, E., Jager, D. and Knuth, A. (2002). "Clinical cancer vaccine trials." Curr. Opin. Immunol. **14**(2): 178-82.

Jahn-Schmid, B., Kelemen, P., Himly, M., Bohle, B., Fischer, G., Ferreira, F. and Ebner, C. (2002). "The T cell response to art v 1, the major mugwort pollen allergen, is dominated by one epitope." J. Immunol. **169**: 6005-6011.

Jakobsen, I. B., Gao, X., Easteal, S. and Chelvanayagam, G. (1998). "Correlating sequence variation with HLA-A allelic families: implications for T cell receptor binding specificities." Immunol. Cell Biol. **76**(2): 135-42.

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**(10): 2663-7.

Janeway, C., A. (2001). Immunobiology 5 : the immune system in health and disease. New York, Garland ; Edinburgh : Churchill Livingstone c2001.

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**(6342): 326-9.

Jazwinska, E. C., Dunckley, H., Gatenby, P. A. and Serjeantson, S. W. (1989). "Gm allogenotype distribution and lack of HLA-DR, Gm interaction in SLE and CREST/PSS." Immunol. Cell Biol. **67 (Pt 4)**: 261-5.

Jiang, S., Borthwick, N. J., Morrison, P., Gao, G. F. and Steward, M. W. (2002). "Virus-specific CTL responses induced by an H-2K(d)-restricted, motif-negative 15-mer peptide from the fusion protein of respiratory syncytial virus." J. Gen. Virol. **83**(Pt 2): 429-38.

Johansen, T. E., McCullough, K., Catipovic, B., Su, X. M., Amzel, M. and Schneck, J. P. (1997). "Peptide binding to MHC class I is determined by individual pockets in the binding groove." Scand. J. Immunol. **46**(2): 137-46.

Johnson, S. C. (1967). "Hierarchical clustering schemes." Psychometrika **32**(3): 241-54.

Jojic, V., Jojic, N., Meek, C., Geiger, D., Siepel, A., Haussler, D. and Heckerman, D. (2004). "Efficient approximations for learning phylogenetic HMM models from data." Bioinformatics **20 Suppl 1**: I161-I168.

Jones, E. Y. (1997). "MHC class I and class II structures." Curr. Opin. Immunol. **9**(1): 75-9.

Jouyban, A., Majidi, M. R., Jalilzadeh, H. and Asadpour-Zeynali, K. (2004). "Modeling drug solubility in water-cosolvent mixtures using an artificial neural network." Farmaco **59**(6): 505-12.

Joyce, S. and Nathenson, S. G. (1994). "Methods to study peptides associated with MHC class I molecules." Curr. Opin. Immunol. **6**(1): 24-31.

Jung, G., Fleckenstein, B., von der Mulbe, F., Wessels, J., Niethammer, D. and Wiesmuller, K. H. (2001). "From combinatorial libraries to MHC ligand motifs, T-cell superagonists and antagonists." Biologicals **29**(3-4): 179-81.

Kambayashi, T., Assarsson, E., Chambers, B. J. and Ljunggren, H. G. (2001). "IL-2 down-regulates the expression of TCR and TCR-associated surface molecules on CD8(+) T cells." Eur. J. Immunol. **31**(11): 3248-54.

Kampf, D., Malchus, R., Alexander, M. and Hoppe, I. (1979). "HLA-antigens in systemic lupus erythematosus (SLE)." Arch. Dermatol. Res. **264**(3): 345-50.

Kangueane, P., Sakharkar, M. K., Lim, K. S., Hao, H., Lin, K., Chee, R. E. and Kolatkar, P. R. (2000). "Knowledge-based grouping of modeled HLA peptide complexes." Hum. Immunol. **61**(5): 460-6.

Kaslow, R. A., Carrington, M., Apple, R., Park, L., Munoz, A., Saah, A. J., Goedert, J. J., Winkler, C., O'Brien, S. J., Rinaldo, C., Detels, R., Blattner, W., Phair, J., Erlich, H. and Mann, D. L. (1996). "Influence of combinations of human major histocompatibility complex genes on the course of HIV-1 infection." Nat. Med. **2**(4): 405-11.

Kast, W. M. and Melief, C. J. (1991). "Fine peptide specificity of cytotoxic T lymphocytes directed against adenovirus-induced tumours and peptide-MHC binding." Int. J. Cancer Suppl. **6**: 90-4.

Kast, W. M., Brandt, R. M., Sidney, J., Drijfhout, J. W., Kubo, R. T., Grey, H. M., Melief, C. J. and Sette, A. (1994). "Role of HLA-A motifs in identification of potential CTL epitopes in human papillomavirus type 16 E6 and E7 proteins." J Immunol **152**(8): 3904-12.

Kastenholz, M. A., Pastor, M., Cruciani, G., Haaksma, E. E. and Fox, T. (2000). "GRID/CPCA: a new computational tool to design selective ligands." J. Med. Chem. **43**(16): 3033-44.

Kawashima, I., Hudson, S. J., Tsai, V., Southwood, S., Takesako, K., Appella, E., Sette, A. and Celis, E. (1998). "The multi-epitope approach for immunotherapy for cancer: identification of several CTL epitopes from various tumor-associated antigens expressed on solid epithelial tumors." Hum. Immunol. **59**(1): 1-14.

Kawashima, I., Tsai, V., Southwood, S., Takesako, K., Sette, A. and Celis, E. (1999). "Identification of HLA-A3-restricted cytotoxic T lymphocyte epitopes from carcinoembryonic antigen and HER-2/neu by primary in vitro immunization with peptide-pulsed dendritic cells." Cancer Res. **59**(2): 431-5.

Kawashima, S. and Kanehisa, M. (2000). "AAindex: amino acid index database." Nucleic Acids Res. **28**(1): 374.

Kelly, H., McCann, V. J., Kay, P. H. and Dawkins, R. L. (1985). "Susceptibility to IDDM is marked by MHC supratypes rather than individual alleles." Immunogenetics **22**(6): 643-51.

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.

Kesmir, C., Nussbaum, A. K., Schild, H., Detours, V. and Brunak, S. (2002). "Prediction of proteasome cleavage motifs by neural networks." Protein Eng. **15**(4): 287-96.

Khan, A. R., Baker, B. M., Ghosh, P., Biddison, W. E. and Wiley, D. C. (2000). "The structure and stability of an HLA-A*0201/octameric tax peptide complex with an empty conserved peptide-N-terminal binding site." J. Immunol. **164**(12): 6398-405.

Khandelwal, A., Narayanan, R. and Gopalakrishnan, B. (2003). "3-D-QSAR CoMFA and CoMSIA studies on tetrahydrofuroyl-L-phenylalanine derivatives as VLA-4 antagonists." Bioorg. Med. Chem. **11**(19): 4235-44.

Khanna, R., Burrows, S. R., Moss, D. J. and Silins, S. L. (1996). "Peptide transporter (TAP-1 and TAP-2)-independent endogenous processing of Epstein-Barr virus (EBV) latent membrane protein 2A: implications for cytotoxic T-lymphocyte control of EBV-associated malignancies." J. Virol. **70**(8): 5357-62.

Khanna, R., Burrows, S. R., Neisig, A., Neefjes, J., Moss, D. J. and Silins, S. L. (1997). "Hierarchy of Epstein-Barr virus-specific cytotoxic T-cell responses in individuals carrying different subtypes of an HLA allele: implications for epitope-based antiviral vaccines." J. Virol. **71**(10): 7429-35.

Khong, H. T. and Rosenberg, S. A. (2002). "The Waardenburg syndrome type 4 gene, SOX10, is a novel tumor-associated antigen identified in a patient with a dramatic response to immunotherapy." Cancer Res. **62**(11): 3020-3.

Kidera, A., Konishi, Y., POka, M., Ooi, T. and Scheraga, H. A. (1985). "Statistical analysis of the physical properties of the 10 naturally occurring amino acids." J. Protein Chem. **4**: 23-55.

Klebe, G., Abraham, U. and Mietzner, T. (1994). "Molecular similarity indices in a comparative analysis (CoMSIA) of drug molecules to correlate and predict their biological activity." J. Med. Chem. **37**(24): 4130-46.

Klebe, G. (1998). "Comparative molecular similarity indices analysis: CoMSIA." Perspect. Drug. Disc. Design **12-14**: 87-104.

Klebe, G. and Abraham, U. (1999). "Comparative molecular similarity index analysis (CoMSIA) to study hydrogen-bonding properties and to score combinatorial libraries." J Comput Aided Mol Des **13**(1): 1-10.

Kloetzel, P. M. and Ossendorp, F. (2004). "Proteasome and peptidase function in MHC-class-I-mediated antigen presentation." Curr. Opin. Immunol. **16**(1): 76-81.

Kockum, I., Wassmuth, R., Holmberg, E., Michelsen, B. and Lernmark, A. (1994). "Inheritance of MHC class II genes in IDDM studied in population-based affected and control families." Diabetologia **37**(11): 1105-12.

Koeller, D. and Ozato, K. (1986). "Evaluation of the structure function relationships of MHC class I antigens by molecular genetic techniques." Year Immunol **2**: 195-204.

Konnai, S., Nagaoka, Y., Takesima, S., Onuma, M. and Aida, Y. (2003). "Technical note: DNA typing for ovine MHC DRB1 using polymerase chain reaction-restriction fragment length polymorphism (PCR-RFLP)." J. Dairy Sci. **86**(10): 3362-5.

Kono, K., Rongcun, Y., Charo, J., Ichihara, F., Celis, E., Sette, A., Appella, E., Sekikawa, T., Matsumoto, Y. and Kiessling, R. (1998). "Identification of HER2/neu-derived peptide epitopes recognized by gastric cancer-specific cytotoxic T lymphocytes." Int. J. Cancer **78**(2): 202-8.

Koppers-Lalic, D., Rychlowski, M., van Leeuwen, D., Rijsewijk, F. A., Ressing, M. E., Neefjes, J. J., Bienkowska-Szewczyk, K. and Wiertz, E. J. (2003). "Bovine herpesvirus 1 interferes with TAP-dependent peptide transport and intracellular trafficking of MHC class I molecules in human cells." Arch. Virol. **148**(10): 2023-37.

Korangy, F., Ormandy, L. A., Bleck, J. S., Klempnauer, J., Wilkens, L., Manns, M. P. and Greten, T. F. (2004). "Spontaneous tumor-specific humoral and cellular immune responses to NY-ESO-1 in hepatocellular carcinoma." Clin

Cancer Res **10**(13): 4332-41.

Koretzky, G. A., Kohmetscher, M. and Ross, S. (1993). "CD45-associated kinase activity requires lck but not T cell receptor expression in the Jurkat T cell line." J. Biol. Chem. **268**(12): 8958-64.

Koskimies, S. and Eklund, B. (1997). "MHC genes and histocompatibility. A review." Ann Chir Gynaecol **86**(2): 171-9.

Krangel, M. S., McMurry, M. T., Hernandez-Munain, C., Zhong, X. P. and Carabana, J. (2000). "Accessibility control of T cell receptor gene rearrangement in developing thymocytes. The TCR alpha/delta locus." Immunol. Res. **22**(2-3): 127-35.

Krensky, A. M. and Clayberger, C. (1996). "Structure of HLA molecules and immunosuppressive effects of HLA derived peptides." Int Rev Immunol **13**(3): 173-85.

Krogh, A., Brown, M., Mian, I. S., Sjolander, K. and Haussler, D. (1994). "Hidden Markov models in computational biology. Applications to protein modeling." J. Mol. Biol. **235**(5): 1501-31.

Kruger, S., Schroers, R., Rooney, C., Gahn, B. and Chen, S. (2003). "Identification of a naturally processed HLA-DR-restricted T-helper epitope in Epstein-Barr virus nuclear antigen type 1." J. Immunother. **26**: 212-221.

Kubayashi, H., Omiya, R., Ruiz, M., Huarte, E., Sarobe, P., Lasarte, J., Herraiz, M., Sangro, B., Prieto, J., Borras-Cuesta, F. and Celis, E. (2002). "Identification of an antigenic epitope for helper T lymphocytes from carcinoembryonic antigen." Clin. Cancer Res. **8**: 3219-3225.

Kubinyi, H. and Kehrhahn, O. H. (1976). "Quantitative structure-activity relationships. 3.1 A comparison of different Free-Wilson models." J. Med. Chem. **19**(8): 1040-9.

Kubo, R. T., Sette, A., Grey, H. M., Appella, E., Sakaguchi, K., Zhu, N. Z., Arnott, D., Sherman, N., Shabanowitz, J., Michel, H. and et al. (1994). "Definition of specific peptide motifs for four major HLA-A alleles." J. Immunol. **152**(8): 3913-24.

Kunick, C., Lauenroth, K., Wieking, K., Xie, X., Schultz, C., Gussio, R., Zaharevitz, D., Leost, M., Meijer, L., Weber, A., Jorgensen, F. S. and Lemcke, T. (2004). "Evaluation and comparison of 3D-QSAR CoMSIA models for CDK1, CDK5, and GSK-3 inhibition by paullones." J. Med. Chem. **47**(1): 22-36.

Kuo, C. L., Assefa, H., Kamath, S., Brzozowski, Z., Slawinski, J., Saczewski, F., Buolamwini, J. K. and Neamati, N. (2004). "Application of CoMFA and CoMSIA 3D-QSAR and docking studies in optimization of mercaptobenzenesulfonamides as HIV-1 integrase inhibitors." J. Med. Chem. **47**(2): 385-99.

Kurata, A. and Berzofsky, J. A. (1990). "Analysis of peptide residues interacting with MHC molecule or T cell receptor. Can a peptide bind in more than one way to the same MHC molecule?" J. Immunol. **144**(12): 4526-35.

Kurokohchi, K., Akatsuka, T., Pendleton, C. D., Takamizawa, A., Nishioka, M., Battegay, M., Feinstone, S. M. and Berzofsky, J. A. (1996). "Use of recombinant protein to identify a motif-negative human cytotoxic T-cell epitope presented by HLA-A2 in the hepatitis C virus NS3 region." J. Virol. **70**(1): 232-40.

Kuttler, C., Nussbaum, A. K., Dick, T. P., Rammensee, H. G., Schild, H. and Hadeler, K. P. (2000). "An algorithm for the prediction of proteasomal cleavages." J. Mol. Biol. **298**(3): 417-29.

Lamas, J. R., Brooks, J. M., Galocha, B., Rickinson, A. B. and Lopez de Castro, J. A. (1998). "Relationship between peptide binding and T cell epitope selection: a study with subtypes of HLA-B27." Int. Immunol. **10**(3): 259-66.

Lauemoller, S. L., Holm, A., Hilden, J., Brunak, S., Holst Nissen, M., Stryhn, A., Ostergaard Pedersen, L. and Buus, S. (2001). "Quantitative predictions of peptide binding to MHC class I molecules using specificity matrices and anchor-stratified calibrations." Tissue Antigens **57**(5): 405-14.

Lauer, G., Ouchi, K., Chung, R., Nguyen, N., Duy, C., Purkis, D., Reiser, M., Kim, A., Lucas, M., Klenerman, P. and Walker, B. (2002). "Comprehensive analysis of CD8+-T-cell responses against hepatitis C virus reveals multiple unpredicted specificities." J. Virol. **76**: 6104-6113.

Lawlor, D. A., Warren, E., Taylor, P. and Parham, P. (1991). "Gorilla class I major histocompatibility complex alleles: comparison to human and chimpanzee class I." J. Exp. Med. **174**(6): 1491-509.

Lee, K. O., Park, H. J., Kim, Y. H., Seo, S. Y., Lee, Y. S., Moon, S. H., Kim, N. J., Choi, N. S. and Suh, Y. G. (2004). "CoMFA and CoMSIA 3D QSAR studies on pimarane cyclooxygenase-2 (COX-2) inhibitors." Arch Pharm Res **27**(5): 467-70.

Lehner, T., Walker, P., Smerdon, R., Childerstone, A., Bergmeier, L. A. and Haron, J. (1990). "Identification of T- and B-cell epitopes in synthetic peptides derived from a Streptococcus mutans protein and characterization of their antigenicity and immunogenicity." Arch. Oral Biol. **35 Suppl**: 39S-45S.

Letvin, N. L. and Walker, B. D. (2001). "HIV versus the immune system: another apparent victory for the virus." J Clin Invest **107**(3): 273-5.

Levenstien, M. A., Yang, Y. and Ott, J. (2003). "Statistical significance for hierarchical clustering in genetic association and microarray expression studies." BMC Bioinformatics **4**(1): 62.

Levitsky, V., Liu, D., Southwood, S., Levitskaya, J., Sette, A. and Masucci, M. G. (2000). "Supermotif peptide binding and degeneracy of MHC: peptide recognition in an EBV peptide-specific CTL response with highly restricted TCR usage." Hum. Immunol. **61**(10): 972-84.

Li, A. H., Moro, S., Forsyth, N., Melman, N., Ji, X. D. and Jacobson, K. A. (1999). "Synthesis, CoMFA analysis, and receptor docking of 3,5-diacyl-2, 4-dialkylpyridine derivatives as selective A3 adenosine receptor antagonists." J. Med. Chem. **42**(4): 706-21.

Lilly, F., Boyse, E. A. and Old, L. J. (1964). "Genetic Basis of Susceptibility to Viral Leukaemogenesis." Lancet **14**: 1207-9.

Lilly, F. (1971). "The influence of H-2 type on Gross virus leukemogenesis in mice." Transplant Proc. **3**: 1239-1241.

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**(2): 221-30.

Lin, W., Yuan, X., Yuen, P., Wei, W. I., Sham, J., Shi, P. and Qu, J. (2004). "Classification of in vivo autofluorescence spectra using support vector machines." J Biomed Opt **9**(1): 180-6.

Little, C. C. and Tyzzer, E. E. (1916). "Further experimental studies on the inheritance of susceptibility to a transplantable tumor, carcinoma (JWA) of the Japanese waltzing mouse." J. Med. Res. **33**(393-453).

Liu, H. X., Zhang, R. S., Yao, X. J., Liu, M. C., Hu, Z. D. and Fan, B. T. (2004a). "Prediction of the isoelectric point of an amino acid based on GA-PLS and SVMs." J Chem Inf Comput Sci **44**(1): 161-7.

Liu, H., Rhodes, M., Wiest, D. L. and Vignali, D. A. (2000). "On the dynamics of TCR:CD3 complex cell surface expression and downmodulation." Immunity **13**(5): 665-75.

Liu, Q., Zhu, Y. S., Wang, B. H. and Li, Y. X. (2003). "A HMM-based method to predict the transmembrane regions of beta-barrel membrane proteins." Comput Biol Chem **27**(1): 69-76.

Liu, Y., Zhu, P. and Hu, Y. M. (2004b). "[Proliferation of specific cytotoxic T lymphocytes induced by immunoglobulin heavy chain framework region-derived antigenic nonapeptides]." Zhonghua Yi Xue Za Zhi **84**(2): 97-102.

Liu, Z., Dominy, B. N. and Shakhnovich, E. I. (2004c). "Structural mining: self-consistent design on flexible protein-peptide docking and transferable binding affinity potential." J. Am. Chem. Soc. **126**(27): 8515-28.

Ljunggren, H., G. and Thorpe, C., J. (1996). "Principles of MHC class I mediated antigen presentation and T cell selection." Histology and Histopath **11**: 267-274.

Logean, A., Sette, A. and Rognan, D. (2001). "Customized versus universal scoring functions: application to class I MHC-peptide binding free energy predictions." Bioorg. Med. Chem. Lett. **11**(5): 675-9.

Logean, A. and Rognan, D. (2002). "Recovery of known T-cell epitopes by computational scanning of a viral genome." J Comput Aided Mol Des **16**(4): 229-43.

Lopez-Botet, M. and Bellon, T. (1999). "Natural killer cell activation and inhibition by receptors for MHC class I." Curr. Opin. Immunol. **11**(3): 301-7.

Lowe, J., Stock, D., Jap, B., Zwickl, P., Baumeister, W. and Huber, R. (1995). "Crystal structure of the 20S proteasome from the archaeon T. acidophilum at 3.4 A resolution." Science **268**(5210): 533-9.

Lu, J. and Celis, E. (2000). "Use of two predictive algorithms of the world wide web for the identification of tumor-reactive T-cell epitopes." Cancer Res. **60**(18): 5223-7.

Luescher, I. F., Vivier, E., Layer, A., Mahiou, J., Godeau, F., Malissen, B. and Romero, P. (1995). "CD8 modulation of T-cell antigen receptor-ligand interactions on living cytotoxic T lymphocytes." Nature **373**(6512): 353-6.

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 **55**(12): 797-810.

Lyman, M., Lee, H., Kang, B., Kang, H. and Kim, B. (2002). "Capsid-specific T lymphocytes recognise three distinct H-2Db-restricted regions of the BeAn strain of Theiler's virus and exhibit different cytokine profiles." J. Virol. **76**: 3125-3134.

Maby, E., Le Bouquin Jeannes, R., Liegeois-Chauvel, C., Gourevitch, B. and Faucon, G. (2004). "Analysis of auditory evoked potential parameters in the presence of radiofrequency fields using a support vector machines method." Med. Biol. Eng. Comput. **42**(4): 562-8.

Madden, D. R., Gorga, J. C., Strominger, J. L. and Wiley, D. C. (1991). "The structure of HLA-B27 reveals nonamer self-peptides bound in an extended conformation." Nature **353**(6342): 321-5.

Madden, D. R., Gorga, J. C., Strominger, J. L. and Wiley, D. C. (1992). "The three-dimensional structure of HLA-B27 at 2.1 A resolution suggests a general mechanism for tight peptide binding to MHC." Cell **70**(6): 1035-48.

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.

Madden, D. R. (1995). "The three-dimensional structure of peptide-MHC complexes." Annu. Rev. Immunol. **13**: 587-622.

Madnaka, K. and Yvonne Jones, E. (1999). "MHC superfamily structure and the immune system." Curr. Opin. Immunol. **9**: 745-753.

Mahler, M., Bluthner, M. and Pollard, K. M. (2003). "Advances in B-cell epitope analysis of autoantigens in connective tissue diseases." Clin Immunol **107**(2): 65-79.

Mallios, R. R. (1993). "Predicting the probability of helper T cell immunodominant sites through discriminant analysis." Ann Clin Biochem **30 (Pt 2)**: 152-6.

Mallios, R. R. (1994). "Multiple regression analysis suggests motifs for class II MHC binding." J. Theor. Biol. **166**(2): 167-72.

Mallios, R. R. (1997). "An iterative algorithm for converting a class II MHC binding motif into a quantitative predictive model." Comput Appl Biosci **13**(3): 211-5.

Mallios, R. R. (1998). "Iterative stepwise discriminant analysis: a meta-algorithm for detecting quantitative sequence motifs." J. Comput. Biol. **5**(4): 703-11.

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**(6): 432-9.

Mallios, R. R. (2001). "Predicting class II MHC/peptide multi-level binding with an iterative stepwise discriminant analysis meta-algorithm." Bioinformatics **17**(10): 942-8.

Mallios, R. R. (2003). "A consensus strategy for combining HLA-DR binding algorithms." Hum. Immunol. **64**(9): 852-6.

Maloy, W. L. (1987). "Comparison of the primary structure of class I molecules." Immunol. Res. **6**(1-2): 11-29.

Mamitsuka, H. (1998). "Predicting peptides that bind to MHC molecules using supervised learning of hidden Markov models." Proteins **33**(4): 460-74.

Manfredi, A. A., Yuen, M. H., Raftery, M. A. and Conti-Tronconi, B. M. (1993). "An assay for simultaneous multiple determinations of peptide binding to MHC class II molecules." Anal. Biochem. **211**(2): 267-73.

Manici, S., Sturniolo, T., Imro, M. A., Hammer, J., Sinigaglia, F., Noppen, C., Spagnoli, G., Mazzi, B., Bellone, M., Dellabona, P. and Protti, M. P. (1999). "Melanoma cells present a MAGE-3 epitope to CD4(+) cytotoxic T cells in association with histocompatibility leukocyte antigen DR11." J. Exp. Med. **189**(5): 871-6.

Mann, D. L., Rogentine, G. N., Jr., Fahey, J. L. and Nathenson, S. G. (1968). "Solubilization of human leucocyte membrane isoantigens." Nature **217**(134): 1180-1.

Marchal-Bras-Goncalves, R., Rouas-Freiss, N., Connan, F., Choppin, J., Dausset, J., Carosella, E. D., Kirszenbaum, M. and Guillet, J. (2001). "A soluble HLA-G protein that inhibits natural killer cell-mediated cytotoxicity." Transplant. Proc. **33**(3): 2355-9.

Marchand, M., van Baren, N., Weynants, P., Brichard, V., Dreno, B., Tessier, M. H., Rankin, E., Parmiani, G., Arienti, F., Humblet, Y., Bourlond, A., Vanwijck, R., Lienard, D., Beauduin, M., Dietrich, P. Y., Russo, V., Kerger, J., Masucci, G., Jager, E., De Greve, J., Atzpodien, J., Brasseur, F., Coulie, P. G., van der Bruggen, P. and Boon, T. (1999). "Tumor regressions observed in patients with metastatic melanoma treated with an antigenic peptide encoded by gene MAGE-3 and presented by HLA-A1." Int. J. Cancer **80**(2): 219-30.

Marchand, M., Punt, C. J., Aamdal, S., Escudier, B., Kruit, W. H., Keilholz, U., Hakansson, L., van Baren, N., Humblet, Y., Mulders, P., Avril, M. F., Eggermont, A. M., Scheibenbogen, C., Uiters, J., Wanders, J., Delire, M., Boon, T. and Stoter, G. (2003). "Immunisation of metastatic cancer patients with MAGE-3 protein combined with adjuvant SBAS-2: a clinical report." Eur. J. Cancer **39**(1): 70-7.

Marchini, M., Antonioli, R., Lleo, A., Barili, M., Caronni, M., Origgi, L., Vanoli, M. and Scorza, R. (2003). "HLA class II antigens associated with lupus nephritis in Italian SLE patients." Hum. Immunol. **64**(4): 462-8.

Margalit, H. and Altuvia, Y. (2003). "Insights from MHC-bound peptides." Novartis Found. Symp. **254**: 77-90; discussion 91-101, 216-22, 250-2.

Marsh, S. G. (2003). "HLA nomenclature and the IMGT/HLA sequence database." Novartis Found. Symp. **254**: 165-73; discussion 173-6, 216-22, 250-2.

Marsh, S. G. (2004). "Nomenclature for factors of the HLA system, update October 2003." Eur. J. Immunogenet. **31**(1): 53-4.

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**(10): 4946-57.

Martelli, P. L., Fariselli, P., Krogh, A. and Casadio, R. (2002). "A sequence-profile-based HMM for predicting and discriminating beta barrel membrane proteins." Bioinformatics **18 Suppl 1**: S46-53.

Maruyama, T., Shimada, A., Kasuga, A., Kasatani, T., Ozawa, Y., Ishii, M., Takei, I., Suzuki, Y., Kobayashi, A., Takeda, S. and et al. (1994). "Analysis of MHC class II antigens in Japanese IDDM by a novel HLA-typing method, hybridization protection assay." Diabetes Res. Clin. Pract. **23**(2): 77-84.

Mata, M., Travers, P. J., Liu, Q., Frankel, F. R. and Paterson, Y. (1998). "The MHC class I-restricted immune response to HIV-gag in BALB/c mice selects a single epitope that does not have a predictable MHC-binding motif and binds to Kd through interactions between a glutamine at P3 and pocket D." J. Immunol. **161**(6): 2985-93.

Matsumoto, C. and Awata, T. (1994). "[Non-MHC susceptibility genes in Japanese subjects with insulin-dependent diabetes mellitus (IDDM)]." Nippon Rinsho **52**(10): 2758-61.

Matsumura, M., Fremont, D. H., Peterson, P. A. and Wilson, I. A. (1992a). "Emerging principles for the recognition of peptide antigens by MHC class I molecules." Science **257**(5072): 927-34.

Matsumura, M., Saito, Y., Jackson, M. R., Song, E. S. and Peterson, P. A. (1992b). "In vitro peptide binding to soluble empty class I major histocompatibility complex molecules isolated from transfected Drosophila melanogaster cells." J. Biol. Chem. **267**(33): 23589-95.

Matsunami, K., Miyagawa, S., Nakai, R., Yamada, M. and Shirakura, R. (2000a). "Protection against natural killer-mediated swine endothelial cell lysis by HLA-G and HLA-E." Transplant. Proc. **32**(5): 939-40.

Matsunami, K., Miyagawa, S., Nakai, R., Yamada, M. and Shirakura, R. (2000b). "The regulation of natural killer-mediated swine endothelial cell lysis by HLA-G (G1 and G3)." Transplant. Proc. **32**(7): 2087-8.

Matter, H. and Schwab, W. (1999). "Affinity and selectivity of matrix metalloproteinase inhibitors: a chemometrical study from the perspective of ligands and proteins." J. Med. Chem. **42**(22): 4506-23.

Mayr, A. (1999). "[Historical review of smallpox, the eradication of smallpox and the attenuated smallpox MVA vaccine]." Berl Munch Tierarztl Wochenschr **112**(9): 322-8.

McKenzie, L. M., Pecon-Slattery, J., Carrington, M. and O'Brien, S. J. (1999). "Taxonomic hierarchy of HLA class I allele sequences." Genes Immun. **1**(2): 120-9.

McNeil, A. J., Yap, P. L., Gore, S. M., Brettle, R. P., McColl, M., Wyld, R., Davidson, S., Weightman, R., Richardson, A. M. and Robertson, J. R. (1996). "Association of HLA types A1-B8-DR3 and B27 with rapid and slow progression of HIV disease." Qjm **89**(3): 177-85.

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**(4): 1276-87.

Mehta, M. M., Chablani, U. A., Contractor, N. M., Bhatia, H. M., Singhal, B. S., Mondkar, V. P. and Desai, A. D. (1986). "HLA-A & HLA-B antigens in multiple sclerosis, motor neuron disease & Duchenne muscular dystrophy." Indian J. Med. Res. **83**: 519-21.

Meister, G. E., Roberts, C. G., 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**(6): 581-91.

Meloen, R. H., Langeveld, J. P., Schaaper, W. M. and Slootstra, J. W. (2001). "Synthetic peptide vaccines: unexpected fulfillment of discarded hope?" Biologicals **29**(3-4): 233-6.

Mendez-Samperio, P. and Jimenez-Zamudio, L. (1991). "Peptide competition at the level of MHC-binding sites using T cell clones from a rheumatoid arthritis patient." J. Autoimmun. **4**(5): 795-806.

Meng, W. S., von Grafenstein, H. and Haworth, I. S. (2000). "Water dynamics at the binding interface of four different HLA-A2-peptide complexes." Int. Immunol. **12**(7): 949-57.

Merriman, T. R. and Todd, J. A. (1995). "Genetics of autoimmune disease." Curr. Opin. Immunol. **7**(6): 786-92.

Michielin, O. and Karplus, M. (2002). "Binding free energy differences in a TCR-peptide-MHC complex induced by a peptide mutation: a simulation analysis." J. Mol. Biol. **324**(3): 547-69.

Middleton, D. (1999). "History of DNA typing for the human MHC." Rev Immunogenet **1**(2): 135-56.

Mielke, P. W., Jr. and Berry, K. J. (2002). "Multivariate multiple regression analyses: a permutation method for linear models." Psychol Rep **91**(1): 3-9.

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**(8): 753-6.

Mizumachi, K. and Kurisaki, J. (2003). "Localization of T cell epitope regions of chicken ovomucoid recognized by mice." Biosci. Biotechnol. Biochem. **67**(4): 712-9.

Mohler, K. M. and Streilein, J. W. (1989). "Lymphokine production by MLR-reactive reaction lymphocytes obtained from normal mice and mice rendered tolerant of class II MHC antigens." Transplantation **47**(4): 625-33.

Moingeon, P. (2001). "Cancer vaccines." Vaccine **19**(11-12): 1305-26.

Mondal, S., Jaishankar, S. P. and Ramakumar, S. (2003). "Role of context in the relationship between form and function: structural plasticity of some PROSITE patterns." Biochem. Biophys. Res. Commun. **305**(4): 1078-84.

Monneret, G., Boumiza, R., Gravel, S., Cossette, C., Bienvenu, J., Rokach, J. and Powell, W. S. (2004). "Effects of prostaglandin D2 and 5-lipoxygenase products on the expression of CD203c and CD11b by basophils." J. Pharmacol. Exp. Ther.

Moore, A., Medarova, Z., Potthast, A. and Dai, G. (2004). "In vivo targeting of underglycosylated MUC-1 tumor antigen using a multimodal imaging probe." Cancer Res. **64**(5): 1821-7.

Mukasa, A., Born, W. K. and O'Brien, R. L. (1999). "Inflammation alone evokes the response of a TCR-invariant mouse gamma delta T cell subset." J. Immunol. **162**(8): 4910-3.

Mullbacher, A. (1997). "Hypothesis: MHC class I, rather than just a flagpole for CD8+ T cells is also a protease in its own right." Immunol. Cell Biol. **75**(3): 310-7.

Murthy, V. S. and Kulkarni, V. M. (2002). "3D-QSAR CoMFA and CoMSIA on protein tyrosine phosphatase 1B inhibitors." Bioorg. Med. Chem. **10**(7): 2267-82.

MyIntyre, C., Rees, R., Platts, K., Cooke, C., Smith, M., Mulcahy, K. and Murray, A. (1996). "Identification of peptide epitopes of MAGE-1, -2, -3 that demonstrate HLA-A3-specific binding." Cancer Immunol. Immunother. **42**: 246-250.

Myshkin, E. and Wang, B. (2003). "Chemometrical classification of ephrin ligands and Eph kinases using GRID/CPCA approach." J Chem Inf Comput Sci **43**(3): 1004-10.

Nadasdi, L. and Medzihradszky, K. (1981). "A study of the applicability of QSAR calculation for peptide hormones." Biochem. Biophys. Res. Commun. **99**(2): 451-7.

Nag, B., Mukku, P. V., Arimilli, S., Phan, D., Deshpande, S. V. and Winkelhake, J. L. (1994). "Antigenic peptide binding to MHC class II molecules at increased peptide concentrations." Mol. Immunol. **31**(15): 1161-8.

Nakai, K., Kidera, A. and Kanehisa, M. (1988). "Cluster analysis of amino acid indices for prediction of protein structure and function." Protein Eng. **2**(2): 93-100.

Nathenson, S. G. and Davis, D. A. (1966). "Solubilization and partial purification of mouse histocompatibility antigens from a membranous lipoprotein fraction." Proc Natl Acad Sci **56**: 476-483.

Nathenson, S. G., Geliebter, J., Pfaffenbach, G. M. and Zeff, R. A. (1986). "Murine major histocompatibility complex class-I mutants: molecular analysis and structure-function implications." Annu. Rev. Immunol. **4**: 471-502.

Naumann, T. and Matter, H. (2002). "Structural classification of protein kinases using 3D molecular interaction field analysis of their ligand binding sites: target family landscapes." J. Med. Chem. **45**(12): 2366-78.

Neumann, F., Wagner, C., Kubuschok, B., Stevanovic, S., Rammensee, H. G. and Pfreundschuh, M. (2004). "Identification of an antigenic peptide derived from the cancer-testis antigen NY-ESO-1 binding to a broad range of HLA-DR subtypes." Cancer Immunol. Immunother. **53**(7): 589-99.

Newman, A. H., Izenwasser, S., Robarge, M. J. and Kline, R. H. (1999). "CoMFA study of novel phenyl ring-substituted 3alpha-(diphenylmethoxy)tropane analogues at the dopamine transporter." J. Med. Chem. **42**(18): 3502-9.

Newman, M. J., Livingston, B., McKinney, D. M., Chesnut, R. W. and Sette, A. (2002). "T-lymphocyte epitope identification and their use in vaccine development for HIV-1." Front. Biosci. **7**: d1503-15.

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**(9): 1388-97.

Nijenhuis, M., Schmitt, S., Armandola, E. A., Obst, R., Brunner, J. and Hammerling, G. J. (1996). "Identification of a contact region for peptide on the TAP1 chain of the transporter associated with antigen processing." J. Immunol. **156**(6): 2186-95.

Nino-Vasquez, J. J., Allicotti, G., Borras, E., Wilson, D. B., Valmori, D., Simon, R., Martin, R. and Pinilla, C. (2004). "A powerful combination: the use of positional scanning libraries and biometrical analysis to identify cross-reactive T cell epitopes." Mol. Immunol. **40**(14-15): 1063-74.

Norinder, U. (1991). "Theoretical amino acid descriptors. Application to bradykinin potentiating peptides." Peptides **12**(6): 1223-7.

Nussbaum, A. K., Dick, T. P., Keilholz, W., Schirle, M., Stevanovic, S., Dietz, K., Heinemeyer, W., Groll, M., Wolf, D. H., Huber, R., Rammensee, H. G. and Schild, H. (1998). "Cleavage motifs of the yeast 20S proteasome beta subunits deduced from digests of enolase 1." Proc. Natl. Acad. Sci. U. S. A. **95**(21): 12504-9.

Nussbaum, A. K., Kuttler, C., Hadeler, K. P., Rammensee, H. G. and Schild, H. (2001). "PAProC: a prediction algorithm for proteasomal cleavages available on the WWW." Immunogenetics 53(2): 87-94.

Nussbaum, A. K., Kuttler, C., Tenzer, S. and Schild, H. (2003). "Using the World Wide Web for predicting CTL epitopes." Curr. Opin. Immunol. **15**(1): 69-74.

O'Callaghan, C. A. (2000). "Molecular basis of human natural killer cell recognition of HLA-E (human leucocyte antigen-E) and its relevance to clearance of pathogen-infected and tumour cells." Clin. Sci. (Lond.) **99**(1): 9-17.

Ojcius, D. M., Abastado, J. P., Godeau, F. and Kourilsky, P. (1992). "Peptide binding to MHC class I proteins measured with a novel fluorescent technique." C R Acad Sci III **315**(9): 337-41.

Ono, S. J., Issa-Chergui, B., Colle, E., Guttmann, R. D., Seemayer, T. A. and Fuks, A. (1988). "IDDM in BB rats. Enhanced MHC class I heavy-chain gene expression in pancreatic islets." Diabetes **37**(10): 1411-8.

Paabo, S., Severinsson, L., Andersson, M., Martens, I., Nilsson, T. and Peterson, P. A. (1989). "Adenovirus proteins and MHC expression." Adv. Cancer Res. **52**: 151-63.

Pamer, E. G., Harty, J. T. and Bevan, M. J. (1991). "Precise prediction of a dominant class I MHC-restricted epitope of Listeria monocytogenes." Nature **353**(6347): 852-5.

Panagiotopoulos, C., Qin, H., Tan, R. and Verchere, C. (2003). "Identification of a b-cell-specific HLA class I restricted epitope in type 1 diabetes." Diabetes **52**: 2647-2651.

Panigada, M., Sturniolo, T., Besozzi, G., Boccieri, M. G., Sinigaglia, F., Grassi, G. G. and Grassi, F. (2002). "Identification of a promiscuous T-cell epitope in Mycobacterium tuberculosis Mce proteins." Infect. Immun. **70**(1): 79-85.

Pantuck, A. J., van Ophoven, A., Gitlitz, B. J., Tso, C. L., Acres, B., Squiban, P., Ross, M. E., Belldegrun, A. S. and Figlin, R. A. (2004). "Phase I trial of antigen-specific gene therapy using a recombinant vaccinia virus encoding MUC-1 and IL-2 in MUC-1-positive patients with advanced prostate cancer." J. Immunother. **27**(3): 240-53.

Papac, D. I., Hoyes, J. and Tomer, K. B. (1994). "Epitope mapping of the gastrin-releasing peptide/anti-bombesin monoclonal antibody complex by proteolysis followed by matrix-assisted laser desorption ionization mass spectrometry." Protein Sci. **3**(9): 1485-92.

Parker, K. C., Bednarek, M. A., Hull, L. K., Utz, U., Cunningham, B., Zweerink, H. J., Biddison, W. E. and Coligan, J. E. (1992a). "Sequence motifs important for peptide binding to the human MHC class I molecule, HLA-A2." J. Immunol. **149**(11): 3580-7.

Parker, K. C., DiBrino, M., Hull, L. and Coligan, J. E. (1992b). "The beta 2-microglobulin dissociation rate is an accurate measure of the stability of MHC class I heterotrimers and depends on which peptide is bound." J. Immunol. **149**(6): 1896-904.

Parker, C. E., Papac, D. I., Trojak, S. K. and Tomer, K. B. (1996). "Epitope mapping by mass spectrometry: determination of an epitope on HIV-1 IIIB p26 recognized by a monoclonal antibody." J. Immunol. **157**(1): 198-206.

Parkhurst, M. R., Salgaller, M. L., Southwood, S., Robbins, P. F., Sette, A., Rosenberg, S. A. and Kawakami, Y. (1996). "Improved induction of melanoma-reactive CTL with peptides from the melanoma antigen gp100 modified at HLA-A*0201-binding residues." J. Immunol. **157**(6): 2539-48.

Parkhurst, M. R., Fitzgerald, E. B., Southwood, S., Sette, A., Rosenberg, S. A. and Kawakami, Y. (1998). "Identification of a shared HLA-A*0201-restricted T-cell epitope from the melanoma antigen tyrosinase-related protein 2 (TRP2)." Cancer Res. **58**(21): 4895-901.

Paschen, A., Eichmuller, S. and Schadendorf, D. (2004). "Identification of tumor antigens and T-cell epitopes, and its clinical application." Cancer Immunol. Immunother. **53**(3): 196-203.

Pascolo, S., Schirle, M., Guckel, B., Dumrese, T., Stumm, S., Kayser, S., Moris, A., Wallwiener, D., Rammensee, H. G. and Stevanovic, S. (2001). "A MAGE-A1 HLA-A A*0201 epitope identified by mass spectrometry." Cancer Res. **61**(10): 4072-7.

Passoni, L., Scardino, A., Bertazzoli, C., Gallo, B., Coluccia, A. M., Lemonnier, F. A., Kosmatopoulos, K. and Gambacorti-Passerini, C. (2002). "ALK as a novel lymphoma-associated tumor antigen: identification of 2 HLA-A2.1-restricted CD8+ T-cell epitopes." Blood **99**(6): 2100-6.

Pastor, M. and Cruciani, G. (1995). "A novel strategy for improving ligand selectivity in receptor-based drug design." J. Med. Chem. **38**(23): 4637-47.

Pate, M. E., Turner, M. K., Thornhill, N. F. and Titchener-Hooker, N. J. (2004). "Principal component analysis of nonlinear chromatography." Biotechnol. Prog. **20**(1): 215-22.

Payette, P. J. and Davis, H. L. (2001). "History of vaccines and positioning of current trends." Curr Drug Targets Infect Disord **1**(3): 241-7.

Pazmany, L., Mandelboim, O., Vales-Gomez, M., Davis, D. M., Reyburn, H. T. and Strominger, J. L. (1996). "Protection from natural killer cell-mediated lysis by HLA-G expression on target cells." Science **274**(5288): 792-5.

Pearlstein, R. A., Vaz, R. J., Kang, J., Chen, X. L., Preobrazhenskaya, M., Shchekotikhin, A. E., Korolev, A. M., Lysenkova, L. N., Miroshnikova, O. V., Hendrix, J. and Rampe, D. (2003). "Characterization of HERG potassium channel inhibition using CoMSiA 3D QSAR and homology modeling approaches." Bioorg. Med. Chem. Lett. **13**(10): 1829-35.

Pelte, C., Cherepnev, G., Wang, Y., Schoenemann, C., Volk, H. D. and Kern, F. (2004). "Random screening of proteins for HLA-A*0201-binding nine-amino acid peptides is not sufficient for identifying CD8 T cell epitopes recognized in the context of HLA-A*0201." J. Immunol. **172**(11): 6783-9.

Perkins, R., Fang, H., Tong, W. and Welsh, W. J. (2003). "Quantitative structure-activity relationship methods: perspectives on drug discovery and toxicology." Environ. Toxicol. Chem. **22**(8): 1666-79.

Peter, J. F. and Tomer, K. B. (2001). "A general strategy for epitope mapping by direct MALDI-TOF mass spectrometry using secondary antibodies and cross-linking." Anal. Chem. **73**(16): 4012-9.

Peters, B., Tong, W., Sidney, J., Sette, A. and Weng, Z. (2003). "Examining the Independent Binding Assumption for Binding of Peptide Epitopes to MHC-I Molecules." Bioinformatics **19**: 1765-1772.

Petrone, P. M. and Garcia, A. E. (2004). "MHC-peptide binding is assisted by bound water molecules." J. Mol. Biol. **338**(2): 419-35.

Pinilla, C., Appel, J. R., Blanc, P. and Houghten, R. A. (1992). "Rapid identification of high affinity peptide ligands using positional scanning synthetic peptide combinatorial libraries." BioTechniques **13**(6): 901-5.

Pituch-Noworolska, A., Dziatkowiak, H. and Zembala, M. (1991). "MHC class II determinants on peripheral blood monocytes from newly diagnosed IDDM patients." Endokrynol Pol **42**(4): 507-12.

Ploegh, H. L., Orr, H. T. and Strominger, J. L. (1981). "Major histocompatibility antigens: the human (HLA-A, -B, -C) and murine (H-2K, H-2D) class I molecules." Cell **24**(2): 287-99.

Pohlmann, T., Bockmann, R. A., Grubmuller, H., Uchanska-Ziegler, B., Ziegler, A. and Alexiev, U. (2004). "Differential peptide dynamics is linked to major histocompatibility complex polymorphism." J. Biol. Chem. **279**(27): 28197-201.

Posch, P. E., Borrego, F., Brooks, A. G. and Coligan, J. E. (1998). "HLA-E is the ligand for the natural killer cell CD94/NKG2 receptors." J. Biomed. Sci. **5**(5): 321-31.

Prod'homme, V., Retiere, C., Imbert-Marcille, B. M., Bonneville, M. and Hallet, M. M. (2003). "Modulation of HLA-A*0201-restricted T cell responses by natural polymorphism in the IE1(315-324) epitope of human cytomegalovirus." J. Immunol. **170**(4): 2030-6.

Purushottamachar, P. and Kulkarni, V. M. (2003). "3D-QSAR of N-myristoyltransferase inhibiting antifungal agents by CoMFA and CoMSIA methods." Bioorg. Med. Chem. **11**(16): 3487-97.

Qian, B. and Goldstein, R. A. (2004). "Performance of an iterated T-HMM for homology detection." Bioinformatics.

Raichurkar, A. V. and Kulkarni, V. M. (2003). "Understanding the antitumor activity of novel hydroxysemicarbazide derivatives as ribonucleotide reductase inhibitors using CoMFA and CoMSIA." J. Med. Chem. **46**(21): 4419-27.

Rammensee, H. G., Friede, T. and Stevanoviic, S. (1995). "MHC ligands and peptide motifs: first listing." Immunogenetics **41**(4): 178-228.

Rammensee, H., Bachmann, J., Emmerich, N. P., Bachor, O. A. and Stevanovic, S. (1999). "SYFPEITHI: database for MHC ligands and peptide motifs." Immunogenetics **50**(3-4): 213-9.

Rani, R., Sood, A., Lazaro, A. M. and Stastny, P. (1999). "Associations of MHC class II alleles with insulin-dependent diabetes mellitus (IDDM) in patients from North India." Hum. Immunol. **60**(6): 524-31.

Rau, L., Cohan, N. and Robert, J. (2001). "MHC-restricted and -unrestricted CD8 T cells: an evolutionary perspective." Transplantation **72**: 1830-1835.

Reche, P. A., Glutting, J. P. and Reinherz, E. L. (2002). "Prediction of MHC class I binding peptides using profile motifs." Hum. Immunol. **63**(9): 701-9.

Reche, P. A. and Reinherz, E. L. (2003). "Sequence variability analysis of human class I and class II MHC molecules: functional and structural correlates of amino acid polymorphisms." J. Mol. Biol. **331**(3): 623-41.

Reid, S. W., McAdam, S., Smith, K. J., Klenerman, P., O'Callaghan, C. A., Harlos, K., Jakobsen, B. K., McMichael, A. J., Bell, J. I., Stuart, D. I. and Jones, E. Y. (1996). "Antagonist HIV-1 Gag peptides induce structural changes in HLA B8." J. Exp. Med. **184**(6): 2279-86.

Restifo, N. P., Esquivel, F., Kawakami, Y., Yewdell, J. W., Mule, J. J., Rosenberg, S. A. and Bennink, J. R. (1993a). "Identification of human cancers deficient in antigen processing." J. Exp. Med. **177**(2): 265-72.

Restifo, N. P., Kawakami, Y., Marincola, F., Shamamian, P., Taggarse, A., Esquivel, F. and Rosenberg, S. A. (1993b). "Molecular mechanisms used by tumors to escape immune recognition: immunogenetherapy and the cell biology of major histocompatibility complex class I." J. Immunother. **14**(3): 182-90.

Rhodes, D. A. and Trowsdale, J. (1999). "Genetics and molecular genetics of the MHC." Rev Immunogenet **1**(1): 21-31.

Ritz, U. and Seliger, B. (2001). "The transporter associated with antigen processing (TAP): structural integrity, expression, function, and its clinical relevance." Mol Med **7**(3): 149-58.

Rivoltini, L., Kawakami, Y., Sakaguchi, K., Southwood, S., Sette, A., Robbins, P. F., Marincola, F. M., Salgaller, M. L., Yannelli, J. R., Appella, E. and et al. (1995). "Induction of tumor-reactive CTL from peripheral blood and tumor-infiltrating lymphocytes of melanoma patients by in vitro stimulation with an immunodominant peptide of the human melanoma antigen MART-1." J. Immunol. **154**(5): 2257-65.

Robbins, P. F. and Kawakami, Y. (1996). "Human tumor antigens recognized by T cells." Curr. Opin. Immunol. **8**(5): 628-36.

Robinson, J., Waller, M., Parham, P., de Groot, N., Bontrop, R., Kennedy, L., Stoehr, P. and Marsh, S. (2003). "IMGT/HLA and IMGT/MHC: sequence databases for the study of the major histocompatibility complex." Nucleic Acids Res. **31**: 311-314.

Robinson, K., Bellaby, T., Chan, W. C. and Wakelin, D. (1995). "High levels of protection induced by a 40-mer synthetic peptide vaccine against the intestinal nematode parasite Trichinella spiralis." Immunology **86**(4): 495-8.

Rognan, D., Scapozza, L., Folkers, G. and Daser, A. (1994). "Molecular dynamics simulation of MHC-peptide complexes as a tool for predicting potential T cell epitopes." Biochemistry **33**(38): 11476-85.

Rognan, D., Lauemoller, 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**(22): 4650-8.

Rohren, E. M., Pease, L. R., Ploegh, H. L. and Schumacher, T. N. (1993). "Polymorphisms in pockets of major histocompatibility complex class I molecules influence peptide preference." J. Exp. Med. **177**(6): 1713-21.

Rojo, S., Garcia, F., Villadangos, J. A. and Lopez de Castro, J. A. (1993). "Changes in the repertoire of peptides bound to HLA-B27 subtypes and to site-specific mutants inside and outside pocket B." J. Exp. Med. **177**(3): 613-20.

Rongcun, Y., Salazar-Onfray, F., Charo, J., Malmberg, K. J., Evrin, K., Maes, H., Kono, K., Hising, C., Petersson, M., Larsson, O., Lan, L., Appella, E., Sette, A., Celis, E. and Kiessling, R. (1999). "Identification of new HER2/neu-derived peptide epitopes that can elicit specific CTL against autologous and allogeneic carcinomas and melanomas." J. Immunol. **163**(2): 1037-44.

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): 1-21.

Ross, S. E., Schraven, B., Goldman, F. D., Crabtree, J. and Koretzky, G. A. (1994). "The association between CD45 and lck does not require CD4 or CD8 and is independent of T cell receptor stimulation." Biochem. Biophys. Res. Commun. **198**(1): 88-96.

Rotzschke, O., Falk, K., Stevanovic, S., Jung, G., Walden, P. and Rammensee, H. G. (1991). "Exact prediction of a natural T cell epitope." Eur. J. Immunol. **21**(11): 2891-4.

Rouas-Freiss, N., Marchal, R. E., Kirszenbaum, M., Dausset, J. and Carosella, E. D. (1997). "The alpha1 domain of HLA-G1 and HLA-G2 inhibits cytotoxicity induced by natural killer cells: is HLA-G the public ligand for natural killer cell inhibitory receptors?" Proc. Natl. Acad. Sci. U. S. A. **94**(10): 5249-54.

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**(5): 929-37.

Salter, R. and Cresswell, P. (1986). "Impaired assembly and transport of HLA-A and -B antigens in a mutant TxB cell hybrid." EMBO **5**: 943-949.

Sandberg, M., Eriksson, L., Jonsson, J., Sjostrom, M. and Wold, S. (1998). "New Chemical Descriptors Relevant for the Design of Biologically Active Peptides. A Multivariate Characterization of 87 Amino Acids." J. Med. Chem. **41**: 2481-2491.

Saper, M. A., Bjorkman, P. J. and Wiley, D. C. (1991). "Refined structure of the human histocompatibility antigen HLA-A2 at 2.6 A resolution." J Mol Biol. **219**: 277-319.

Saren, A., Pascolo, S., Stevanovic, S., Dumrese, T., Puolakkainen, M., Sarvas, M., Rammensee, H. and Vuola, J. (2002). "Identification of Chlamydia pneumoniae-derived mouse CD8 epitopes." Infect. Immun. **70**: 3339-3343.

Saric, T., Beninga, J., Graef, C. I., Akopian, T. N., Rock, K. L. and Goldberg, A. L. (2001). "Major histocompatibility complex class I-presented antigenic peptides are degraded in cytosolic extracts primarily by thimet oligopeptidase." J. Biol. Chem. **276**(39): 36474-81.

Sarobe, P., Huarte, E., Lasarte, J., Cerio, A., Garcia, N., Borras-Cuesta, F. and Prieto, J. (2001). "Characterization of an immunologically conserved epitope from hepatitis C virus E2 glycoprotein recognized by HLA-A2 restricted cytotoxic T lymphocytes." J. Hepatol. **34**: 321-329.

Sasaki, H., Xu, X. C., Smith, D. M., Howard, T. and Mohanakumar, T. (1999). "HLA-G expression protects porcine endothelial cells against natural killer cell-mediated xenogeneic cytotoxicity." Transplantation **67**(1): 31-7.

Saxena, A. K. and Prathipati, P. (2003). "Comparison of MLR, PLS and GA-MLR in QSAR analysis." SAR QSAR Environ Res **14**(5-6): 433-45.

Sbai, H., Mehta, A. and DeGroot, A. S. (2001). "Use of T cell epitopes for vaccine development." Curr Drug Targets Infect Disord **1**(3): 303-13.

Scapozza, L. (1995). "Molecular dynamics and structure-based drug design for predicting non-natural nonapeptide binding to a class I MHC protein." Acta Crystallogr D Biol Crystallogr **51**(Pt 4): 541-9.

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**(19): 1880-4.

Schafroth, H. D. and Floudas, C. A. (2004). "Predicting peptide binding to MHC pockets via molecular modeling, implicit solvation, and global optimization." Proteins **54**(3): 534-56.

Schapira, M., Totrov, M. and Abagyan, R. (1999). "Prediction of the binding energy for small molecules, peptides and proteins." J. Mol. Recognit. **12**(3): 177-90.

Schneider, J., Brichard, V., Boon, T., Meyer zum Buschenfelde, K. H. and Wolfel, T. (1998). "Overlapping peptides of melanocyte differentiation antigen Melan-A/MART-1 recognized by autologous cytolytic T lymphocytes in association with HLA-B45.1 and HLA-A2.1." Int. J. Cancer **75**(3): 451-8.

Schol, D. J., Meulenbroek, M. F., Snijdewint, F. G., von Mensdorff-Pouilly, S., Verstraeten, R. A., Murakami, F., Kenemans, P. and Hilgers, J. (1998). "'Epitope fingerprinting' using overlapping 20-mer peptides of the MUC1 tandem repeat sequence." Tumour Biol **19 Suppl 1**: 35-45.

Scholkopf, S., Burges, C. J. C. and Smola, A. J. (1999). Advances in kernel methods: support vector learning. Cambridge, MA, MIT Press.

Schonbach, C., Koh, J. L., Sheng, X., Wong, L. and Brusic, V. (2000). "FIMM, a database of functional molecular immunology." Nucleic Acids Res. **28**(1): 222-4.

Schreurs, M. W., Eggert, A. A., de Boer, A. J., Vissers, J. L., van Hall, T., Offringa, R., Figdor, C. G. and Adema, G. J. (2000). "Dendritic cells break tolerance and induce protective immunity against a melanocyte differentiation antigen in an autologous melanoma model." Cancer Res. **60**(24): 6995-7001.

Schueler-Furman, O., Elber, R. and Margalit, H. (1998). "Knowledge-based structure prediction of MHC class I bound peptides: a study of 23 complexes." Fold Des **3**(6): 549-64.

Schueler-Furman, O., Altuvia, Y., Sette, A. and Margalit, H. (2000). "Structure-based prediction of binding peptides to MHC class I molecules: application to a broad range of MHC alleles." Protein Sci. **9**(9): 1838-46.

Schueler-Furman, O., Altuvia, Y. and Margalit, H. (2001). "Examination of possible structural constraints of MHC-binding peptides by assessment of their native structure within their source proteins." Proteins **45**(1): 47-54.

Schulze, K., Medina, E., Chhatwal, G. S. and Guzman, C. A. (2003). "Identification of B- and T-cell epitopes within the fibronectin-binding domain of the SfbI protein of Streptococcus pyogenes." Infect. Immun. **71**(12): 7197-201.

Schumacher, T. N., Heemels, M. T., Neefjes, J. J., Kast, W. M., Melief, C. J. and Ploegh, H. L. (1990). "Direct binding of peptide to empty MHC class I molecules on intact cells and in vitro." Cell **62**(3): 563-7.

Scognamiglio, P., Accapezzato, D., Casciaro, M. A., Cacciani, A., Artini, M., Bruno, G., Chircu, M. L., Sidney, J., Southwood, S., Abrignani, S., Sette, A. and Barnaba, V. (1999). "Presence of effector CD8+ T cells in hepatitis C virus-exposed healthy seronegative donors." J. Immunol. **162**(11): 6681-9.

Scott, J. E. and Dawson, J. R. (1995). "MHC class I expression and transport in calnexin-deficient cell line." J. Immunol. **155**: 143-8.

Segal, M. R., Cummings, M. P. and Hubbard, A. E. (2001). "Relating amino acid sequence to phenotype: analysis of peptide-binding data." Biometrics **57**(2): 632-42.

Selassie, C. D. (2003). Burger's medicinal chemistry and drug discovery, John Wiley & Sons, Inc.

Serwold, T., Gonzalez, F., Kim, J., Jacob, R. and Shastri, N. (2002). "ERAAP customizes peptides for MHC class I molecules in the endoplasmic reticulum." Nature **419**(6906): 480-3.

Sette, A., Buus, S., Appella, E., Smith, J. A., Chesnut, R., Miles, C., Colon, S. M. and Grey, H. M. (1989a). "Prediction of major histocompatibility complex binding regions of protein antigens by sequence pattern analysis." Proc. Natl. Acad. Sci. U. S. A. **86**(9): 3296-300.

Sette, A., Buus, S., Colon, S., Miles, C. and Grey, H. (1989b). "Structural analysis of peptides capable of binding to more than one Ia antigen." J. Immunol. **142**: 35-40.

Sette, A., Vitiello, A., Reherman, B., Fowler, P., Nayersina, R., Kast, W. M., Melief, C. J., Oseroff, C., Yuan, L., Ruppert, J. and et al. (1994). "The relationship between class I binding affinity and immunogenicity of potential cytotoxic T cell epitopes." J. Immunol. **153**(12): 5586-92.

Sette, A. and Sidney, J. (1998). "HLA supertypes and supermotifs: a functional perspective on HLA polymorphism." Curr. Opin. Immunol. **10**(4): 478-82.

Sette, A. and Sidney, J. (1999). "Nine major HLA class I supertypes account for the vast preponderance of HLA-A and -B polymorphism." Immunogenetics **50**(3-4): 201-12.

Sette, A., Livingston, 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**(3-4): 271-276.

Sette, A., Keogh, E., Ishioka, G., Sidney, J., Tangri, S., Livingston, B., McKinney, D., Newman, M., Chesnut, R. and Fikes, J. (2002). "Epitope identification and vaccine design for cancer immunotherapy." Curr Opin Investig Drugs **3**(1): 132-9.

Sezerman, U., Vajda, S., Cornette, J. and DeLisi, C. (1993). "Toward computational determination of peptide-receptor structure." Protein Sci. **2**(11): 1827-43.

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.

Shaffer, R. E., Small, G. W. and Arnold, M. A. (1996). "Genetic algorithm-based protocol for coupling digital filtering and partial least-squares regression: application to the near-infrared analysis of glucose in biological matrices." Anal. Chem. **68**(15): 2663-75.

Shastri, N., Serwold, T. and Gonzalez, F. (1995). "Presentation of endogenous peptide/MHC class I complexes is profoundly influenced by specific C-terminal flanking residues." J. Immunol. **155**: 4339-4346.

Shields, M. J., Hodgson, W. and Ribaudo, R. K. (1999). "Differential association of beta2-microglobulin mutants with MHC class I heavy chains and structural analysis demonstrate allele-specific interactions." Mol. Immunol. **36**(9): 561-73.

Shigematsu, H., Shimoda, S., Nakamura, M., Matsushita, S., Nishimura, Y., Sakamoto, N., Ichiki, Y., Niho, Y., Gershwin, M. E. and Ishibashi, H. (2000). "Fine specificity of T cells reactive to human PDC-E2 163-176 peptide, the immunodominant autoantigen in primary biliary cirrhosis: implications for molecular mimicry and cross-recognition among mitochondrial autoantigens." Hepatology **32**(5): 901-9.

Sidney, J., del Guercio, M. F., Southwood, S., Engelhard, V. H., Appella, E., Rammensee, H. G., Falk, K., Rotzschke, O., Takiguchi, M., Kubo, R. T. and et al. (1995). "Several HLA alleles share overlapping peptide specificities." J. Immunol. **154**(1): 247-59.

Sidney, J., Grey, H. M., Kubo, R. T. and Sette, A. (1996a). "Practical, biochemical and evolutionary implications of the discovery of HLA class I supermotifs." Immunol Today **17**(6): 261-6.

Sidney, J., Grey, H. M., Southwood, S., Celis, E., Wentworth, P. A., del Guercio, M. F., Kubo, R. T., Chesnut, R. W. and Sette, A. (1996). "Definition of an HLA-A3-like supermotif demonstrates the overlapping peptide-binding repertoires of common HLA molecules." Hum. Immunol. **45**(2): 79-93.

Sidney, J., Southwood, S., Mann, D. L., Fernandez-Vina, M. A., Newman, M. J. and Sette, A. (2001). "Majority of peptides binding HLA-A*0201 with high affinity crossreact with other A2-supertype molecules." Hum. Immunol. **62**(11): 1200-16.

Sidney, J., Southwood, S., Pasquetto, V. and Sette, A. (2003). "Simultaneous prediction of binding capacity for multiple molecules of the HLA B44 supertype." J. Immunol. **171**(11): 5964-74.

Siebert, K. J. (2001). "Quantitative structure-activity relationship modeling of peptide and protein behavior as a function of amino acid composition." J. Agric. Food Chem. **49**(2): 851-8.

Silver, M. L., Guo, H. C., Strominger, J. L. and Wiley, D. C. (1992). "Atomic structure of a human MHC molecule presenting an influenza virus peptide." Nature **360**(6402): 367-9.

Singh, S. P., Mehra, N. K., Dingley, H. B., Pande, J. N. and Vaidya, M. C. (1983). "Human leukocyte antigen (HLA)-linked control of susceptibility to pulmonary tuberculosis and association with HLA-DR types." J. Infect. Dis. **148**(4): 676-81.

Singh, H. and Raghava, G. P. (2001). "ProPred: prediction of HLA-DR binding sites." Bioinformatics **17**(12): 1236-7.

Smith, S. G. (1999). "The polyepitope approach to DNA vaccination." Curr Opin Mol Ther **1**(1): 10-5.

Sneath, P. H. (1966). "Relations between chemical structure and biological activity in peptides." J. Theor. Biol. **12**(2): 157-95.

Song, R. and Harding, C. V. (1996). "Roles of proteasomes, transporter for antigen presentation (TAP), and beta 2-microglobulin in the processing of bacterial or particulate antigens via an alternate class I MHC processing pathway." J. Immunol. **156**(11): 4182-90.

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**(7): 3363-73.

Sperandio Da Silva, G. M., Sant'Anna, C. M. and Barreiro, E. J. (2004). "A novel 3D-QSAR comparative molecular field analysis (CoMFA) model of imidazole and quinazolinone functionalized p38 MAP kinase inhibitors." Bioorg. Med. Chem. **12**(12): 3159-66.

Steere, A. C., Falk, B., Drouin, E. E., Baxter-Lowe, L. A., Hammer, J. and Nepom, G. T. (2003). "Binding of outer surface protein A and human lymphocyte function-associated antigen 1 peptides to HLA-DR molecules associated with antibiotic treatment-resistant Lyme arthritis." Arthritis. Rheum. **48**(2): 534-40.

Stefaniak, B., Cholewinski, W. and Tarkowska, A. (2004). "Prediction of left ventricular ejection fraction in patients with coronary artery disease based on an analysis of perfusion patterns at rest. Assessment by an artificial neural network." Nucl Med Rev Cent East Eur **7**(1): 7-12.

Stephansson, E. A., Koskimies, S. and Lokki, M. L. (1993). "HLA antigens and complement C4 allotypes in patients with chronic biologically false positive (CBFP) seroreactions for syphilis: a follow-up study of SLE patients and CBFP reactors." Lupus **2**(2): 77-81.

Stevens, J., Wiesmuller, K. H., Walden, P. and Joly, E. (1998). "Peptide length preferences for rat and mouse MHC class I molecules using random peptide libraries." Eur. J. Immunol. **28**(4): 1272-9.

Stoltze, L., Schirle, M., Schwarz, G., Schroter, C., Thompson, M. W., Hersh, L. B., Kalbacher, H., Stevanovic, S., Rammensee, H. G. and Schild, H. (2000). "Two new proteases in the MHC class I processing pathway." Nat. Immunol. **1**(5): 413-8.

Stone, J. D., Howlett, S., Byth, K. F., Holmes, N. and Alexander, D. R. (1997). "T-cell antigen receptor signal transduction in CD45-/- thymocytes." Biochem. Soc. Trans. **25**(2): 302S.

Stryhn, A., Pedersen, L. O., Romme, T., Holm, C. B., Holm, A. and Buus, S. (1996). "Peptide binding specificity of major histocompatibility complex class I resolved into an array of apparently independent subspecificities: quantitation by peptide libraries and improved prediction of binding." Eur. J. Immunol. **26**(8): 1911-8.

Stuber, G., Dillner, J., Modrow, S., Wolf, H., Szekely, L., Klein, G. and Klein, E. (1995). "HLA-A0201 and HLA-B7 binding peptides in the EBV-encoded EBNA-1, EBNA-2 and BZLF-1 proteins detected in the MHC class I stabilization assay. Low proportion of binding motifs for several HLA class I alleles in EBNA-1." Int. Immunol. **7**(4): 653-63.

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**(6): 555-61.

Sudo, T., Kamikawaji, N., Kimura, A., Date, Y., Savoie, C. J., Nakashima, H., Furuichi, E., Kuhara, S. and Sasazuki, T. (1995). "Differences in MHC class I self peptide repertoires among HLA-A2 subtypes." J. Immunol. **155**(10): 4749-56.

Sugita, Y., Wada, H., Fujita, S., Nakata, T., Sato, S., Noguchi, Y., Jungbluth, A. A., Yamaguchi, M., Chen, Y. T., Stockert, E., Gnjatic, S., Williamson, B., Scanlan, M. J., Ono, T., Sakita, I., Yasui, M., Miyoshi, Y., Tamaki, Y., Matsuura, N., Noguchi, S., Old, L. J., Nakayama, E. and Monden, M. (2004). "NY-ESO-1 expression and immunogenicity in malignant and benign breast tumors." Cancer Res. **64**(6): 2199-204.

Suh, W. K., Mitchell, E. K., Yang, Y., Peterson, P. A., Waneck, G. L. and Williams, D. B. (1996). "MHC class I molecules form ternary complexes with calnexin and TAP and undergo peptide-regulated interaction with TAP via their extracellular domains." J. Exp. Med. **184**(2): 337-48.

Suhrbier, A., Schmidt, C. and Fernan, A. (1993). "Prediction of an HLA B8-restricted influenza epitope by motif." Immunology **79**(1): 171-3.

Sun, Y., Sijts, A. J., Song, M., Janek, K., Nussbaum, A. K., Kral, S., Schirle, M., Stevanovic, S., Paschen, A., Schild, H., Kloetzel, P. M. and Schadendorf, D. (2002). "Expression of the proteasome activator PA28 rescues the presentation of a cytotoxic T lymphocyte epitope on melanoma cells." Cancer Res. **62**(10): 2875-82.

Sun, D., Zhang, Y., Wei, B., Peiper, S., Shao, H. and Kaplan, H. (2003). "Encephalitogenic activity of truncated myelin oligodendrocyte glycoprotein (MOG) peptides and their recognition by CD8+ MOG-specific T cells on oligomeric MHC class I molecules." Int. Immunol. **15**: 261-268.

Sun, H. (2004). "A universal molecular descriptor system for prediction of logP, logS, logBB, and absorption." J Chem Inf Comput Sci **44**(2): 748-57.

Sundaram, R., Beebe, M. and Kaumaya, P. T. (2004). "Structural and immunogenicity analysis of chimeric B-cell epitope constructs derived from the gp46 and gp21 subunits of the envelope glycoproteins of HTLV-1." J. Pept. Res. **63**(2): 132-40.

Sung, M. H., Zhao, Y., Martin, R. and Simon, R. (2002). "T-cell epitope prediction with combinatorial peptide libraries." J. Comput. Biol. **9**(3): 527-39.

Suzuki, E., Kessler, M., Montgomery, K. E. and Arai, A. C. (2004). "Divergent effects of the purinoceptor antagonists suramin and PPNDS on AMPA receptors." Mol. Pharmacol.

Sylvester-Hvid, C., Nielsen, M., Lamberth, K., Roder, G., Justesen, S., Lundegaard, C., Worning, P., Thomadsen, H., Lund, O., Brunak, S. and Buus, S. (2004). "SARS CTL vaccine candidates; HLA supertype-, genome-wide scanning and biochemical validation." Tissue Antigens **63**(5): 395-400.

Szmuness, W., Stevens, C. E., Harley, E. J., Zang, E. A., Taylor, P. E. and Alter, H. J. (1981a). "The immune response of healthy adults to a reduced dose of hepatitis B vaccine." J. Med. Virol. **8**(2): 123-9.

Szmuness, W., Stevens, C. E., Oleszko, W. R. and Goodman, A. (1981b). "Passive-active immunisation against hepatitis B: immunogenicity studies in adult Americans." Lancet **1**(8220 Pt 1): 575-7.

Takiguchi, M. (1994). "[Structure of MHC molecules and binding peptides]." Nippon Rinsho **52**(11): 2817-23.

Tatsumi, T., Kierstead, L., Ranieri, E., Gesualdo, L., Schena, F., Finke, J., Bukowski, R., Brusic, V., Sidney, J., Sette, A., Logan, T., Kasamon, Y., Slingluff, C., Kirkwood, J. and Storkus, W. (2003). "MAGE-6 encodes HLA-DRB1*0401-presented epitopes recognized by CD4+ T cells from patients with melanoma or renal cell carcinoma." Clinical Cancer Research **9**: 947-954.

Teoh, C. Y. and Davies, K. J. A. (2004). "Potential roles of protein oxidation and the immunoproteasome in MHC class I antigen presentation: the 'PrOxI' hypothesis." Arch. Biochem. Biophys. **423**(1): 88-96.

Terajima, M., Cruz, J., Raines, G., Kilpatrick, E., Kennedy, J., Rothman, A. and Ennis, F. (2003). "Quantitation of CD8+ T cell responses to newly identified HLA-A*0201-restricted T cell epitopes conserved among vaccinia and variola (smallpox) viruses." J. Exp. Med. **197**: 927-932.

Terp, G. E., Cruciani, G., Christensen, I. T. and Jorgensen, F. S. (2002). "Structural differences of matrix metalloproteinases with potential implications for inhibitor selectivity examined by the GRID/CPCA approach." J. Med. Chem. **45**(13): 2675-84.

Thissen, U., Ustun, B., Melssen, W. J. and Buydens, L. M. (2004). "Multivariate calibration with least-squares support vector machines." Anal. Chem. **76**(11): 3099-105.

Thomsen, M., Morling, N., Platz, P., Ryder, L. P. and Svejgaard, A. (1979). "HLA and disease." Transplant Proc **11**(1): 633-7.

Threlkeld, S. C., Wentworth, P. A., Kalams, S. A., Wilkes, B. M., Ruhl, D. J., Keogh, E., Sidney, J., Southwood, S., Walker, B. D. and Sette, A. (1997). "Degenerate and promiscuous recognition by CTL of peptides presented by the MHC class I A3-like superfamily: implications for vaccine development." J. Immunol. **159**(4): 1648-57.

Tibbs, C., Donaldson, P., Underhill, J., Thomson, L., Manabe, K. and Williams, R. (1996). "Evidence that the HLA DQA1*03 allele confers protection from chronic HCV-infection in Northern European Caucasoids." Hepatology **24**(6): 1342-5.

Toes, R. E., Hoeben, R. C., van der Voort, E. I., Ressing, M. E., van der Eb, A. J., Melief, C. J. and Offringa, R. (1997). "Protective anti-tumor immunity induced by vaccination with recombinant adenoviruses encoding multiple tumor-associated cytotoxic T lymphocyte epitopes in a string-of-beads fashion." Proc. Natl. Acad. Sci. U. S. A. **94**(26): 14660-5.

Toh, H., Savoie, C. J., Kamikawaji, N., Muta, S., Sasazuki, T. and Kuhara, S. (2000). "Changes at the floor of the peptide-binding groove induce a strong preference for proline at position 3 of the bound peptide: molecular dynamics simulations of HLA-A*0217." Biopolymers **54**(5): 318-27.

Tomita, Y., Tomida, S., Hasegawa, Y., Suzuki, Y., Shirakawa, T., Kobayashi, T. and Honda, H. (2004). "Artificial neural network approach for selection of susceptible single nucleotide polymorphisms and construction of prediction model on childhood allergic asthma." BMC Bioinformatics **5**(1): 120.

Tomkinson, B. (1999). "Tripeptidyl peptidases: enzymes that count." Trends Biochem. Sci. **24**(9): 355-9.

Tong, J. C., Tan, T. W. and Ranganathan, S. (2004). "Modeling the structure of bound peptide ligands to major histocompatibility complex." Protein Sci. **13**(9): 2523-32.

Topalian, S. L., Rivoltini, L., Mancini, M., Markus, N. R., Robbins, P. F., Kawakami, Y. and Rosenberg, S. A. (1997). "Human CD4+ T cells specifically recognize a shard melanoma-associated antigen encoded by the tyrosinase gene." Proceedings of the National Academy of Sciences **91**: 9461-9465.

Touloukian, C. E., Leitner, W. W., Robbins, P. F., Li, Y. F., Kang, X., Lapointe, R., Hwu, P., Rosenberg, S. A. and Restifo, N. P. (2002). "Expression of a "self-"antigen by human tumor cells enhances tumor antigen-specific CD4(+) T-cell function." Cancer Res. **62**(18): 5144-7.

Townsend, A., Ohlen, C., Bastin, J., Ljunggren, H. G., Foster, L. and Karre, K. (1989). "Association of class I major histocompatibility heavy and light chains induced by viral peptides." Nature **340**(6233): 443-8.

Townsend, A., Elliott, T., Cerundolo, V., Foster, L., Barber, B. and Tse, A. (1990). "Assembly of MHC class I molecules analyzed in vitro." Cell **62**(2): 285-95.

Traversari, C., van der Bruggen, P., Luescher, I. F., Lurquin, C., Chomez, P., Van Pel, A., De Plaen, E., Amar-Costesec, A. and Boon, T. (1992). "A nonapeptide encoded by human gene MAGE-1 is recognized on HLA-A1 by cytolytic T lymphocytes directed against tumor antigen MZ2-E." J. Exp. Med. **176**(5): 1453-7.

Trojan, A., Urosevic, M., Hummerjohann, J., Giger, R., Schanz, U. and Stahel, R. (2003). "Immune reactivity against a novel HLA-A3 restricted influenza virus peptide identified by predictive algorithms and interferon-r quantitative PCR." J. Immunother. **26**: 41-46.

Tsai, V., Southwood, S., Sidney, J., Sakaguchi, K., Kawakami, Y., Appella, E., Sette, A. and Celis, E. (1997). "Identification of subdominant CTL epitopes of the GP100 melanoma-associated tumor antigen by primary in vitro immunization with peptide-pulsed dendritic cells." J. Immunol. **158**(4): 1796-802.

Tsang, K. Y., Palena, C., Gulley, J., Arlen, P. and Schlom, J. (2004). "A human cytotoxic T-lymphocyte epitope and its agonist epitope from the nonvariable number of tandem repeat sequence of MUC-1." Clin Cancer Res **10**(6): 2139-49.

Tzakos, A. G., Fuchs, P., van Nuland, N. A., Troganis, A., Tselios, T., Deraos, S., Matsoukas, J., Gerothanassis, I. P. and Bonvin, A. M. (2004). "NMR and molecular dynamics studies of an autoimmune myelin basic protein peptide and its antagonist: structural implications for the MHC II (I-Au)-peptide complex from docking calculations." Eur. J. Biochem. **271**(16): 3399-413.

Udaka, K., Wiesmuller, K. H., Kienle, S., Jung, G. and Walden, P. (1995). "Tolerance to amino acid variations in peptides binding to the major histocompatibility complex class I protein H-2Kb." J. Biol. Chem. **270**(41): 24130-4.

Udaka, K., Wiesmuller, K. H., Kienle, S., Jung, G., Tamamura, H., Yamagishi, H., Okumura, K., Walden, P., Suto, T. and Kawasaki, T. (2000). "An automated prediction of MHC class I-binding peptides based on positional scanning with peptide libraries." Immunogenetics **51**(10): 816-28.

Udaka, K., Mamitsuka, H., Nakaseko, Y. and Abe, N. (2002). "Empirical evaluation of a dynamic experiment design method for prediction of MHC class I-binding peptides." J. Immunol. **169**(10): 5744-53.

Uehara, H., Coligan, J. E. and Nathenson, S. G. (1981a). "Amino acid sequence of the carboxyl-terminal hydrophilic region of the H-2Kb MHC alloantigen. Completion of the entire primary structure of the H-2Kb molecule." Biochemistry **20**: 5940-5945.

Uehara, H., Coligan, J. E. and Nathenson, S. G. (1981b). "Isolation and sequence analysis of the intramembranous hydrophobic segment of the H-2Kb murine histocompatibility antigen." Biochemistry **20**: 5936-5939.

Ulbrecht, M., Hofmeister, V., Yuksekdag, G., Ellwart, J. W., Hengel, H., Momburg, F., Martinozzi, S., Reboul, M., Pla, M. and Weiss, E. H. (2003). "HCMV glycoprotein US6 mediated inhibition of TAP does not affect HLA-E dependent protection of K-562 cells from NK cell lysis." Hum. Immunol. **64**(2): 231-7.

Ullenhag, G. J., Fagerberg, J., Strigard, K., Frodin, J. E. and Mellstedt, H. (2004). "Functional HLA-DR T cell epitopes of CEA identified in patients with colorectal carcinoma immunized with the recombinant protein CEA." Cancer Immunol. Immunother. **53**(4): 331-7.

Unno, M., Mizushima, T., Morimoto, Y., Tomisugi, Y., Tanaka, K., Yasuoka, N. and Tsukihara, T. (2002). "The Structure of the Mammalian 20S Proteasome at 2.75 A Resolution." structure **10**: 609.

Urbani, S., Uggeri, J., Matsuura, Y., Miyamura, T., Penna, A., Boni, C. and Ferrari, C. (2001). "Identification of immunodominant hepatitis C virus (HCV)-specific cytotoxic T-cell epitopes by stimulation with endogenously synthesized HCV antigens." Hepatology **33**(6): 1533-43.

Vajda, S. and Camacho, C. J. (2004). "Protein-protein docking: is the glass half-full or half-empty?" Trends Biotechnol. **22**(3): 110-6.

Valiante, N. M. and Parham, P. (1996). "NK cells and CTL: opposite sides of the same coin." Chem. Immunol. **64**: 146-63.

van der Bruggen, P., Traversari, C., Chomez, P., Lurquin, C., De Plaen, E., Van den Eynde, B., Knuth, A. and Boon, T. (1991). "A gene encoding an antigen recognized by cytolytic T lymphocytes on a human melanoma." Science **254**(5038): 1643-7.

van der Burg, S. H., Ras, E., Drijfhout, J. W., Benckhuijsen, W. E., Bremers, A. J., Melief, C. J. and Kast, W. M. (1995). "An HLA class I peptide-binding assay based on competition for binding to class I molecules on intact human B cells. Identification of conserved HIV-1 polymerase peptides binding to HLA-A*0301." Hum. Immunol. **44**(4): 189-98.

van der Burg, S. H., Visseren, M. J., Brandt, R. M., Kast, W. M. and Melief, C. J. (1996). "Immunogenicity of peptides bound to MHC class I molecules depends on the MHC-peptide complex stability." J. Immunol. **156**(9): 3308-14.

van der Voet, H. and Franke, J. P. (1985). "A discussion of principal component analysis." J Anal Toxicol **9**(4): 185-8.

van Eden, W., de Vries, R. R., Mehra, N. K., Vaidya, M. C., D'Amaro, J. and van Rood, J. J. (1980). "HLA segregation of tuberculoid leprosy: confirmation of the DR2 marker." J. Infect. Dis. **141**(6): 693-701.

van Endert, P. M. (2001). "Designing peptide vaccines for cellular cross-presentation." Biologicals **29**(3-4): 285-8.

van Rood, J. J., Eernisse, J. G. and van Leeuwen, A. (1958). "Leukocyte antibodies in sera of pregnant women." Nature **181**: 1735-1736.

Vapnik, v. (1998). Statistical learning theory. New York, Wiley-Interscience.

Vasmatzis, G., Zhang, C., Cornette, J. L. and DeLisi, C. (1996). "Computational determination of side chain specificity for pockets in class I MHC molecules." Mol. Immunol. **33**(16): 1231-9.

Venter, M., Rock, M., Puren, A. J., Tiemessen, C. T. and Crowe, J. E., Jr. (2003). "Respiratory syncytial virus nucleoprotein-specific cytotoxic T-cell epitopes in a South African population of diverse HLA types are conserved in circulating field strains." J. Virol. **77**(13): 7319-29.

Vierboom, M. P., Nijman, H. W., Offringa, R., van der Voort, E. L., van Hall, T., van der Broek, L., Fleuren, G. J., Kenemans, P., Kast, W. M. and Melief, C. J. (1997). "Tumor eradication by wild-type p53-specific cytotoxic T lymphocytes." J. Exp. Med. **186**: 695-704.

Vijh, S., Pilip, I. M. and Pamer, E. G. (1998). "Effect of antigen-processing efficiency on in vivo T cell response magnitudes." J. Immunol. **160**: 3971-3977.

Vinayagam, A., Koenig, R., Moormann, J., Schubert, F., Eils, R., Glatting, K. H. and Suhai, S. (2004). "Applying Support Vector Machines for Gene ontology based gene function prediction." BMC Bioinformatics **5**(1): 116.

Vitiello, A., Sette, A., Yuan, L., Farness, P., Southwood, S., Sidney, J., Chesnut, R. W., Grey, H. M. and Livingston, B. (1997). "Comparison of cytotoxic T lymphocyte responses induced by peptide or DNA immunization: implications on immunogenicity and immunodominance." Eur. J. Immunol. **27**(3): 671-8.

Vonderheide, R. H., Hahn, W. C., Schultze, J. L. and Nadler, L. M. (1999). "The telomerase catalytic subunit is a widely expressed tumor-associated antigen recognized by cytotoxic T lymphocytes." Immunity **10**(6): 673-9.

Wagner, C., Neumann, F., Kubuschok, B., Regitz, E., Mischo, A., Stevanovic, S., Friedrich, M., Schmidt, W., Rammensee, H. G. and Pfreundschuh, M. (2003). "Identification of an HLA-A*02 restricted immunogenic peptide derived from the cancer testis antigen HOM-MEL-40/SSX2." Cancer Immun **3**: 18.

Wang, B., Chen, H., Jiang, X., Zhang, M., Wan, T., Li, R., Zhou, X., Wu, Y., Yang, F., Yu, Y., Wang, X., Yang, R. and Cao, X. (2004). "Identification of an HLA-A*0201-restricted CD8+ T-cell epitope SSp-1 of SARS-CoV spike protein." Blood: Ahead of print.

Wang, M., Yang, J., Liu, G. P., Xu, Z. J. and Chou, K. C. (2004). "Weighted-support vector machines for predicting membrane protein types based on pseudo amino acid composition." Protein Eng Des Sel.

Wang, R. F., Johnston, S. L., Southwood, S., Sette, A. and Rosenberg, S. A. (1998). "Recognition of an antigenic peptide derived from tyrosinase-related protein-2 by CTL in the context of HLA-A31 and -A33." J. Immunol. **160**(2): 890-7.

Wauben, M. H., van der Kraan, M., Grosfeld-Stulemeyer, M. C. and Joosten, I. (1997). "Definition of an extended MHC class II-peptide binding motif for the autoimmune disease-associated Lewis rat RT1.BL molecule." Int. Immunol. **9**(2): 281-90.

Welsh, K. and Bunce, M. (1999). "Molecular typing for the MHC with PCR-SSP." Rev Immunogenet **1**(2): 157-76.

Welsh, K. I. (1989). "MHC antigens: new methods in tissue typing." Curr. Opin. Immunol. **1**(6): 1178-83.

Wertheimer, A. M., Miner, C., Lewinsohn, D. M., Sasaki, A. W., Kaufman, E. and Rosen, H. R. (2003). "Novel CD4+ and CD8+ T-cell determinants within the NS3 protein in subjects with spontaneously resolved HCV infection." Hepatology **37**(3): 577-89.

Westbrook, J., Feng, Z., Jain, S., Bhat, T. N., Thanki, N., Ravichandran, V., Gilliland, G. L., Bluhm, W., Weissig, H., Greer, D. S., Bourne, P. E. and Berman, H. M. (2002). "The Protein Data Bank: unifying the archive." Nucleic Acids Res. **30**(1): 245-8.

Westerdahl, H., Wittzell, H., von Schantz, T. and Bensch, S. (2004). "MHC class I typing in a songbird with numerous loci and high polymorphism using motif-specific PCR and DGGE." Heredity **92**(6): 534-42.

Wettstein, P. J., van Bleek, G. M. and Nathenson, S. G. (1993). "Differential binding of a minor histocompatibility antigen peptide to H-2 class I molecules correlates with immune responsiveness." J. Immunol. **150**(7): 2753-60.

Wiertz, E. J., Mukherjee, S. and Ploegh, H. L. (1997). "Viruses use stealth technology to escape from the host immune system." Mol Med Today **3**(3): 116-23.

Willcox, B. E., Thomas, L. M. and Bjorkman, P. J. (2003). "Crystal structure of HLA-A2 bound to LIR-1, a host and viral major histocompatibility complex receptor." Nat. Immunol. **4**(9): 913-9.

Williams, R. D., Hing, S. N., Greer, B. T., Whiteford, C. C., Wei, J. S., Natrajan, R., Kelsey, A., Rogers, S., Campbell, C., Pritchard-Jones, K. and Khan, J. (2004). "Prognostic classification of relapsing favorable histology Wilms tumor using cDNA microarray expression profiling and support vector machines." Genes Chromosomes Cancer **41**(1): 65-79.

Wilson, I. A. and Garcia, K. C. (1997). "T-cell receptor structure and TCR complexes." Curr. Opin. Struct. Biol. 7(6): 839-48.

Wold, S. (1995). PLS for multivariate linear modelling. Chemometric methods in molecular design. H. van de Waterbeemd. VCH, Weinheim: 195-218.

Wold, S., Hellberg, S., Lundstedt, T., Sjostrom, M. and Wold, H. (1987). Proc. Symp. on PLS Model Building: Theory and Application. Germany, Frankfurt am Main.

Wu, B., Elst, L., Carlier, V., Jacquemin, M. and Saint-Remy, J. (2002a). "The Dermatophagoides pteronyssinus group 2 allergen contains a universally immunogenic T cell epitope." J. Immunol. **169**: 2430-2435.

Wu, L. C., Tuot, D. S., Lyons, D. S., Garcia, K. C. and Davis, M. M. (2002b). "Two-step binding mechanism for T-cell receptor recognition of peptide MHC." Nature **418**(6897): 552-6.

Xing, L., Welsh, W. J., Tong, W., Perkins, R. and Sheehan, D. M. (1999). "Comparison of estrogen receptor alpha and beta subtypes based on comparative molecular field analysis (CoMFA)." SAR QSAR Environ Res **10**(2-3): 215-37.

Xing, P. X., Poulos, G. and McKenzie, I. F. (2001). "Breast cancer in mice: effect of murine MUC-1 immunization on tumor incidence in C3H/HeOuj mice." J. Immunother. **24**(1): 10-8.

Xing, L. and Glen, R. C. (2002). "Novel methods for the prediction of logP, pK(a), and logD." J Chem Inf Comput Sci **42**(4): 796-805.

Xing, L., Glen, R. C. and Clark, R. D. (2003). "Predicting pK(a) by molecular tree structured fingerprints and PLS." J Chem Inf Comput Sci **43**(3): 870-9.

Xue, C. X., Zhang, R. S., Liu, H. X., Liu, M. C., Hu, Z. D. and Fan, B. T. (2004a). "Support vector machines-based quantitative structure-property relationship for the prediction of heat capacity." J Chem Inf Comput Sci **44**(4): 1267-74.

Xue, C. X., Zhang, R. S., Liu, H. X., Yao, X. J., Liu, M. C., Hu, Z. D. and Fan, B. T. (2004b). "An accurate QSPR study of O-H bond dissociation energy in substituted phenols based on support vector machines." J Chem Inf Comput Sci **44**(2): 669-77.

Xue, C. X., Zhang, R. S., Liu, M. C., Hu, Z. D. and Fan, B. T. (2004c). "Study of the quantitative structure-mobility relationship of carboxylic acids in capillary electrophoresis based on support vector machines." J Chem Inf Comput Sci **44**(3): 950-7.

Yamano, T., Murata, S., Shimbara, N., Tanaka, N., Chiba, T., Tanaka, K., Yui, K. and Udono, H. (2002). "Two distinct pathways mediated by PA28 and hsp90 in major histocompatibility complex class I antigen processing." J. Exp. Med. **196**(2): 185-96.

Yamashita, F., Wanchana, S. and Hashida, M. (2002). "Quantitative structure/property relationship analysis of Caco-2 permeability using a genetic algorithm-based partial least squares method." J. Pharm. Sci. **91**(10): 2230-9.

Yao, X. J., Panaye, A., Doucet, J. P., Zhang, R. S., Chen, H. F., Liu, M. C., Hu, Z. D. and Fan, B. T. (2004). "Comparative study of QSAR/QSPR correlations using support vector machines, radial basis function neural networks, and multiple linear regression." J Chem Inf Comput Sci **44**(4): 1257-66.

Yao, Z., Hartung, K., Deicher, H. G., Brunnler, G., Bettinotti, M. P., Keller, E., Paul, C., Gawron, C., Mikschl, S. and Albert, E. (1993). "DNA typing for HLA-DPB1-alleles in German patients with systemic lupus erythematosus using the polymerase chain reaction and DIG-ddUTP-labelled oligonucleotide probes. Members of SLE Study Group." Eur. J. Immunogenet. **20**(4): 259-66.

Yao, Z., Seelig, H. P., Ehrfeld, H., Renz, M., Hartung, K., Deicher, H., Keller, E., Nevinny-Stickel, C. and Albert, E. D. (1994). "HLA class II genes and antibodies against recombinant U1-nRNP proteins in patients with systemic lupus erythematosus. SLE Study Group." Rheumatol. Int. **14**(2): 63-9.

Yellen-Shaw, A. J., Wherry, E. J., Dubois, G. C. and Eisenlohr, L. C. (1997). "Point mutation flanking a CTL epitope ablates in vitro and in vivo recognition of a full-length viral protein." J. Immunol. **158**: 3227-3234.

Yokoyama, W. M., Daniels, B. F., Seaman, W. E., Hunziker, R., Margulies, D. H. and Smith, H. R. (1995). "A family of murine NK cell receptors specific for target cell MHC class I molecules." Semin Immunol **7**(2): 89-101.

Yoon, H., Chung, M. K., Min, S. S., Lee, H. G., Yoo, W. D., Chung, K. T., Jung, N. P. and Park, S. N. (1998). "Synthetic peptides of human papillomavirus type 18 E6 harboring HLA-A2.1 motif can induce peptide-specific cytotoxic T-cells from peripheral blood mononuclear cells of healthy donors." Virus Res. **54**(1): 23-9.

York, I. A., Chang, S. C., Saric, T., Keys, J. A., Favreau, J. M., Goldberg, A. L. and Rock, K. L. (2002). "The ER aminopeptidase ERAP1 enhances or limits antigen presentation by trimming epitopes to 8-9 residues." Nat. Immunol. **3**(12): 1177-84.

Yoshizawa, K. and Yano, A. (1984). "Mouse T lymphocytes proliferative responses specific for human MHC products in mouse anti-human xenogeneic MLR." J. Immunol. **132**(6): 2820-9.

Young, A. C., Zhang, W., Sacchettini, J. C. and Nathenson, S. G. (1994). "The three-dimensional structure of H-2Db at 2.4 A resolution: implications for antigen-determinant selection." Cell **76**(1): 39-50.

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**(3): 137-48.

Zarour, H., Maillere, B., Brusic, V., Coval, K., Williams, E., Poivelle-Moratille, S., Castelli, F., Land, S., Bennouna, J., Logan, T. and Kirkwook, J. (2002). "EY-ESO-1 119-143 is a promiscuous major histocompatibility complex class II T-helper epitope recognized by Th1- and Th2- type tumor-reactive CD4+ T cells." Cancer Res. **62**: 213-218.

Zehbe, I., Mytilineos, J., Wikstrom, I., Henriksen, R., Edler, L. and Tommasino, M. (2003). "Association between human papillomavirus 16 E6 variants and human leukocyte antigen class I polymorphism in cervical cancer of Swedish women." Hum. Immunol. **64**(5): 538-42.

Zeng, G., Li, Y., El-Gamil, M., Sidney, J., Sette, A., Wang, R. F., Rosenberg, S. A. and Robbins, P. F. (2002). "Generation of NY-ESO-1-specific CD4+ and CD8+ T cells by a single peptide with dual MHC class I and class II specificities: a new strategy for vaccine design." Cancer Res. **62**(13): 3630-5.

Zeng, J., Treutlein, H. R. and Rudy, G. B. (2001). "Predicting sequences and structures of MHC-binding peptides: a computational combinatorial approach." J. Comput-Aided. Mol. Des. **15**: 573-586.

Zerbini, A., Pilli, M., Soliani, P., Ziegler, S., Pelosi, G., Orlandini, A., Cavallo, C., Uggeri, J., Scandroglio, R., Crafa, P., Spagnoli, G. C., Ferrari, C. and Missale, G. (2004). "Ex vivo characterization of tumor-derived melanoma antigen encoding gene-specific CD8+cells in patients with hepatocellular carcinoma." J. Hepatol. **40**(1): 102-9.

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

Zhang, C., Cornette, J. L. and Delisi, C. (1997). "Consistency in structural energetics of protein folding and peptide recognition." Protein Sci. **6**(5): 1057-64.

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. U. S. A. **90**(6): 2217-21.

Zhang, Z. and Wood, W. I. (2003). "A profile hidden Markov model for signal peptides generated by HMMER." Bioinformatics **19**(2): 307-8.

Zhao, B., Mathura, V. S., Rajaseger, G., Moochhala, S., Sakharkar, M. K. and Kangueane, P. (2003a). "A novel MHCp binding prediction model." Hum. Immunol. **64**(12): 1123-43.

Zhao, H., Pfeiffer, R. and Gail, M. H. (2003b). "Haplotype analysis in population genetics and association studies." Pharmacogenomics **4**(2): 171-8.

Zhao, Y. and Chalt, B. T. (1994). "Protein epitope mapping by mass spectrometry." Anal. Chem. **66**(21): 3723-6.

Zhao, Y., Jona, J., Chow, D. T., Rong, H., Semin, D., Xia, X., Zanon, R., Spancake, C. and Maliski, E. (2002). "High-throughput logP measurement using parallel liquid chromatography/ultraviolet/mass spectrometry and sample-pooling." Rapid Commun. Mass Spectrom. **16**(16): 1548-55.

Zhao, Y., Pinilla, C., Valmori, D., Martin, R. and Simon, R. (2003c). "Application of support vector machines for T-cell epitopes prediction." Bioinformatics **19**(15): 1978-84.

Zheng, D., O'Keefe, G., Li, L., Johnson, L. W. and Ewald, S. J. (1999). "A PCR method for typing B-L beta II family (class II MHC) alleles in broiler chickens." Anim. Genet. **30**(2): 109-19.

Zinkernagel, R. M. (1986). "Biological role of major transplantation antigens in T cell self-recognition." Experientia **42**(9): 970-2.