Identifying Major Histocompatibility Complex Supertypes

Pingping Guan¹, Irini A. Doytchinova², and Darren R. Flower²

¹ John Innes Centre, Norwich, NR4 7UH, UK
² The Jenner Institute, University of Oxford, Compton, Berkshire, RG20 7NN,
UKDarren.flowern@jenner.ac.uk

Summary. Human leukocyte antigen (HLA) recognizes antigenic fragments and presents them to T cells. HLA is polymorphic. There are over 2000 different HLA alleles at present and the number is constantly increasing. However, antigen binding studies are limited to a small proportion of these alleles; the binding specificities of most alleles are unknown. Several research groups have attempted to partition different HLA alleles into groups. In this chapter previous classifications are reviewed and we present two chemometric approaches to classifying class I HLA alleles. The program GRID is used to calculate interaction energy between protein molecules and defined chemical probes. These interaction energy values are imported into another program GOLPE and used for principal component analysis (PCA) calculation, which groups HLA alleles into supertypes. Amino acids that are involved in the classification are displayed in the loading plots of the PCA model. Another method, hierarchical clustering based on comparative molecular similarity indices (CoMSIA) is also applied to classify HLA alleles and the results are compared with those of the PCA models.

10.1 Introduction

Major histocompatibility complex (MHC) molecules are polymorphic membrane glycoproteins [Zinkernagel 1986]. Human MHCs are also called human leukocyte antigen, often abbreviated to HLA [Clark & Forman 1984]. There are two classes of HLA, class I and class II. Class I HLA is present on most nucleated cells, including 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) such as macrophages, B cells and dendritic cells. Partly as a result of their importance in mediating tissue rejection, sequencing has identified MHC proteins as amongst
the most polymorphic of all human gene products. According to the international ImMunoGeneTics information system (IMGT), there are over 2000 different HLA class I and II alleles and a significant number of new alleles are discovered every year [Robinson et al. 2003]. In Chapter 9, Borghans et al. explore the nature and origin of MHC diversity in more detail.

MHCs exhibit much polymorphic amino acid variation, and seemingly trivial alterations in the identity of binding site amino acid residues give rise to differences in peptide selectivity exhibited during peptide binding. Peptide binding assays are the most widely-used way of identifying T cell epitopes and measuring the affinities of peptides binding to MHC. Such assays include direct binding and the quantitative measurement of radio- or fluorescence- labeled peptides bound to the MHC molecules [Chen & Parham 1989, Schumacher & Heemels 1990, Cerottini & Luescher 1991, Christinck & Luscher 1991, Kast & Melief 1991, Mendez-Samperio & Jimenez-Zamudio 1991, Stuber & Dillner 1995, Wauben & van der Kraan 1997, Levitsky & Liu 2000]. Several databases have been set up to store peptide binding affinity data, such as MHCPEP [Brusic et al. 1998], MHCBN [Bhasin & Singh 2003], and AntiJen [Blythe et al. 2002, McSparren et al. 2003, Toseland et al. 2005].

Many HLA alleles have been demonstrated to bind peptides with similar anchor residues [Southwood et al. 1998]. This has led to the concept of MHC supertypes: the idea that MHCs with distinct sequences can be classified into separate groups, each of which displays equivalent, if not necessarily identical, specificities when binding peptides. The celerity of experimental research will be greatly accelerated if one could identify a procedure able to cluster HLA alleles with similar specificities. Several research groups have sought to classify HLA alleles in this way, using a wide variety of different methods. Examples of such disparate methodologies include sequence analysis [Lawlor & Warren 1991], structural analysis [Chelvanayagam 1997], use of geometrical similarity matrix methods [Cano & Fan 1998], and motif search [Sette & Sidney 1998, Lund et al. 2004].

We have recently developed and applied chemometric GRID/CPCA and hierarchical clustering methods to the identification of MHC supertypes [Doytchinova et al. 2004b]. Within vaccinology, HLA classification, using bioinformatics methods, can potentially reduce the overall experimental burden by rendering unnecessary the individual study of every allele. It can thus accelerate the discovery of both epitope-based vaccines, and other immunotherapies, that are targeted at multiple alleles. In the remainder of this chapter, we will explore attempts, both ours and those of others, to address the problem of finding and populating MHC supertypes.

10.1.1 Evolutionary Analysis

An early attempt to classify MHC molecules is from protein sequence studies [Lawlor & Warren 1991]. 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 are generated for A, B and C genes and it is found that HLA-A alleles are divided into five families: A2, A3, A9, A10 and A19. Two divergent groups of HLA-C alleles are 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 is 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 position 62, 97 and 114 in the binding site are important in the classification [Jakobsen & Gao 1998].

Another HLA grouping based on evolutionary analysis was undertaken by McKenzie et al. in 1999 [McKenzie & Pecon-Slattery 1992]. In their study, phylogenetic trees were built 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, A*4301 and the other A19 members. HLA-B and HLA-C did not form any consistent clusters.

10.1.2 Structural analysis

The binding of peptides to MHC molecules is influenced by the interactions between the side chains of bound peptides and the binding pockets within the peptide binding site. In contrast to data driven models, which rely on the accumulation of significant quantities of binding data, an important approach seeks a structural understanding of peptide binding by analysing the structure of MHC receptor binding sites. These allow connections to be identified between different MHC alleles at the functional level. Any significant similarity apparent between binding sites should also be mirrored in the overall peptide selectivities exhibited by different MHCs. Comparative investigation of such relationships should allow the prediction of similarities in peptide selectivity and the effective grouping of different alleles.

Kurata and Berzofsky studied the interaction of peptide analogs with the MHC binding site and their concomitant interactions with the T cell receptor (TCR). It was identified that the same peptide can bind to class II allele I-Ed 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 regions of the peptide in different conformations [Kurata & Berzofsky 1990]. Similarly, Gopalakrishnan and Roques simulated the interactions between a peptide and the H-2Kd binding site using the molecular dynamics 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 & Roques 1992].
In 1996, Chelvanayagam studied binding pockets and grouped HLAs according to the amino acid composition in each pocket [Chelvanayagam 1997]. HLA molecules within one group have the same or similar amino acids in a particular binding pocket are expected to bind to the same peptide. The analysis was used to classify HLA molecules that have not been studied experimentally and thus to 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].

10.1.3 Geometrical similarity matrix

Cano et al. clustered the HLA-A and HLA-B alleles by constructing similarity matrices [Cano & Fan 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 flexibility 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 as listed in table 10.4.

10.1.4 Sequence and binding motif approach

Another way of classifying HLA molecules is to group alleles with similar binding motifs together. Class I HLA molecules have been classified into superfamilies by Sette and Sidney using this approach [Sidney et al. 1996a]. Sidney et al. defined four supertypes by examining reported cross-reactive epitopes [Sidney et al. 1996a]. They then compared the sequences corresponding to the MHC binding pockets B and F. Experimentally confirmed binding motifs of the alleles were also examined, and those with similar motifs are grouped into one supertype [Southwood et al. 1998]. The supertypes identified in the paper are listed in table 10.4. 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 & Sidney 1998, Sette et al. 1989]. Recently A*2902 and A*3002 are
added to the A1 supertype [Sidney et al. 2005]. Sette and Sidney carried out further analysis in 1999 and defined a total of nine supertypes including the previously defined supertypes [Sette et al. 1989]. The nine supertypes were estimated to cover 99% of the world population (table 10.4) [Sette & Livingston 2001]. Table 10.4 lists the supertypes and alleles within each supertype.

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

It should be noted that class II HLA molecules have also been classified using a sequence approach. Chelvanayagam defined the HLA-DR roadmap by allele binding specificities and the polymorphic residues inside the binding site. The important residues were identified by studying the crystal structures of known HLA-DR-peptide complexes [Chelvanayagam 1997]. HLA-DP [Castelli & Buhot 2002] and DQ [Baas & Gao 1999] supertypes have also been defined based on a combination of binding studies to define motifs together with 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 & Reinherz 2003].

10.2 GRID/CPCA AND Hierarchical Clustering

Class I HLA supertypes have been defined by Doytchinova et al. using GRID/CPCA combined with hierarchical clustering based on comparative molecular similarity indices (CoMSIA) fields. The GRID program identifies the energetically favoured or disfavoured regions on molecules with known three-dimensional structures. Many molecules can be included in one calculation [Cruciani & Watson 1994]. A selection of chemical probes is included in the program; each probe represents atoms or functional groups with different properties. GRID calculates the interaction energy between selected chemical probes and each of the molecules. Molecular interaction fields (MIFs) between different chemical probes and a set of different HLA proteins were calculated in GRID, and these were used to build PCA and CPCA models in GOLPE.

The program Generating Optimal Linear Partial least square Estimations (GOLPE) [Cruciani & Watson 1994] has one module for PCA calculation. PCA decomposes a 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, 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 to the PCs and identifies variables that are important in distinguishing groups of observations.

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 consensus PCA (CPCA). The advantage of CPCA compared to PCA is that it compares the relative importance of each block in the calculation and makes a 'consensus' clustering of the objects. CPCA uses the same underlying 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 & Hellberg 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 also returns the weight matrix that can illustrate the contribution of each probe to the overall scores.

Cluster analysis is a process of grouping of observations into subsets or clusters, the grouping is dependent on the similarities between each observation. Commonly used clustering methods are hierarchical clustering and k-means clustering, etc. Hierarchical clustering based on the agglomerative method is used in HLA classification, in which observations are separated into n clusters at the beginning of the clustering; each cluster contains one observation. The distance between two clusters is proportional to the similarities of the observations. Clusters with the shortest distance are merged and the distance between the new cluster and others is computed. These steps are repeated until there is only one cluster left. The cluster analysis used is as implemented in Sybyl6.9, complete linkage clustering is used in distance computation, in which the maximum distance between data points in two clusters is used. The clustering process makes use of the five molecular interaction fields calculated by CoMSIA.

A total of 783 class I HLA sequences were found in the IMGT/HLA database and were included in the classification. The sequences were selected on the basis of the differences at protein sequence level. The classification is defined according to the scores plots of the CPCA model and the dendrograms obtained from hierarchical clustering. 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. Amino acid fingerprints are identified from the loading plots of the CPCA models, the fingerprints are the basis of the classification and can be used for future classification.

Three HLA-A clusters are defined from the scores plot of the CPCA model and hierarchical clustering, A2, A3 and A24 (Fig 1, Fig 2). In the scores plot, 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. Therefore, the CPCA analysis revealed three clusters as demonstrated in the 3D scores plot: the A3 cluster on the top right of figure 10.4, 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 (Table 10.4).

Hierarchical clustering analysis using CoMSIA fields also defined three clusters (Fig. 2). 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 the A2 cluster. The A24 cluster was well distinguished and included A*23 and A*24 alleles. 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.

The loading plot of the HLA-A model highlighted position 9, 97, 114 and position 116 (Fig 10.3). 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.

The scores plot of the HLA-B CPCA model reveals that the HLA-B molecules are divided into three clusters (Fig 10.4, Table 10.4): 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 (B*13, B*44, B*47, B*49, B*51, B*52, B*53, B*5607, B*57, B*58 and B*5901). Similar clusters are found using hierarchical clustering method, in which three clusters (B7, B27 and B44) are identified (Fig 5).

The PC1 loading plot showed that two areas were important in the classification (Fig 10.6). Position 63 and 66 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.

Results the HLA-C model is in figures 10.7 and 10.8, in which HLA-C molecules were divided into two clusters. Cw*01, Cw*03, Cw*07, Cw*08, Cw*12 and Cw*16 are grouped into one cluster, and Cw*02, Cw*03, Cw*04, Cw*05, Cw*06, Cw*15, Cw*17 and Cw*18 are in the second cluster. Some of the Cw*03, Cw*07 and Cw*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, with only eight amino acids mis-placed Cw*0308, Cw*0310, Cw*0701, Cw*0706, Cw*0716, Cw*0718, Cw*1208 and Cw*1404 (Table 10.4).
The PC2 loading plots showed that positions 70, 74, 77 and 81 of the HLA-C molecules are involved in the classification (Fig 9). 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.

Class I HLA classification using GRID/CPCA and hierarchical clustering based on CoMSIA fields exhibit, on average, a 77% consensus. HLA-A classification by both methods was 88% identical. 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 properly into clusters. They require future re-classification using other, more sensitive techniques. A closer look at the protein sequence level showed that these outliers do not significantly resemble 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.

The GRID/CPCA procedure grouped all class I HLA-A, B and C alleles into several supertypes. Of these alleles, A*0201, A*0202, A*0204, A*0206 and A*0207 had been grouped into the A2 supertype by binding studies [del Guercio & Sidney 1995, Southwood et al. 1998, Sudo & Kamikawaji 1995, Sidney et al. 1996a, Sidney et al. 1996b] 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 (L2L9) was closer to A*0201, therefore it was possible that A*0204 is 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. Our work was also compared with the classification by Lund et al.,
which produced a set of five distinct 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, although 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 [Southwood et al. 1998, Sidney et al. 1996a, Doolan & Hoffman 1997, Lamas et al. 1998, 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 found scattered within the 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 the B7 or the B44 superfamilies in the GRID/CPCA analysis. The B58 cluster in Sette’s classification can be found 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 B58 and B62 cluster were in the B7 or B27 cluster in the current analysis.

Although there is no previous HLA-C classification available for comparison, we can nonetheless make the interesting observation that the NK cell inhibitor receptor KIR2DL can be divided into two groups based on their HLA-C specificity. KIR2DL1 recognised HLA-Cw*2, Cw*4, Cw*5 and Cw*6, all of which possessed Asn77, whereas KIR2DL2 recognised HLA-Cw*1, Cw*3, Cw*7 and Cw*8, which had Ser at position 77 [Fan & Long 2001]. The specificity of KIR2DL was in agreement with our HLA-C classification, which suggested that position 77 was important in substrate binding: HLA-C molecules with the same residue at position 77 tend to share the same specificity.

A hierarchical clustering study based on HLA binding pockets has also been carried out, in which HLA-A molecules are classified according to molecular specificities of each of the six binding pockets. Three clusters have been defined according specificities of pocket A (table 10.4). The first cluster is consisted of A1 (A*01 and A*11), A*0208, A*16, A*20, A*29 and A*56, most of the A3 (A*03, A*30 A*31 A*32 and A*36), A*6810, A*6813, A*6814 alleles, A*7401-09 and A*8001. The second cluster includes A*25, A*26, A*33, A*34 and most A*68 alleles. Most of the A*02 alleles are present in the third cluster together with A*23, A*24, A*29, A*4301, A*6601, A*6604, A*6801-09, A*6815-23, A*6901. Two residues lining pocket A are identified to be important in the classification: position 63 and 66. Alleles in the first and second clusters all have polar amino acid Asn at position 66, while alleles in the third cluster have basic amino acid Lys at this position. Alleles in the first cluster have acidic Glu63 but those in cluster 2 have Asn63.
Two clusters are identified from the hierarchical clustering based on CoMSIA fields of pocket B (table 10.4). Cluster B1 has A*01, A*0201-33, A*0236-60, A*23, A*24, A*25, A*26, A*30, A*31, A*32, A*33, A*4301, A*6803-05, A*74 and A*8001. The second cluster includes A*0234-35, A*2424, A*29, A*34, A*0301-10, A*1101-04, A*6601-01, A*6601-02, A*6604, A*68 and A*6901. The classification is based on a single amino acid at position 70. Alleles with basic amino acid His70 are in cluster 1, while those with Gln70 are in cluster 2.

The HLA-A alleles are separated into two clusters according to specificities in pocket C (table 10.10). The first cluster include mainly A*01, A*0211, A*0235, A*0248, A*03, A*11, A*23, A*24, A*25, A*26, A*29, A*30, A*31, A*32, A*34, A*36, A*66, A*6801-04, A*6806-14, A*6816-19, A*6821-23, A*6901 and A*7401-05, A*7408-09, while most of A*02, A*1106, A*2428, A*2430, A*2603, A*2606, A*3009, A*6805, A*6815, A*6820, A*7406 and A*8001 are in the second cluster. One main feature of pocket C is position 74, alleles within cluster C1 all possess acidic amino acid Asp74, whereas alleles within cluster C2 have basic His74.

There are four clusters defined for pocket D (table 10.11). Cluster D1 includes all A*01 except A*0106, A*0249, A*1108 and A*36. The second cluster also has a small group of alleles including A*0310, A*1101-07, A*2417, A*2905, A*3402 and A*6801, A*6803-05, A*6807-23. The third cluster is consisted of A*0106, A*0201-40, A*0242-48, A*0250-51, A*0253-60, A*0301, A*0304-09, A*2301-09, A*2402-16, A*2418-38, A*2901-07, A*3103-06, A*3204, A*3402-04, A*6802, A*6808, A*6815, A*6901 and A*8001. The fourth cluster includes A*25, A*26, A*3004, A*3006, A*3202, A*3401 and A*3405,A*4301 and A*66. A*0241, A*0252, A*3001-03, A*3007-12, A*3101-02, A*3105, A*3107-09, A*32, A*33 and A*74 are grouped in the fifth cluster. Amino acids fingerprint for this classification is consisted of position 114 and 156. Alleles with basic amino acids such as Arg or His at position 114 are grouped into the first three clusters and alleles with Phe or Gln are in the last cluster. The first three clusters are further separated by polymorphism at position 156. Alleles with Arg, Gln/Trp and Leu are grouped into cluster 1, 2 and 3, respectively.

Four clusters are identified for pocket E (table 10.12). The first cluster is composed of serotype A*01, A*03 (A*0302, A*0307, A*0310), A*1101-02, A*1104-07, A*1109-14, A*2612, A*2618, A*29, A*31, A*32, A*33, A*36, A*6801, A*6803-05, A*6808-11, A*6813, A*6814, A*6816, A*6818-23, A*74 and A*8001. The second cluster is consisted of A*0301, A*0304-06, A*0308-09, together with A*1103, A*1108, A*2504, A*2608, A*2905, A*3204 and A*34. The third cluster includes most of A*02, A*23, A*24, A*30, A*6802, A*6806, A*6807, A*6815, A*6817 and A*6901. The last cluster has A*0203, A*0213, A*0226, A*0238, A*2418, A*25, A*26, A*3401, A*3405, A*4301 and A*66. The classification can be explained by two amino acids at position 116 and 152. Alleles with acid Asp at position 116 are grouped in cluster 1 and 2, while alleles in cluster 3 and 4 have bulky amino acid His or Tyr.

Only two clusters are found for pocket F (table 10.4). The first one consists of A*01, A*0301-10, A*11, A*2417, A*2501-04, A*26, A*29, A*31, A*32, A*33, A*34, A*36, A*4301, A*6801, A*6803, A*6808-14, A*6816, A*6818-23, A*74 and A*8001, and the second cluster has A*02, A*23, A*24, A*2602, A*30, A*6802, A*6806-
07, A*6815, A*6817 and A*6901. One position is identified to be important in the clustering process, position 116. Alleles in the first cluster possess negatively charged amino acid Asp at position 116, while those in the second cluster have aromatic Tyr at this position.

Compared with classifications on the whole binding site, the pocket classification considers one pocket at a time, therefore one allele may be classified into different groups in different classifications. For example, pocket B of some A*02 and A*03 alleles favour aliphatic amino acids therefore they are in the same group. However, pocket F of the A*03 alleles favours charged amino acids while A*02 alleles accept small aliphatic amino acids and they are in different clusters. In contrast to the three amino acids fingerprint from whole binding site classification, eight amino acids (position 63, 66, 70, 74, 114, 116, 152, 156) are identified in the pocket classification, indicating that more amino acids of the binding site are important in peptide specificity. However, as peptide binding motifs are only available for a small group of alleles, therefore the current classifications can not be validated.

10.3 Class II HLA Classification

Class II HLA alleles have also been classified by clustering. Doytchinova et al. applied hierarchical clustering using CoMSIA fields and non-hierarchical clustering based on z-scores. The hierarchical clustering follows the same procedure as class I classification described above. Nonhierarchical clustering uses five z descriptors to describe hydrophobicity, steric bulk, polarity and electronic effects of the HLA molecules. K-means clustering is applied to the set and the initial number of k seeds is equal to the number of clusters obtained from the hierarchical clustering. The known crystal structures of class II HLA are used as templates in homology modelling. Like the class I HLA classification, the class II alleles are grouped into twelve families. HLA-DR alleles are classified into DR1, DR3, DR4, DR5 and DR9 supertypes (Table 10.4). Hierarchical clustering groups DRB1*01-11, DRB1*1501-11 and DRB1*1601-08 in DR1, DRB1*0701-07, DRB1*0301-25, DRB3*0101-10, DRB3*0301-03, DRB1*0422 and DRB1*1107 in DR3, DRB1*0401, 03-48, DRB1*1113, 17, 26, 34, 42, DRB1*1309, DRB1*1401-48, DRB1*1001, DRB4*0101-06 in DR4, DRB1*0402, 12, 15, 25, 36, 37, 47, DRB1*1101-47, DRB1*1201-09, DRB1*1301-62, DRB1*1403, 16, 22, 25, 37, 40, DRB1*0801-25 in DR5, DRB1*0901-02, DRB5*0101-12, DRB5*0202-05 in DR9. Nonhierarchical clustering classifies DRB1*01-11, DRB1*1501-11 and DRB1*1601-08 in DR1, DRB3*0101-10, DRB3*0201-18, DRB3*0301-03, DRB1*1333, DRB1*1447 in DR3, DRB1*0401, 03-48, DRB1*1107, 13, 17, DRB1*1401-48, DRB3*0215, DRB1*1001, DRB4*0101-06, DRB1*0301-25 in DR4, DRB*0402, 15, 25, 36, 47, DRB1*1101-47, DRB1*1201-09DRB1*1301-62, DRB1*1403, 16, 17, 21, 22, 24, 25, 27, 29, 30, 37, 40, 41, 48, DRB1*0801-25, DRB5*0101-12, DRB5*0202-05 in DR5, DRB1*0901-02, DRB1*0701-07 in DR9. The DQ alleles are divided into DQ1 (DQB1*0501-03, DQB1*0601-21), DQ2 (DQB1*0201-03) and DQ3 (DQB1*0301-13, DQB1*0401, DQB1*0402 and DQA1*0301-03) (table 10.4). Four families are discovered for HLA-DP alleles, they are DPw1, DPw2 (DPB1*0201, DPB1*0202, DPB1*32, 33, 41, 46,
HLA-DR classification is due to polymorphism at position 9, 70 and 74 of the beta chain. Alleles with Trp9 are found in DR1 and those with Lys/Gln9 are in DR9. Alleles with Glu9/Asp70 are grouped in DR5. Alleles that have the combination of Glu9/Gln70 and Gln/Arg74 are in DR3 and those with Glu9, Gln/Arg70 and Glu/Ala74 are in DR4. Only two positions are responsible for HLA-DQ classification, position 71 and 86 of the beta chain. All alleles with Val86 are grouped in DQ1 cluster, while those with Glu86/Lys71 are in DQ2 and those with Glu86 and Thr/Asp71 are in DQ3. The DP classification is mainly based on polymorphism at position 69 and 84 of the beta chain. Alleles with Asp84 are grouped into DPw1/2 and those with Gly/Val84 are in DPw4/6. Alleles in DPw1/2 are separated by amino acid differences at position 69, those with Lys69 are grouped into DPw1 and those with Glu69 are in DPw2. DPw4 and DPw6 are separated by Lys69 and Glu69, respectively. The classification by hierarchical and non hierarchical clustering have a consensus of more than 85%.

A possible limitation of the GRID/CPCA technique is that it relies on accurate molecular structures. As the number of unique HLA sequences greatly exceeds the number of unique solved MHC crystal structures, protein structures used in these studies have been derived by homology modelling. Although HLA molecules are structurally similar, there may be some differences in the binding site conformation, and potentially this limitation is confounding. However, compared with HLA classifications based on peptide binding motifs, chemometric methods have some advantages (table 10.4). GRID/CPCA and hierarchical clustering are more flexible as they only require 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 classify that small number of HLA molecules with sufficient binding data. Most of the motifs include only anchor residues of the peptide, therefore only part of the peptide binding site interaction is studied. GRID/CPCA method takes the whole binding site into consideration and identifies important positions involved in the classification. Also, motif based classifications use a haphazard mishmash of differently derived experimental binding data, which may be biased and inconsistent. GRID/CPCA classification only uses sequence information, albeit manifest as homology modelled 3D structures, and thus minimises data inconsistency.
10.4 Discussion

As we have seen, HLA alleles can be classified into supertypes using only their sequence information. Some have sought insights into MHC supertypes from a sequence perspective, others from the perspective of structural data. The classification we outlined here 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. The HLA-B fingerprint is Asn63 and Leu81 for B7, Glu63 and Leu81 for B27 and Ala81 for B44. The HLA-C fingerprint is Ser/Gly77 for C1 and Asn77 for C4 supertype. The important positions for Class II DR supertype classification are position 9, 70 and 74, and position 71 and 86 are identified as the fingerprint for DQ clusters. These 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 that will help in identifying good MHC binders and testable potential epitopes.

The veracity and pace of vaccine identification would be enhanced greatly were one able to group HLA alleles into effective supertypes. An accurate and sufficiently extensive classification would render experimental work much more efficient, since one could look at a few supertypes rather than at thousands of separate alleles. This would thus greatly expedite the discovery of epitope based vaccines targeted at multiple alleles. Supertype definitions have already shown utility in epitope based vaccine discovery. Epitopes taken from hepatitis B virus infected patients have been shown to cross react with alleles in the A2, A3 and B7 superfamilies [Bertoni & Sidney 1997]. Epitopes isolated from Epstein-Barr virus reacted with several alleles of the B*44 family [Khanna & Burrows 1997]. Epitopes have been shown to cross-react with the A24 family [Burrows & Elkington 2003]. Many viral and tumour antigen derived vaccine candidates have also been shown to be able to bind multiple alleles. Sette et al. predicted 223 potential cancer peptides of CEA, Her-2/neu, P53 and MAGE antigens using a T cell epitope prediction algorithm, among which 115 were cross-reactive with peptides of the A2 supertype. 43 peptides were tested for immunogenecity and 73% were positive [Sette & Livingston 2001]. Recently a protein sequence scan has been carried out to search for T cell epitopes within the sequence of the SARS virus, based on the nine HLA supertypes in Sette’s analysis [Sylvestre-Hvid et al. 2004]. Fifteen predicted epitopes for each supertype were identified and tested experimentally: 75% of the predicted epitopes were found to be high affinity peptides (IC50 < 500nM) and about 112 candidate epitopes were obtained.

All supertypes are theoretically derived, even the experimental supertypes promulgated by Sette. His supertypes were derived from the comparison of binding motifs. Motifs are, at best, an inadequate description of peptide specificity. Possessing a certain verisimilitude, they can only give rise to a partial and largely incomplete definition of supertypes, limited by the lack of data for most MHC molecules. Structural supertypes represent an encouraging solution to this problem, unencumbered by lim-
iterations imposed by the availability of binding data. Modern methods in particular, such as the GRID/CPCA method we outline here, allow us to propose supertype definitions solely based on sequence and structural data. In one seamless movement we can progress from HLA sequencing to structure to supertype classification to binding specificity to epitope prediction. The clinical potential of such a process are tantalizing. Moreover, the same fundamental methodology can be used to address the issue of identifying benign HLA mismatches in tissue rejection, such as kidney transplants, bone marrow donation, and the like. Such problems require a robust, reliable and, preferably, transparent measure of structural similarity between HLA molecules in order to suggest which pairs of alleles will present the same peptides or be equally invisible to antibody surveillance. The GRID/CPCA method offers the possibility of effectively addressing all these problems and many more. All that is required is the requisite investment of time and resource in order to realize this potential coupled, of course, to the willingness of experimentalists to exploit these techniques.
Identifying Major Histocompatibility Complex Supertypes

| Clusters | Alleles                                      |
|----------|---------------------------------------------|
| 1        | HLA-A3, HLA-A11, HLA-31, and HLA-33        |
| 2        | HLA-B7, HLA-B35, HLA-B51, HLA-B53 and HLA-B54 |
| 3        | HLA-A29, HLA-B44 and HLA-B61               |

Table 10.1. Three clusters are identified by Cano's geometrical similarity matrix analysis

| Supertype | Alleles                                      |
|-----------|---------------------------------------------|
| 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 |
| B44       | B37, B41, B44, B45, B47, B49, B50, B60, B61 |

Table 10.2. The four supertypes defined by Sette's group.

| Supertype | MHC alleles                                      |
|-----------|-------------------------------------------------|
| A1        | 0101, 2501, 2601, 02, 2902, 3001, 3201          |
| A2        | 0201-07, 6802, 6901                             |
| A24       | 2301, 2402-04, 3001-03                          |
| A3        | 0301, 1101, 3101, 3301, 6801                    |
| B44       | 37, 4001, 4002 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 10.3. Nine supertypes defined by Sette and Sidney
| Supertype | Motif-based | Hierarchical clustering | Consensus PCA | Fingerprint |
|-----------|-------------|-------------------------|---------------|-------------|
| A1        | 0101 2501 2601,02 3201 |                       |               |             |
| A2        | 0201-07 6802 6901 | 0201 – 60 2501 – 04 2601 – 18 3401, 05 4301 6601 – 04 6802, 15, 23 6901 | 0201 – 60 without 04, 17, 57 6802, 15, 23 6901 | Tyr9/Phe9 Arg97 His114 and Tyr116 |
| A24       | 2310 2402-04 3001-03 | 2301-09 2402 – 38 | 231-09 2402 – 38 | Ser9 Met97 |
| A3        | 0301 1101 3101 3301 6801 | 0101 – 09 1101 – 14 2901 – 07 3001 – 12 3101 – 09 3201 – 07 3402 – 04 3601 – 04 6801 – 22 without 02, 15 | 0101 - 09 1101 – 14 2501 – 04 2601 – 18 2901 – 07 3101 – 09 3201 – 07 3301 – 06 3401 – 05 | Tyr9/Phe9/Ser9 Ile97/Met97 Glu114 and Asp116 |

Table 10.4. 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.
| Supertype | Motif-based | Hierarchical clustering | Consensus PCA | Fingerprint |
|-----------|-------------|-------------------------|---------------|-------------|
| B44       | 37          | 0802                    | 0802          | Ala81       |
|           | 40012       | 1301 - 1311 without 09  | 1301 - 1311 without 09 |             |
|           | 4006        | 1513, 16, 17, 23, 24, 36, 67 | 1513, 16, 17, 23, 24, 36, 67 |             |
|           | 41          | 1809                    | 1809          |             |
|           | 44          | 2701, 02                | 3805          |             |
|           | 45          | 3801 - 3809 without 03  |               |             |
|           | 47          | 4013, 4019              |               |             |
|           | 49          | 4402 - 4433 without 09, 31 |               |             |
|           | 50          | 4704                    | 5101 - 34     |             |
|           |             | 4901 - 03               | 5201 - 05     |             |
|           |             | 5101 - 34               | 5301 - 09     |             |
|           |             | 5201 - 05               | 5607          |             |
|           |             | 5301 - 09 without 03, 05 |               |             |
|           |             | 5701 - 09               | 5701 - 09     |             |
|           |             | 5801 - 07               | 58 - 07       |             |
|           |             | 5901                    |               |             |
| B27       | 1401 - 02   | 0713                    | 0727          | Glu63       |
|           | 1503, 09, 10, 18 | 1309               | 2701 - 25 without 08, 12, 18 | Leu81       |
|           | 2701 - 08   | 1501 - 1575 without these in B7 | 3701 - 04 |             |
|           | 3801.02     | and B44                | 3801 - 09     |             |
|           | 3901 - 04   | 1812                    | 4013, 19, 28  |             |
|           | 4801, 02    | 2703 - 2725             |               |             |
|           | 7301        | 3513, 16, 28            |               |             |
|           |             | 3701 - 05               |               |             |
|           |             | 3803                    |               |             |
|           |             | 3902, 08, 13, 22, 23    |               |             |
|           |             | 4001 - 44 without 08, 13, 19, 25 |       |             |
|           |             | 4101 - 06               |               |             |
|           |             | 4409, 31                |               |             |
|           |             | 4501 - 06               |               |             |
|           |             | 4601, 02                |               |             |
|           |             | 4701 - 03               |               |             |
|           |             | 4801 - 07               |               |             |
|           |             | 5001 - 04               |               |             |
|           |             | 5608                    |               |             |
|           |             | 6702                    |               |             |
|           |             | 7805                    |               |             |

**Table 10.5.** A list of the HLA-B molecules in the scores plot (Part 1).
| Supertype | Motif-based | Hierarchical clustering | Consensus PCA | Fingerprint |
|-----------|-------------|-------------------------|---------------|-------------|
| B7        | 07          | 0702 – 31 without 13    | 0702 – 31 without 13 | Asn63       |
|           | 35          | 0801 – 17 without 02    | 0801 – 17 without 02 | Leu81       |
|           | 51          | 1401 – 06               | 1309          |             |
|           | 53          | 1502, 08, 09, 10, 11, 15, 18, 21, 29, 37, | 1401 – 06 |             |
|           | 54          | 44, 51, 52, 55, 64, 72 | 1501 – 75 without 13, 16, 17 |             |
|           | 55          | 1801 – 18 without 09, 12 | 23, 24, 36, 67 |             |
|           | 56          | 2723                    | 1801 – 18 without 09 |             |
|           | 67          | 3501 – 45 without 13, 16, 28 | 2708, 12, 18 |             |
|           | 78          | 3901 – 27 without 02, 08, 13, 22, 23 | 3501 – 45 |             |
|           |             |                         | 3705          |             |
|           |             |                         | 3904          |             |
|           |             |                         | 4101 – 06     |             |
|           |             |                         | 4201 – 04     |             |
|           |             |                         | 4409          |             |
|           |             |                         | 4501 – 06     |             |
|           |             |                         | 4601, 02      |             |
|           |             |                         | 4702          |             |
|           |             |                         | 4801 – 07     |             |
|           |             |                         | 5001 – 04     |             |
|           |             |                         | 5401, 02      |             |
|           |             |                         | 5501 – 10     |             |
|           |             |                         | 5601 – 11 without 08 |             |
|           |             |                         | 5601 – 11 without 08 |             |
|           |             |                         | 5607          |             |
|           |             |                         | 6701          |             |
|           |             |                         | 7301          |             |
|           |             |                         | 7801 – 05     |             |
|           |             |                         | 8101          |             |
|           |             |                         | 8201, 02      |             |
|           |             |                         | 8301          |             |
| B58       | 1516, 17    |                         |               |             |
|           | 5701, 02    |                         |               |             |
|           | 58          |                         |               |             |
| B62       | 1301 – 02   |                         |               |             |
|           | 1501, 02, 06, 12, 13, 14, 19, 21 |               |             |
|           | 4601        |                         |               |             |
|           | 52          |                         |               |             |

**Table 10.6.** A list of the HLA-B molecules in the scores plot (Part 2).
| Supertype | Motif-based | Hierarchical clustering | Consensus PCA | Fingerprint |
|-----------|-------------|-------------------------|---------------|-------------|
| C1        | No data     | 0102 - 09, 0302 - 16 without 07, 08, 10, 15, 0702 - 18 without 01, 06, 07, 09, 16, 18, 0801 - 09, 1202 - 07 without 04, 05, 08, 1402 - 05, 1601, 04 | 0102 - 09, 0302 - 16 without 07, 08, 15, 0701 - 18 without 07, 09 | Ser77/Gly77 |
| C2        | No data     | 0202 - 06, 0307, 08, 10, 15, 0401 - 10, 0501 - 06, 0602 - 09, 0701, 06, 07, 09, 16, 18, 1204, 05, 08, 1502 - 11, 1602, 1701 - 03, 1801, 02 | 0202 - 06, 0307, 15, 0401 - 10, 0501 - 06, 0602 - 09, 0707, 09, 1204, 05, 1404, 1502 - 11, 1602, 1701 - 03, 1801, 02 | Asn77 |

Table 10.7. A list of the HLA-C molecules in each cluster. The important residues in defining the clusters were listed in the last column.
### Table 10.8. Supertypes defined by pocket analysis: pocket a

| Cluster | Alleles | Fingerprint |
|---------|---------|-------------|
| A1      | 0101-09 0208 0216 0220 0229 0256 0301-10 1101-09 1112-14 2309 3001-12 3101 3103-06 3019 3201-04 3206-07 3601-04 6002-03 6810 6813-14 7401-09 8001 | Asn66 Glu63 |
| A2      | 0255 1110-11 2424 2501-04 2601-08 3301 3303-06 3402-04 4301 6601 6604 6801-09 6815-23 6901 | Asn66 Asn63 |
| A3      | 0201-07 0209-19 0221-28 0230-54 0257-60 2301-06 2402-38 2607 2901-07 3007 3102 3107-08 3205 3401 3405 7409 | Lys66 |

### Table 10.9. Supertypes defined by pocket analysis: pocket b

| Supertype | alleles | Fingerprint |
|-----------|---------|-------------|
| B1        | 0101-08 0201-33 0236-60 2301-06 2401-23 2425-38 2501-04 2601-08 3002-10 3012 3101-09 3201-07 3301-06 4301 6803-05 7401-09 8001 | His70 |
| B2        | 0234-35 2424 2901-07 3001 3011 3401-05 0301-10 1101-14 6601-02 6604 6801-23 6901 | Gln70 |
| Supertype | Alleles                                                                 | Fingerprint |
|-----------|-------------------------------------------------------------------------|-------------|
| C1        | 0101-09 0211 0235 0248 0301-10 1101-05 1107-14 2301-09 2402-27 2429 2431-38 2501-04 2601-18 2901-07 3001-12 3101-09 3201-07 3401-05 3601-04 4301 6601-04 6801-04 6806-14 6816-19 6821-23 6901 7401-05 7408-09 | Asp74       |
| C2        | 0201-10 0212-34 0236-47 0249-60 1106 2428 2430 2603 2606 3009 6805 6815 6820 7406 8001 | His74       |

Table 10.10. Supertypes defined by pocket analysis: pocket c
| Supertype | Alleles                      | Fingerprint        |
|-----------|------------------------------|--------------------|
| D1        | 0101-03 0107-08 0249 1108 3601-04 | Arg/His114  
                |                              | Arg156              |
| D2        | 0310 1101-07 1109-14 2417 2905 3402 6801 6803-05 6807-23 | Arg114  
                |                              | Gln/Trp156          |
| D3        | 0106 0201-40 0242-48 0250-51 0253-60 0301 0304-09 2301-09 2402-16 2418-38 2901-07 3103-06 3204 3402-04 6802 6808 6815 6901 8001 | Arg114  
                |                              | Leu156              |
| D4        | 2501-04 2601-18 3004 3006 3202 3401 3405 4301 6601-04 | Phe/Gln114  
                |                              | Gln/Trp156          |
| D5        | 0241 0252 3001-03 3007-12 3101-02 3105 3107-09 3201-03 3205 3207 3301-06 7401-09 | Phe/Gln114  
                |                              | Leu156              |

Table 10.11. Supertypes defined by pocket analysis: pocket d
| Supertype | Alleles                                                                 | Fingerprint       |
|-----------|-------------------------------------------------------------------------|-------------------|
| E1        | 0101-09 0302 0307 0310 1101-02 1104-07 1109-14 2612 2618 2901-04 2906-07 3101-09 3201-03 3205-07 3301-06 3601-04 6801 6803-05 6808-11 6813-14 6816 6818-23 7401-09 8001 | Asp116 Val/Ala152 |
| E2        | 0301 0304-06 0308-09 1103 1108 2504 2608 2905 3204 3402-04                                                             | Asp116 Glu152     |
| E3        | 0201-02 0204-12 0214-25 0227-37 0239-60 2301-09 2402-17 2419-38 3001-12 6802 6806 6807 6815 6817 6901         | Tyr/His116 Val152 |
| E4        | 0203 0213 0226 0238 2418 2501-03 2601-11 2613-17 3401 3405 4301 6601-04                                            | Tyr/His116 Glu152 |

**Table 10.12.** Supertypes defined by pocket analysis: pocket e
| Supertype | Alleles | Fingerprint |
|-----------|---------|-------------|
| F1        | 0101-09 0301-10 1101-14 2417 2501-04 2601-09 2901-07 3101-09 3202-07 3301-06 3401-05 3601-04 4301 6601-04 6801 6803 6808-14 6816 6818-23 7401-09 8001 | Asp116 |
| F2        | 0201-60 2301-06 2309 2402-38 2602 3001-12 6802 6806-07 6815 6817 6901 | Tyr116 |

**Table 10.13.** Supertypes defined by pocket analysis: pocket f
| Supertype | Hierarchical clustering | Non-hierarchical clustering | Common alleles | Fingerprint | Known supers types | Known motifs |
|-----------|------------------------|-----------------------------|----------------|-------------|-------------------|-------------|
| DR1       | DR1 (DRB1*0101-11)     | DR1 (DRB1*0101-11)          | 11             | Trp<sup>99</sup> | DRB1*0101 (9, 50-52) | YFW LA AG - LA |
|           | DR2 (DRB1*1501-11,    | DR2 (DRB1*1501-11,         | 13             |             | DRB1*1501 (53, 54) | LVI FYI - IL GSP |
|           | DRB1*1601-08)         | DRB1*1601-08)               | 7              |             |                   |             |
|           | DR7 (DRB1*0701-07)    | -                           | 7              |             |                   |             |
|           | DR3 (DRB1*0301-25)    | -                           | -              | Glu<sup>99</sup> | DRB1*0301 (55-57) | LIF D KR - YLF |
|           | DR52 (DRB3*0101-10,   | -                           | 10             | Gln<sup>70a</sup> |                   |             |
|           | DRB3*0201-18,         | -                           | 18             | Gln<sup>Arg</sup><sup>74g</sup> |                   |             |
|           | DRB3*0301-03)         | -                           | 3              |             |                   |             |
|           | DRB1*0422 DRB1*1107   | -                           | -              |             |                   |             |
|           | DR4 (DRB1*0401, 03-48 without the alleles from DR5 supertype) | DR4 (DRB1*0401, 03-48 without the alleles from DR5 supertype) | 38             | Glu<sup>99</sup> | DRB1*0401 (58-60) | FY no RK NS pol<sup>b</sup> pol W chang<sup>c</sup> alien<sup>d</sup> |
|           | DR5 (DRB1*1113, 17, 26, 34, 42) | DR5 (DRB1*1113, 17, 26, 34, 42) | 2              | Gln/Arg<sup>70a</sup> |                   |             |
|           | DR6 (DRB1*1309,      | DR6 (DRB1*1309,            | 31             | Glu/Ala<sup>24</sup> |                   |             |
|           | DRB1*1401-48 without the alleles from DR5 supertype) | DRB1*1401-48 without the alleles from DR5 supertype) | 5              |             |                   |             |
|           | -                      | -                           | -              |             |                   |             |
|           | DR10 (DRB1*1001)     | -                           | 5              |             |                   |             |
|           | DR53 (DRB4*0101-06)  | -                           | -              |             |                   |             |
|           |                         | -                           | 2              |             |                   |             |

Table 10.14. DR supertypes and fingerprints. Results of The content of hierarchical and non-hierarchical clustering are compared and fingerprints are defined (Part 1).
| Supertype | Hierarchical clustering | Non-hierarchical clustering | Common alleles | Fingerprint | Known supertypes<sup>a</sup> | Known motifs |
|-----------|-------------------------|-----------------------------|----------------|-------------|-----------------------------|--------------|
| DR5       | DR4 (DRB1*0402, 12, 15, 25, 36, 37, 47) | DR4 (DRB1*0402, 15, 25, 36, 47) | 5 9 12 25 | Glu<sup>99</sup> Asp<sup>705</sup> | DR<sup>RS</sup> | DRB1*0402 (61) VIL no DE NQ RK pol ali H DRB1*1101 (59, 64) WYF LVI RK – – DRB1*1201 (51) IL L MN VYF – YFM |
|           | DR5 (DRB1*1101-47, DRB1*1201-09) | DR5 (DRB1*1101-47, DRB1*1201-09) | – – | – | “D” | – – |
|           | DR6 (DRB1*1301-62, DRB1*1403, 16, 22, 25, 27, 40) | DR6 (DRB1*1301-62, DRB1*1403, 16, 17, 21, 22, 24, 25, 27, 29, 30, 37, 40, 41, 48) | – – | – | – | – – |
|           | DR8 (DRB1*0801-25) | DR8 (DRB1*0801-25) | 12, DRB5*0202-05 | – | – | – – |
|           | – | – | – | – | – | – – |
| DR9       | DR9 (DRB1*0901, 02) | DR9 (DRB1*0901, 02) | 2 | Lys/Gln<sup>99</sup> | – | DRB5*0101 (53, 54) FY QV – – RK |
|           | DR51 (DRB5*0101-12) | DR7 (DRB1*0701-07) | – | – | – | – – |
|           | DRB5*0202-05 | – | – | – | – | – – |
| Sum       | 347 | 347 | 285 (82%) | – | – | – – |

Table 10.15. DR supertypes and fingerprints. Results of The content of hierarchical and non-hierarchical clustering are compared and fingerprints are defined (Part 2).
| Supertype | Hierarchical clustering | Non-hierarchical clustering | Common alleles | Fingerprint |
|-----------|-------------------------|-----------------------------|----------------|-------------|
| DQ1       | DQB1*0501-03 DQB1*0601-21 | DQB1*0501-03 DQB1*0601-21 | 45 300 | Val^{86β} |
| DQ2       | DQB1*0201-03 | DQB1*0201-03 | 45 | Glu^{86β} Lys^{71β} |
| DQ3       | DQB1*0301-13 DQB1*0401, 02 | DQB1*0301-13 DQB1*0401, 02 | 195 30 | Glu^{86β} Thr/Asp^{71β} |
| DQA1*03c  | DQA1*0301-03 | – | – | Arg^{53α} |
| Sum       | 738 | 738 | 615 (83%) |

Table 10.16. HLA-DQ supertypes and fingerprints. Results of the content of hierarchical and non-hierarchical clustering are compared and fingerprints are defined.
Table 10.17. HLA-DP supertypes and fingerprints. Results of The content of hierarchical and non-hierarchical clustering are compared and fingerprints are defined.
| Motif based                              | GRID/CPCA approach                  |
|-----------------------------------------|--------------------------------------|
| Considers part of the binding site      | Considers the whole binding site     |
| Requires the binding motif of the alleles | Sequence information only            |
| Can only classify alleles with known motifs | Able to classify all HLA alleles     |

**Table 10.18.** The advantages of chemometric methods over motif based classification.
Fig. 10.1. 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.
Fig. 10.2. The HLA-A classification defined by hierarchical clustering. 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.
Fig. 10.3. 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).
Fig. 10.4. The 3D scores plot of the CPCA analysis for HLA-B molecules. Three clusters were identified in the plot: B7, B27 and B44.
Fig. 10.5. HLA clusters produced using hierarchical clustering. A hierarchical tree was produced for the 447 HLA-B alleles. Each leaf represents one allele.

Fig. 10.6. 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)
Fig. 10.7. 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.
Fig. 10.8. The hierarchical tree obtained from hierarchical clustering, in which the HLA-C molecules were classified into C1 and C4 clusters. Each leaf represented one HLA-C allele. Results of the analysis were in accordance with the GRID/CPCA classification.
Fig. 10.9. 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.