Identifiying Human MHC Supertypes Using Bioinformatic Methods

Irini A. Doytchinova, Pingping Guan, and Darren R. Flower[1]

Classification of MHC molecules into supertypes in terms of peptide-binding specificities is an important issue, with direct implications for the development of epitope-based vaccines with wide population coverage. In view of extremely high MHC polymorphism (948 class I and 633 class II HLA alleles) the experimental solution of this task is presently impossible. In this study, we describe a bioinformatics strategy for classifying MHC molecules into supertypes using information drawn solely from three-dimensional protein structure. Two chemometric techniques–hierarchical clustering and principal component analysis–were used independently on a set of 783 HLA class I molecules to identify supertypes based on structural similarities and molecular interaction fields calculated for the peptide binding site. Eight supertypes were defined: A2, A3, A24, B7, B27, B44, C1, and C4. The two techniques gave 77% consensus, i.e., 605 HLA class I alleles were classified in the same supertype by both methods. The proposed strategy allowed "supertype fingerprints" to be identified. Thus, the A2 supertype fingerprint is Tyr^{9}/Phe^{9}, Arg^{97}, and His^{114} or Tyr^{116}; the A3-Tyr^{9}/Phe^{9}/Ser^{9}, Ile^{97}/Met^{97} and Glu^{114} or Asp^{116}; the A24-Ser^{9} and Met^{97}; the B7-Asn^{63} and Leu^{81}; the B27-Glu^{63} and Leu^{81}; for B44-Ala^{81}; the C1-Ser^{77}; and the C4-Asn^{77}. *The Journal of Immunology,* 2004, 172: 4314–4323.

Molecules of the MHC bind small peptides and present them on the surface of cells for recognition by TCR-bearing T cells. Formation of the ternary complex between TCR and the peptide-MHC complex is the key molecular recognition event at the heart of the adaptive and memory cellular immune responses. MHCs exhibit different selectivities for peptide, and TCRs, in their turn, exhibit different specificities for peptide-MHC complexes. Thus, the combination of selectivities displayed by MHC and TCR determines the nature and scope of peptide recognition by the immune system, and thus the detection of foreign proteins and pathogens. The T cell repertoire is large, diverse, and, in terms of its dominant recognition properties, also dependent on environmental factors, particularly infection and vaccination histories. The selectivity for peptides exhibited by an MHC molecule is, by contrast, determined solely by its molecular structure. It undergoes no somatic hypermutation, affinity maturation, or thymic selection and is, as part of a restricted set of alleles, an inherited characteristic of individual organisms. However, in humans, as in most other species, the MHC is both polygenic (there are several MHC class I and MHC class II genes) and polymorphic (there are multiple alleles of each gene) (1). Each class of MHC is represented by several loci: HLA-A, HLA-B, and HLA-C for class I and HLA-DR, HLA-DQ, and HLA-DP for class II. All MHC loci are codominant: both maternally and paternally inherited sets of alleles are expressed. The linked set of MHC alleles found on one chromosome is called a haplotype. MHCs exhibit extreme polymorphism: within the human population there are, at each genetic locus, a great number of alleles–the last IMGT/HLA database release (April 2003) lists 948 class I and 633 class II molecules (2), many of which are represented at high frequency (>1%). MHC alleles may differ by as many as 30-aa substitutions. Such an uncommon degree of polymorphism implies a selective pressure to create and maintain it. Different polymorphic MHC alleles, of both class I and class II, have different peptide specificities: each allele binds peptides exhibiting particular sequence patterns.

The majority of polymorphic sites in class I alleles are found in the α1 and α2 domains of the mature protein. These domains form the peptide binding site (3, 4). Analysis of x-ray structures of HLA/peptide complexes identified ~35 residues within the binding site as being critical for peptide binding and therefore for T cell recognition (5–7). About 20 of them are polymorphic. Six binding pockets, denoted A through F, are formed within the binding site. Some of the pockets are nonpolar and can form hydrophobic contacts, but others contain polar atoms that can make hydrogen bonds with peptide side chains.

The peptide side chains at certain positions bind tightly to certain pockets and form primary anchors (8). The combination of two or more anchors is called a motif. It has been found experimentally that certain alleles can recognize very similar motifs (9), and thus be grouped into HLA supertypes. The classification of HLA molecules into supertypes (superfamilies) in terms of their structural features and abilities to bind peptides with common motifs (supermotifs) is an important issue, with direct implications for the development of epitope-based vaccines (10–14). Determining the peptide specificity of even a single allele is by no means a trivial undertaking, requiring prodigious experimental work (13, 14) and sophisticated theoretical analysis (15, 16). To undertake this rigorous task for over 1500 alleles is simply too intensive, in terms of labor, time, and resources, to be practical. The only viable alternative is to seek a bioinformatic approach.

At the present time, there are two groups of in silico classification: whole sequence-based (17, 18) and pocket-based (19–21). Both types of methods, although informed by structural knowledge, make use of protein sequences. The difference between these methods is the number of residues involved in the analyses. The pocket-based classifications have more direct, practical application to vaccine development. Motif-based categorization is a variety of

Edward Jenner Institute for Vaccine Research, Compton, Berkshire, United Kingdom

Received for publication September 18, 2003. Accepted for publication January 7, 2004.

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked *advertisement* in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

[1] Address correspondence and reprint requests to Dr. Darren R. Flower, Edward Jenner Institute for Vaccine Research, High Street, Compton, Berkshire, RG 7NN, U.K. E-mail address: darren.flower@jenner.ac.uk

0022-1767/04/$02.00

pocket-based classification, where only pockets accepting peptide anchors are considered (9–14).

In this study, we describe a bioinformatics strategy for classifying the HLA class I molecules into supertypes, using information drawn solely from three-dimensional (3D)[2] protein structures. Two chemometric techniques–hierarchical clustering and principal component (PC) analysis (PCA)–were used independently on a set of 783 HLA class I molecules to identify supertypes based on structural similarities and molecular interaction fields (MIFs) calculated for the peptide binding site on HLA class I molecules. The supertypes derived by both techniques were compared and analyzed to identify a "supertype fingerprint," i.e., a 3D structural basis for supertype classification. The identified fingerprints revealed the striking observation that only one to three amino acids are sufficient for the classification of an allele to a particular supertype.

Materials and Methods

Protein structures

The protein sequences of 783 HLA class I molecules were collected from the IMGT/HLA database (2): 229 molecules for HLA-A, 447 for HLA-B, and 107 for HLA-C. The protein molecules were modeled based on x-ray data for three reference proteins retrieved from the Research Collaboratory for Structural Bioinformatics (RCSB) protein database (22): molecule A*0201 (Protein Data Bank code: 1I4F) for the A locus (23), B*0801 (Protein Data Bank code: 1AGD) for the B locus (24), and Cw*0401 (Protein Data Bank code: 1IM9) for the C locus (25). For these proteins, all cofactors, counterions, ligands, and water molecules were removed before the modeling of the remaining 780 HLA molecules. Side chain placement of polymorphic residues was performed using SCRWL2.8 (26). As the polymorphic amino acids and the amino acids involved in the binding site are localized among the first 180 residues, the modeled molecules consisted of α1 and α2 domains only (residues 1–180). MHC nomenclature is becoming complex (2), with the recent inclusion of subsidiary numbers beyond the familiar four digit code, to include, inter alia, nucleic acid difference in both exon and introns which have no effect on the amino acid sequence. In such cases, only the first molecule (*xxxx01) was considered so that a set of all unique protein sequences were retained.

All chemometric modeling studies were performed on a SGI workstation (Silicon Graphics, Mountain View, CA) using the programs Sybyl (version 6.9; Tripos, St. Louis, MO), GRID (version 21; Molecular Discovery, Oxford, U.K.), and Golpe (version 4.5.12; Multivariate Infometric Analysis, Perugia, Italy). Modeled protein molecules were aligned within each locus using the initial x-ray structure as a template. Hydrogen atoms were added and Kollman charges were calculated for each molecule. For GRID/Golpe, the molecules from each locus were added as a multiple target and the GRID MIFs with different probes were calculated.

Binding site definition

Based on the x-ray structures, three different MHC binding sites were defined: one site per locus. The HLA-A binding site includes the following 35 residues: 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, and 171 (5). The HLA-B binding site consisted of 37 residues: 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, and 171 (6). The HLA-C binding site comprises 32 residues: 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, and 171 (7).

Hierarchical clustering

Clustering is the process of dividing a set of entities into subsets in which the members of each subset are similar to each other but different from members of other subsets (27). There have been numerous cluster methods described (28). In the present study, a hierarchical clustering using the agglomerative algorithm (27) was applied. According to this algorithm, the clusters are built from the bottom up, first by merging individual items into clusters, and then by merging clusters into superclusters, until the final merge brings all items into a single cluster.

This method was applied as implemented in Sybyl 6.9 (Tripos). The distance between the clusters was calculated by the complete-linkage method, i.e., the distance between the most distant pair of data points in both clusters is taken into account. The last two or three levels were considered for the supertypes definition.

Comparative similarity indices analysis (CoMSIA) fields

The hierarchical clustering of the HLA class I molecules studied here is based on five properties generated by CoMSIA (29–31): steric bulk, electrostatic potential, hydrophobicity, hydrogen-bond donor, and acceptor abilities. CoMSIA is a "3D-grid" method. In the 3D-grid methods, a probe group is moved through a regular 3D grid of points in a region of interest around the target molecule. At each point, an interaction energy or similarity index between the probe and the target molecule is calculated using an empirical function (29). In CoMSIA, similarity indices are calculated using a Gaussian-type distance dependence between a probe and the atoms of the protein molecules.

In this study, CoMSIA was used as implemented in Sybyl 6.9 (Tripos). The common probe has a 1Å radius, charge +1, hydrophobicity +1, hydrogen-bond donor and acceptor properties +1. The region was defined on the atomic positions of the amino acids belonging to the binding site. Previous studies (32–34) indicated that 3D-quantitative structure-activity relationship models based on all fields give better results than models based on each field separately or models based on different field combinations. Because of this, all five CoMSIA fields were included in the clustering process.

Principal component analysis

The PCA is a multivariate data analysis designed to represent large, multidimensional data sets in a limited, but visually interpretable, number of dimensions, usually two to five, usually referred to as PCs, such that an overview of the data is obtained. This overview may reveal groups of observations, trends, and outliers. It also uncovers the relationships between observations and variables and between the variables themselves (35). The results from PCA can be visualized on different plots–scores and loadings. Score plots (observation projections on PC) visualize groupings of the protein molecules based on their protein-probe interaction pattern. Proteins with similar interaction patterns should be clustered. Loading plots (variable projections on PC) visualize the structural features on the molecules responsible for different or similar interaction patterns.

PCA and consensus PCA (CPCA) were used as implemented in the Golpe program (Multivariate Infometric Analysis). CPCA explains the variance on a superlevel, which expresses the "consensus" of all probes in the analysis. To focus the analysis only on attractive protein-probe interactions, a maximum cutoff value of 0 kcal/mol was used. Absolute values smaller than 0.01, or with a SD below 0.01, were set to zero. Initially, PCA was applied on MIFs generated for each probe. The sum of the explained variance by the first two PCs was considered as a criterion for the probe importance. Probes with >35% explained variance remained in the final analysis. They were analyzed by CPCA. Block unscaled weights scaling was used to normalize the importance of probe interactions.

Molecular interaction fields

In this paper, the PCA was based on the MIFs computed by the GRID program (version 21; Molecular Discovery) using particular probes. The GRID MIFs method is also a 3D-grid method, but instead of similarity indices, interaction energies are calculated at each point of the grid. These energies may be displayed as 3D contours around the target molecule using computer graphics. Contours at large negative energies indicate energetically favorable binding regions for the particular probe while those at large positive energies correspond to regions from which it would be repelled.

MIFs were calculated using the GRID force field and 13 probes (Table I), 2Å grid spacing, and dynamic side chain treatment (GRID directive MOVE = 1). The GRID box was defined to include only the amino acids of the binding site. The GRID probes were chosen to represent all relevant interactions (hydrophobic, charge-charge, hydrogen-bond donor/acceptor) and to cover the chemical groups presented by naturally occurring amino acids.

Results

HLA-A supertypes

Hierarchical clustering. The hierarchical clustering on CoMSIA fields for 229 HLA-A molecules is given in Fig. 1, *upper left panel*. Three well-defined clusters were distinguished. The first

[2] Abbreviations used in this paper: 3D, three dimensional; PC, principal component; PCA, PC analysis; MIF, molecular interaction field; CoMSIA, comparative similarity indices analysis; CPCA, consensus PCA.

Table I. *GRID probes used in this study*[a]

Name	Chemical group	HLA-A	HLA-B	HLA-C
OH2	Water	X	X	X
DRY	The hydrophobic probe	X	X	X
H	Hydrogen	X	X	
C3	Methyl CH3 group	X		
C1=	sp2 CH aromatic or vinyl	X		
N:#	sp N with lone pair	X	X	X
N:=	sp2 N with lone pair	X	X	
N1	Neutral flat NH, e.g. amide		X	
$N2^+$	sp3 amine NH2 cation	X	X	
O1	Alkyl hydroxy OH group			X
OH	Phenol or carboxy OH	X	X	X
O	sp2 carbonyl oxygen	X	X	X
S1	Neutral SH group			X

[a] In the last three columns are given the probes used in the final CPCA for each locus.

cluster, on the left side of the dendrogram, consisted of *02, *25, *26, *3401, *3405, *4301, *66, *6802, *6815, *6823, and *6901 alleles. For more clarification, the alleles from each supertype are listed in Table II. Data from the immunological literature classifies alleles *02, *6802, and *6901 molecules as the *A2 supertype* (9–12). For consistency with the literature, we adopted this name for the cluster. The central cluster was classified as *A24 supertype*, because it includes alleles *23 and *24 (Table II). This is a small, but quite well distinguished, cluster. It is closer to the A3 sypertype than to A2. The last cluster is the biggest HLA-A supertype, consisting of a great variety of alleles. As alleles *03, *11, *31, and *33 were classified here (Table II), we named it the *A3 supertype* in accordance with literature precedent (9–12).

Principal component analysis. Golpe PCA was applied to multivariate GRID descriptors to uncover differences in the binding sites with respect to their probe interaction pattern. After superimposing the 3D structures of 229 HLA-A molecules using as a template the A*0201 allele (Protein Data Bank code: 1I4F), their binding sites were characterized by the interaction energies to 13 functional groups (probes). The variance explained by the first two PCs for every probe was calculated by Golpe PCA and those with >35% of the variance explained by the first two PCs were selected for CPCA (Table I). The CPCA gave a satisfactory two-component model. The first PC explains 25% of the variance, while the second adds 17%.

The CPCA scores plot is given in Fig. 1, *upper right panel*. Each point represents a single HLA-A molecule. Three clusters were defined by the first two components. On the *upper middle/right part* of the plot are molecules defined as the A3 supertype by the hierarchical clustering (Table II). Compared with the hierarchical clustering, 28 more molecules were classified here: A*25, A*26, A*4301, and A*66. According to the hierarchical classification they belong to the A2 supertype. On the *middle lower part* are clustered molecules from the A2 supertype. Molecules from the A24 supertype are situated in the *upper left quadrant*.

The CPCA loadings plots contain information about the structural features responsible for the discrimination between HLA-A supertypes. They are represented as isocontour plots for each PC. Positive contours indicate regions with positive contributions to the PC, negative contours correspond to negative PC values. The interaction fields are localized in the vicinity of four positions: 9, 97, 114, and 116. The C3/PC1 plot is shown in Fig. 1, *lower left panel*. The other plots are not shown because they are similar. The C3 probe accounts for steric and hydrophobic interactions. The black area near position 9 indicates bulky and/or hydrophobic regions favored by the A3 supertype. The gray areas near positions 97, 114, and 116 indicate regions where bulky/hydrophobic amino acids are disfavored by A24. Indeed, A3 molecules have Phe and Tyr at position 9, whereas all A24 proteins have Ser. Furthermore, A24 supertype has the unique combination of Met^{97}, His^{114}, and Tyr^{116}.

HLA-B supertypes

Hierarchical clustering. Three clusters are distinguished for the HLA-B locus after hierarchical clustering on 447 HLA-B molecules (Fig. 2, *upper left panel*). These three clusters define three B supertypes. The first cluster from the left was defined as the *B44 supertype* as B*44 is the serotype with the largest of alleles here (Table III). The B44 supertype also includes molecules B*0802, B*13, several B*15, B*1809, B*2701, B*2702, B*38, B*4013, B*4019, B*44, B*4704, B*49, B*51, B*52, B*53, B*57, B*58, and B*5901. The middle cluster was christened the *B27 supertype* because most B*27 alleles were included there. Other molecules belonging to this supertype are B*0713, B*1309, B*15, B*1812, B*27, B*3513, B*3516, B*3528, B*37, B*3803, several B*39, B*40, B*41, B*4409, B*4431, B*45, B*46, B*47, B*48, B*50, B*5608, B*6702, and B*7805 (Table III). The last HLA-B cluster includes the following alleles: B*07, B*08, B*14, several B*15, B*18, B*2723, B*35, B*39, B*4008, B*4025, B*42, B*4806, B*5303, B*5305, B*54, B*55, B*56, B*6701, B*7301, B*78, B*81, B*82, and B*83 (Table III). The largest serotype here is B*07 and this cluster was named the *B7 supertype*.

Principal component analysis. All HLA-B molecules were aligned using as a template B*0801 (Protein Data Bank code: 1AGD). Six probes were found to explain >35% of the variance in the first two PCs (Table II). Initially, no consensus in the probe-protein interaction patterns was found by CPCA. To reduce the number of field variables, a region selection tool in Golpe was applied. The amino acids of the binding site were used as a template, and all data points within 4Å were retained. This gave three well-defined clusters on the CPCA scores plot with 37% of the variance explained by the first two PCs (Fig. 2, *upper right quadrant*). The cluster on the *upper right quadrant* consisted mainly of B*27 molecules (Table III). The cluster on the *lower right quadrant* includes alleles from the B44 supertype and the big cluster on the *left side* comprises molecules belonging to the hierarchical B7 supertype. The comparison between the classifications made by hierarchical clustering and CPCA showed that 305 of 447 molecules (68%) are classified in the same supertype. The results for each supertype are: 173 common for B7, 24 for B27, and 108 for B44.

The CPCA loading plots identify two important areas of interaction (Fig. 2, *lower left panel*). The former area is localized near positions 63 and 66. Both residues participate in pockets A and B within the peptide binding site. Position 63 is polymorphic, 66 is conserved. There are two optional amino acids for position 63: Glu and Asn (only B*0810 has with Asp^{63}). The latter area of interaction is near positions 77 and 81, both parts of pocket F. Asn, Ser, and Asp are found at 77, Leu and Ala at 81.

HLA-C supertypes

Hierarchical clustering. Two clusters were generated after hierarchical clustering on CoMSIA fields for 107 HLA-C molecules (Fig. 3, *upper left panel*, Table IV). The cluster on the *left side* of the dendrogram consists of Cw*02, Cw*0307, 08, 10, 15, Cw*04, Cw*05, Cw*06, Cw*0701, 06, 07, 09, 16, 18, Cw*1204, 05, 08, Cw*15, Cw*1602, Cw*1701, 02, 03, Cw*1801 and 02. We named this cluster supertype C4. The cluster on the *right side* includes Cw*01, Cw*0302, 03, 04, 05, 06, 09, 11, 12, 13, 14, 16, Cw*0702, 03, 04, 05, 08, 10, 11, 12, 13, 14, 15, 17, Cw*08, Cw*1202, 03, 06, 07, Cw*14, Cw*1601 and 04. We called this cluster *supertype C1*.

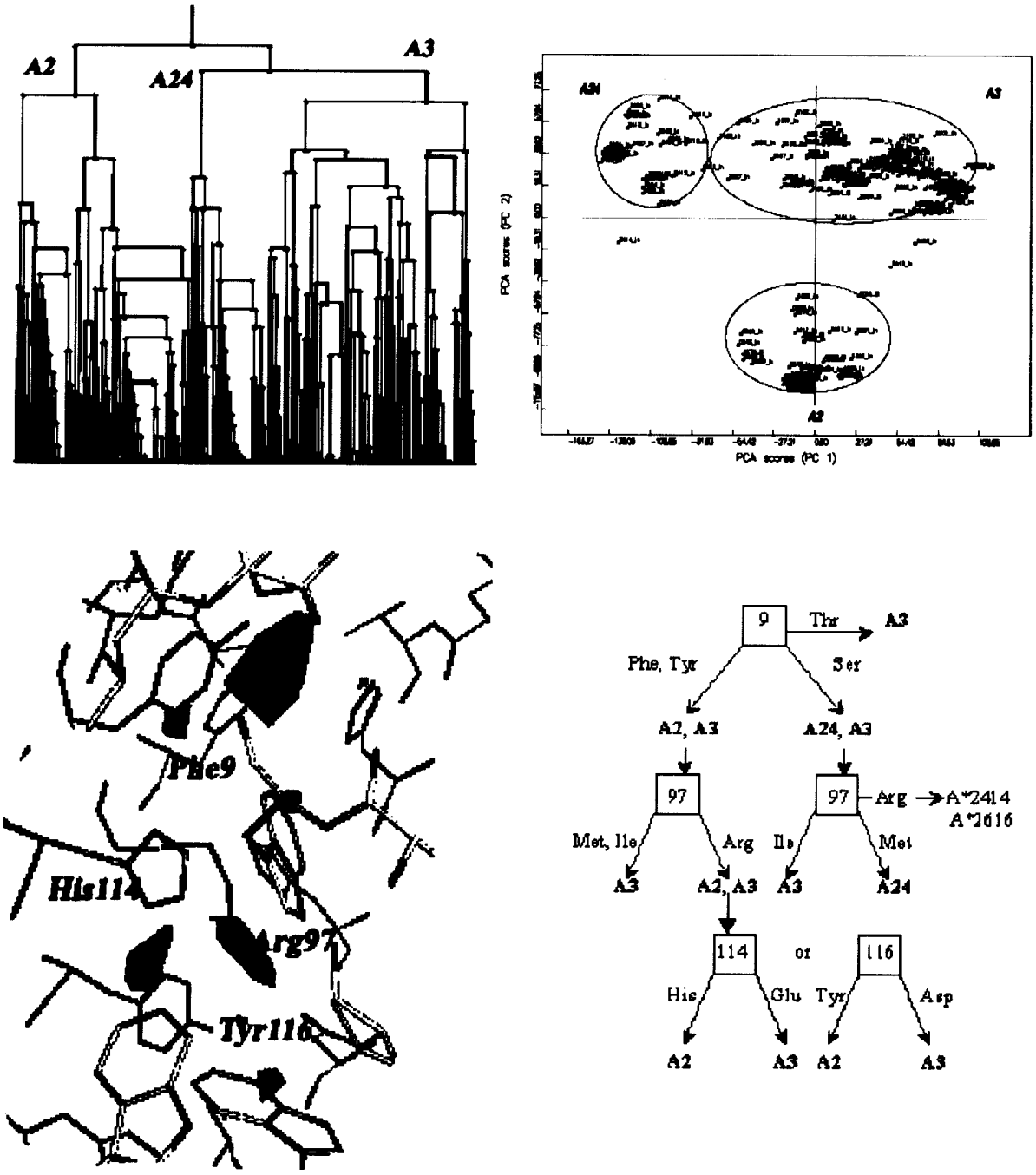

FIGURE 1. HLA-A supertypes. *Upper left panel*, Hierarchical clustering. *Upper right panel*, CPCA scores plot. *Lower left panel*, CPCA loadings plot of PC1 for C3 probe. X-ray structure of HLA-A*0201 is used to show the binding site amino acids close to the fields. The black and gray areas describe positions where differences between alleles exist. *Lower right panel*, Guide for HLA-A supertype determination.

Principal component analysis. Allele Cw*0401 (Protein Data Bank code: 1IM9) was used as a template in the alignment of HLA-C molecules. Eight probes were selected for the CPCA on HLA-C molecules (Table II). Two clusters could be defined in the two-dimensional CPCA scores plot (Fig. 3, *upper right panel*). The first PC explains 21% of the variance, the second 15%. The clusters are quite diffuse with many outliers. PC2 is more important for the clustering and that is why the loading plots were created on PC2. The clusters are in a good agreement with the hierarchical classification (Table IV). Only eight alleles were misclassified: Cw*0308, Cw*0310, Cw*0701, Cw*0706, Cw*0716, Cw*0718, Cw*1208, and Cw*1404.

All loading plots highlight two areas (one positive and one negative) inside pocket F: positions 70, 74, 77, and 81 (Fig. 3, *lower left panel*), of which only position 77 is polymorphic. Ser and Asn are the optional amino acids at this position. The cluster which makes positive contributions to PC2 (C1 supertype) favors interactions with a water molecule, i.e., Ser^{77} is preferred. The cluster making negative contributions (C4 supertype) favors a less polar residue here, like Asn^{77}.

Discussion

Two independent chemometric techniques were applied to the 3D structures of 783 HLA molecules to classify them into supertypes.

Table II. *HLA-A supertypes*

Supertype	Motif-based[a]	Hierarchical Clustering	CPCA	Fingerprint
A1	0101			
	2501			
	2601, 02			
	3201			
A2	0201-07	0201-60	0201-60 without 04, 17, 57	Tyr^{9}/Phe^{9}
	6802	2501-04	6802, 15, 23	Arg^{97}
	6901	2601-18	6901	His^{114} and Tyr^{116}
		3401, 05		
		4301		
		6601-04		
		6802, 15, 23		
		6901		
A24	2301	2301-09	2301-09	Ser^{9}
	2402-04	2402-38	2402-38	Met^{97}
	3001-03			
A3	0301	0101-09	0101-09	Tyr^{9}/Phe^{9}/Ser^{9}
	1101	0301-10	0301-10	Ile^{97}/Met^{97}
	3101	1101-14	1101-14	Glu^{114} and Asp^{116}
	3301	2901-07	2501-04	
	6801	3001-12	2601-18	
		3101-09	2901-07	
		3201-07	3001-12	
		3301-06	3101-09	
		3402-04	3201-07	
		3601-04	3301-06	
		6801-22 without 02, 15	3401-05	
		7401-09	3601-04	
		8001	4301	
			6601-04	
			6801-23 without 02, 15	
			7401-09	
			8001	

[a] Refs. 9–12.

The former technique–hierarchical clustering using CoMSIA fields–classifies the molecules on the basis of structural similarities in terms of steric bulk, electrostatic potential, hydrophobicity, and hydrogen-bond donor and acceptor abilities. The latter technique uses CPCA on probe-ligand interaction fields and classifies the proteins according to their interaction patterns. Seventy-seven percent of all molecules were classified in the same supertype by both techniques. Where data is available, our classification seems in good agreement with motif-based supertype classification (10–12, 36).

The CPCA loading plots indicate the differences between the supertypes. Taking into account these differences, guides were created for supertype determination within each locus. Thus, every known or new HLA class I molecule could be classified on the basis of only one to three residues. The combination of these residues for each supertype we called a supertype fingerprint. As five different properties (steric bulk, electrostatic potential, hydrophobicity, hydrogen-bond donor, and acceptor affinities) are considered at the same time in the hierarchical clustering, the guides were designed to follow the hierarchical classification.

Four HLA-A supertypes are defined by Sette et al. (12), based on common binding motifs: A1, A2, A3, and A24. The A1 supertype consists of A*0101, A*2501, A*2601, A*2602, and A*3201 (12). The A2 supertype includes A*0201-07, A*6802, and A*6901 (9). The A3 supertype comprises A*0301, A*1101, A*3101, A*3301, and A*6801 (36). The A24 supertype was defined to include A*2301, A*2402-04, and A*3001-03 (12). Our previous quantitative structure-activity relationship (QSAR) studies confirmed and extended the supermotifs for A2 and A3 supertypes (37, 38). The A*6802 molecule was found to be an intermediate allele lying between the A2 and A3 supertypes: in anchor position 2 it is closer to A3 and in anchor position 9 it is closer to A2 (38).

The hierarchical clustering and PCA classification made in the present study on 229 HLA-A molecules gave an 88% consensus, i.e., 201 HLA-A molecules are classified in the same manner by both methods. Good agreement exists with the motif-based supertypes, except for of A*0101, A*2501, A*2601, 02, A*3001–03, and A*3201 molecules. They were classified as A1 or A24 supertypes by Sette et al. (12). Our classification puts them into the A2 or A3 supertypes (Table II).

The CPCA loadings plots (Fig. 1, *lower left panel*) draw attention to four positions in the HLA-A alleles: 9, 97, 114, and 116. The amino acid at position 9 takes part in the formation of pocket B, one of the most specific pockets of the binding site. All A2 supertype, and most of the A3 supertype, alleles have Phe or Tyr at this position, making pocket B shallow, while A24 supertype alleles have Ser there. Position 9 of the MHC molecules corresponds to the primary anchor position 2 of the binding peptide. The binding motifs for A2 and A3 supertypes have common anchors at position 2: Leu, Ile, Met, Vat, and Ala, in contrast with A24 motif where Tyr is preferred. Residue 97 forms part of pockets C and E, corresponding to peptide positions 6 and 7, respectively. The amino acid at position 114 forms pockets D and E. Finally, residue 116 is part of pocket F, where the peptide C-terminal binds. The great difference between A2 and A3 binding motifs concerns anchors at position 9 (C-terminal): A2 alleles prefer Val, while A3 prefers Arg. Bulky residues like Tyr^{116} in A2 alleles restrict the size of pocket F and thus peptides bearing small hydrophobic amino acids are preferred there. In contrast, small negatively

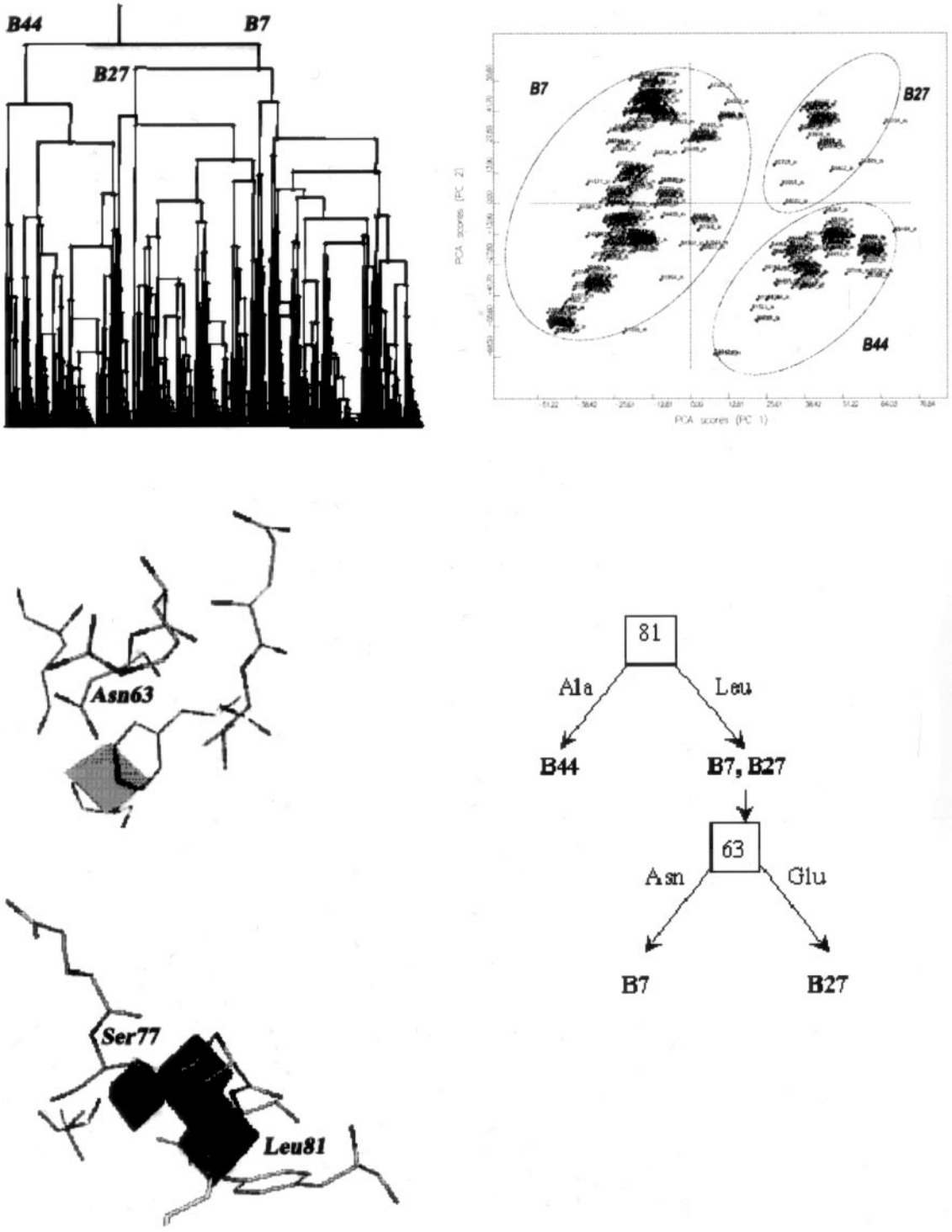

FIGURE 2. HLA-B supertypes. *Upper left panel*, Hierarchical clustering. *Upper right panel*, CPCA scores plot. *Lower left panel*, CPCA loadings plots of PC1 for DRY probe. X-ray structure of HLA-B*0801 is used to show the binding site amino acids close to the fields. The black and gray areas describe positions where differences between alleles exist. *Lower right panel*, Guide for HLA-B supertype determination.

charged residues like Asp^{116} in A3 alleles are more compatible with larger, positively charged amino acids such as Arg.

Considering positions 9, 97, 114, and 116, we identified a guide for HLA-A supertype fingerprint determination (Fig. 1, *lower right panel*). Starting at residue 9, Phe, Tyr or Thr here indicates A2 or A3 supertypes, Ser leads to A24 and certain A3 alleles. Next is position 97, which discriminates between A3 (Met^{97} or Ile^{97}) and A2 (Arg^{97}), and between A3 (Ile^{97}) and A24 (Met^{97}). When position 97 is Arg, an additional step is required: residue 114 or 116. His^{114} and Tyr^{116} are characteristic of A2 molecules, while Glu^{114} or Asp^{116} selects for A3. Hence, the fingerprint for the A2 supertype is Tyr^{9}/Phe^{9}, Arg^{97} and His^{114} or Tyr^{116}; for A3–Tyr^{9}/Phe^{9}/Ser9, Ile^{97}/Met^{97} and Glu^{114} or Asp^{116}; for A24–Ser^{9} and Met^{97}.

For HLA-B, hierarchical clustering and CPCA define three supertypes: B7, B27, and B44 with a consensus of 68%. The most questionable is the B27 supertype (only 24 common alleles). Most of the alleles defined as B27 by hierarchical clustering are classified as B7 or B44 by the CPCA. Sette et al. (12) define five motif-based

Table III. *HLA-B supertypes*

Supertype	Motif-Based[a]	Hierarchical Clustering	CPCA	Fingerprint
B44	37 40012 4006 41 44 45 47 49 50	0802 1301-1311 without 09 1513, 16, 17, 23, 24, 36, 67 1809 2701, 02 3801-3809 without 03 4013, 4019 4402-4433 without 09, 31 4704 4901-03 5101-34 5201-05 5301-09 without 03, 05 5701-09 5801-07 5901	0802 1301-1311 without 09 1513, 16, 17, 23, 24, 36, 67 1809 3805 4402-33 without 09 4701-04 without 02 4901-03 5101-34 5201-05 5301-09 5607 5701-09 58-07 5901	Ala[81]
B27	1401-02 1503, 09, 10, 18 2701-08 3801, 02 3901-04 4801, 02 7301	0713 1309 1501-1575 without these in B7 and B44 1812 2703-2725 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	0727 2701-25 without 08, 12, 18 3701-04 3801-09 4013, 19, 28	Glu[63] Leu[81]
B7	07 35 51 53 54 55 56 67 78	0702-31 without 13 0801-17 without 02 1401-06 1502, 08, 09, 10, 11, 15, 18, 21, 29, 37, 44, 51, 52, 55, 64, 72 1801-18 without 09, 12 2723 3501-45 without 13, 16, 28 3901-27 without 02, 08, 13, 22, 23 4008, 25 4201-04 4806 5303, 05 5401, 02 5501-12 5601-11 without 08 6701 7301 7801-04 8101 8201, 02 8301	0702-31 without 0727 0801-17 without 02 1309 1401-06 1501-75 without 13, 16, 17, 23, 24, 36, 67 1801-18 without 09 2708, 12, 18 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 5607 6701, 02 7301 7801-05 8101 8201, 02 8301	Asn[63] Leu[81]
B58	1516, 17 5701, 02 58			
B62	1301-02 1501, 02, 06, 12, 13, 14, 19, 21 4601 52			

[a] Refs. 10–12, 36.

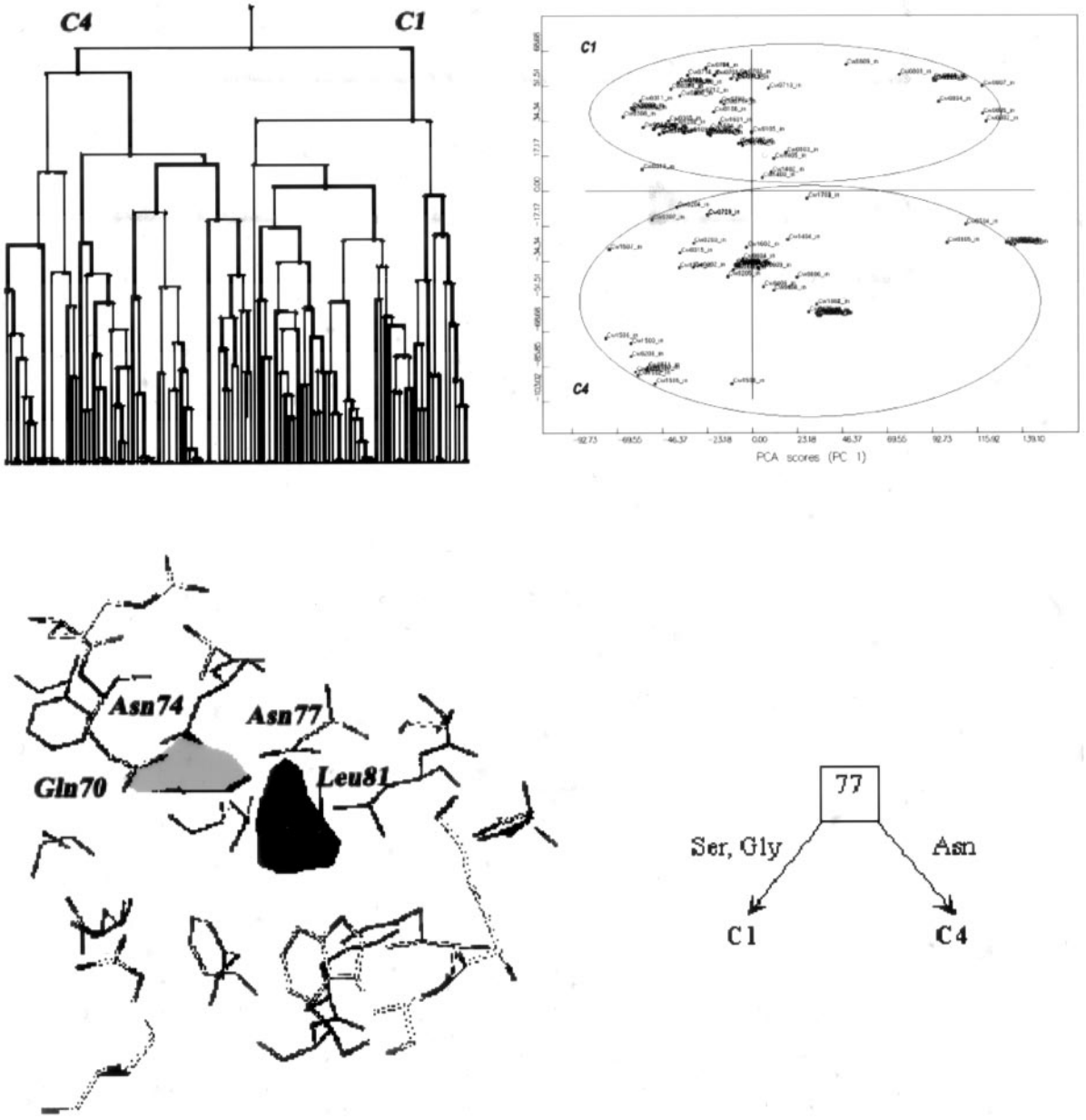

FIGURE 3. HLA-C supertypes. *Upper left panel*, Hierarchical clustering. *Upper right panel*, CPCA scores plot. *Lower left panel*, CPCA loadings plot of PC2 for H_2O probe. X-ray structure of HLA-Cw*0401 is used to show the binding site amino acids close to the fields. The black and gray areas describe positions where differences between alleles exist. *Lower right panel*, Guide for HLA-C supertype determination.

HLA-B supertypes: B7, B27, B44, B58, and B62 (Table III). Sette's B58 supertype was clustered as a part of B44 in the present study, and B62 as part of B7, B27, or B44.

The positions highlighted in the CPCA loading plots allowed us to draw a guide for HLA-B supertype fingerprint determination (Fig. 2, *lower right*), analogous to the HLA-A guide. Starting at position 81, the HLA-B molecules could be divided into two groups: the former group has Ala, the latter has Leu. All B44 alleles have Ala^{81}, while B7 and B27 have Leu^{81}. The next is position 63. B27 alleles have Glu^{63} and B7 has Asn^{63}. Thus, the fingerprint for B7 affiliation is Asn^{63} and Leu^{81}; for B27–Glu^{63} and Leu^{81}; for B44-Ala^{81}.

At the present time, information about HLA-C supertypes is not available in the literature. The present study identifies two supertypes, which we call C1 and C4, respectively (Table IV, Fig. 3, *lower left* and *right panels*). Only position 77 was found to be important for this classification. Molecules with Ser^{77} belong to C1 supertype, these with Asn^{77} to C4. Position 77 is a key position in protein molecules from the C locus. It takes part in the formation of pockets C and F, and the killer Ig-like receptor binding site (7). There is 93% consensus between the hierarchical and CPCA-based classifications for HLA-C molecules (Table IV).

The small number of supertype-determining MHC sequence positions delineated by our analysis all equate to solvent-exposed residues contributing to the peptide binding site. If, for simplicity, we assume that, over time, random mutations will occur at the nucleic acid level–so called single nucleotide polymorphisms–then an MHC will, in the absence of other selective pressures, accumulate point mutations throughout its sequence. Some will be "silent" mutations, which do not change amino acid identity, and others, often called mis-sense mutations, which will alter the phenotypic amino acid. For solvent inaccessible residues buried in the

Table IV. *HLA-C supertypes*

Supertype	Motif-Based	Hierarchical Clustering	CPCA	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 7, 15 0701-18 without 07, 09 0801-09 1202-08 without 04, 05 1402-05 1601, 04	Ser^{77}/Gly^{77}
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	Asn^{77}

protein core, such changes will only rarely be advantageous and, on the basis of stability, be selected against. Some mutations will have little or no functional consequence, but those interfering with, say, binding to β_2-microglobulin, or to TCRs, or to bound peptides will affect function. In terms of peptide binding, certain amino acid mutations will have a significant effect on the peptide binding site microenvironment, altering residue size, hydrophobicity, or charge. Within the context of the cellular Ag presentation pathway, class I MHCs will select for changes that moderate the size, shape, and physical-chemical environment within the characteristic binding sites exhibited by different supertypes. MHC supertypes are thought to arise through both divergent and convergent molecular evolution. Such processes will concentrate acceptable mutations at certain sequence positions where conflicting constraints can best be balanced. Our analysis, though it contains no overt phylogenetic component, and indeed uses no direct amino acid sequence information in calculating similarities between binding sites, nonetheless allows us to identify crucial sequence positions for the determination of peptide specificity.

Two principal advantages of the present classification should be highlighted. First, this approach is based on the whole binding site, not on particular pockets. There is a lot of experimental data to show that peptide anchors are necessary, but not sufficient, for good binding. There are several secondary anchors (positions 1, 3, 6, and 7) which are also responsible for high affinity (39). Considering the whole binding site provides us with a wider, and clearer, overview of the interactions within it. Second, the approach classifies all currently known HLA molecules, an experimental task that is presently impossible. The resulting supertype fingerprints are both simple and easy to apply for every known, or newly discovered, HLA class I molecule, and, potentially, for MHCs from other species. The classification of alleles into supertypes is useful for defining promiscuous peptides able to bind to numerous alleles (40, 41) and for identifying proper population coverage of peptide-based vaccines (42–44). The greatly increased scope of our analysis opens the way to a broad definition of MHC specificities, ensuring cross-reactive peptide vaccines able to cover the whole human population can be designed.

Acknowledgments

The Edward Jenner Institute for Vaccine Research would like to thank its sponsors: GlaxoSmithKline, Medical Research Council, Biotechnology and Biological Sciences Research Council, and U.K. Department of Health.

References

1. Janeway, Jr., C. A., P. Travers, M. Walport, and J. D. Capra, eds. 1999. *Immunobiology: The Immune System in Health and Disease*. Current Biology Publications, London, p. 135.
2. Robinson, J., M. J. Waller, P. Parham, N. de Groot, R. Bontrop, L. J. Kennedy, P. Stoehr, and S. G. E. Marsh. 2003. IMGT/HLA and IMGT/MHC: sequence databases for the study of the major histocompatibility complex. *Nucleic Acids Res. 31:311*.
3. Bjorkman, P. J., M. A. Saper, B. Samraouri, W. S. Bennett, J. L. Strominger, and D. C. Wiley. 1987. Structure of the human class I histocompatibility antigen, HLA-A2. *Nature 329:506*.
4. Bjorkman, P. J., M. A. Saper, B. Samraouri, W. S. Bennett, J. L. Strominger, and D. C. Wiley. 1987. The foreign antigen binding site and T cell recognition regions of class I histocompatibility antigens. *Nature 329:512*.
5. Saper, M. A., P. J. Bjorkman, and D. C. Wiley. 1991. Refined structure of the human histocompatibility antigen HLA-A2 at 2.6 A resolution. *J. Mol. Biol. 219:277*.
6. Smith, K. J., S. W. Reid, K. Harlos, A. J. McMichael, D. I. Stuard, J. I. Bell, and E. Y. Jones. 1996. Bound water structure and polymorphic amino acids act together to allow the binding of different peptides to MHC class I HLA-B53. *Immunity 4:215*.
7. Fan, Q. R., and D. C. Wiley. 1999. Structure of human histocompatibility leukocyte antigen (HLA)-Cw4, a ligand for the KIR2D natural killer cell inhibitory receptor. *J. Exp. Med. 190:113*.
8. Falk, K., O. Rötzschke, S. Stefanovic, G. Jung, and H.-G. Rammensee. 1991. Allele-specific motifs revealed by sequencing of self-peptides eluted from MHC molecules. *Nature 351:290*.
9. Del Guercio, M. F., J. Sidney, G. Hermanson, C. Perez, H. M. Grey, R. T. Kubo, and A. Sette. 1995. Binding of a peptide antigen to multiple HLA alleles allows definition of an A2-like supertype. *J. Immunol. 154:685*.
10. Sidney, J., H. M. Grey, R. T. Kubo, and A. Sette. 1996. Practical, biochemical and evolutionary implications of the discovery of HLA class I supermotifs. *Immunol. Today 17:261*.
11. Sette, A., and J. Sidney. 1998. HLA supertypes and supermotifs: a functional perspective on HLA polymorphism. *Curr. Opin. Immunol. 10:478*.
12. Sette, A., and J. Sidney. 1999. Nine major HLA class I supertypes account for the vast preponderance of HLA-A and -B polymorphism. *Immunogenetics 50:201*.
13. Sette, A., B. Livingstone, D. McKinney, E. Appella, J. Fikes, J. Sidney, M. Newman, and R. Chesnut. 2001. The development of multi-epitope vaccines: epitope identification, vaccine design and clinical evaluation. *Biologicals 29:271*.
14. Sette, A., M. Newman, B. Livingston, D. McKinney, J. Sidney, G. Ishioka, S. Tangri, J. Alexander, J. Fikes, and R. Chestnut. 2002. Optimizing vaccine design for cellular processing, MHC binding and TCR recognition. *Tissue Antigens 59:443*.
15. Flower, D. R., I. A. Doytchinova, K. Paine, P. Taylor, M. J. Blythe, D. Lamponi, C. Zygouri, P. Guan, H. McSparron, and H. Kirkbride. 2002. Computational vaccine design. In *Drug Design: Cutting Edge Approaches*, D. R. Flower, ed. RSC Publications, Cambridge, p. 136.
16. Flower, D. R., and I. A. Doytchinova. 2003. Immunoinformatics and the prediction of immunogenicity. *Appl. Bioinformatics 1:167*.
17. Cano, P., B. Fan, and S. Stass. 1998. A geometric study of the amino acid sequence of class I HLA molecules. *Immunogenetics 48:324*.

18. McKenzie, L. M., J. Pecon-Slattery, M. Carrington, and S. J. O'Brien. 1999. Taxonomic hierarchy of HLA class I allele sequences. *Genes Immun. 1:120*.
19. Chelvanayagam, G. 1996. A roadmap for HLA-A, HLA-B, and HLA-C peptide binding specificities. *Immunogenetics 45:15*.
20. Zhang, C., A. Anderson, and C. DeLisi. 1998. Structural principles that govern the peptide-binding motifs of class I MHC molecules. *J. Mol. Biol. 281:929*.
21. Zhao, B., A. E. H. Png, E. C. Ren, P. R. Kolatkat, V. S. Methura, M. K. Sakharkar, and P. Kangueane. 2003. Compression of functional space in HLA-A sequence diversity. *Hum. Immunol. 64:718*.
22. Berman, H. M., J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne. 2000. The Protein Data Bank. *Nucleic Acids Res. 28:235*.
23. Hillig, R. C., P. G. Coulie, V. Stroobant, W. Saenger, A. Ziegler, and M. Huelsmeyer. 2001. High resolution structure of HLA-A*0201 in complex with a tumor-specific antigenic peptide encoded by the Mage-A4 gene. *J. Mol. Biol. 310:1167*.
24. Reid, S., S. McAdam, K. J. Smith, P. Klenerman, C. A. O'Callaghan, K. Harlos, B. K. Jakobsen, A. J. McMichael, J. I. Bell, D. I. Stuart, and E. Y. Jones. 1996. Antagonist HIV-1 Gag peptides induce structural changes in HLA B8. *J. Exp. Med. 184:2279*.
25. Fan, Q. R., E. O. Long, and D. C. Wiley. 2001. Crystal structure of the human natural killer cell inhibitory receptor Kir2Dl1 bound to its MHC ligand HLA-Cw4. *Nat. Immunol. 2:452*.
26. Bower, M., F. E. Cohen, and R. L. Dunbrack, Jr. 1997. Sidechain prediction from a backbone-dependent rotamer library: a new tool for homology modeling. *J. Mol. Biol. 267:1268.*
27. Barnard, J. M., and G. M. Downs. 1992. Clustering of chemical structures on the basis of two-dimensional similarity measures. *J. Chem. Inf. Comput. Sci. 32:644*.
28. Brown, R. D., and Y. C. Martin. 1996. Use of structure-activity data to compare structure-based clustering methods and descriptors for use in compound selection. *J. Chem. Inf. Comput. Sci. 36:572*.
29. Klebe, G., U. Abraham, and T. Mietzner. 1994. Molecular similarity indices in a comparative analysis (CoMSIA) of drug molecules to correlate and predict their biological activity. *J. Med. Chem. 37:4130*.
30. Klebe, G., and U. Abraham. 1999. Comparative molecular similarity index analysis (CoMSIA) to study hydrogen-bonding properties and to score combinatorial libraries. *J. Comput. Aided Mol. Des. 13:1*.
31. Böhm, M., J. Stürzebecher, and G. Klebe. 1999. Three-dimensional quantitative structure-activity relationship analyses using comparative molecular field analysis and comparative molecular similarity indices analysis to elucidate selectivity differences of inhibitors binding to trypsin, thrombin, and factor Xa. *J. Med. Chem. 42:458*.
32. Doytchinova, I., and D. R. Flower. 2001. Towards the quantitative prediction of T-cell epitopes: CoMFA and CoMSIA studies of peptides with affinity to class I MHC molecule HLA-A*0201. *J. Med. Chem. 44:3572*.
33. Doytchinova, I. A., and D. R. Flower. 2002. A comparative molecular similarity index analysis (CoMSIA) study identifies an HLA-A2 binding supermotif. *J. Comput. Aided Mol. Des. 16:535*.
34. Guan, P., I. A. Doytchinova, and D. R. Flower. 2003. A comparative molecular similarity indices (CoMSIA) study of peptides binding to the HLA-A3 superfamily. *Bioorg. Med. Chem. 11:2307*.
35. Eriksson, L., E. Johansson, N. Kettaneh-Wold, and S. Wold. 2001. Multi- and megavariate data analysis. Umetrics Academy, Umeå, p. 43.
36. Sidney, J., H. M. Grey, S. Southwood, E. Celis, P. A. Wentworth, M. F. del Guercio, R. T. Kubo, R. W. Chesnut, and A. Sette. 1996. Definition of an HLA-A3-like supermotif demonstrates the overlapping peptide binding repertoires of common HLA molecules. *Hum. Immunol. 45:79*.
37. Guan, P., I. A. Doytchinova, and D. R. Flower. 2003. HLA-A3-supermotif defined by quantitative structure-activity relationship analysis. *Protein Eng. 16:11*.
38. Doytchinova, I. A., and D. R. Flower. 2003. The HLA-A2-supermotif: a QSAR definition. *Org. Biomol. Chem. 1:2648*.
39. Ruppert, J., J. Sidney, E. Celis, R. T. Kubo, H. M. Grey, and A. Sette. 1993. Prominent role of secondary anchor residues in peptide binding to HLA-A*0201 molecules. *Cell 74:929*.
40. Sidney, J., S. Southwood, D. L. Mann, M. A. Fernandez-Vina, M. J. Newman, and A. Sette. 2001. Majority of peptides binding HLA-A*0201 with high affinity crossreact with other A2-supertype molecules. *Hum. Immunol. 62:1200*.
41. Brusic, V., N. Petrovsky, G. Zhang, and V. B. Bajic. 2002. Prediction of promiscuous peptides that bind HLA class I molecules. *Immunol. Cell Biol. 80:280*.
42. Gulukota, K., and C. DeLisi. 1996. HLA allele selection for designing peptide vaccines. *Genet. Anal. 13:81*.
43. Schipper R. F., C. A. van Els, J. D'Amaro, and M. Oudshoorn. 1996. Minimal phenotype panels: a method for achieving maximum population coverage with a minimum of HLA antigens. *Hum. Immunol. 51:95.*
44. Dawson, D. V., M. Ozgur, K. Sari, M. Ghanayem, and D. D. Kostyu. 2001. Ramifications of HLA class I polymorphism and population genetics for vaccine development. *Genet. Epidemiol. 20:87*.