

# In Silico Identification of Supertypes for Class II MHCs<sup>1</sup>

Irina A. Doytchinova and Darren R. Flower<sup>2</sup>

The development of epitope-based vaccines, which have wide population coverage, is greatly complicated by MHC polymorphism. The grouping of alleles into supertypes, on the basis of common structural and functional features, addresses this problem directly. In the present study we applied a combined bioinformatics approach, based on analysis of both protein sequence and structure, to identify similarities in the peptide binding sites of 2225 human class II MHC molecules, and thus define supertypes and supertype fingerprints. Two chemometric techniques were used: hierarchical clustering using three-dimensional Comparative Similarity Indices Analysis fields and nonhierarchical *k*-means clustering using sequence-based *z*-descriptors. An average consensus of 84% was achieved, i.e., 1872 of 2225 class II molecules were classified in the same supertype by both techniques. Twelve class II supertypes were defined: five DRs, three DQs, and four DPs. The HLA class II supertypes and their fingerprints given in parenthesis are DR1 (Trp<sup>9β</sup>), DR3 (Glu<sup>9β</sup>, Gln<sup>70β</sup>, and Gln/Arg<sup>74β</sup>), DR4 (Glu<sup>9β</sup>, Gln/Arg<sup>70β</sup>, and Glu/Ala<sup>74β</sup>), DR5 (Glu<sup>9β</sup>, Asp<sup>70β</sup>), and DR9 (Lys/Gln<sup>9β</sup>); DQ1 (Ala/Gly<sup>86β</sup>), DQ2 (Glu<sup>86β</sup>, Lys<sup>71β</sup>), and DQ3 (Glu<sup>86β</sup>, Thr/Asp<sup>71β</sup>); DPw1 (Asp<sup>84β</sup> and Lys<sup>69β</sup>), DPw2 (Gly/Val<sup>84β</sup> and Glu<sup>69β</sup>), DPw4 (Gly/Val<sup>84β</sup> and Lys<sup>69β</sup>), and DPw6 (Asp<sup>84β</sup> and Glu<sup>69β</sup>). Apart from the good agreement between known binding motifs and our classification, several new supertypes, and corresponding thematic binding motifs, were also defined. *The Journal of Immunology*, 2005, 174: 7085–7095.

**M**ajor histocompatibility complex proteins are glycoproteins that bind, within the cell, small peptide fragments, or epitopes, derived, through proteolysis, from both host and pathogen proteins, and present them at the cell surface for interaction by T cells. Recognition by the immune system of peptide-bound MHCs is fundamental to the mechanism by which the host identifies and responds to foreign Ags. MHC class I molecules, available on most cell types, present peptides from protein synthesized within the cell (endogenous processing pathway), and only a subset of macrophages are able to present peptides derived from phagocytosed material via class I molecules (1). MHC class II molecules, expressed on a restricted number of cell types, such as B cells and macrophages, can present peptides derived from endocytosed extracellular proteins (exogenous processing pathway) (2).

A principal feature of MHC molecules is their allelic polymorphism: the July 2004 ImMunoGeneTics/HLA database release lists 1114 class I and 707 class II molecules (3). Such polymorphism presumably enhances the probability of mounting an immune response by at least a subset of individuals within a population, ultimately increasing the chance of group survival against infection (4). Unlike many proteins, MHC alleles have arisen under a specific and discernible evolutionary pressure, adapting to a fitness landscape mediated by geographically constrained infectious disease. Moreover, any poly-epitope vaccine targeting the whole population would, on the same basis, need to bind a range of HLA molecules. Gulucota and DeLisi (5) found that three to six class I HLA alleles, depending on

the ethnic group, would cover ~90% of the population. Indeed, because of linkage disequilibrium (the joint probability of a given allelic pair is not generally equal to the product of their individual probabilities,  $P_{ij} \neq P_i P_j$ ), it is not necessarily optimal to choose the alleles with the highest individual frequencies.

The peptide binding site of MHC molecules is composed of a single protein chain for class I and two separate chains in class II. X-ray data reveal that the walls of the cleft are formed by two antiparallel helices and the floor is formed by an eight-stranded  $\beta$ -sheet (6–10). In MHC class I molecules, the ends of the cleft are closed off, generally allowing only short peptides of 8–11 aa to bind. In contrast, the cleft in class II is open-ended, allowing much longer peptides to bind, even though only 9 aa occupy the site itself. Both clefts have binding pockets, corresponding to primary and secondary anchor positions on the binding peptide. The combination of two or more anchors is called a motif. It has been found that certain class I alleles can recognize similar motifs (11–14) and thus be grouped into HLA “supertypes”, binding common “super-motifs”. The classification of MHC molecules into supertypes, based on structural features and/or peptide specificity, is of prime importance in the development of epitope-based vaccines (15, 16). The experimental determination of motifs for every allele is prohibitively expensive in terms of labor, time, and resources. The only comprehensive, yet practical, alternative is a bioinformatic approach.

Chemometric methods are widely used and extensively validated in computational chemistry for structural classifications (17, 18). Recently, we proposed a “three-dimensional (3D)<sup>3</sup> supertype fingerprint” approach which classifies alleles on the basis of information from the structure of the binding sites using two chemometric techniques: principal component analysis and hierarchical clustering (18). We applied this approach to class I MHC molecules belonging to the HLA-A, HLA-B, and HLA-C loci and showed that only 1–3 aa are sufficient for an allele to be classified within a particular supertype.

Edward Jenner Institute for Vaccine Research, Compton, Berkshire, United Kingdom  
Received for publication September 8, 2004. Accepted for publication February 28, 2005.

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.

<sup>1</sup> This work was supported by GlaxoSmithKline, Medical Research Council, Biotechnology and Biological Sciences Research Council, and United Kingdom Department of Health.

<sup>2</sup> Address correspondence and reprint requests to Dr. Darren R. Flower, Edward Jenner Institute for Vaccine Research, Compton, Berkshire, U.K., RG20 7NN. E-mail address: darren.flower@jenner.ac.uk

<sup>3</sup> Abbreviations used in this paper: 3D, three dimensional; CoMSIA, Comparative Similarity Indices Analysis; RSP, restrictive supertype pattern; JRA, juvenile rheumatoid arthritis.

In the present study, a combined two-dimensional-3D approach was applied to class II HLA molecules belonging to the DR, DQ, and DP loci, identifying a consensus supertype classification. In contrast to class I, supertypes and supermotifs for class II MHC molecules have not been widely studied. There are only a few classifications for class II molecules: three for HLA-DR molecules (3, 19, 20) and one each for HLA-DQ (21) and HLA-DP (22). Here, we use two clustering techniques—hierarchical and nonhierarchical—applied to both the sequences of HLA class II proteins and their 3D structures. Clustering is a data analysis technique that, when applied to a set of heterogeneous items, identifies homogeneous subgroups as defined by a given model or measure of similarity. For a detailed review see Ref. 23.

In hierarchical clustering, the data set is analyzed iteratively: at each step either a pair of clusters is merged (agglomerative) or a single cluster is divided (divisive). Determining the number of “natural” clusters is among the most difficult problems in clustering and, to date, no general solution has been found. Agglomerative hierarchical clustering was applied to HLA class II molecules using similarity fields generated by Comparative Similarity Indices Analysis (CoMSIA) (24, 25). CoMSIA is widely used in 3D molecular design to model the interactions between small molecules and proteins (26–32).

Nonhierarchical methods generate a specific number of disjoint, flat, unconnected clusters. *K*-means clustering is a nonhierarchical method in which the dataset is partitioned into *k* clusters by choosing an initial set of *k* seed compounds to act as initial cluster centers. Each compound is assigned to its nearest cluster and cluster membership is iteratively refined by shifting compounds between clusters until stability is achieved, i.e., no compounds are moved from one cluster to another. The *k*-means method was applied to a set of *z*-scales, as defined by Hellberg et al. (33) and extended by Sandberg et al. (34), which describe the most important properties of each amino acid within the HLA class II binding site. In the field of quantitative structure–activity relationships, *z*-descriptors are used to model the interactions between peptides and proteins (35–39).

Based on the consensus of the classifications, made by the hierarchical and nonhierarchical methods, twelve class II supertypes were defined: five DR, three DQ, and four DP. Fingerprints for each supertype were also identified. In a similar way to our existing analysis of class I supertypes (40), it was found that 1–3 aa are sufficient to distinguish between class II supertypes.

## Materials and Methods

### Protein structure modeling

The protein sequences of HLA class II molecules were collected from the ImMunoGeneTics/HLA database (3). Only the first 80 aa of the  $\alpha$ -chains and the first 90 residues of the  $\beta$ -chain were required as they formed the binding site. Sequences missing >10 aa at the N terminus were removed. All possible combinations of  $\alpha$ - and  $\beta$ -chains for DQ and DP were modeled. This generates 738 DQ molecules (18 DQA  $\times$  41 DQB) and 1140 DP molecules (12 DPA  $\times$  95 DPB). Because the HLA-DR  $\alpha$ -chain has yet to

exhibit binding site polymorphism, only the HLA-DR  $\beta$ -chains were varied, generating 347 HLA-DRB structures. The rotameric placement of amino acid substitutions was modeled using SCRWL2.8 (41), with invariant amino acids held rigid. The x-ray structures of 1PYW (42) for DRB and 1JK8 (43) for DQ were used as templates for the backbone and invariant side chains. As x-ray data for DP molecules are not available, a homology model of the DP  $\alpha$ 1 $\beta$ 1 domain (DPA1\*0103/DPB1\*0101) was derived using the structure of HLA-DR1 (DRB\*0101) (PDB: 1PYW) as a template. The de facto standard protein homology modeling system Modeler was used (44). We applied standard settings for all parameters (45). Deletions were made at positions 23 and 24 in the  $\beta$ -chain. The resulting DP structure was used as the basis for side chain placement using SCRWL2.8. In the multivariant cases only the first molecule (\*xxx01) was considered. For use in Sybyl, all homology models were aligned to the starting structure. Hydrogen atoms and Kollman charges were added for each molecule. To avoid the influence of polymorphism outside the binding site (“polymorphic noise”), only the amino acids forming the binding site (“polymorphic signal”) were considered, with conserved binding site amino acids also omitted. The amino acids forming the binding sites were selected on the basis of x-ray data of peptide-class II protein complexes (Refs. 3, 9, 10, 21, 22, 42, 43, 46, 47) (Table I).

### Hierarchical clustering on CoMSIA fields

CoMSIA was used as implemented in Sybyl version 6.9 (Tripos, 2004). The 3D structures of proteins belonging to the same locus were aligned: the x-ray structure 1PYW was used as a template for DR (42), the x-ray structure 1JK8 (43) was used for DQ, and the modeled DP structure was used for DP molecules. The amino acids outside the binding site were excluded. The grid had a resolution of 2.0 Å and extended beyond the molecular dimensions by 4.0 Å in all directions. At each grid point, a similarity index between the probe and the target molecule is calculated using a Gaussian-type distance-dependent function. Similarity indices fields were generated with an attenuation factor  $\alpha = 0.3$ . The attenuation factor shows the steepness of the Gaussian-type function. The probe used had a 1 Å radius, charge + 1, hydrophobicity + 1, hydrogen-bond donor + 1, and acceptor properties + 1. The agglomerative hierarchical clustering (23) option of Sybyl version 6.9 was applied to CoMSIA fields. According to this technique 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. The distance between the clusters was calculated using the complete-linkage method, i.e., using the distance between the most distant pair of data points in both clusters. The last four levels of the hierarchy were considered for supertype definition.

### Nonhierarchical clustering on *z*-scores

The protein sequences for each class II locus were aligned. As in the CoMSIA study, amino acids outside the binding site were excluded. Each amino acid was described by five *z*-descriptors:  $z_1$  (hydrophobicity),  $z_2$  (steric bulk),  $z_3$  (polarity),  $z_4$ , and  $z_5$  (electronic effects) (34). An X-matrix was formed for each locus. Rows corresponded to the number of proteins and columns equaled five times the number of polymorphic amino acids in the binding site. The X-matrices were imported into MDL QSAR version 2.2. *K*-means clustering was applied, with the initial set of *k* seeds equal to the number of clusters generated by the hierarchical clustering. The members of the clusters generated by the hierarchical and nonhierarchical clustering were compared, and the commonly clustered members were calculated as a percentage of all alleles for every locus.

Table I. Definition of amino acids HLA comprising the three class II binding sites considered in the analysis<sup>a</sup>

Locus	$\alpha$ -Chain	$\beta$ -Chain	Refs.
HLA-DR	— <sup>b</sup>	9, 11, 13, 26, 28, 30, 38, 47, 56, 57, 60, 61, 67, 70, 71, 74, 77, 78, 81, 82, 85, 86, 89, 90	3, 9, 10, 42, 46, 47
HLA-DQ	8, 22, 31, 52, 53, 58, 61, 66, 72, 73	9, 13, 26, 28, 30, 37, 38, 47, 55, 56, 57, 67, 70, 71, 74, 77, 84, 85, 86, 87	21, 43, 47
HLA-DP	11, 31, 66, 72, 73	8, 9, 35, 36, 55, 56, 65, 69, 72, 76, 84, 85, 86, 87	22, 47, 48, 49

<sup>a</sup> Conserved amino acids are omitted.

<sup>b</sup>  $\alpha$ -Chain of HLA-DR is not polymorphic in the binding site.

## Results

Systematic models of combinatorially generated class II human MHC dimers were analyzed using a 3D technique, which applies hierarchical clustering to CoMSIA fields, and a two-dimensional technique, which uses  $z$ -descriptors and  $k$ -means clustering. Together, these clustering methods generated a robust grouping of MHCs based on the similarities of their binding sites. Structural models were built using three templates: two taken from x-ray data (42, 43) and one generated using homology modeling (44, 45). The DP-modeled structure had a root mean squared deviation from the DRB\*0101 structure (1PYW) of 0.28 Å for all invariant atoms. The three templates were then used to generate, in a combinatorial fashion, all modeled dimers through side chain placement using rotamers (41). Invariant side chains were held fixed. As a result, the models created are best viewed as a conservative estimate: an attempt to reduce error by overlaying as much of the generated models as possible. This approach is well suited to the statistical nature of the method, which looks for overall correlations in CoMSIA fields, and is tolerant of small misplacements of individual side chains (17, 18). This is confirmed by the high agreement seen with the  $k$ -means clustering, which makes no use of structural information and is based on sequence information alone.

Making use of similarity fields generated by CoMSIA, agglomerative hierarchical clustering was applied to amino acids forming the binding sites on HLA class II molecules (Table I). CoMSIA is a 3D grid method, in which a probe is placed at all lattice points in a regular 3D grid in and around the target molecule (24, 25). At each point, a similarity index between probe and target is calculated using a Gaussian-type distance dependent function. Five similarity fields were calculated: steric bulk, electrostatic potential, local hydrophobicity, and hydrogen-bond donor and acceptor abilities. In hierarchical clustering, each level defines a partition of the data set into clusters. However, in general it is not clear which level is best in terms of splitting the data set into a “natural” number of clusters, so that each cluster contains the most appropriate compounds (23). In the present study, the number of clusters was selected to be in a good agreement with previous classifications and known binding motifs, where these are available. Usually, the last four levels were considered for the supertype definition. The number of clusters defined in the hierarchical clustering was used as the input  $k$  cluster number in the nonhierarchical  $k$ -means clustering.

Nonhierarchical  $k$ -means clustering was applied to a set of  $z$ -properties describing each amino acid of the HLA class II binding site. These  $z$ -scales, as defined by Hellberg et al. (33), reflect the most important properties of amino acids and are referred to as “principal properties”. These scales were derived by principal component analysis from a data matrix consisting of a large number of physicochemical variables, such as m.w.,  $pK_a$ 's,  $^{13}C$  NMR-shifts, etc. The first principal component (PC) reflects amino acid hydrophobicity, the second PC reflects their size, and the third, their polarity. The three PCs are labeled:  $z_1$ -,  $z_2$ - and  $z_3$ -scales, respectively. More recently, Sandberg et al. (34) extended the three  $z$ -scales to five, adding  $z_4$  and  $z_5$ , which account for electronic effects of the amino acids. With these  $z$ -scales, it is possible to quantify numerically the structural variations within a series of related peptides, by arranging the  $z$ -scales according to the amino acid sequence. In the present study the five  $z$ -scales were used to describe the polymorphic amino acid sequences of the binding site of HLA class II molecules (Table I).

The common members of the clusters derived by both methods were expressed as a percentage of all alleles for every locus. Supertype fingerprints were defined on the basis of common amino

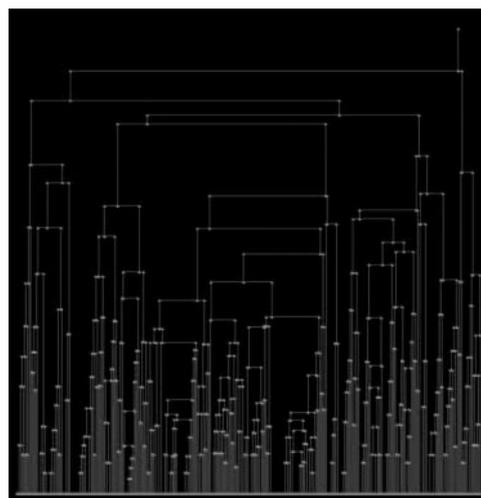
acids found within the multiple sequence alignment of alleles belonging to one supertype.

### HLA-DR supertypes

**Hierarchical clustering.** The hierarchical clustering, using CoMSIA fields, of HLA-DRB is plotted in Fig. 1, and the detailed content of each cluster is shown in Table II. At the second level of the hierarchy there are three clusters: two small clusters flanking one huge central cluster. At this level, the clustering is associated with polymorphism around pocket 9. The leftmost cluster is composed of structures with Trp<sup>9β</sup>, the middle cluster has Glu<sup>9β</sup>, and the rightmost has Lys/Gln<sup>9β</sup>. As the structures in the middle cluster were still quite diverse, mainly around binding pocket 4, ( $\beta$ 70,  $\beta$ 71, and  $\beta$ 74), this group was further subdivided. At the fourth level there are five roughly equally sized clusters, which corresponds well with known DR binding motifs. The clusters were defined as super-types and named after the lowest serotype included.

The first cluster, which we called the DR1 supertype, includes DR1 (DRB1\*0101–11), DR2 (DRB1\*1501–11, DRB1\*1601–08), and DR7 (DRB1\*0701–07). The main structural feature for this supertype is the presence of Trp<sup>9β</sup>. Residues  $\beta$ 9,  $\beta$ 30,  $\beta$ 38,  $\beta$ 57, and  $\beta$ 61 are part of pocket 9, which accommodates peptide residue 9. Trp<sup>9β</sup> makes the pocket shallow and allows the binding of small nonpolar residues: Leu, Ala, Gly, and Pro. The binding motif for HLA-DRB1\*0101 favors Leu or Ala at position 9 (Table II) and that for HLA-DRB1\*1501, favors Gly, Ser, Pro, or Thr (Table II). Trp<sup>9β</sup> is the fingerprint residue for the DR1 supertype.

The second cluster, called DR3 supertype, comprises DR3 (DRB1\*0301–25), DR52 (DRB3\*0101–10, DRB3\*0201–18, DRB3\*0301–03), DRB1\*0422, and DRB1\*1107. The common features for this cluster are Gln<sup>70β</sup>, Lys<sup>71β</sup>, and Gln/Arg<sup>74β</sup>. Residues  $\beta$ 70,  $\beta$ 71, and  $\beta$ 74 are the polymorphic residues forming pocket 4. This pocket binds one of the main anchor residues for MHC class II molecules (9, 10, 19). Amino acid side chain charges are important for interaction between TCRs and peptide-DR complexes (48). Due to Lys<sup>71β</sup> the total charge in pocket 4, for this supertype, is positive, which corresponds well to the preference for



**FIGURE 1.** Hierarchical clustering on CoMSIA fields for HLA-DRB molecules. The detailed content of each cluster is shown in Table II. At the second level of the hierarchy, the clusters are associated with polymorphism around pocket 9: the *leftmost* (DR1) is composed of structures with Trp<sup>9β</sup>, the *middle clusters* (DR3, DR5, and DR4) have Glu<sup>9β</sup>, and the *rightmost* (DR9) has Lys/Gln<sup>9β</sup>. At the fourth level, the clustering is associated with polymorphism in pocket 4 ( $\beta$ 70,  $\beta$ 71, and  $\beta$ 74).

Table II. DR supertypes and fingerprints<sup>a</sup>

Supertype	Hierarchical Clustering	Nonhierarchical Clustering	Common Alleles	Fingerprint	Known Supertypes <sup>b</sup>	Known Motifs				
						p1	p4	p6	p7	p9
DR1	DR1 (DRB1*0101–11)	DR1 (DRB1*0101–11)	11	Trp <sup>9β</sup>		DRB1*0101 (Refs. 50–52)				
	DR2 (DRB1*1501–11, DRB1*1601–08)	DR2 (DRB1*1501–11, DRB1*1601–08)	13			YFW	LA	AG	-	LA
DR3	DR7 (DRB1*0701–07)		7			DRB1*1501 (Refs. 53 and 54)				
	DR3 (DRB1*0301–25)			Glu <sup>9β</sup>	DR RSP	LVI	FYI	-	IL	GSP
	DR52 (DRB3*0101–10, DRB3*0201–18, DRB3*0301–03)	DR52 (DRB3*0101–10, DRB3*0201–18, DRB3*0301–03)	10 18	Gln <sup>70β</sup> Gln/Arg <sup>74β</sup>	“R”	DRB1*0301 (Refs. 55–57)				
			3			LIF	D	KR	-	YLF
DR4	DRB1*0422 DRB1*1107	DRB1*1333 DRB1*1447								
	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>9β</sup> Gln/Arg <sup>70β</sup> Glu/Ala <sup>74</sup>	DR RSP	DRB1*0401 (Refs. 58–60)				
					“A”	FY	no RK	NS	pol <sup>c</sup> chg <sup>d</sup> ali	pol ali <sup>c</sup> K
	DR5 (DRB1*1113, 17, 26, 34, 42)	DR5 (DRB1*1107, 13, 17)	2			W				
DR6	DR6 (DRB1*1309, DRB1*1401–48 without the alleles from DR5 supertype)	DR6 (DRB1*1401–48 without the alleles from DR5 supertype)	31			DRB1*0404 (Ref. 61)				
						VIL	no RK	NT	pol	pol
									chg ali	ali K
DR5		DRB3*0215				DRB1*0405 (Refs. 61–63)				
	DR10 (DRB1*1001)	DR10 (DRB1*1001)	1			FY	VIL	NS	pol	DE
	DR53 (DRB4*0101–06)	DR53 (DRB4*0101–06)	5						chg ali	
	DR3 (DRB1*0301–25)	DR3 (DRB1*0301–25)								
DR5	DR4 (DRB1*0402, 12, 15, 25, 36, 37, 47)	DR4 (DRB1*0402, 15, 25, 36, 47)	5	Glu <sup>9β</sup>	DR RSP	DRB1*0402 (Ref. 61)				
				Asp <sup>70β</sup>	“D”	VIL	no DE	NQ	RK	pol ali
	DR5 (DRB1*1101–47, DRB1*1201–09)	DR5 (DRB1*1101–47, DRB1*1201–09)	42							H
	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)	61			DRB1*1101 (Refs. 59 and 64)				
DR8			6			WYF	LVI	RK	-	-
						DRB1*1201 (Ref. 51)				
	DR8 (DRB1*0801–25)	DR8 (DRB1*0801–25)	25			IL	LMN	VYF	-	YFM
DR9	DR9 (DRB1*0901, 02)	DR9 (DRB1*0901, 02)	2	Lys/Gln <sup>9β</sup>		DRB5*0101 (Refs. 53 and 54)				
	DR51 (DRB5*0101–12, DRB5*0202–05)	DR51 (DRB5*0101–12, DRB5*0202–05)				FY	QV	-	-	RK
Sum	347	347	285 (82%)							

<sup>a</sup> The content of the clusters derived by hierarchical and nonhierarchical clustering is compared and the common alleles are presented as percent of all DR alleles. The supertype fingerprints were defined on the basis of common amino acids from the binding site. The classification defined in the present study was compared with the categorization of Ou et al. (20). Binding motifs for each supertype are listed where these are known.

<sup>b</sup> Ref. 20.

<sup>c</sup> Polar.

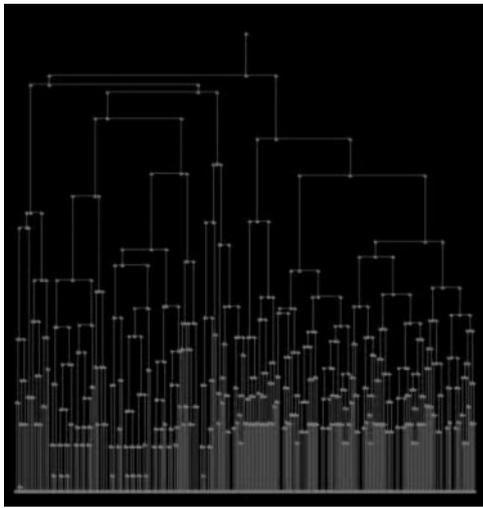
<sup>d</sup> Charged.

<sup>e</sup> Aliphatic.

negatively charged Asp and Glu at position 4 in the DRB1\*0301 binding motif (55–57). DR3 supertype fingerprint residues are Gln<sup>70β</sup>, Lys<sup>71β</sup>, and Gln/Arg<sup>74β</sup>.

The third cluster consists of DR4 (DRB1\*0402, 12, 15, 25, 36, 37, 47), DR5 (DRB1\*1101–47, DRB1\*1201–09), DR6 (DRB1\*1301–62, DRB1\*1403, 16, 22, 25, 27, 40), and DR8 (DRB1\*0801–25). As most DR4 alleles go into another cluster, this supertype was named after the next numerically lowest sero-

type DR5. The common feature for all MHC molecules belonging to this supertype is Asp<sup>70β</sup>. To distinguish certain DR51 alleles with Asp<sup>70β</sup> which belong to another supertype, Glu<sup>9β</sup> is added to the DR5 supertype fingerprint. It is probable that Asp<sup>70β</sup> is negatively charged in the pocket. When Glu<sup>71β</sup> is next to Asp<sup>70β</sup>, as in DRB1\*0402, negatively charged residues—Asp and Glu—at peptide position 4 are detrimental to binding (61). If Arg or Lys are available at position β71, the pocket becomes both positively and



**FIGURE 2.** Hierarchical clustering on CoMSIA fields for HLA-DQ molecules. The detailed content of each cluster is shown in Table III. At the first level of the hierarchy, the clusters are associated with polymorphism around pocket 1: the *left clusters* (DQ2, DQ3, and DQA\*03) contain Val<sup>86β</sup> and the *right clusters* (DQ1) contain Glu<sup>86β</sup>. At the third level, the clustering is associated with polymorphism in pocket 4: cluster DQ2 has Lys<sup>71β</sup>, while cluster DQ3 has Thr/Asp<sup>71β</sup>. Cluster DQA\*03 contains alleles with the additional amino acid Arg<sup>53α</sup>, in contrast to other DQ α-chains.

negatively charged and can accommodate neutral amino acids like Leu, Val, and Ile, as in the DRB1\*1101 binding motif (59, 64).

The fourth cluster includes DR4 (DRB1\*0401, 03–48 without alleles from the DR5 supertype), DR5 (DRB1\*1113, 17, 26, 34, 42), DR6 (DRB1\*1309, DRB1\*1401–48), DR10 (DRB1\*1001), and DR53 (DRB4\*0101–06). This cluster was called the DR4 supertype. The DR4 supertype fingerprint is Gln/Arg<sup>70β</sup>, Arg/Lys<sup>71β</sup>, and Glu/Ala<sup>74β</sup>. Arg or Lys at β71 position makes pocket 4 positively charged and residues like Arg and Lys at peptide position 4 become detrimental for MHC binding, as is evident from binding motifs for DRB1\*0401 and DRB1\*0404 (58–61).

The last cluster, which we call the DR9 supertype, is composed of DR9 (DRB1\*0901 and DRB1\*0902) and DR51 (DRB5\*0101–12, DRB5\*0202–05), and has the fingerprint Lys/Gln<sup>9β</sup>. Lys/Gln<sup>9β</sup> coexists with Asp<sup>11β</sup>. Both residues take part in the formation of binding pocket 9. Asp<sup>11β</sup> makes the pocket negatively

charged and peptides with Arg and Lys at position 9 are preferred as is evident from the DRB5\*0101 binding motif (53, 54).

**Nonhierarchical clustering.** The contents of clusters derived by nonhierarchical clustering are given in Table II. The major discrepancies concern alleles DR3 (DRB1\*0301–25), DR7 (DRB1\*0701–07), and DR51 (DRB5\*0101–12, DRB5\*0202–05). The first was classified as DR3 by hierarchical clustering and as DR4 by nonhierarchical. The second was considered part of the DR1 supertype by hierarchical clustering and as part of DR9 by nonhierarchical. The third belongs to the DR9 supertype according to hierarchical clustering and to DR5 according to nonhierarchical. Despite these minor differences, 82% (285 of 347) of the DR alleles were classified in the same supertype by both clustering methods.

#### HLA-DQ supertypes

**Hierarchical clustering.** The hierarchical clustering of HLA-DQ molecules is shown in Fig. 2 and the cluster contents are listed in Table III. Two clusters exist at the first level. The structural differences here concern the polymorphic region 84–87 of the β-chain. Alleles from the left cluster contain Gln<sup>84β</sup>, Leu<sup>85β</sup>, Glu<sup>86β</sup>, and Leu<sup>87β</sup> (DQB1\*02, 03, and 04), whereas the right cluster members have Glu<sup>84β</sup>, Val<sup>85β</sup>, Ala<sup>86β</sup>, or Gly<sup>86β</sup>, and Tyr<sup>87β</sup> or Phe<sup>87β</sup> (DQB1\*05 and 06), respectively. Position β86 is part of pocket 1, together with the amino acids at positions 24, 31, and 52 from the α-chain. X-ray data for DQ8 (DQA1\*0301/DQB1\*0302) indicates that pocket 1 is lined by two positively charged side chains—His<sup>24α</sup> and Arg<sup>52α</sup>—inside the entrance and two negatively charged residues—Glu<sup>31α</sup> and Glu<sup>86β</sup>—deeper in the pocket (43). Together, they form a hydrogen-bonding network. The replacement of Glu<sup>86β</sup> with Ala or Gly destroys this network and the side chain of Arg<sup>52α</sup> might reorient closer to the pocket entrance. This might explain the pronounced intolerance of positively charged amino acids at position 1 according to the DQA1\*0102/DQB1\*0602 binding motif (65). The cluster with a fingerprint Ala/Gly<sup>86β</sup> was called the DQ1 supertype. It includes the serotypes: DQ1 (DQB1\*0611, 12), DQ5 (DQB1\*0501, 02, 03), and DQ6 (DQB1\*0601–05, 09) and the rest of molecules containing DQB1\*05 or 06.

At the third level of the DQ dendrogram (Fig. 2), the left cluster is divided into three smaller subclusters. The first cluster contains alleles beginning DQB1\*02, the second contains alleles starting DQB1\*03 and 04, and the third comprises alleles commencing

Table III. DQ supertypes and fingerprints<sup>a</sup>

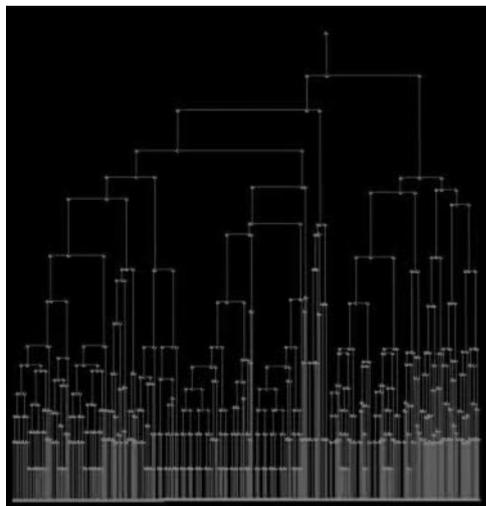
Supertype	Hierarchical Clustering	Nonhierarchical Clustering	Common Alleles	Fingerprint	Known Motifs					
					p1	p3	p4	p6	p7	p9
DQ1	DQB1*0501–03 DQB1*0601–21	DQB1*0501–03 DQB1*0601–21	45 300	Val <sup>86β</sup>	DQA1*0102/DQB1*0602 (Ref. 65) no+ <sup>b</sup> no+ no+ LIV - AST noPG noP no- <sup>c</sup>					
DQ2	DQB1*0201–03	DQB1*0201–03	45	Glu <sup>86β</sup> Lys <sup>71β</sup>	DQA1*0501/DQB1*0201 (Refs. 66–70) FWYLI - DE PDE ED FYW					
DQ3	DQB1*0301–13 DQB1*0401, 02	DQB1*0301–13 DQB1*0401, 02	195 30	Glu <sup>86β</sup> Thr/Asp <sup>71β</sup>	DQA1*0301/DQB1*0301 (Refs. 71–73) TSE AGS noDE - RVD DE no+ WFL					
DQA1*03 <sup>d</sup>	DQA1*0301–03			Arg <sup>53α</sup>	DQA1*0301/DQB1*0301 (Refs. 71–73) TSE AGS noDE - RVD DE no+ WFL					
Sum	738	738	615 (83%)							

<sup>a</sup> The content of the clusters derived by hierarchical and nonhierarchical clustering is compared and the common alleles are presented as percent of all DQ alleles. The supertype fingerprints were defined on the basis of common amino acids from the binding site. Binding motifs for each supertype are listed where these are known.

<sup>b</sup> No basic.

<sup>c</sup> No acidic.

<sup>d</sup> Considered as outliers.



**FIGURE 3.** Hierarchical clustering on CoMSIA fields for HLA-DP molecules. The detailed content of each cluster is shown in Table IV. At the first level of the hierarchy, the clusters are associated with polymorphism in pocket 1: the *left clusters* (DPw1 and DPw6) contain Asp<sup>84β</sup> and the *right clusters* (DPw2 and DPw4) contain Gly/Val<sup>84β</sup>. At the second level, the clustering is connected with polymorphism in pocket 4: clusters DPw1 and DPw4 contain Lys<sup>69β</sup>, while clusters DPw2 and DPw6 have Glu<sup>69β</sup>.

DQA1\*03 (Table III). DQA1\*03 chains contain an additional Arg after Arg<sup>52α</sup>, which is the main difference compared with other DQ α-chains. As these alleles were not classified as a separate cluster by the nonhierarchical clustering, we decided to define them as

outliers rather than as a separate supertype. The alleles from the other two clusters differ in several positions at the binding site and correspond well to known motifs.

The cluster containing the serotype DQ2 (DQB1\*0201, 02, and 03) was labeled the DQ2 supertype. Its fingerprint is Glu<sup>86β</sup> and Lys<sup>71β</sup>. The amino acid at position 71 takes part in the formation of pockets 4 and 7. DQ2 is the only serotype with Lys at position 71. This gave the pockets a strongly basic character, a factor accounting for the almost absolute requirement for acidic residues, principally at peptide position 7, but also at position 4 (67). A computer model of DQ2 with HLA class I α 46–60 peptide in the binding cleft indicates that Lys<sup>71β</sup> makes a direct salt-bridged hydrogen bond to the glutamic acid or aspartic acid at position 7 of the peptide (67).

The cluster, named DQ3, includes several serotypes: DQ3 (DQB1\*0306), DQ4 (DQB1\*0401, 02), DQ7 (DQB1\*0301, 04), DQ8 (DQB1\*0302, 05), DQ9 (DQB1\*0303), and the rest of alleles carrying DQB1\*03 or 04. Its fingerprint is Glu<sup>86β</sup> and Thr<sup>71β</sup> (DQB1\*03) or Asp<sup>71β</sup> (DQB1\*04). Peptide positions 4 and 7 must be aliphatic amino acids (Table III), although Godkin et al. (72) found that Arg is also tolerated.

**Nonhierarchical clustering.** The nonhierarchical clustering did not classify DQA1\*03 alleles in a separate cluster. For the rest of the DQ molecules, we found an 83% (615 of 738) agreement with the hierarchical classification (Table III).

#### HLA-DP supertypes

**Hierarchical clustering.** The DP dendrogram is plotted in Fig. 3 and the cluster members are listed in Table IV. Like the DR locus,

Table IV. DP supertypes and fingerprints<sup>a</sup>

Supertype	Hierarchical Clustering	Nonhierarchical Clustering	Common	Fingerprint	Known Motifs			
					p1	p4	p6	p9
DPw1	DPw1 (DPB1*0101)	DPw1 (DPB1*0101)	12	Asp <sup>84β</sup>			N/A	
	DPw3 (DPB1*0301) DPw5 (DPB1*0501) DPB1*14, 20, 25, 26, 27, 31, 35, 36, 38, 45, 50, 52, 56, 57, 63, 65, 67, 68, 70, 76, 78, 79, 84, 85, 87, 89, 90, 91, 92, 97, 98	DPw3 (DPB1*0301) DPB1*11, 14, 25, 26, 31, 35, 45, 50, 52, 56, 57, 65, 67, 68, 69, 70, 76, 78, 79, 84, 90, 92	12 240	Lys <sup>69β</sup>				
DPw2	DPw2 (DPB1*0201 and 0202)	DPw2 (DPB1*0201 and 0202)	24	Gly/Val <sup>84β</sup>	DPA1*0103/DPB1*0201 (Ref. 74)			
	DBP1*32, 33, 41, 46, 47, 48, 71, 81, 86, 95	DBP1*32, 33, 41, 46, 47, 48, 71, 81, 86, 95	120	Glu <sup>69β</sup>	YLF DSQ YFW LVI			
DPw4	DPw4 (DPB1*0401 and 0402)	DPw4 (DPB1*0401 and 0402)	24	Gly/Val <sup>84β</sup>	DPA1*0103/DPB1*0401 (Ref. 22)			
	DPB1*15, 18, 23, 24, 28, 34, 39, 40, 49, 51, 53, 59, 60, 62, 66, 72, 73, 74, 75, 77, 80, 83, 94, 96, 99	DPB1*15, 18, 23, 24, 28, 34, 39, 40, 49, 51, 53, 59, 60, 62, 66, 72, 73, 74, 75, 77, 80, 82, 83, 94, 96, 99	312	Lys <sup>69β</sup>	FYL FYL FYWL FYLVM DPA1*0103/DPB1*0402 (Ref. 22)			
DPw6	DPw6 (DPB1*0601)	DPw5 (DPB1*0501)		Asp <sup>84β</sup>	FYL FYL FYWL FLVM			
	DPB1*08, 09, 10, 11, 13, 16, 17, 19, 21, 22, 29, 30, 37, 44, 54, 55, 58, 69, 88, 93	DPw6 (DPB1*0601) DPB1*08, 09, 10, 13, 16, 17, 19, 20, 21, 22, 27, 29, 30, 36, 37, 38, 44, 54, 55, 58, 63, 85, 87, 88, 89, 91, 93, 97, 98	12 216	Glu <sup>69β</sup>	N/A			
Sum	1140	1140	972 (85%)					

<sup>a</sup> The content of the clusters derived by hierarchical and nonhierarchical clustering is compared and the common alleles are presented as percent of all DP alleles. The supertype fingerprints were defined on the basis of common amino acids from the binding site. Binding motifs for each supertype are listed where these are known.

DP clustering depends on the polymorphism of the  $\beta$ -chain only. Two large clusters are apparent at the first level of the hierarchy. The left cluster comprises alleles with Asp<sup>84 $\beta$</sup> , Glu<sup>85 $\beta$</sup> , Ala<sup>86 $\beta$</sup> , and Val<sup>87 $\beta$</sup> . The right one includes alleles with Gly<sup>84 $\beta$</sup>  or Val<sup>84 $\beta$</sup> , Gly<sup>85 $\beta$</sup> , Pro<sup>86 $\beta$</sup> , and Met<sup>87 $\beta$</sup> . All four residues play a major role in forming the contact area between  $\alpha$ - and  $\beta$ -chains and only position  $\beta$ 84 and partly  $\beta$ 85 are involved in forming the surface of pocket 1 (46). Further division appears at the second level of the hierarchy, and it is connected with position  $\beta$ 69, which is involved in forming pockets 4 and 6 (47). Clustering at the third level is contingent upon positions  $\beta$ 55 and  $\beta$ 56. Position  $\beta$ 55 is part of pocket 9. At this level of the hierarchy, DPB1\*0401 and 0402 split into separate clusters. As recent experimental data indicate that they probably belong to the same supertype (22), the second level of the dendrogram was chosen for definition of the DP superotypes.

The first cluster includes DPw1 (DPB1\*0101), DPw3 (DPB1\*0301), DPw5 (DPB1\*0501), DPB1\*14, 20, 25, 26, 27, 31, 35, 36, 38, 45, 50, 52, 56, 57, 63, 65, 67, 68, 70, 76, 78, 79, 84, 85, 87, 89, 90, 91, 92, 97, and 98. It was called DPw1 and its fingerprint is Asp<sup>84 $\beta$</sup>  and Lys<sup>69 $\beta$</sup> . The MHC molecules forming this supertype have a negatively charged pocket 1 and a positively charged pocket 4. To the best of our knowledge, no binding motif is currently available for any of the alleles in this supertype. One may imagine that peptides with complementary charges—positively charged amino acids at position 1 and negatively charged residues at position 4—should, in general, bind well to members of this supertype.

The second cluster involves DPw6 (DPB1\*0601), DPB1\*08, 09, 10, 11, 13, 16, 17, 19, 21, 22, 29, 30, 37, 44, 54, 55, 58, 69, 88, and 93. This supertype was called DPw6. All alleles have Asp<sup>84 $\beta$</sup>  and Glu<sup>69 $\beta$</sup> , except DPB1\*1101 and DPB1\*6901 which have Arg<sup>69 $\beta$</sup> . Unfortunately, no binding motif was available for any alleles from this supertype. Again, peptides with complementary charges—positively charged amino acids at positions 1 and 4—may be supposed to bind well to this supertype.

The third cluster, called the DPw2 supertype, consisted of DPw2 (DPB1\*0201 and 0202), DBP1\*32, 33, 41, 46, 47, 48, 71, 81, 86, and 95. Its fingerprint is Gly<sup>84 $\beta$</sup>  or Val<sup>84 $\beta$</sup>  and Glu<sup>69 $\beta$</sup> . Alleles of the supertype have a deep, nonpolar pocket 1 capable of accepting bulky amino acids, as is evident from the available binding motif (Table IV) (74). Due to a negatively charged Glu<sup>69 $\beta$</sup> , pocket 4 of HLA-DP2 showed high affinity for peptides with positively charged residues at this position (47).

DPw4 (DPB1\*0401 and 0402), DPB1\*15, 18, 23, 24, 28, 34, 39, 40, 49, 51, 53, 59, 60, 62, 66, 72, 73, 74, 75, 77, 80, 83, 94, 96, and 99 form the last cluster, which we call the DPw4 supertype. Its fingerprint is Gly<sup>84 $\beta$</sup>  or Val<sup>84 $\beta$</sup>  and Lys<sup>69 $\beta$</sup> ; only DRB1\*1501 and 74 have Arg<sup>69 $\beta$</sup> . Again, consistent with our results, known motifs for DPB1\*0401 and 0402 indicate preferences for bulky aromatic amino acids at positions 1 and 4 (Table IV) (22).

**Nonhierarchical clustering.** The contents of clusters derived by nonhierarchical clustering are listed in Table IV. Among several minor differences, DPB1\*11 and 69 (Arg<sup>69 $\beta$</sup> ) are clustered into the DPw1 supertype (Lys<sup>69 $\beta$</sup> ), as opposed to DPw6 (Glu<sup>69 $\beta$</sup> ), which was identified by hierarchical clustering. 85% (972 of 1140) of the DP molecules were clustered into the same supertype by both methods.

## Discussion

The extreme polymorphism, apparent within higher vertebrates, confounds the study of epitope binding by MHCs, particularly from an experimental perspective: no existing technique is fast or reliable enough to determine peptide specificities on an appropriate scale. MHC polymorphism greatly complicates epitope-based vac-

cine development, particularly in regard to population coverage. One initial approach to the problem has been to characterize the binding specificity of five to nine of the most common HLA alleles and to develop a mixture of several epitopes to cover the general population (75). Latter, the logical framework of this approach has been inverted. Instead of developing a single epitope for each of the common HLA alleles, attempts were made to identify epitopes capable of binding multiple HLA types (12–14, 16). The grouping of alleles into superotypes, based on common structural and functional features, is useful in addressing such attempts. Sette et al. (16) found that by focusing only on the HLA class I A1, A2, A3, A24, and B7 superotypes, 100% population coverage is achieved (76). The strategy of epitope selection based on HLA superotypes has been validated in different disease settings worldwide (77–82). However, while HLA class I superotypes have been widely explored, class II superotypes are still in the relatively early stages of investigation.

In this study, we have applied a combined bioinformatics approach, using both protein sequence and structural data, to 2225 HLA class II molecules, to detect similarities in their peptide binding sites and to define supertype fingerprints. Two chemometric techniques were used: hierarchical clustering on 3D CoMSIA fields and nonhierarchical *k*-means clustering on sequence-based *z*-descriptors. The former method classifies the molecules on the basis of binding site similarities, in terms of steric bulk, electrostatic potential, local hydrophobicity, and hydrogen-bond-donor and acceptor abilities. The latter method uses five principal properties (*z*-scales) of the amino acids and classifies the proteins according to their sequence-based binding site similarities.

An average consensus of 84% was achieved, i.e., 1872 of 2225 class II molecules were classified in the same supertype by both techniques. Twelve class II superotypes were defined: five DRs, three DQs, and four DPs. The DR superotypes are DR1 (fingerprint Trp<sup>9 $\beta$</sup> ), DR3 (Glu<sup>9 $\beta$</sup> , Gln<sup>70 $\beta$</sup> , and Gln/Arg<sup>74 $\beta$</sup> ), DR4 (Glu<sup>9 $\beta$</sup> , Gln/Arg<sup>70 $\beta$</sup> , and Glu/Ala<sup>74 $\beta$</sup> ), DR5 (Glu<sup>9 $\beta$</sup> , Asp<sup>70 $\beta$</sup> ), and DR9 (Lys/Gln<sup>9 $\beta$</sup> ). The DQ superotypes are DQ1 (Ala/Gly<sup>86 $\beta$</sup> ), DQ2 (Glu<sup>86 $\beta$</sup> , Lys<sup>71 $\beta$</sup> ), and DQ3 (Glu<sup>86 $\beta$</sup> , Thr/Asp<sup>71 $\beta$</sup> ) and the DP superotypes are DPw1 (Asp<sup>84 $\beta$</sup>  and Lys<sup>69 $\beta$</sup> ), DPw2 (Gly/Val<sup>84 $\beta$</sup>  and Glu<sup>69 $\beta$</sup> ), DPw4 (Gly/Val<sup>84 $\beta$</sup>  and Lys<sup>69 $\beta$</sup> ), and DPw6 (Asp<sup>84 $\beta$</sup>  and Glu<sup>69 $\beta$</sup> ). Apart from the good agreement between known binding motifs and our classification, several new superotypes have been defined and thematic binding motifs have been outlined for them. In the following, we discuss the congruence of our systematic structural analysis of binding with extant data on the biology of class II human MHCs, rather than making unsupported speculations.

HLA-DR molecules account for >90% of the HLA class II isotypes expressed on APCs (83). Although the HLA-DRA locus is monomorphic, >300 alleles have been described for the HLA-DRB1 locus (3). X-ray data indicate that 12 hydrogen bonds exist between conserved DR atoms and main-chain atoms of the bound peptide (9). As they do not involve the side chains of the peptide, these hydrogen bonds are likely to play a common role in peptide binding to HLA-DR.

Five binding pockets, pockets 1, 4, 6, 7, and 9 (named after the corresponding positions on the binding peptide), were found to be common for most DR proteins (9, 10). Specificity of pocket 1 is modulated by a Gly/Val<sup>86 $\beta$</sup>  dimorphism. DR proteins with Gly<sup>86 $\beta$</sup>  show strong preferences for large hydrophobic side chains (Trp, Tyr, Phe) at peptide position 1, whereas Val<sup>86 $\beta$</sup>  restricts the pocket size and alters the preferences to small hydrophobic side chains (Val and Ala) at this position. The main difference in the preferences concern bulky aromatic residues—Trp, Tyr, and Phe—which are not accepted at pocket 1 when it contains Val<sup>86 $\beta$</sup> . However, the medium sized hydrophobic amino acids Leu and Ile are

well accepted in all DR molecules and peptide position 1 could not be considered as an anchor able to distinguish between different DR alleles.

Pocket 4 is formed by polymorphic amino acids at positions  $\beta 13$ ,  $\beta 26$ ,  $\beta 28$ ,  $\beta 70$ ,  $\beta 71$ ,  $\beta 74$ , and  $\beta 78$ . Residues at positions  $\beta 70$ ,  $\beta 71$ , and  $\beta 74$  play a significant role both in protein binding and T cell recognition (Refs. 4, 9, 10, 19). Residues  $\beta 71$  and  $\beta 74$  also take part in the formation of pockets 6 and 7 (9, 84). Ou et al. (19) made a functional categorization of DR alleles on the basis of pocket 4 polymorphism, associating each group with certain autoimmune diseases. Good agreement was found between this categorization and our classification. The DR3 supertype corresponds to the functional DR restrictive supertype pattern (RSP) "R". It contains the pattern Gln<sup>70 $\beta$</sup> , Lys<sup>71 $\beta$</sup> , and Arg/Gln<sup>74 $\beta$</sup>  and the overall charge within pocket 4 is positive, which requires negatively charged amino acids Asp and Glu at position 4 of the binding peptide (Table II, motif DRB1\*0301). This supertype is associated with two autoimmune diseases: systematic lupus erythematosus and Hashimoto's thyroiditis (19, 83). The DR4 supertype corresponds to DR RSP "A" (19). Its pattern, Gln/Arg<sup>70 $\beta$</sup> , Arg/Lys<sup>71 $\beta$</sup>  and Glu/Ala<sup>74 $\beta$</sup> , is close to that of DR RSP "R", differing only in position  $\beta 74$ . When Ala appears at  $\beta 74$ , pocket 4 increases in size and can accommodate larger amino acids such as Phe, Trp, and Ile (Table II, motifs DRB1\*0401, 04, 05). Unfortunately, no binding motif is available for any allele bearing Glu<sup>74 $\beta$</sup> , but one could suppose that small polar residues, like Ser and Thr, will be accepted. This supertype is associated with a susceptibility to rheumatoid arthritis (19, 83). The DR5 supertype corresponds to DR RSP "D" with pattern Asp<sup>70 $\beta$</sup> , Glu/Arg<sup>71 $\beta$</sup> , and Leu/Ala<sup>74 $\beta$</sup>  (19). The main feature here is the negatively charged Asp at position  $\beta 70$ , which restricts the accommodation of negatively charged amino acids at peptide position 4 (Table II, motif DRB1\*0402). Juvenile rheumatoid arthritis (JRA), pemphigus vulgaris, and allergic bronchopulmonary aspergillosis are autoimmune diseases associated with this supertype (19).

Residues  $\beta 9$ ,  $\beta 30$ ,  $\beta 37$ ,  $\beta 38$ ,  $\beta 57$ , and  $\beta 61$  are involved in the formation of pocket 9 (9, 84). The polymorphism at  $\beta 9$  determines the pocket size and hence binding motif preferences at this position. The clustering at the first and second level of the DR dendrogram (Fig. 1) is associated with the  $\beta 9$  polymorphism. Trp<sup>9 $\beta$</sup>  is the fingerprint for the DR1 supertype, Lys/Gln<sup>9 $\beta$</sup>  for DR9, and Glu<sup>9 $\beta$</sup>  for DR3, DR4, and DR5. Small amino acids (Ala, Val, Gly, Ser, Thr, Pro) are accepted in pocket 9 of the DR1 supertype (Table II, motifs DRB1\*0101, 1501). Glu<sup>9 $\beta$</sup> , in combination with Asp<sup>57 $\beta$</sup> , makes this pocket negatively charged, facilitating the accommodation of positively charged amino acids, such as Lys (motifs DRB1\*0401, 0404) and His (motif DRB1\*0402). In most MHC class II alleles, Asp<sup>57 $\beta$</sup>  makes a salt-bridged hydrogen bond with Arg<sup>76 $\alpha$</sup> , allowing the pocket to also accommodate aliphatic and polar amino acids (43). In cases where Asp<sup>57 $\beta$</sup>  is replaced by Ser (DRB1\*0405) or Ala (DQ8), the hydrogen bonding network is destroyed and Arg<sup>76 $\alpha$</sup>  can strongly attract negatively charged amino acids (Asp, Glu) available at position 9 of the binding peptide (motif DRB1\*0405). Lys/Gln<sup>9 $\beta$</sup>  always coexists with Asp<sup>11 $\beta$</sup>  and Asp/Gly<sup>30 $\beta$</sup> . Vogt et al. (53) suggested that the positively charged anchor residue R and K (motif DRB5\*0101) may form a salt bridge with Asp at position 11 and/or position 30 of the DRB5\*0101 molecule.

During the last 10 years, interest in HLA-DQ proteins has increased because certain DQ alleles are associated with susceptibility to type 1 diabetes and celiac disease (85, 86). The x-ray structure of DQ8 (DQA1\*0301/DQB1\*0302) complexed with an immunodominant peptide from insulin was solved (43). Several DQ binding motifs have been defined (65–73). The initial hypoth-

esis was that class II molecules with non-Asp<sup>57 $\beta$</sup>  (i.e., DQ2, DQ8, I-A<sup>g7</sup>) preferentially bind peptides with negatively charged anchor residue at peptide position 9, such as peptides from insulin  $\beta$ -chain, gliadin, glutenin, and present them to islet-infiltrating T cells or mucosal T cells (87–90). As was discussed above, the molecular explanation for this phenomenon is that Asp<sup>57 $\beta$</sup>  forms a salt bridge with Arg<sup>76 $\alpha$</sup> , whereas in non-Asp<sup>57 $\beta$</sup>  molecules Arg<sup>76 $\alpha$</sup>  is free to interact with the negatively charged peptide anchor at position 9 (43). However, recent data does not support this hypothesis: not all non-Asp<sup>57 $\beta$</sup>  class II molecules have a preference for negatively charged anchor residues at peptide position 9 and should thus be associated with susceptibility to type 1 diabetes and celiac disease (69). For example, in the Japanese population the class II molecule DQA1\*0301/DQB1\*0401, which has the same  $\alpha$ -chain as DQ8, but has a  $\beta$ -chain containing an Asp<sup>57 $\beta$</sup> , is associated with increased susceptibility to type 1 diabetes (43). Other exceptions include molecules DQA1\*0201/DQB1\*0201 and DQ9 (DQA1\*0301/DQB1\*0303). The former does not contain Asp<sup>57 $\beta$</sup>  but is neutral-protective to type 1 diabetes (43), while the latter does contain Asp<sup>57 $\beta$</sup>  yet is associated with susceptibility to celiac disease (73).

The DQ classification defined in the present study is based on two important amino acids from the  $\beta$ -chain: positions  $\beta 71$  and  $\beta 86$ . Residue  $\beta 71$  participates in the formation of pockets 4 and 7, while residue  $\beta 86$  is part of pocket 1. Pocket 1 is a deep, very polar pocket in HLA-DQ molecules, formed by two positively and two negatively charged amino acids, which form a hydrogen bonding network. Replacement of Glu<sup>86 $\beta$</sup>  with Ala or Gly will destroy this network and leave Arg<sup>52 $\alpha$</sup>  free to contact the side chain of peptide position 1 (43). This is consistent with strong intolerance for positively charged amino acids at position 1 for the DQ1 supertype (Table III, motif for DQA1\*0102/DQB1\*0602). Ala/Gly<sup>86 $\beta$</sup>  coexists with Phe/Tyr<sup>87 $\beta$</sup> . The last residue is also part of pocket 1 and the Phe/Tyr→Leu replacement increases the pocket size. Large hydrophobic amino acids (Trp, Tyr, Phe) at position 1 are well accepted by alleles bearing Glu<sup>86 $\beta$</sup> /Leu<sup>87 $\beta$</sup>  and belong to superotypes DQ2 and DQ3 (Table III, motifs DQA1\*0501/DQB1\*0201, DQA1\*0301/DQB1\*0301), whereas alleles with Ala/Gly<sup>86 $\beta$</sup>  and Phe/Tyr<sup>87 $\beta$</sup>  (supertype DQ1) prefer medium sized hydrophobic or polar amino acids (Leu, Ile, Thr, Ser) (Table III, motif DQA1\*0102/DQB1\*0602).

DQ pocket 4 is significantly deeper than the corresponding pocket 4 in DR molecules (43). Lys<sup>71 $\beta$</sup>  accounts for the strong basic character of this pocket in DQ2 supertype molecules. Lys<sup>71 $\beta$</sup>  makes a salt bridge with acidic residues at position 7 of the binding peptide (67). Asp and Glu are preferred amino acids at positions 4 and 7 of the DQ2 binding motif (Table III). In the DQ3 supertype, Lys<sup>71 $\beta$</sup>  is replaced by Thr<sup>71 $\beta$</sup> , which coexists with Glu<sup>74 $\beta$</sup> . The last amino acid makes the pocket negatively charged and acidic residues (Asp and Glu) are not observed at this peptide position (motif DQA1\*0301/DQB1\*0301).

DQ alleles beginning DQA1\*03 differ from other DQ alleles in having an additional Arg residue after Arg<sup>52 $\alpha$</sup> . This affects the architecture of pocket 1 (21) and determines a preference for small to medium sized amino acids at peptide position 1, including aliphatic or negatively charged side chains (Table III, motif DQA1\*0301/DQB1\*0301). DQA1\*03 alleles were classified as outliers and not as a separate supertype.

Apart from type 1 diabetes and celiac disease, HLA-DQ alleles are strongly associated with either protection or susceptibility to other autoimmune diseases. Susceptibility to multiple sclerosis has been suggested for individuals with DQA1\*0102/DQB1\*0602 (91, 92); pemphigus vulgaris is associated with DQB1\*0503 (93); rheumatoid arthritis with DQ3 (DQA1\*03/DQB1\*03 and DQA1\*03/DQB1\*04)

and DQ5 (DQA1\*0101/DQB1\*0501) (94); systemic sclerosis with DQA1\*0501 (95); and protection against type 1 diabetes with DQA1\*0102/DQB1\*0602 (96). Although these associations concern single HLA-DQ alleles, one could draw a more general conclusion, connecting susceptibility to multiple sclerosis, pemphigus vulgaris, or rheumatoid arthritis as well as protection against type 1 diabetes with alleles from DQ1 supertype.

In contrast to HLA-DR and DQ, HLA-DP molecules have not been studied extensively, as they have been viewed as less important in immune responses than DRs and DQs. Moreover, currently, no x-ray data exist for any peptide/HLA-DP complex. However, it is now known that HLA-DP proteins contribute to the risk of graft-vs-host disease (97, 98), and that some DP alleles are associated with chronic beryllium disease (99), sarcoidosis (100), and JRA (101). Both the  $\alpha$ - and  $\beta$ -chains of HLA-DP are polymorphic, allowing multiple combinations, but only a few DP molecules are abundant globally. For example, DPA1\*0103/DPB1\*0401 and 0402 are overrepresented, carried by ~76% of individuals in the Caucasian population (22).

The HLA-DP classification, made in this study, is based on two key amino acids of the DP  $\beta$ -chain: positions  $\beta$ 69 and  $\beta$ 84. These positions correspond to DR/DQ  $\beta$ 71 and  $\beta$ 86. Both are important for DQ classification, while only  $\beta$ 71 takes part in the DR classification. Positions  $\beta$ 84 and, to a lesser extent,  $\beta$ 85 take part in the formation of pocket 1. Almost half (40 of 95) of the  $\beta$ -chains have Gly/Val<sup>84 $\beta$</sup>  and Gly<sup>85 $\beta$</sup> , the other half (55 of 95) have Asp<sup>84 $\beta$</sup>  and Glu<sup>85 $\beta$</sup> . The chemical nature of the two pairs is very different and this determines the strong differences in the pockets formed by them. Pocket 1 with Gly/Val<sup>84 $\beta$</sup>  is deep and nonpolar and could accept large hydrophobic amino acids like Phe, Tyr, and Leu (Table IV). Pocket 1 with Asp<sup>84 $\beta$</sup>  is more shallow and negatively charged. Because no binding motif is available for alleles with Asp<sup>84 $\beta$</sup> , one might suppose positively charged amino acids, such as Arg and Lys, may be favored here. Position  $\beta$ 84 was found to be a key amino acid in Castelli's HLA-DP classification (22). They defined three superotypes, based on positions  $\beta$ 11 and  $\beta$ 84, in contrast to the four identified by our analysis.

Glu/Lys dimorphism exists at position DP  $\beta$ 69. Additionally, there are four alleles (DPB1\*11, 15, 69, and 74) with Arg<sup>69 $\beta$</sup> . Because Lys and Arg are similar, these alleles were grouped into Lys<sup>69 $\beta$</sup>  clusters. Position  $\beta$ 69 affects the shape and charge distribution of pockets 4 and 6 (47). Pockets 4 and 6 with Glu<sup>69 $\beta$</sup>  show high affinity for positive polar residues like Arg, Lys, Gln, and Asn or nonpolar aromatic residues (Phe, Trp, Tyr, and His), but reduced affinity for large nonpolar aliphatic residues (Table IV, motif DPA1\*0103/DPB1\*0201). Because Glu<sup>69 $\beta$</sup>  is associated with sarcoidosis, one could suppose a connection between susceptibility to this disease and alleles from DPw2 and DPw6 superotypes (100). The susceptibility of JRA is strongly associated with DPB1\*0201 allele (101). By analogy, a relation between JRA and the DPw2 superotype could be supposed. Pockets 4 and 6 with Lys/Arg<sup>69 $\beta$</sup>  have reduced amino acid selectivity, with aromatic residues most preferred (motifs DPA1\*0103/DPB1\*0401 and 0402). Additionally, Lys<sup>69 $\beta$</sup>  favors the binding of large residues endowed with the capacity to form hydrogen bonds (such as Arg) with residue Gln<sup>60 $\alpha$</sup>  (47).

Analysis of our classification of HLA class II proteins into superotypes reveals several general trends. First,  $\beta$ -chain polymorphism within the peptide binding site plays the leading role in the overall polymorphism of human MHC. The key polymorphic positions revealed to be important for our and other superotype definitions (22) all belong to  $\beta$ -chains. Second, despite the extraordinary diversity of HLA proteins, common structural features and similarities could be detected and used as fingerprints for their

identification and classification into superotypes. The number of amino acids involved in the superotype fingerprints is strikingly small, i.e., one to three. Finally, the classifications proposed here are based on key amino acids with very different, even opposite, properties. For example, position Glu/Lys<sup>69 $\beta$</sup>  for HLA-DP alleles could be considered as a key position, because of the opposite properties of Glu and Lys. However, position Gly/Val<sup>86 $\beta$</sup>  could not be a key position for DR classification, because of the similar properties of Gly and Val.

The MHC is among the most polymorphic of human proteins, and this has greatly complicated the discovery of epitope vaccines. Superotype analysis is one approach taken to address this confounding problem. We have previously identified class I superotypes using computational methods (40), which we now complement with our present analysis of human class II superotypes. The veracity of this analysis is confirmed, as far as possible, by reference to known peptide binding motifs. Although such motifs are an imperfect, or at least incomplete, representation of binding (102, 103), they have clear utility as an approximation to peptide specificity. All superotypes are theoretically derived. Superotypes, based on "binding motifs", may possess a certain verisimilitude, but are, at best, only a partial definition of superotypic membership, limited by the lack of available data for most MHC molecules. Indeed, all work based on the analysis of experimental work, including our own (104–107), is necessarily limited by the paucity and haphazard nature of extant experimental binding studies. The approach presented here is complementary to such analysis and to existing superotype analyses (3, 19–22). However, our approach is fundamentally different, at a conceptual and technical level, from other, earlier attempts to cluster alleles into superotypes using structural approach.

We have discussed such data as exists which supports and verifies our analysis, rather than speculating in a specious and uncorroborated manner. In the context of human class II MHC, this data is, unfortunately, only partial. Further demonstration of the accuracy of our classification will come in either of two ways: through the accumulation of further motifs in the literature or by the exploration of the peptide specificity repertoire of MHC molecules through systematic study. The utility of the method, though obvious to us, will again require independent, external validation for a sufficiently large number of peptides and alleles that its accuracy can be shown to work to statistical significance. We see superotype definition as a grand challenge with significant scientific and utilitarian merit: it is difficult, and thus exciting, and is also truly valuable, as a pivotal tool in the drive to develop new and better vaccines.

## Disclosures

The authors have no financial conflict of interest.

## References

- Pfeifer, J. D., M. J. Wick, R. L. Roberts, K. Findlay, S. J. Normark, and C. V. Harding. 1993. Phagocytic processing of bacterial antigens for class I MHC presentation to T cells. *Nature* 361: 359–362.
- Rötzschke, O., and K. Falk. 1994. Origin, structure and motifs of naturally processed MHC class II ligands. *Curr. Opin. Immunol.* 6: 45–51.
- 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: 311–314.
- Chelvanayagam, G. 1997. A roadmap for HLA-DR peptide binding specificities. *Hum. Immunol.* 58: 61–69.
- Gulucota, K., and C. DeLisi. 1996. HLA allele selection for designing peptide vaccines. *Genet. Anal. Biomol. Eng.* 13: 81–86.
- Bjorkman, P. J., M. A. Saper, B. Samraoui, W. S. Bennett, J. L. Strominger, and D. C. Wiley. 1987. Structure of the human class I histocompatibility antigen, HLA-A2. *Nature* 329: 506–512.
- Bjorkman, P. J., M. A. Saper, B. Samraoui, 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: 512–518.

8. Saper, M. A., P. J. Bjorkman, and D. C. Wiley. 1991. Refined structure of the human histocompatibility antigen HLA-A2 at 2.6 Å resolution. *J. Mol. Biol.* 219: 277–319.
9. Stern, L. J., J. H. Brown, T. S. Jardetzky, J. C. Gorga, R. G. Urban, J. L. Strominger, and D. C. Wiley. 1994. Crystal structure of the human class II MHC protein HLA-DR1 complexed with an influenza virus peptide. *Nature* 368: 215–221.
10. Dessen, A., C. M. Lawrence, S. Cupo, D. M. Zaller, and D. C. Wiley. 1997. X-ray crystal structure of HLA-DR4 (DRA\*0101, DRB1\*0401) complexed with a peptide from human collagen II. *Immunity* 7: 473–481.
11. 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: 685–693.
12. 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: 261–266.
13. Sette, A., and J. Sidney. 1998. HLA supertypes and supermotifs: a functional perspective on HLA polymorphism. *Curr. Opin. Immunol.* 10: 478–482.
14. 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: 201–212.
15. 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: 271–276.
16. 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: 443–451.
17. Naumann, T., and H. Matter. 2002. Structural classification of protein kinases using 3D molecular interaction field analysis of their ligand binding sites: target family landscapes. *J. Med. Chem.* 45: 2366–2378.
18. Myshkin, E., and B. Wang. 2003. Chemometrical classification of ephrin ligands and Eph kinases using GRID/CPCA approach. *J. Chem. Inf. Comput. Sci.* 43: 1004–1010.
19. Ou, D., L. A. Mitchell, and A. J. Tingle. 1998. A new categorization of HLA DR alleles on a functional basis. *Hum. Immunol.* 59: 665–676.
20. Lund, O., M. Nielsen, C. Kesmir, A. G. Petersen, C. Lundegaard, P. Worning, C. Sylvestre-Hvid, K. Lamberth, G. Roder, S. Justesen, S. Buus, and S. Brunak. 2004. Definition of supertypes for HLA molecules using clustering of specificity matrices. *Immunogenetics* 55: 797–810.
21. Baas, A., X. Gao, and G. Chelvanayagam. 1999. Peptide binding motifs and specificities for HLA-DQ molecules. *Immunogenetics* 50: 8–15.
22. Castelli, F. A., C. Buhot, A. Sanson, H. Zarour, S. Pouvelle-Moraitille, C. Nonn, H. Gahery-Segard, J.-G. Guillet, A. Menez, B. Georges, and B. Maillere. 2002. HLA-DP4, the most frequent HLA II molecule, defines a new supertype of peptide-binding specificity. *J. Immunol.* 169: 6928–6934.
23. Downs, G. M., and J. M. Barnard. 2002. Clustering methods and their uses in computational chemistry. In *Reviews in Computational Chemistry*, Vol. 18. K. B. Lipkowitz and D. B. Boyd, eds. Wiley, Hoboken, p. 1.
24. 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: 4130–4146.
25. 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: 1–10.
26. Welch, W. 2002. Quantitative relationships between ryanoids, receptor affinity and channel conductance. *Front. Biosci.* 7: d1727–d1742.
27. Bordas, B., T. Komives, and A. Lopata. 2003. Ligand-based computer-aided pesticide design: a review of applications of the CoMFA and CoMSIA methodologies. *Pest. Manag. Sci.* 59: 393–400.
28. Kamath, S., and J. K. Buolamwini. 2003. Receptor-guided alignment-based comparative 3D-QSAR studies of benzylidene malonitrile tyrosinase as EGFR and HER-2 kinase inhibitors. *J. Med. Chem.* 46: 4657–4668.
29. Fleischer, R., and M. Wiese. 2003. Three-dimensional quantitative structure-activity relationship analysis of propafenone-type multidrug resistance modulators: influence of variable selection on test set predictivity. *J. Med. Chem.* 46: 4988–5004.
30. Kunick, C., K. Lauenroth, K. Wiekling, X. Xie, C. Schultz, R. Gussio, D. Zaharevitz, M. Leost, L. Meijer, A. Weber, et al. 2004. Evaluation and comparison of 3D-QSAR CoMSIA models for CDK1, CDK5, and GSK-3 inhibition by paullones. *J. Med. Chem.* 47: 22–36.
31. Kuo, C. L., H. Assefa, S. Kamath, Z. Brzozowski, J. Slawinski, F. Saczewski, J. K. Buolamwini, and N. Neamati. 2004. Application of CoMFA and CoMSIA 3D-QSAR and docking studies in optimization of mercaptobenzenesulfonamides as HIV-1 integrase inhibitors. *J. Med. Chem.* 47: 385–399.
32. Doytchinova, I. A., P. Guan, and D. R. Flower. 2004. Quantitative structure-activity relationships and the prediction of MHC supermotifs. *Methods* 34: 444–453.
33. Hellberg, S., M. Sjöström, and S. Wold. 1986. The prediction of bradykinin potentiating potency of pentapeptides: an example of a peptide quantitative structure-activity relationship. *Acta Chem. Scand. B.* 40: 135–140.
34. Sandberg, M., L. Eriksson, J. Jonsson, M. Sjustrum, and S. Wold. 1998. New chemical descriptors relevant for the design of biologically active peptides: a multivariate characterization of 87 amino acids. *J. Med. Chem.* 41: 2481–2491.
35. Freyhult, E. K., K. Andersson, and M. G. Gustafsson. 2003. Structural modeling extends QSAR analysis of antibody-lysozyme interactions to 3D-QSAR. *Bio-phys. J.* 84: 2264–2272.
36. Eriksson, L., H. Antti, J. Gottfries, E. Holmes, E. Johansson, F. Lindgren, I. Long, T. Lundstedt, J. Trygg, and S. Wold. 2004. Using chemometrics for navigating in the large data sets of genomics, proteomics, and metabonomics (gpm). *Anal. Bioanal. Chem.* 380: 419–429.
37. Long, L. P., Andersson, E. Seifert, and T. Lundstedt. 2004. Multivariate analysis of five GPCR receptor classes. *Chemometr. Intell. Lab. 73*: 95–104.
38. Cruciani, G., M. Baroni, E. Carosati, M. Clementi, R. Valigi, and S. Clementi. 2004. Peptide studies by means of principal properties of amino acids derived from MIF descriptors. *J. Chemometr.* 18: 146–155.
39. Siebert, K. J. 2003. Modeling protein functional properties from amino acid composition. *J. Agr. Food Chem.* 51: 7792–7797.
40. Doytchinova, I. A., P. Guan, and D. R. Flower. 2004. Identifying human major histocompatibility complex supertypes using bioinformatics methods. *J. Immunol.* 172: 4314–4323.
41. Bower, M., F. E. Cohen, and R. L. Dunbrack, Jr. 1997. Side chain prediction from a backbone-dependent rotamer library: a new tool for homology modeling. *J. Mol. Biol.* 267: 1268–1282.
42. Zavala-Ruiz, Z., E. J. Sundberg, J. D. Stone, D. B. De Oliveira, I. C. Chan, J. Svendsen, R. A. Mariuzzi, and L. J. Stern. 2003. Exploration of the P6/P7 region of the peptide-binding site of the human class II major histocompatibility complex protein HLA-DR1. *J. Biol. Chem.* 278: 44904–44912.
43. Lee, K. H., K. W. Wucherpfennig, and D. C. Wiley. 2001. Structure of a human insulin peptide-HLA-DQ8 complex and susceptibility to type I diabetes. *Nat. Immunol.* 2: 501–507.
44. Sali, A. and T. L. Blundell. 1993. Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol.* 234: 779–815.
45. Marti-Renom, M. A., A. Stuart, A. Fiser, R. Sánchez, F. Melo, and A. Sali. 2000. Comparative protein structure modeling of genes and genomes. *Annu. Rev. Bio-phys. Biomol. Struct.* 29: 291–325.
46. Li, Y., H. Li, R. Martin, and R. A. Mariuzza. 2000. Structural basis for the binding of an immunodominant peptide from myelin basic protein in different registers by two HLA-DR2 protein. *J. Mol. Biol.* 304: 177–188.
47. Reche, P. A., and E. L. Reinherz. 2003. Sequence variability analysis of human class I and class II MHC molecules: functional and structural correlates of amino acid polymorphisms. *J. Mol. Biol.* 331: 623–641.
48. Diaz, G., M. Amicosante, D. Jaraquemada, R. H. Butler, M. V. Guillen, M. Sanchez, C. Nombela, and J. Arroyo. 2003. Functional analysis of HLA-DP polymorphism: a crucial role for DPβ residues 9, 11, 35, 55, 69 and 84–87 in T cell allorecognition and peptide binding. *Int. Immunol.* 15: 565–576.
49. Berretta, F., R. H. Butler, G. Diaz, N. Sanarico, J. Arroyo, M. Fraziano, G. Aichinger, K. W. Wucherpfennig, V. Colizzi, C. Saltini, and M. Amicosante. 2003. Detailed analysis of the effects of Glu/Lys β69 human leukocyte antigen-DP polymorphism on peptide-binding specificity. *Tissue Antigens* 62: 459–471.
50. Hammer, J., B. Takacs, and F. Sinigaglia. 1992. Identification of a motif for HLA-DR1 binding peptides using M13 display libraries. *J. Exp. Med.* 176: 1007–1013.
51. Falk, K., O. Rötzschke, S. Stevanović, G. Jung, and H.-G. Rammensee. 1994. Pool sequencing of natural HLA-DR, DQ, and DP ligands reveals detailed peptide motifs, constraints of processing, and general rules. *Immunogenetics* 39: 230–242.
52. Chic, R. M., R. G. Urban, W. S. Lane, J. C. Gorga, L. J. Stern, D. A. Vignali, and J. L. Strominger. 1992. Predominant naturally processed peptides bound to HLA-DR1 are derived from MHC-related molecules and are heterogeneous in size. *Nature* 358: 764–768.
53. Vogt, A. B., H. Kropshofer, H. Kalbacher, M. Kalbus, H.-G. Rammensee, J. E. Coligan, and R. Martin. 1994. Ligand motifs of HLA-DRB5\*0101 and DRB1\*1501 molecules delineated from self-peptides. *J. Immunol.* 153: 1665–1673.
54. Wucherpfennig, K. W., A. Sette, S. Southwood, C. Oseroff, M. Matsui, J. L. Strominger, and D. A. Hafler. 1994. Structural requirements for binding of an immunodominant myelin basic protein peptide to DR2 isotypes and for its recognition by human T cell clones. *J. Exp. Med.* 179: 279–290.
55. Malcherek, G., K. Falk, O. Rötzschke, H.-G. Rammensee, S. Stevanović, V. Gnau, G. Jung, and A. Melms. 1993. Natural peptide ligand motifs of two HLA molecules associated with myasthenia gravis. *Int. Immunol.* 5: 1229–1237.
56. Geluk, A., K. E. van Meijgaarden, A. A. M. Janson, J. W. Drijfhout, R. H. Meleno, R. R. P. de Vries, and T. H. M. Ottenhoff. 1992. Functional analysis of DR17(DR3)-restricted mycobacterial T-cell epitopes reveals DR17-binding motif and enables the design of allele-specific competitor peptides. *J. Immunol.* 149: 2864–2871.
57. Geluk, A., K. E. van Meijgaarden, S. Southwood, C. Oseroff, J. W. Drijfhout, R. R. P. de Vries, T. H. M. Ottenhoff, and A. Sette. 1994. HLA-DR3 molecules can bind peptides carrying two alternative specific submotifs. *J. Immunol.* 152: 5742–5748.
58. Sette, A., J. Sidney, C. Oseroff, M. F. del Guercio, S. Southwood, T. Arrhenius, M. F. Powell, S. M. Colon, F. C. A. Gaeta, and H. M. Grey. 1993. HLA DR4w4-binding motifs illustrate the biochemical basis of degeneracy and specificity in peptide-DR interactions. *J. Immunol.* 151: 3163–3170.
59. Hammer, J., P. Valsasini, K. Tolba, D. Bolin, J. Higelin, B. Takacs, and F. Sinigaglia. 1993. Promiscuous and allele-specific anchors in HLA-DR-binding peptides. *Cell* 74: 197–203.

60. Hill, C. M., A. Liu, K. W. Marshall, J. Mayer, B. Jorgensen, B. Yuan, R. M. Cubbon, E. A. Nichols, L. S. Wicker, and J. B. Rothbard. 1994. Exploration of requirements for peptide binding to HLA DRB1\*0101 and DRB1\*0401. *J. Immunol.* 152: 2890–2898.
61. Friede, T., V. Gnau, G. Jung, W. Keilholz, S. Stevanović, and H.-G. Rammensee. 1996. Natural ligand motifs of closely related HLA-DR4 molecules predict features of rheumatoid arthritis associated peptides. *Mol. Basis Dis.* 1316: 85–101.
62. Matsushita, S., K. Takahashi, M. Motoki, K. Komoriya, S. Ikagawa, and Y. Nishimura. 1994. Allele specificity of structural requirement for peptides bound to HLA-DRB1\*0405 and -DRB1\*0406 complexes: implication for the HLA-associated susceptibility to methimazole-induced insulin autoimmune syndrome. *J. Exp. Med.* 180: 873–883.
63. Kinouchi, R., H. Kobayashi, K. Sato, S. Kimura, and M. Katagiri. 1994. Peptide motifs of HLA-DR4/DR53 (DRB1\*0405/DRB4\*0101) molecules. *Immunogenetics* 40: 376–378.
64. Newcomb, J. R., and P. Cresswell. 1993. Characterization of endogenous peptides bound to purified HLA-DR molecules and their absence from invariant chain-associated  $\alpha$ - $\beta$ -dimers. *J. Immunol.* 150: 499–507.
65. Ettinger, R. A., and W. W. Kwok. 1998. A peptide binding motif for HLA-DQA1\*0102/DQB1\*0602, a class II MHC molecule associated with dominant protection in insulin-dependent diabetes mellitus. *J. Immunol.* 160: 2365–2373.
66. Johansen, B. H., F. Varddal, J. A. Eriksen, E. Thorsby, and L. M. Sollid. 1995. Identification of a putative motif for binding of peptides to HLA-DQ2. *Int. Immunol.* 8: 177–182.
67. Varddal, F., B. H. Johansen, T. Friede, C. J. Thorpe, S. Stevanovic, J. E. Eriksen, K. Sletten, E. Thorsby, H.-G. Rammensee, and L. M. Sollid. 1996. The peptide binding motif of the disease associated HLA-DQ ( $\alpha$ 1\*0501,  $\beta$ 1\*0201) molecule. *Eur. J. Immunol.* 26: 2764–2772.
68. van de Wal, Y., Y. M. C. Kooy, J. W. Drijfhout, R. Amons, G. K. Papadopoulos, and F. Koning. 1997. Unique peptide binding characteristics of the disease-associated DQ( $\alpha$ 1\*0501,  $\beta$ 1\*0201) vs the non-disease-associated DQ( $\alpha$ 1\*0201,  $\beta$ 1\*0202) molecule. *Immunogenetics* 46: 484–492.
69. Quarsten, H., G. Paulsen, B. H. Johansen, C. J. Thorpe, A. Holm, S. Buus, and L. M. Sollid. 1998. The P9 pocket of HLA-DQ2 (non-Asp<sup>57</sup>) has no particular preference for negatively charged anchor residues found in other type 1 diabetes-predisposing non-Asp<sup>57</sup> MHC class II molecules. *Int. Immunol.* 10: 1229–1236.
70. Ihle, J., B. Fleckenstein, C. Terreaux, H. Beck, E. D. Albert, and G. E. Dannecker. 2003. Differential peptide binding motif for three juvenile arthritis associated HLA-DQ molecules. *Clin. Exp. Rheumatol.* 21: 257–262.
71. Sidney, J., C. Oseroff, M.-F. del Guercio, S. Southwood, J. I. Krieger, G. Y. Ishioka, K. Sakaguchi, E. Appella, and A. Sette. 1994. Definition of a DQ3.1-specific binding motif. *J. Immunol.* 152: 4516–4525.
72. Godkin, A., T. Friede, M. Davenport, S. Stevanovic, A. Willis, D. Jewell, A. Hill, and H.-G. Rammensee. 1996. Use of eluted peptide sequence data to identify the binding characteristics of peptides to the insulin-dependent diabetes susceptibility allele HLA-DQ8 (DQ 3.2). *Int. Immunol.* 9: 905–911.
73. Moustakas, A. K., Y. van de Wal, J. Routsias, Y. M. C. Kooy, P. van Veelen, L. W. Drijfhout, F. Koning, and G. K. Papadopoulos. 1999. Structure of celiac disease-associated HLA-DQ8 and non-associated HLA-DQ9 alleles in complex with two disease-specific epitopes. *Int. Immunol.* 12: 1157–1166.
74. Chicz, R. M., D. F. Graziano, M. Trucco, J. L. Strominger, and J. C. Gorga. 1997. HLA-DP2: self peptide sequences and binding properties. *J. Immunol.* 159: 4935–4942.
75. Kubo, R. T., A. Sette, H. M. Grey, E. Appella, K. Sakaguchi, N. Z. Zhu, D. Arnott, N. Sherman, J. Shabanowitz and H. Michel. 1994. Definition of specific peptide motifs for four major HLA-A alleles. *J. Immunol.* 152: 3913–3924.
76. Longmate, J., J. York, C. La Rosa, R. Krishnan, M. Zhang, D. Senitzer and D. J. Diamond. 2001. Population coverage by HLA class I-restricted cytotoxic T-lymphocyte epitopes. *Immunogenetics* 52: 165–173.
77. Doolan, D. L., S. Southwood, R. Chesnut, E. Appella, E. Gomez, A. Richards, Y. I. Higashimoto, A. Maewal, J. Sidney, R. A. Gramzinski, et al. 2000. HLA-DR-promiscuous T cell epitopes from *Plasmodium falciparum* pre-erythrocytic-stage antigens restricted by multiple HLA class II alleles. *J. Immunol.* 165: 1123–1137.
78. Propato, A., E. Schiaffella, E. Vicenzi, V. Francavilla, L. Baloni, M. Paroli, L. Finocchi, N. Tanigaki, S. Ghezzi, R. Ferrara, et al. Spreading of HIV-specific CD8<sup>+</sup> T-cell repertoire in long-term nonprogressors and its role in the control of viral load and disease activity. *Hum. Immunol.* 62: 561–576.
79. Chang, K. M., N. H. Gruener, S. Southwood, J. Sidney, G. R. Pape, F. V. Chisari and A. Sette. 1999. Identification of HLA-A3 and -B7-restricted CTL response to hepatitis C virus in patients with acute and chronic hepatitis C. *J. Immunol.* 162: 1156–1164.
80. Bertoni, R., J. Sidney, P. Fowler, R. W. Chesnut, F. V. Chisari and A. Sette. 1997. Human histocompatibility leukocyte antigen-binding supermotifs predict broadly cross-reactive cytotoxic T lymphocyte responses in patients with acute hepatitis. *J. Clin. Invest.* 100: 503–513.
81. Kawashima, I., S. J. Hudson, V. Tsai, S. Southwood, K. Takesako, E. Appella, A. Sette and E. Celis. 1998. The multi-epitope approach for immunotherapy for cancer: identification of several CTL epitopes from various tumor-associated antigens expressed on solid epithelial tumors. *Hum. Immunol.* 59: 1–14.
82. Fikes, J. D., and A. Sette. 2003. Design of multi-epitope, analogue-based cancer vaccines. *Expert Opin. Biol. Ther.* 3: 985–993.
83. Hammer, J., T. Sturniolo, and F. Sinaglia. 1997. HLA class II peptide binding specificity and autoimmunity. *Adv. Immunol.* 66: 67–100.
84. Ghosh, P., M. Amaya, E. Mellins, and D. C. Wiley. 1995. The structure of an intermediate in class II MHC maturation: CLIP bound to HLA-DR3. *Nature* 378: 457–462.
85. Todd, J. A., J. I. Bell, and H. O. McDevitt. 1987. HLA-DQ  $\beta$  gene contributes to susceptibility and resistance to insulin-dependent diabetes mellitus. *Nature* 329: 599–604.
86. Sillid, L. M., G. Markussen, J. Ek, H. Gierde, F. Varddal, and E. Thorsby. 1989. Evidence for a primary association of coeliac disease to a particular HLA-DQ  $\alpha/\beta$  heterodimer. *J. Exp. Med.* 169: 345–350.
87. Morel, P. A., J. S. Dorman, J. A. Todd, H. O. McDevitt, and M. Trucco. 1988. Aspartic acid at position 57 of the HLA-DQ  $\beta$  chain protects against type I diabetes. A family study. *Proc. Natl. Acad. Sci. USA* 85: 8111–8115.
88. Ronningen, K. S., T. Iwe, T. S. Halstensen, A. Spurkland, and E. Thorsby. 1989. The amino acid at position 57 of the HLA-DQ  $\beta$  chain and susceptibility to develop insulin-dependent diabetes mellitus. *Hum. Immunol.* 26: 215–225.
89. Lundin, K. E. A., H. Scott, T. Hansen, G. Paulsen, T. S. Halstensen, O. Fausa, E. Thorsby, and L. M. Sollid. 1993. Gliadin-specific HLA-DQ2 ( $\alpha$ 1\*0501,  $\beta$ 1\*0201) restricted T cells isolated from the small intestinal mucosa of celiac disease patients. *J. Exp. Med.* 178: 187–196.
90. Lundin, K. E. A., H. Scott, O. Fausa, E. Thorsby, and L. M. Sollid. 1994. T cells from the small intestinal mucosa of a DR4, DQ7/DR4, DQ8 celiac disease preferentially recognise gliadin when presented by DQ8. *Hum. Immunol.* 41: 285–291.
91. Spurkland, A., E. G. Celius, I. Knusten, A. Beiske, E. Thorsby, F. Varddal. 1997. The HLA-DQ ( $\alpha$ 1\*0102,  $\beta$ 1\*0602) heterodimer may confer susceptibility to multiple sclerosis in the absence of the HLA-DR ( $\alpha$ 1\*01,  $\beta$ 1\*1501) heterodimer. *Tissue Antigens* 50: 15–22.
92. Arcos-Burgos, M., G. Palacio, J. L. Sanchez, A. C. Londoño, C. S. Uribe, M. Jimenez, A. Villa, J. M. Anaya, M. L. Bravo, N. Jaramillo, C. Espinal, J. J. Builes, M. Moreno, and I. Jimenez. 1999. Multiple sclerosis: association to HLA DQ $\alpha$  in a tropical population. *Exp. Clin. Immunogenet.* 16: 131–138.
93. Delgado, J. C., A. Hameed, J. J. Yunis, K. Bhol, A. I. Rojas, S. B. Rehman, A. A. Khan, M. Ahmad, C. A. Alper, A. R. Ahmed, and E. J. Yunis. 1997. Pemphigus vulgaris autoantibody response is linked to HLA-DQB1\*0503 in Pakistani patients. *Hum. Immunol.* 57: 110–119.
94. Zanelli, E., F. C. Breedveld, and R. R. P. de Vries. 2000. HLA association with autoimmune disease: a failure to protect? *Rheumatology* 39: 1060–1066.
95. Lambert, N. C., P. C. Evans, T. L. Hashizumi, S. Maloney, T. Gooley, D. E. Furst, and J. L. Nelson. 2000. Cutting edge: persistent fetal microchimerism in T lymphocytes is associated with HLA-DQA1\*0501: implications in autoimmunity. *J. Immunol.* 164: 5545–5548.
96. Sanjeevi, C. B., M. Landin-Olsson, I. Kockum, G. Dahlquist, and A. Lernmark. 1995. Effects of the second HLA-DQ haplotype on the association with childhood insulin-dependent diabetes mellitus. *Tissue Antigens* 45: 148–152.
97. Petersdorf, E. W., A. G. Smith, E. M. Mickelson, G. M. Longton, C. Anasetti, S. Y. Choo, P. J. Martin, and J. A. Hansen. 1993. The role of HLA-DPB1 disparity in the development of acute graft-versus-host disease following unrelated donor marrow transplantation. *Blood* 81: 1923–1932.
98. Moreau, P., and A. Cesbron. 1994. HLA-DP and allogeneic bone marrow transplantation. *Bone Marrow Transplant.* 13: 675–681.
99. Fontenot, A. P., and B. L. Kotzin. 2003. Chronic beryllium disease: immune-mediated destruction with implications for organ-specific autoimmunity. *Tissue Antigens* 62: 449–458.
100. Lympny, P. A., M. Petrek, A. M. Southcott, A. J. Newman Taylor, K. I. Welsh, and R. M. du Bois. 1996. HLA-DPB polymorphism: Glu 69 association with sarcoidosis. *Eur. J. Immunogenet.* 23: 353–359.
101. Begovich, A. B., T. L. Bugawan, B. S. Nepom, W. Klitz, G. T. Nepom, and H. A. Erlich. 1989. A specific HLA-DP $\beta$  allele is associated with pauciarticular juvenile rheumatoid arthritis but not adult rheumatoid arthritis. *Proc. Natl. Acad. Sci. USA* 86: 9489–9493.
102. Flower, D. R. 2003. Towards in silico prediction of immunogenic epitopes. *Trends Immunol.* 24: 667–674.
103. Doytchinova, I. A., V. A. Walshe, N. A. Jones, S. E. Gloster, P. Borrow, and D. R. Flower. 2004. Coupling in silico and in vitro analysis of peptide-MHC binding: a bioinformatics approach enabling prediction of superbinding peptides and anchorless epitopes. *J. Immunol.* 172: 7495–7502.
104. 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: 535–544.
105. Doytchinova, I., and D. R. Flower. 2003. The HLA-A2-supermotif: a QSAR definition. *Org. Biomol. Chem.* 1: 2648–2654.
106. Guan, P., I. A. Doytchinova, and D. R. Flower. 2003. A comparative molecular similarity indices (CoMSIA) study of peptide binding to the HLA-A3 superfamily. *Bioorg. Med. Chem.* 11: 2307–2311.
107. Guan, P., I. A. Doytchinova, and D. R. Flower. 2003. HLA-A3 supermotif defined by quantitative structure-activity relationship analysis. *Protein Eng.* 16: 11–18.