Immunogenetics (2004) 55:797–810
DOI 10.1007/s00251-004-0647-4

ORIGINAL PAPER

Ole Lund · Morten Nielsen · Can Kesmir · Anders Gorm Petersen · Claus Lundegaard · Peder Worning · Christina Sylvester-Hvid · Kasper Lamberth · Gustav Røder · Sune Justesen · Søren Buus · Søren Brunak

Definition of supertypes for HLA molecules using clustering of specificity matrices

Received: 30 October 2003 / Revised: 12 January 2003 / Published online: 13 February 2004

Abstract Major histocompatibility complex (MHC) proteins are encoded by extremely polymorphic genes and play a crucial role in immunity. However, not all genetically different MHC molecules are functionally different. Sette and Sidney (1999) have defined nine HLA class I supertypes and showed that with only nine main functional binding specificities it is possible to cover the binding properties of almost all known HLA class I molecules. Here we present a comprehensive study of the functional relationship between all HLA molecules with known specificities in a uniform and automated way. We have developed a novel method for clustering sequence motifs. We construct hidden Markov models for HLA class I molecules using a Gibbs sampling procedure and use the similarities among these to define clusters of specificities. These clusters are extensions of the previously suggested ones. We suggest splitting some of the alleles in the A1 supertype into a new A26 supertype, and some of the alleles in the B27 supertype into a new B39 supertype. Furthermore the B8 alleles may define their own supertype. We also use the published specificities for a number of HLA-DR types to define clusters with similar specificities. We report that the previously observed specificities of these class II molecules can be clustered into nine classes, which only partly correspond to the serological classification. We show that classification of HLA molecules may be done in a uniform and automated way. The definition of clusters allows for selection of representative HLA molecules that can cover the HLA specificity space better. This makes it possible to target most of the known HLA alleles with known specificities using only a few peptides, and may be used in construction of vaccines. Supplementary material is available at http://www.cbs.dtu.dk/researchgroups/immunology/supertypes.html.

Keywords HLA · Supertype · Classification · Class I · Class II

O. Lund (✉) · M. Nielsen · C. Kesmir · A. G. Petersen · C. Lundegaard · P. Worning · S. Brunak
Center for Biological Sequence Analysis, BioCentrum-DTU,
Technical University of Denmark,
Building 208, 2800 Lyngby, Denmark
e-mail: lund@cbs.dtu.dk
Tel.: +45-4525-2425
Fax: +45-4593-1585

C. Kesmir
Theoretical Biology/Bioinformatics,
Utrecht University,
Utrecht, The Netherlands

C. Sylvester-Hvid · K. Lamberth · G. Røder · S. Justesen · S. Buus
Department of Experimental Immunology,
Institute of Medical Microbiology and Immunology,
University of Copenhagen,
Panum Building 18.3.22, Blegdamsvej 3B,
2200 Copenhagen N, Denmark

Introduction

A number of computational methods have been developed to identify T-cell epitopes. These can be used to design epitope vaccines. Ishioka and co-workers (1999) list a number of advantages of epitope-based vaccines: they can be more potent, be controlled better, induce subdominant epitopes (for example against tumor antigens where there is tolerance against dominant epitopes), target multiple conserved epitopes in rapidly mutating pathogens like HIV and HCV, be analogued to break tolerance, and overcome safety concerns associated with entire organisms or proteins. Epitope-based vaccines have also been shown to confer protection in animal models (see references in Rodriguez et al. 1998; Sette and Sidney 1999).

One potential drawback of epitope-based vaccines is that the HLA genes are extremely polymorphic. Each of the corresponding molecules has a different specificity. If a vaccine needs to contain a unique peptide for each of these molecules it will need to comprise hundreds of peptides. One way to counter this is to select sets of a few HLA molecules that together have a broad distribution in the human population. Gulukota and DeLisi (1996) compiled lists with three, four, and five alleles that give

the maximal coverage of different ethnic groups. One complication they had to deal with is that HLA alleles are in linkage disequilibrium; i.e., that the joint probability of an allelic pair may not be equal to the product of their individual frequencies [$P(a)P(b) \neq P(ab)$]. This means that it is not necessarily optimal to choose the alleles with the highest individual frequencies. Moreover, Gulukota and DeLisi (1996) find that populations like the Japanese, Chinese and Thais can be covered by fewer alleles than the North American Black population, which turns out to be very diverse. Thus, different alleles should be targeted in order to make vaccines for different ethnic groups or geographic regions.

A factor that may reduce the number of epitopes necessary to include in a vaccine is that many of the different HLA molecules have similar specificities. The different HLA molecules have been grouped together in so-called supertypes (del Guercio et al. 1995; Sidney et al. 1995, Sette and Sidney 1999). This means ideally that if a peptide can bind to one allele within a supertype, it can bind to all alleles within that supertype. In practice, however, only some peptides that bind to one allele in a supertype will bind to all alleles within that supertype. A number of different criteria have been used to define these supertypes, including structural similarities, shared peptide binding motifs, identification of cross reacting peptides, and ability to generate methods that can predict cross binding peptides (Sidney et al. 1996). For HLA class I molecules, Sette and Sidney (1999) defined nine supertypes (A1, A2, A3, A24, B7, B27, B44, B58, B62), which were reported to cover most of the HLA-A and -B polymorphism. They argued that the different alleles within each of these supertypes have almost identical peptide binding specificity. They found that while the frequencies at which the different alleles were found in different ethnic groups were very different, the frequencies of the supertypes were quite constant. Assuming Hardy-Weinberg equilibrium (i.e., infinitely large, random mating populations free from outside evolutionary forces), they found that >99.6% of persons in all ethnic groups surveyed possessed at least one allele within at least one of these supertypes. They also showed that the smaller collections of supertypes (A2, A3, B7) and (A1, A2, A3, B7, A24, B44) covered in the range of 83.0–88.5% and 98.1–100.0% of persons in different ethnic groups respectively. Three alleles, *A29*, *B8* and *B46*, were found to be outliers with a different binding specificity than any of the supertypes. These may define supertypes themselves when the specificity of more HLA molecules is known.

Some work has also been done to define supertypes of class II molecules. It has been reported that five alleles from the *DQ* locus (*DQ1*, *DQ2*, *DQ3*, *DQ4*, *DQ5*) covers 95% of most populations (Gulukota and DeLisi 1996). It has also been reported that a number of HLA-DR types share overlapping peptide-binding repertoires (Southwood et al. 1998).

Here we present a novel measure for the difference in the specificities of different HLA molecules and use it to cluster HLA molecules. The result of this analysis is used to revise the class I supertypes. We also use the published specificities for a number of HLA-DR types to define clusters with similar specificities.

Materials and methods

Weight matrices representing the specificities of different HLA class I molecules were constructed. The different class I molecules addressed in this study can be seen in Table 1. HLA ligands were extracted from the SYFPEITHI (Rammensee et al. 1995, 1999) and MHCpep (Brusic et al. 1998) databases. All lines containing amino acid information were treated as sequences and blanks were replaced by "X". For each allele, weight matrices were built using a program implementing a Gibbs Sampling algorithm that estimates the best scoring 9-mer pattern using a Monte Carlo sampling procedure (Lawrence et al. 1993; Nielsen et al. 2004). In brief the best scoring pattern is defined in terms of highest relative entropy (Cover and Thomas 1991) summed over a 9-mer alignment. The program samples possible alignments of the sequences in the input file. For each alignment a weight matrix is calculated as $\log(p_{ap}/q_a)$, where p_{ap} is the estimated frequency of amino acid a at position p in the alignment and q_a the background frequency of amino acid a in the SwissProt Database (Boeckmann et al. 2003). The values for p_{ap} are estimated using sequence weighting and correction for low counts. Sequence weighting is estimated using sequence clustering (Henikoff and Henikoff 1994). The correction for low counts is done using the Blosum weighting scheme in a similar way to that used by PSI-BLAST (Altschul et al. 1997).

In order to define a clustering of HLA molecules, we first calculated difference in specificities (the distance) between each pair of HLA molecules. The distance d_{ij} between two HLA molecules (i, j) was calculated as the sum over each position in the two motifs of one minus the normalized vector products of the amino acids frequency vectors (Lyngs et al. 1999):

$$d_{ij} = \sum\nolimits_p \left(1 - \frac{p_p^i \cdot p_p^j}{\left| p_p^i \right| \left| p_p^j \right|} \right)$$

where $p_a{}^i$, and $p_a{}^j$ are the vectors of 20-amino-acid frequencies at position p in matrix i and j, respectively. The *point* $(\cdot)$ denotes the vector product, and $\|$ denotes the calculation of the euclidian length of the vector. Dividing all distances with the largest distance $d_{ij}^{\max}$ normalizes the distance-matrix.

The distance-matrices were used as input to the program *neighbor* from version 3.5 of the PHYLIP package (http://evolution.genetics.washington.edu/phylip.html), which implements the neighbor-joining method of Saitou and Nei (1987). Default parameters were used. If the lengths of tree branches became negative they were put to zero. To estimate the significance of the neighbor-joining clustering, we employed the bootstrap method (Press et al. 1992). We generate a set of matrices randomly taking out a column N times, with replacement from the original matrix set. Here N is the motif length, which is set to nine throughout the calculation. Each of the N columns in the matrices contains the scores for having each of the 20 amino acids at that position. We calculate a tree for each such matrix set. Repeating this experiment 1,000 times, we can estimate a consensus tree, and corresponding branch bootstrap values. The bootstrap values on branches are the fraction of experiments where one given subset of alleles were connected to all the other alleles with only a single branch; i.e., the fraction on the experiments where the alleles in the given subset clustered together. We further can estimate bootstrap values for suboptimal tree constructions and compare the probability of one tree construction to another.

To visualize the characteristics of the different binding motifs, we used the logo program (Schneider and Stephens 1990). The information content at each position in the sequence motif

Table 1 HLA type, supertype and amino acid motif for all alleles described by Sette and Sidney (1999) and Rammensee et al. (1999). An *arrow* indicates a supertype assignment change; *letters in square brackets* indicate the amino acids that occur at that position

HLA type	Supertype	Amino acid motif
HLA-A1	A1	.[TS][DE].....Y
HLA-A*0101[a]	A1[d]	.[–].....[–]
HLA-A*0102[a]	A1	.[–].....[–]
HLA-A*0201	A2	.[LM].....[VL]
HLA-A*0202	A2	.[AL].....[VL]
HLA-A*0203	A2	.[LV].....[LI]
HLA-A*0204	A2	.[AL].....[VL]
HLA-A*0205	A2	.[LV].....[LS]
HLA-A*0206	A2	.[VQ].....[VS]
HLA-A*0207	A2	.[L-].....[VL]
HLA-A*0209	A2[e]	.[LA].....[V-]
HLA-A*0214	A2[e]	.[QV].....[LV]
HLA-A*0217	A2[e]	.[L-].....[L-]
HLA-A3	A3	K[LY].....[KY]
HLA-A*0301	A3	K[IL].....[K-]
HLA-A*1101	A3	.[YT].....[K-]
HLA-A23	.	.[–].....[–]
HLA-A*2301[a]	A24	.[–].....[–]
HLA-A24	A24[d]	.[YF].....[LF]
HLA-A*2402	A24	.[YF].....[LF]
HLA-A*2403[a]	A24[d]	.[–].....[–]
HLA-A*2404[a]	A24[d]	.[–].....[–]
HLA-A25	A1[d]	.[–].....[–]
HLA-A*2501[a]	A1[d]	.[–].....[–]
HLA-A26	A1[d]→A26	.[–].....[–]
HLA-A*2601	A1[d]→A26	E[TI].....[FY]
HLA-A*2602	A1[d]→A26	[DE]I.....[FY]
HLA-A*2603	A26[e]	E[VL].....[ML]
HLA-A*2604[a]	A1[d]→A26	.[–].....[–]
HLA-A28	A1→A26	.[–].....[–]
HLA-A29	Outlier	.[FN].....[YC]
HLA-A*2902	Outlier	K[E-].....[YL]
HLA-A30[a]	A24[d]	.[–].....[–]
HLA-A*3001	A24→A1	K[TF].....[FL]
HLA-A*3002	A24[d]→A1	R[YV].....[YK]
HLA-A*3003	A24[d]→A1	R[YL].....[Y-]
HLA-A*3004	A1[e]	K[YT].....[YL]
HLA-A31	.	.[–].....[–]
HLA-A*3101	A3	[RK]QL.....[R-]
HLA-A32	.	.[–].....[–]
HLA-A*3201[a]	A1[d]	.[–].....[–]
HLA-A*3301	A3	.[LV].....[RK]
HLA-A*3303	A3[e]	[DE]I.....[R-]
HLA-A*3402	A3[e]	.[V-].....[R-]
HLA-A*3601[a]	A1	.[–].....[–]
HLA-A*4301[a]	A1	.[–].....[–]
HLA-A*6601	A3[e]	[ED]T.....[R-]
HLA-A*6801	A3	E[VT].....[RK]
HLA-A*6802	A2	D[TV].....[VS]
HLA-A*6901	A2	E[TA].....[R-]
HLA-A*7401	.	.[T-].....[V-]
HLA-A*8001[a]	A1	.[–].....[–]
HLA-B07[X]	B7	.[PV].....[LA]
HLA-B*0702	B7	.[PV].....[LA]
HLA-B*0703	B7	.[DP].....[L-]
HLA-B*0704[a]	B7	.[–].....[–]
HLA-B*0705	B7	.[P-].....[FL]
HLA-B08	Outlier	.[LP]K.K...[L-]
HLA-B*0801	Outlier	.[RK].[RK]....
HLA-B*0802	Outlier	.[L-]K.K...[F-]
HLA-B13	.	.[A-].....[–]
HLA-B*1301[a]	B62[d]	.[–].....[–]
HLA-B*1302[a]	B62[d]	.[–].....[–]
HLA-B14	Outlier	.[R-].....[LV]
HLA-B*1401[a]	B27	.[–].....[–]
HLA-B*1402[a]	B27	.[–].....[–]
HLA-B15	B62[e]	.[FM].....[YF]
HLA-B*1501	B62	.[QL].....[YV]
HLA-B*1502	B62	.[LQ].....[YF]
HLA-B*1503	B27[d]	.[QK].....[YV]
HLA-B*1508	B7	.[PV].....[YS]
HLA-B*1506[a]	B62[d]	.[–].....[–]
HLA-B*1509	B27→B39	.[H-].....[LF]
HLA-B*1510	B27[d]→B39	.[H-].....[LF]
HLA-B*1512	B62[d]	.[QL].....[YS]
HLA-B*1513	B62	.[IL].....[W-]
HLA-B*1514[a]	B62[d]	.[–].....[–]
HLA-B*1516	B58	.[TS].....[IV]
HLA-B*1517	B58	.[–].....[–]
HLA-B*1518	B27[d]	.[–].....[–]
HLA-B*1519[a]	B62[d]	.[–].....[–]
HLA-B*1521[a]	B62[d]	.[–].....[–]
HLA-B17	.	.[–].....[–]
HLA-B18	B44[c]	.[E-].....[–]
HLA-B*1801	.	.[–].....[–]
HLA-B22	.	.[–].....[–]
HLA-B27	B27[e]	.[R-].....[–]
HLA-B*2701	B27[d]	R[RQ].....[Y-]
HLA-B*2702	B27	K[R-].....[YF]
HLA-B*2703	B27	[RK]R.....[LY]
HLA-B*2704	B27	R[R-].....[LF]
HLA-B*2705	B27	R[R-]F.....[–]
HLA-B*2706	B27	R[R-].....[LV]
HLA-B*2707	B27	[RK]R.....[LV]
HLA-B*2708[a]	B27[d]	.[–].....[–]
HLA-B*2709	B27[e]	[GR]R.....[–]
HLA-B35	B7[e]	.[P-].....[Y-]
HLA-B*3501[b]	B7	.[PV].....[LY]
HLA-B*3502[a]	B7	.[–].....[–]
HLA-B*3503	B7	.[PM].....[MF]
HLA-B37	.	.[F-].....[T-]
HLA-B*3701	B44	.[DE].....L[I-]
HLA-B*3801	B27→B39	.[HF]D.....[LF]
HLA-B*3802[a]	B27	.[–].....[–]
HLA-B39	B27[e]	.[H-].....[L-]
HLA-B*3901	B27→B39	.[HR].....[L-]
HLA-B*3902	B27	.[–].....[MF]
HLA-B*3903[a]	B27	.[–].....[–]
HLA-B*3904[a]	B27	.[–].....[–]
HLA-B*3905	.	.[–].....[–]
HLA-B*3909	B39[e]	.[RH].....[L-]
HLA-B40	B44[e]	E[F-].....[L-]
HLA-B*4001	B44	.[E-].....[L-]
HLA-B*4002	B44[e]	.[E-].....[L-]
HLA-B*4006	B44	.[E-].....[VA]
HLA-B*4101[a]	B44[c]	.[–].....[–]
HLA-B42	.	.[PL].....[–]
HLA-B44	B44	.[E-].....[–]
HLA-B*4402	B44	.[E-].....[FL]
HLA-B*4403	B44	E[E-].....[FW]
HLA-B*4405	B44[e]	.[E-].....[R-]
HLA-B45	.	.[–].....[–]
HLA-B*4501	B44[c]	.[E-].....[L-]
HLA-B*4601	B62	.[MI].....[YF]
HLA-B*4801	B27[d]	.[QK].....[L-]
HLA-B*4802[a]	B27[d]	.[–].....[–]
HLA-B*4901[a]	B44[c]	.[–].....[–]
HLA-B*5001[a]	B44[c]	.[QK].....[-L]
HLA-B51	B7	.[AP].....[IL]
HLA-B*5101	B7	.[AP].....[LY]
HLA-B*5102	B7[e]	.[PA].....[IV]
HLA-B*5103	.	.[FG].....[YI]
HLA-B52[a]	B62	.[–].....[–]
HLA-B*5201	B62[e]	.[QF].....[VF]

Table 1 (continued)

HLA type	Supertype	Amino acid motif
HLA-B53	B7[e]	.[P-].....[W-]
HLA-B*5301	B7	.[P-].....[FL]
HLA-B*5401	B7	.[P-].....[A-]
HLA-B*5501	B7	.[P-].....[A-]
HLA-B*5502	B7	.[P-].....[AV]
HLA-B*5601	B7	.[P-].....[AL]
HLA-B*5602[a]	B7	.[–].....[–]
HLA-B57	B58	.[AS].....[WT]
HLA-B*5701	B58	.[ST].....[WF]
HLA-B*5702	B58	.[TS].....[WF]
HLA-B58	B58	.[–].....[–]
HLA-B*5801	B58[e]	.[TS].....[WF]
HLA-B*5802	B58[e]	.[ST].....[FM]
HLA-B*6701	B7	.[P-].....[L-]
HLA-B*7301	B27	.[R-].....[P-]
HLA-B*7801	B7	.[GP].....[S-]

[a] Allele is not in SYFPEITHI
[b] X-ray structure exists
[c] Hypothetical supertype assignment according to Sette and Sidney (1999)
[d] Tentative supertype assignment according to Sette and Sidney (1999)
[e] Our supertype assignment

corresponds to the height of a column of letters. The information content I_p on each position p is defined as $\log_2(20)+\Sigma_a[p_{ap}\log_2(p_{ap})]$. The information content is a measure of the degree of conservation and lies within the range of 0 (no conservation — all amino acids are equally probable) and $\log_2(20)=4.3$ (full conservation — only a single amino acid is observed at that position). The height of each letter within the columns is proportional to the frequency p_{ap} of the corresponding amino acid a at that position p. The amino acids are colored according to their properties: acidic [DE], *red*; basic [HKR], *blue*; hydrophobic [ACFILMPVW], *black*; neutral [GNQSTY], *green*.

For most class II molecules relatively few binding peptides are known. We therefore calculated the similarities between different alleles based on published specificity matrices. Specificity matrices for HLA class II molecules were downloaded from the Propred website (http://www.imtech.res.in/raghava/propred/page4.html). The list of alleles is given in Table 2. The matrices were constructed by Singh and Raghava (2001) using the TEPITOPE (http://www.vaccinome.com) method (Hammer et al. 1994; Sturniolo et al. 1999). To test whether the matrices in the Propred server were similar to those in the TEPITOPE program, we ran test sequences through both programs as well as our own implementation using the matrices from Propred. Except for minor round-off errors the calculations gave similar results. The matrix scores were used to estimate the amino acid frequencies at different positions in the motif assuming that the matrix score is proportional to a log-odds score. The odds score is defined as the probability of observing amino acid a in positions p in a binding peptide relative to the probability of observing that amino acids in proteins in general. Thus:

$$p_{ap} = \exp(s_{ap})q_a / (\Sigma_i \exp(s_{ip})q_i),$$

where s_{ap} is the matrix score of amino acid a on position p, and q_a is the background frequency of amino acid.

Table 2 A list of the HLA class II alleles used in this study. The list contains the allele, serotype, pocket profile, and our supertype assignment. The pocket profiles used in assembly of virtual DR matrices are from Sturniolo and co-workers (1999). For each allele, the list of *numbers in square brackets* denotes which pocket specificity has been used to construct the profile for position 1, 4, 6, 7, and 9 (positions 2 and 3 were derived from the DRB1*0401 matrix). The matrix for HLA-DRB1*0421 could not be found at the Propred website (http://www.imtech.res.in/raghava/propred/page4.html)

Allele	Serotype	Pocket profile	Supertype
*HLA-DRB1*0101*	DR1	[1;1;1;1;1]	1
*HLA-DRB1*0102*	DR1	[2;1;1;1;1]	1
*HLA-DRB1*0301*	DR3	[2;3;3;3;2]	3
*HLA-DRB1*0305*	DR3	[1;3;3;3;3]	3
*HLA-DRB1*0306*	DR3	[2;3;3;4;3]	3
*HLA-DRB1*0307*	DR3	[2;3;3;4;3]	3
*HLA-DRB1*0308*	DR3	[2;3;3;4;3]	3
*HLA-DRB1*0309*	DR3	[1;3;3;3;2]	3
*HLA-DRB1*0311*	DR3	[2;3;3;4;3]	3
*HLA-DRB1*0401*	DR4	[1;4;4;4;3]	4
*HLA-DRB1*0402*	DR4	[2;5;4;5;3]	4
*HLA-DRB1*0404*	DR4	[2;6;4;6;3]	4
*HLA-DRB1*0405*	DR4	[1;6;4;6;5]	4
*HLA-DRB1*0408*	DR4	[1;6;4;6;3]	4
*HLA-DRB1*0410*	DR4	[2;6;4;6;5]	4
*HLA-DRB1*0421*	DR4	[1;4;4;4;2]	4
*HLA-DRB1*0423*	DR4	[2;6;4;6;3]	4
*HLA-DRB1*0426*	DR4	[1;4;4;4;3]	4
*HLA-DRB1*0701*	DR7	[1;8;5;8;4]	7
*HLA-DRB1*0703*	DR7	[1;8;5;8;4]	7
*HLA-DRB1*0801*	DR8	[1;9;3;9;5]	8
*HLA-DRB1*0802*	DR8	[1;9;3;9;3]	8
*HLA-DRB1*0804*	DR8	[2;9;3;9;3]	8
*HLA-DRB1*0806*	DR8	[2;9;3;9;5]	8
*HLA-DRB1*0813*	DR8	[1;9;3;6;3]	8
*HLA-DRB1*0817*	DR8	[1;9;3;7;5]	8
*HLA-DRB1*1101*	DR11	[1;7;3;7;3]	11
*HLA-DRB1*1102*	DR11	[2;11;3;11;3]	13
*HLA-DRB1*1104*	DR11	[2;7;3;7;3]	11
*HLA-DRB1*1106*	DR11	[2;7;3;7;3]	11
*HLA-DRB1*1107*	DR11	[2;3;3;3;3]	3
*HLA-DRB1*1114*	DR11	[1;11;3;11;3]	13
*HLA-DRB1*1120*	DR11	[1;11;3;11;2]	13
*HLA-DRB1*1121*	DR11	[2;11;3;11;3]	13
*HLA-DRB1*1128*	DR11	[1;7;3;7;2]	11
*HLA-DRB1*1301*	DR13	[2;11;3;11;2]	13
*HLA-DRB1*1302*	DR13	[1;11;3;11;2]	13
*HLA-DRB1*1304*	DR13	[2;11;3;11;5]	13
*HLA-DRB1*1305*	DR13	[1;7;3;7;2]	11
*HLA-DRB1*1307*	DR13	[1;7;3;9;3]	11
*HLA-DRB1*1311*	DR13	[2;7;3;7;3]	11
*HLA-DRB1*1321*	DR13	[1;7;3;7;5]	11
*HLA-DRB1*1322*	DR13	[2;11;3;11;3]	13
*HLA-DRB1*1323*	DR13	[1;11;3;11;3]	13
*HLA-DRB1*1327*	DR13	[2;11;3;11;2]	13
*HLA-DRB1*1328*	DR13	[2;11;3;11;2]	13
*HLA-DRB1*1501*	DR2	[2;2;2;2;1]	15
*HLA-DRB1*1502*	DR2	[1;2;2;2;1]	15
*HLA-DRB1*1506*	DR2	[2;2;2;2;1]	15
*HLA-DRB5*0101*	DR2	[1;10;6;10;6]	51
*HLA-DRB5*0105*	DR2	[1;10;6;10;6]	51

Results

HLA-A and HLA-B

Table 1 lists the classification of HLA class I types into supertypes by Sette and Sidney (1999). Each of the 150 alleles shown in Table 1 is either described in the Sette and Sidney paper or appears in the SYFPEITHI database (Rammensee et al. 1999).

Log-odds weight matrices were calculated for each allele in the SYFPEITHI database using Gibbs sampling as described in Materials and methods and binding motifs

Fig. 1 Logos displaying the binding motifs for HLA-A molecules. The height of each column of letters is equal to the information content (in bits) at the given positions in the binding motif. The relative height of each letter within each column is proportional to the frequency of the corresponding amino acid at that position

were visualized using sequence logos. The logos showing the specificities for these HLA-A and HLA-B molecules are shown in Figs. 1 and 2. The differences in specificities of the different alleles can be seen on the logos. The logo for *A*0201* for example shows a preference for hydrophobic amino acids both on position 2 and 9, while the logo for *A*1101* shows that this allele only have a preference for hydrophobic amino acids in position 2, but basic amino acids in position 9.

Figures 3 and 4 show clusterings based on the specificities for HLA-A and HLA-B, respectively. For the HLA-A alleles these trees were made only for those alleles where at least five sequences with a length of at least nine amino acids could be found in the SYFPEITHI

Fig. 2 Logos displaying the binding motifs for HLA-B molecules

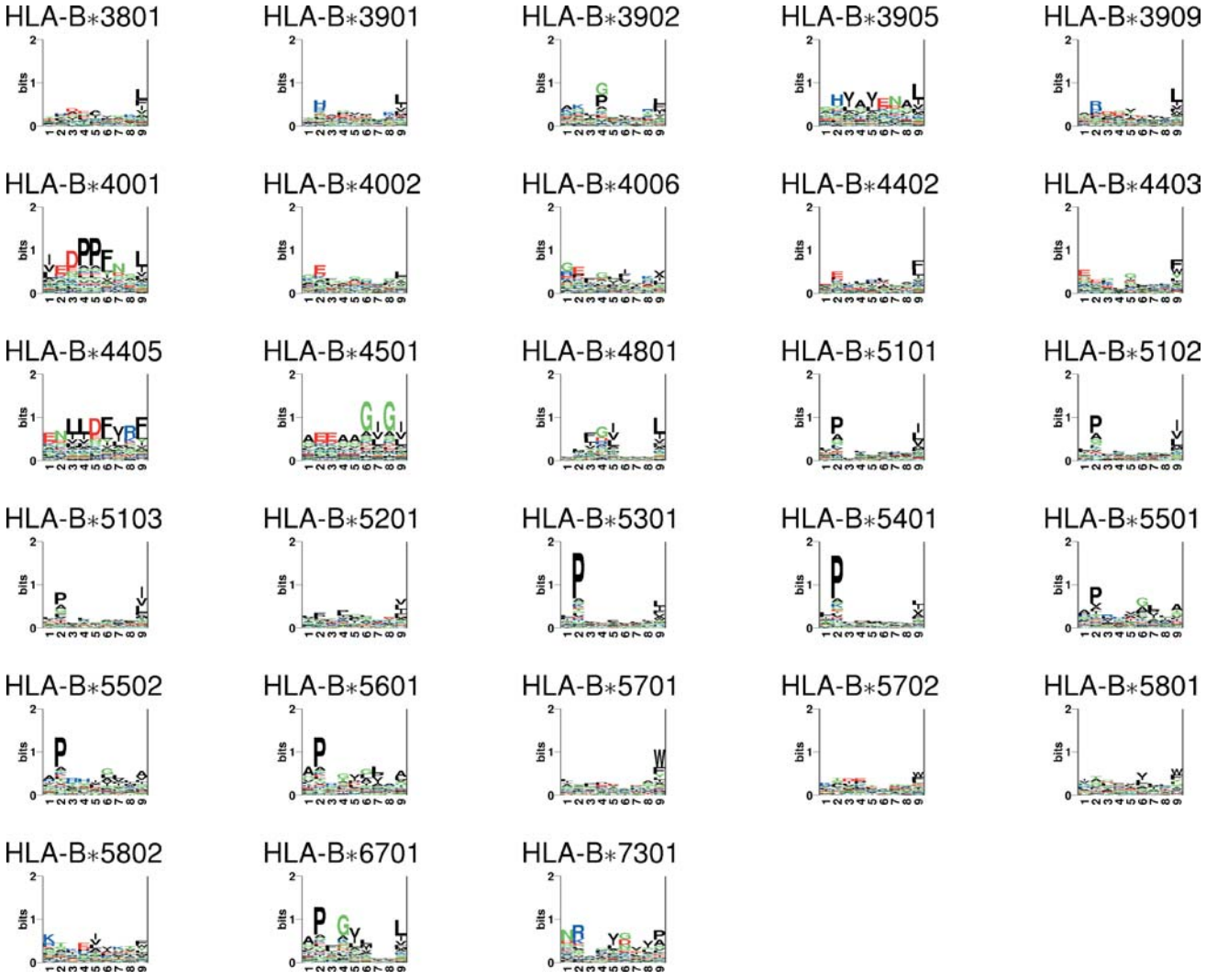

Fig. 2 (continued)

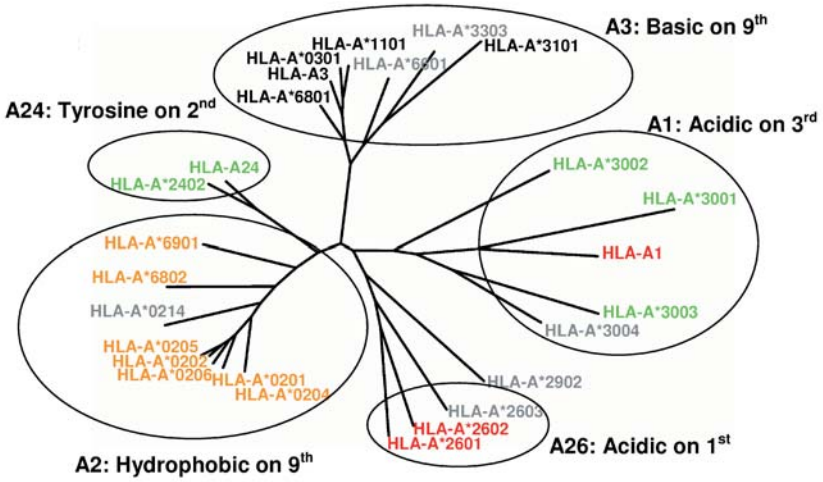

Fig. 3 Tree showing clustering of for HLA-A specificities. The alleles are colored according to the supertype classification by Sette and Sidney (1999): A1, *red*; A2, *orange*; A3, *black*; A24, *green*; A29 and non classified alleles, *grey*

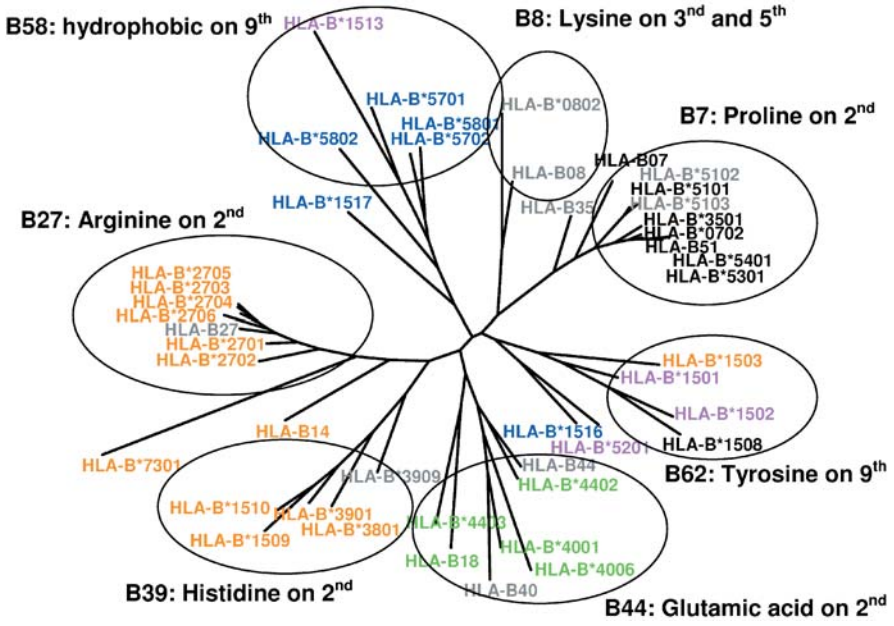

Fig. 4 Tree showing clustering of HLA-B specificities. The alleles are colored according to the supertype classification by Sette and Sidney (1999): B7, *black*; B27, *orange*; B44, *green*; B58, *blue*; B62, *violet*; non classified alleles and outliers (B8 and B46), *grey*

database, and for the HLA-B tree only alleles where at least 15 peptide sequences could be found were included. This means that not all alleles in Table 1 are shown in these figures. The names of the alleles in the trees are colored according to the classification by Sette and Sidney (1999), and the unclassified alleles are shown in grey. The trees were constructed using the bootstrap method described in Materials and methods. By visual inspection of the simple motifs the results shown in Table 1 were extracted. Sette and Sidney (1999) explicitly assigned 109 of the alleles to a supertype. We have assigned an additional 22 alleles/serotypes to a supertype based on the name and specificity listed in Table 1, the information in the SYFPEITHI database, *The HLA facts book* (Marsh et al. 2000), and the logos and trees in Figs. 1, 2, 3, and 4. These are indicated in Table 1. Some of the supertypes defined by Sette and Sidney (1999) seem to contain alleles with specificities, which are quite diverse from the other alleles in the supertype, and in 12 cases we changed the assignment given by Sette and Sidney (1999). We assign six alleles to be outliers and the remaining 13 we cannot classify.

HLA-A supertypes

The tree describing the HLA-A alleles is characterized by five clusters: A1, A2, A3 A24, and A26 cluster. The corresponding branch bootstrap values are 0.37, 0.39, 0.59, 0.98, and 0.38, respectively.

A2 supertype—hydrophobic amino acids in position 9

Our definition of this supertype largely overlaps with the definition by Sette and Sidney (1999). We suggest adding the unassigned *HLA-A*0209*,*HLA-A*0214* and *HLA-A*0217* to the A2 supertype.

A3 supertype—basic amino acids in position 9

We assign A*3303 and A*6601 to the A3 supertype characterized by basic amino acids in position 9. The other alleles in the cluster follow the classification suggested by Sette and Sidney.

A1 supertype—acidic amino acids in position 3

Here our clustering shows a large difference to the A1 supertype defined by Sette and Sidney. The clustering suggests splitting the A1 supertype into two clusters. One cluster is the A1 cluster, which contains the *A1* and *A*3001–4* alleles based on their common preference for acidic amino acids at position 3, and Y or F at position 9. The other cluster is a proposed new A26 supertype. Sette and Sidney (1999) have assigned *HLA-A*3001–3* tentatively to an A24 supertype together with the alleles of *A*24* and *A*2402*. The bootstrap branch value for the A24 cluster suggested by Sette and Sidney (*A24, A*2402, A*3001, A*3002* and *A*3003*) is found to be 0.02. Also the bootstrap value for a cluster containing both the *A2601*, *A*2602* and the *A1* alleles is below 0.01. These numbers stands in contrast to the bootstrap branch values for the A1 cluster and the new A26 clusters, which are 0.37 and 0.38, respectively.

Proposed new A26 supertype—acidic amino acids in position 1

*HLA-A*2601–3* have E/D in position 1 rather than in position 3 in HLA-A1. This difference is consistent with the motif descriptions by Marsh et al. (2000), and we suggest that these alleles form a new supertype. Including *HLA-A*2902* in the A26 supertype leads to a decrease in the branch bootstrap value from 0.38 to 0.12, so we suggest leaving this allele as an outlier.

A24-supertype—tyrosine or hydrophobic in position 2

We assign *A24* and *A*2402* to the A24 supertype. These alleles have a bootstrap value of 0.98.

In summary, we suggest a redefinition of the A1 supertype in that the *HLA-A*2601/2* alleles may form a new separate A26 supertype. The following alleles remain unclassified: *A23*, *A31*, *A32*, *A*7401*.

HLA-B supertypes

The HLA-B supertype tree contains many more alleles than the HLA-A tree. In order to make the clustering analysis more feasible and clear we limit the HLA-B clustering to the alleles where at least 15 peptide sequences are available in either the SYPHEITHI or MHCPEP databases. This limits the analysis to 45 HLA-B alleles out of 99 available.

B7 supertype—proline in position 2

The definition of the B7 supertype by Sette and Sidney (1999) largely corresponds to the B7 cluster in Fig. 4, however with one important exception. Sette and Sidney place the *HLA-B*1508* in the B7 supertype. We find this scenario very unlikely. The bootstrap branch value for the Sette and Sidney B7 cluster is 0.042, whereas the corresponding value for our B7 cluster, excluding the *HLA-B*1508* allele, is 0.66.

New B8 supertype—lysine in position 3 and 5

The B8 alleles were defined as an outlier group by Sette and Sidney (1999) and the specificities of *B*08*, and *B*0802* define a cluster with a corresponding branch bootstrap value of 0.72 in Fig. 4.

B62 supertype—tyrosine in position 9

The B62 cluster shown in Fig. 4 is restricted to contain only the alleles *HLA-B*1503*, *HLA-B*1501*, *HLA-B*1502*, and *HLA-B*1508*. The bootstrap branch value for the cluster is 0.62. If we include the alleles *HLA-B*1516* and *HLA-B*5201* the bootstrap value drops to 0.06. We hence suggest leaving out these two alleles as outliers. The bootstrap branch value for the B62 cluster defined by Sette and Sidney is $<10^{-3}$. This low branch value is due to the misplacement of the *HLA-B*1513* and *HLA-B*5201* alleles in the B62 supertype, and the *HLA-B*1508* allele in the B7 supertype.

B27 supertype—basic in position 2

The definition of the B27 supertype by Sette and Sidney has a branch bootstrap value $<10^{-3}$, whereas the B27 cluster defined in Fig. 4 has a branch value of 0.22. The low branch value for the Sette B27 supertype is due to a misplacement of the *HLA-B*1503* allele. As described above, we suggest placing this allele in the B62 cluster. If we further spit the B27 cluster up in two sub-clusters leaving out the *HLA-B*7301* and the *HLA-B*14* alleles as outliers, we get bootstrap branch values for the remaining of B27 cluster of 0.62. The other alleles form a new *B39* supertype with a bootstrap branch value of 0.41. The B39 cluster contains the alleles of *HLA-B3909, HLA-B*3901, HLA-B*3801, HLA-B*1510*, and *HLA-B*1509*. These alleles have similar B and F pocket residues as defined by Sette and Sidney (1999). The redefined B27 cluster contains the alleles of *HLA-B*2705, HLA-B*2703, HLA-B*2704, HLA-B*2706, HLA-B27, HLA-B*2701,* and *HLA-B*2702*.

B44 supertype—glutamic acid in position 2

The definition of the B44 cluster largely corresponds to the supertype definition of Sette and Sidney. We include the alleles of *HLA-B*40* and *HLA-B*44* in the supertype. The bootstrap branch value for the cluster is 0.36.

B58 supertype—hydrophobic at position 9

The branch bootstrap value for the B58 cluster defined in Fig. 4 is found to be 0.42. If we include the *HLA-B*1517* alleles, this value drops to 0.18, and we thus suggest leaving out these alleles as outliers. The bootstrap value for the Sette and Sidney B58 supertype is 0.156. Leaving out the *HLA-B*1516* and *HLA-B*1517* alleles as outliers, as described above, and including the *HLA-B*1513* allele we obtain the B58 cluster defined in Fig. 4.

In summary, we find good overall consistency between the supertypes defined by Sette and Sidney and the HLA-B tree. We suggest a definition of a novel B8 supertype including the *HLA-B*08* and *HLA-B0802* alleles as well as splitting the B27 supertype into two: a B39 supertype and a B27 supertype. Further we give suggestions as to how some of the alleles could be rearranged so as to increase the likelihood of the clustering. The following HLA-B alleles remain unclassified: *B17*, *B*1801*, *B22*, *B37*, *B*3905*, *B42*, *B45* (two sequences in SYFPEITHI,

both with E in position 2), *B*5301*. Only one or two sequences were found in SYFPEITHI for these alleles except for *B17*, where five sequences were found.

Do cross loci supertypes exist?

The alleles within the supertypes defined by Sette and Sidney (1999) are all encoded by either the *A* or the *B* locus. If we make a tree of all the HLA-A and HLA-B alleles included in the analysis described above, we find no mixing of the HLA-A and HLA-B clusters. Only the outliers *HLA-B*1516* and *HLA-A*2902* mix with a cluster defined by the opposite locus. The *HLA-B*1516* allele clusters within the A1 supertype consistent with a preference for T and S at position 2, and a preference for Y, F, L, and V at position 9. The *HLA-A2902* allele clusters within the B44 supertype consistent with a preference for E at position 2 and a preference for Y in position 9 found in both motifs. The *A*2902* molecule used for elution of peptides is often purified from the EBV transformed cell line SWEIG, which coexpresses *B*4402*, and the apparent similarity may be an experimental artefact caused by cross reactiveness of the antibody used for purification from this cell line. This unrelatedness of HLA-A and HLA-B molecules may be a direct result of evolutionary pressure on the immune system to provide optimal protection against infectious diseases. To have optimal peptide coverage, it is beneficial for the immune system to have a highly diverse set of HLA specificities. A simple way to achieve this could be to have the *HLA-A* and *HLA-B* alleles evolve in an orthogonal manner.

HLA-DR

Sequence logos were constructed to visualize the specificities of different HLA class II molecules (Fig. 5). By visual inspection of these matrices it is clear that some are quite similar. In order to quantify these similarities, we calculated the distance between all pairs of matrices, as described in Materials and methods. These distances were used to construct a tree visualizing the similarities between the peptides that each allele binds (Fig. 6). Based on this tree, we suggest dividing the HLA-DR molecules into nine clusters or supertypes. The clusters may be represented by *DRB1*0101* (1, 0.92), *DRB1*0301* (3, 0.65), *DRB1*0401* (4, 0.45), *DRB1*0701* (7, 1.0), *DRB1*0813* (8, 0.52), *DRB1*1101* (11, 0.32), *DRB1*1301* (13, 0.39), *DRB1*1501* (15, 0.82), and *DRB5*0101* (51, 0.95). Here the numbers in parentheses after each allele name corresponds to the supertype name assigned to each cluster in Fig. 5, and the cluster bootstrap branch value, respectively. The alleles in Fig. 5 are colored according to the serotype. The clustering roughly corresponds to the serotype classification, but with some important exceptions. Note, for example, the mixing of the DR11, and DR13 sequences and that *DRB1*1107* clusters with the DR3 sequences. The bootstrap value for the DR11 and DR13 serotype clusters is, for instance, $<10^{-3}$ and the bootstrap value for the DR3 serotype cluster excluding the *DRB1*1107* allele is 0.03. The matrices were constructed under the assumption that the amino acids at different positions contribute independently (by binding to a pocket in the HLA molecule) to the binding of the peptide. Furthermore, it is also assumed that HLA molecules with the same amino acids in a given pocket will have the same specificity profile (Hammer et al. 1997). Different matrices thus have the same profile at a given position, if the corresponding HLA molecules share the amino acids lining the pocket for that position. In Table 2, it can be seen that *DRB1*1107* and *DRB1*0305* only differ in one binding pocket. This is, hence, consistent with placing the *DRB1*1107* allele in the DR3 supertype. Similarly it seems that the alleles placed in the DR11 and DR13 supertypes in most cases share three out of the five pockets specificities.

Experimental verification of supertypes

To verify the clustering suggested in the above analysis, we constructed weight matrices for all the class I alleles in this study as described in Materials and methods. We then use these weight matrices to predict the binding affinity for sets of peptides where the binding affinity to a specific HLA allele had been measured experimentally as described by Sylvester-Hvid et al. (2002). The alleles for which we had experimental binding information were *HLA-A*0101* (A1), *HLA-A*0202* (A2), *HLA-A*0301* (A3), *HLA-A*1101* (A3), *HLA-A*3101* (A3 outlier), *HLA-B*2705* (B27), *HLA-B*1501* (B62), *HLA-B*5801* (B58), and *HLA-B*0702* (B7). Here the name written in parentheses refers to the supertype classification. We calculate the linear correlation coefficient, also known as Pearson's r (Press et al. 1992), between the prediction score and the log of the measured binding affinity. In this manner we expect to find that alleles with similar specificity to that of the allele used in the experiments will have a positive correlation, and that other alleles will have a correlation close to zero. In general, this calculation supports most the results obtained from the clustering analysis. Below, the results for each supertype are described.

B7

We find that all the alleles, except the *HLA-B*35* allele, classified in this supertype can predict the binding affinity for the *HLA-B*0702* allele data with a correlation greater than 0.60. The calculation also supports the hypothesis that the *HLA-B*1508* is not part of the B7 supertype, since the correlation between the predicted and measured affinities for the allele is only 0.11.

Fig. 5 Logo showing the peptide binding specificity for 50 different HLA class II molecules

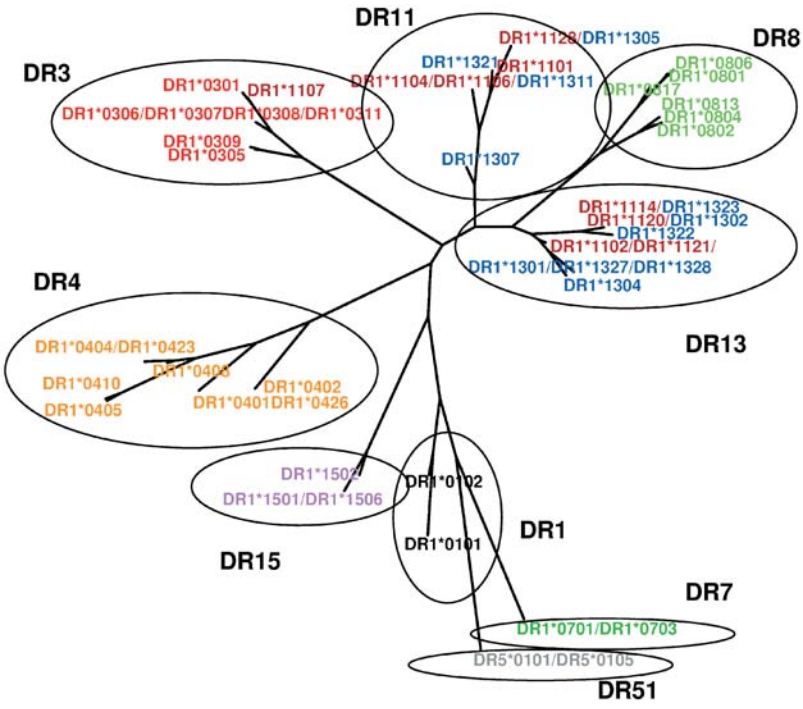

Fig. 6 Tree showing the clustering of 50 different HLA class II molecules based on their peptide binding specificity. The proposed clusters are encircled and labeled

B27

Here the predictions support the suggestion that the B27 must be split into two groups. All the members of the B27 supertype defined in our analysis can predict the binding affinities of the HLA-B*2705 peptides with a correlation greater than 0.45, whereas the largest correlation between the alleles of the new B39 supertype and the HLA-B*2705 data is 0.12.

A1

The HLA-A1 matrix can predict the binding affinities of the HLA-A*0101 data with a Pearson correlation of 0.71, whereas the correlation of the predictions using the HLA-A*2601 and HLA-A*2602 matrix is 0.25 and 0.16, respectively. This result thus supports our earlier finding that the *HLA-A*2601* and *HLA-A*2602* alleles do not belong in the A1 supertype.

A2

All alleles of the A2 supertype can predict the binding affinity of the HLA-A*0201 peptides with a Pearson correlation greater than 0.40.

A3

For the A3 supertype, we have experimental data for three distinct *HLA* alleles (*HLA-A*0301*, *HLA-A*1101*, and *HLA-A*3101*). We perform the calculation for each of the three alleles. We find that the matrices based on *HLA-A3, HLA-A*0301, HLA-A*1101, HLA-A*6801* can predict binding affinities to *HLA-A*0301* and *HLA-A*1101* well. Binding to *HLA-A*3101* is poorly predicted by all matrices. This supports the classification in Fig. 3 that *HLA-A*3101* is only remotely related to the other alleles in the A3 supertype.

B58

The binding affinity to the *HLA-B*5801* allele can be predicted with a correlation greater than 0.4 for all alleles in the B58 supertype. The correlation when using the *HLA-B*1516* matrix is found to be 0.09, supporting the suggestion that this allele is not part of the B58 supertype. An interesting finding is that both *HLA-B*2702* and *HLA-B*2701* can predict the binding of the *HLA-B*5801* peptides with a correlation of 0.4, indicating that the two supertypes might share some part of the peptide-binding repertoire.

B62

The analysis of the data for the *HLA-B*1501* allele revealed some interesting features of the B62 supertype. The correlation between the *HLA-B*1501* data and the prediction using the *HLA-B*5201* matrix is found to be 0.0, whereas the correlation using the *HLA-B1501* matrix is 0.50. This supports the hypothesis that the *HLA-B*5201* allele is not part of the B62 supertype. A surprising

finding is that the correlations between the *HLA-B*1501* data and predictions using the A3 supertype alleles (*HLA-A*1101*, *HLA-A*6801* and *HLA-A3*) give a Pearson correlation close to 0.50, suggesting a significant specificity overlap between the B62 and the A3 supertypes.

Discussion

Specificity matrices were constructed for all HLA-A and -B alleles in the SYFPEITHI database and these were clustered using a nearest neighbor algorithm. Our clustering largely corresponds to that found by Sette and Sidney (1999), even though some marked differences are found. The classification was made in a quantitative way that allows for fast classification of new alleles.

HLA class II molecules were clustered based on the specificity of the molecules. Based on this clustering, eight supertypes were defined. It has been reported that a set of DR molecules, including *DRB1*0101, DRB1*0401, DRB1*0701, DRB5*0101, DRB1*1501, DRB1*0901,* share overlapping peptide binding repertoires (Southwood 1998). Based on our findings this clustering seems to be too broad, and we would rather suggest that these molecules represent the diversity of the HLA-DR molecules.

Our clustering can easily be recalculated if new data become available in the future. The availability of data is expected to increase as the Epitope Immune Database, a large-scale epitope discovery project funded by the NIH, is started.

The clusters define groups of alleles with similar binding specificities. In order to obtain broad coverage of the human population with an epitope-based vaccine, it must be ensured that most people from all ethnic groups have a HLA molecule with specificity for at least one of the peptides in the vaccine. This can, in turn, be obtained making sure that the specificity defined by each cluster is covered by one peptide in the vaccine.

Acknowledgements This work is part of the IHWG and was supported by the Danish National Research Foundation, the Danish MRC (grant 22-01-0272), the 5th Framework Programme of the European Commission (grant QLRT-1999-00173), NIH (grant AI49213-02) and the Bioinformatic Program of the Dutch Science Foundation (NWO PGBMI 015).

References

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25:3389–3402

Boeckmann B, Bairoch A, Apweiler R, Blatter MC, Estreicher A, Gasteiger E, Martin MJ, Michoud K, O'Donovan C, Phan I, Pilbout S, Schneider M (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. Nucleic Acids Res 31:365–370

Brusic V, Rudy G, Harrison LC (1998) MHCPEP, a database of MHC-binding peptides: update 1997. Nucleic Acids Res 26:368–371

Cover TM, Thomas JA (1991) Elements of information theory. Wiley, New York

Eddy SR (1998) Related profile hidden Markov models. Bioinformatics 14:755–763

Guercio MF del, Sidney J, Hermanson G, Perez C, Grey HM, Kubo RT, Sette A (1995) Binding of a peptide antigen to multiple HLA alleles allows definition of an A2-like supertype. J Immunol 1995 154:685–693

Gulukota K, DeLisi C (1996) Related HLA allele selection for designing peptide vaccines. Genet Anal 13:81–86

Hammer J, Bono E, Gallazzi F, Belunis C, Nagy Z, Sinigaglia F (1994) Precise prediction of major histocompatibility complex class II-peptide interaction based on peptide side chain scanning. J Exp Med 180:2353–2358

Hammer J, Sturniolo T, Sinigaglia F (1997) HLA class II peptide binding specificity and autoimmunity. Adv Immunol 66:67–100

Henikoff S, Henikoff JG (1994) Position-based sequence weights. J Mol Biol 243:574–578

Ishioka GY, Fikes J, Hermanson G, Livingston B, Crimi C, Qin M, del Guercio MF, Oseroff C, Dahlberg C, Alexander J, Chesnut RW, Sette A (1999) Utilization of MHC class I transgenic mice for development of minigene DNA vaccines encoding multiple HLA-restricted CTL epitopes. J Immunol 162:3915–3925

Lawrence CE, Altschul SF, Boguski MS, Liu JS, Neuwald AF, Wootton JC (1993) Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. Science 262:208–214

Lyngs, Pedersen CN, Nielsen HR (1999) Metrics and similarity measures for hidden Markov models. Proc Int Conf Intell Syst Mol Biol:178–186

Marsh SGE, Parham P, Barber LD (2000) The HLA facts book. Academic Press, London

Nielsen M, Lundegaard C, Worning P, Lauemr SL, Buus S, Brunak S, Lund O (2004) Improved prediction of MHC class II epitopes using a Gibbs sampling approach. Bioinformatics (in press)

Press WH, Flannery BP, Teukolsky SA, Vetterling, WT (1992) Numerical recipies in C. Cambridge University Press, Cambridge

Rammensee HG, Friede T, Stevanoviic S (1995) MHC ligands and peptide motifs: first listing. Immunogenetics 41:178–228

Rammensee H, Bachmann J, Emmerich NP, Bachor OA, Stevanovic S (1999) SYFPEITHI: database for MHC ligands and peptide motifs. Immunogenetics 50:213–219

Rodriguez F, An LL, Harkins S, Zhang J, Yokoyama M, Widera G, Fuller JT, Kincaid C, Campbell IL, Whitton JL (1998) DNA immunization with minigenes: low frequency of memory cytotoxic T lymphocytes and inefficient antiviral protection are rectified by ubiquitination. J Virol 72:5174–5181

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

Schneider TD, Stephens RM (1990) Sequence logos: a new way to display consensus sequences. Nucleic Acids Res 18:6097–100

Sette A, Sidney J (1999) Nine major HLA class I supertypes account for the vast preponderance of HLA-A and -B polymorphism. Immunogenetics 50:201–212

Sidney J, del Guercio MF, Southwood S, Engelhard VH, Appella E, Rammensee HG, Falk K, Rotzschke O, Takiguchi M, Kubo RT, et al (1995) Several HLA alleles share overlapping peptide specificities. J Immunol 154:247–259

Sidney J, Grey HM, Southwood S, Celis E, Wentworth PA, del Guercio MF, Kubo RT, Chesnut RW, Sette A (1996) Definition of an HLA-A3-like supermotif demonstrates the overlapping peptide binding repertoires of common HLA molecules. Hum Immunol 45:79–93

Singh H, Raghava GP (2001) ProPred: prediction of HLA-DR binding sites. Bioinformatics 17:1236–237

Southwood S, Sidney J, Kondo A, del Guercio MF, Appella E, Hoffman S, Kubo RT, Chesnut RW, Grey HM, Sette A (1998) Several common HLA-DR types share largely overlapping peptide binding repertoires. J Immunol 160:3363–3373

Sturniolo T, Bono E, Ding J, Raddrizzani L, Tuereci O, Sahin U, Braxenthaler M, Gallazzi F, Protti MP, Sinigaglia F, Hammer J (1999) Generation of tissue-specific and promiscuous HLA ligand databases using DNA microarrays and virtual HLA class II matrices. Nat Biotechnol 17:555–561

Sylvester-Hvid C, Kristensen N, Blicher T, Ferre H, Lauemoller SL, Wolf XA, Lamberth K, Nissen MH, Pedersen LO, Buus S (2002) Establishment of a quantitative ELISA capable of determining peptide — MHC class I interaction. Tissue Antigens 59:251–258