Chapter 13

MHC Polymorphism

As a result of fast evolution (thanks to the short generation time of microorganisms), hosts are under constant selection pressure to invent processes that counteract pathogenic invasion. Since the generation time of the vertebrate host is much longer than that of pathogens, obviously the evolution of the host has a much slower pace. The polymorphism of the MHC molecules is one result of this coevolution between the host and its pathogens. The polymorphism is one of the major reasons why mammalian organisms cannot be eradicated by infections by a single pathogen: Pathogens that escape from presentation by the MHC molecules of one individual may not be able to escape the presentation by another individual carrying a different MHC molecule.

13.1 What Causes MHC Polymorphism?

Although there have been extensive debates over the selection pressures leading to the high polymorphism of MHC molecules, there is still not a widely accepted model for a mechanism (see Apanius et al. [1997] for a detailed review). The common view is that MHC polymorphism arises because of the heterozygote advantage. Different MHC molecules bind different peptides, and thus present different parts of a pathogen to T cells. If a host is heterozygous in its MHC loci, it can thus provide a broader immune response, which in turn would make pathogenic adaptation more difficult. This theory, known as the theory of overdominance or heterozygote advantage [Hughes and Nei, 1988, 1989, 1992], is supported by recent studies on HIV-1 patients. Carrington and O'Brien [2003] have reviewed data showing that the degree of MHC heterozygocity correlates with a delayed onset of progress to AIDS.

There exist a number of mathematical models focusing on the heterozy-

gous advantage as the main reason for MHC polymorphism. Work in the general area of population genetics models suggests that the heterozygous advantage is sufficient to explain the high MHC polymorphism observed in several MHC loci [Maruyama and Nei, 1981, Takahata and Nei, 1990, Hughes and Yeager, 1998]. These models assume that all heterozygous individuals would have the same fitness (higher than the homozygous individuals) irrespective of the MHC molecules that they harbor. This is, however, an unrealistic assumption, as it is now well established that different MHC alleles show different degrees of protection to specific pathogens [van Eden et al., 1980, Klein et al., 1994, Hill et al., 1991]. de Boer et al. [2004] show that when the classic population genetics models are corrected for this unrealistic assumption, it is no longer possible to obtain more than 10 alleles per loci. Thus, the heterozygous advantage alone cannot explain the large MHC polymorphism observed in mammalian (and most vertebrate) populations.

An additional mechanism that could enlarge MHC polymorphism is frequency-dependent selection by host-pathogen coevolution. Since it is a good strategy for pathogens to adapt to the most common MHC alleles in a population, the rare alleles would have a selective advantage. This will in time cause the frequency of rare alleles to increase, and the common alleles will become rare. The dynamic picture arising from this scenario resembles the well known principle of red-queen dynamics from ecology [van Valen, 1973]. The studies of the snail-trematode parasite system support that such a frequency-dependent selection can take place in nature, as in this system the parasite evolves to become most virulent in the dominant host genotype. For humans HIV-1 is an example of a rapidly adapting pathogen to most common MHC alleles in the population [Trachtenberg et al., 2003, Scherer et al., 2004].

The relative role of frequency-dependent selection and heterozygote advantage is discussed extensively in the literature [Lewontin et al., 1978, Aoki, 1980, Hughes and Nei, 1988, 1989]. Recently Borghans et al. [2004] and Beltman et al. [2002] have developed a computer simulation model of coevolving hosts and pathogens to study the relative impact of these two mechanisms. This model shows that 1) the frequency-dependent selection scenario alone can account for the existence of at least 50 alleles per MHC loci, and 2) if the host population size is large enough, the MHC polymorphism does not become too dynamic, i.e., a large set of MHC alleles can persist over many host generations even though host MHC frequencies change continuously.

Many other factors such as MHC-dependent mate selection, geographic and social isolation, and strong selection pressures by severe infections (population “bottlenecks”) can influence the degree of MHC polymorphism arising in a population. The chimpanzee species is in this respect very interesting: de Groot et al. [2002] have shown that almost any chimpanzee gene is more polymorphic than human genes, probably because the chimpanzee is an older

species. However, the polymorphism of MHC genes seems to be much lower than in humans, possibly due to a strong selection pressure caused by simian immunodeficiency virus (SIV) infection. Even though host-pathogen evolution seems to be sufficient to explain the large MHC polymorphism [Borghans et al., 2004, Beltman et al., 2002], all other factors mentioned here, together with frequency-dependent selection, generate the MHC polymorphism we observe in many vertebrate populations today.

13.2 MHC Supertypes

The previous section reviewed factors that play a role in generating extremely polymorphic MHC genes. This polymorphism, although very essential to protect a population from invasion by pathogens, generates a major drawback for epitope-based vaccines, which otherwise, from many perspectives, are the most promising vaccine candidates (see chapter 11).

Each MHC molecule 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 3, 4, and 5 alleles which 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., 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 are not functionally different, i.e., they have similar specificities. The different HLA molecules have been grouped together in what is 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] de-

fined nine supertypes (A1, A2, A3, A24, B7, B27, B44, B58, B62) which were reported to cover most of the HLA-A and -B polymorphisms. 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 more than 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, and B44 covered in the range of 83.0 to 88.5% and 98.1 to 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 5 alleles from the DQ locus (DQ1, DQ2, DQ3, DQ4, DQ5) cover 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].

There are recently developed bioinformatical approaches to identification of HLA supertypes [Lund et al., 2004, Reche and Reinhertz, 2004] defining a novel measure for the difference in the specificities of different HLA molecules and using the measure to revise the HLA class I supertypes. In the work of Lund et al. [2004] also MHC supertypes for class II molecules are defined, using published specificities for a number of HLA-DR types. This work will be described in detail below.

13.2.1 A Novel Method to Cluster MHC Binding Specificities

In the first part of this section we will be dealing with how to cluster HLA Class I alleles into supertypes. The basic idea behind the approach is to construct weight matrices of binding peptides as described in chapters 6 and 8, and then use these matrices as a representation of the binding specificity of a given allele. Then all the matrices are compared and clustered by their similarity in the binding space. This is a powerful alternative to clustering based on MHC sequence similarities.

First, a data set of alleles and their binding peptides is needed: The different class I molecules used in this example can be seen in table 13.1. The corresponding HLA ligands were extracted from the SYFPEITHI [Rammensee et al., 1995, 1999] and MHCPEP [Brusic et al., 1998a] databases. All lines con-

taining 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 sampler algorithm that estimates the best scoring 9mer pattern using the Monte Carlo sampling procedure described in chapter 8. In brief, the best scoring pattern is defined in terms of highest relative entropy [Cover and Thomas, 1991] summed over a 9mer alignment. The program samples possible alignments of the sequences in the input file. For each alignment a weight matrix is calculated as $\log(p_{pa}/q_a)$, where p_{pa} is the estimated frequency of amino acid a at position p in the alignment and q_a is the background frequency of amino acid a in SWISS-PROT [Boeckmann et al., 2003]. The values for p_{pa} 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, the difference in specificities (the distance) between each pair of HLA molecules is first calculated. The distance d_{ij} between two HLA molecules (i, j) is calculated as the sum over each position in the two motifs of one minus the normalized vector products of the amino acid's frequency vectors [Lyngsø et al., 1999]:

$$d_{ij} = \sum_p (1 - \frac{p_p^i \cdot p_p^j}{|p_p^i||p_p^j|}) \tag{13.1}$$

p_p^i, and p_p^j are the vectors of 20 amino acid frequencies at position p in matrix i and j, respectively; $\cdot$ denotes the vector product and $\|$ the calculation of the Euclidian length of the vector. Dividing all distances by 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]. A set of matrices were generated by randomly taking out a column N times with replacement from the original matrix set. Here N is the motif length, which is set to 9 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. A tree for each such matrix set is then calculated. 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.

13.2.2 HLA-A and HLA-B

Log-odds weight matrices can be calculated for each allele in the SYFPEITHI database using Gibbs sampling as described above. The resulting matrices can be visualized as sequence logos, and the logos showing the specificities for the HLA-A and HLA-B molecules are listed in figures 13.1 and 13.2. The differences in specificities of the different alleles can be seen on the logos. The logo for A*0201, e.g., shows a preference for hydrophobic amino acids both on positions 2 and 9, while the logo for A*1101 shows that this allele only has a preference for hydrophobic amino acids in position 2, but basic amino acids in position 9.

Table 13.1 lists the classification of HLA class I types into supertypes by Sette and Sidney [1999]. Each of the 150 alleles shown in table 13.1 is either described in the Sette and Sidney paper or appears in the SYFPEITHI database [Rammensee et al., 1999].

Figures 13.3 and 13.4 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 database, and the HLA-B tree only for alleles where at least 15 peptide sequences were included. This means that not all alleles in table 13.1 are shown in these figures. The names of the alleles in the trees are colored according to the classification of Sette and Sidney [1999], and the unclassified alleles are shown in gray. The trees were constructed using the bootstrap method.

By visual inspection of the simple motifs the results shown in table 13.1 were extracted. Sette and Sidney [1999] explicitly assigned 109 of the alleles to a supertype. We have assigned 23 additional alleles/serotypes to a supertype based on the name and specificity listed in table 13.1, the information in the SYFPEITHI database, the HLA facts book [Marsh et al., 2000], and the logos and trees in figures 13.1 and 13.3. These are marked with an “o” in table 13.1. 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 eight cases we changed the assignment given by Sette and Sidney [1999]. We assign six alleles to be outliers and the remaining thirteen we cannot classify.

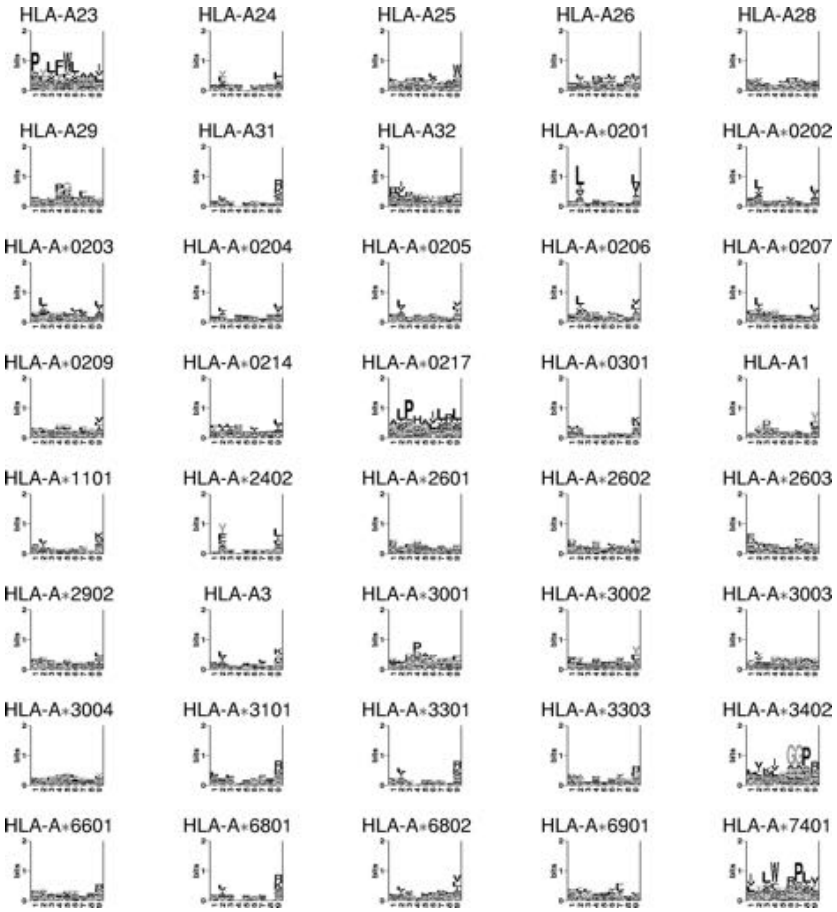
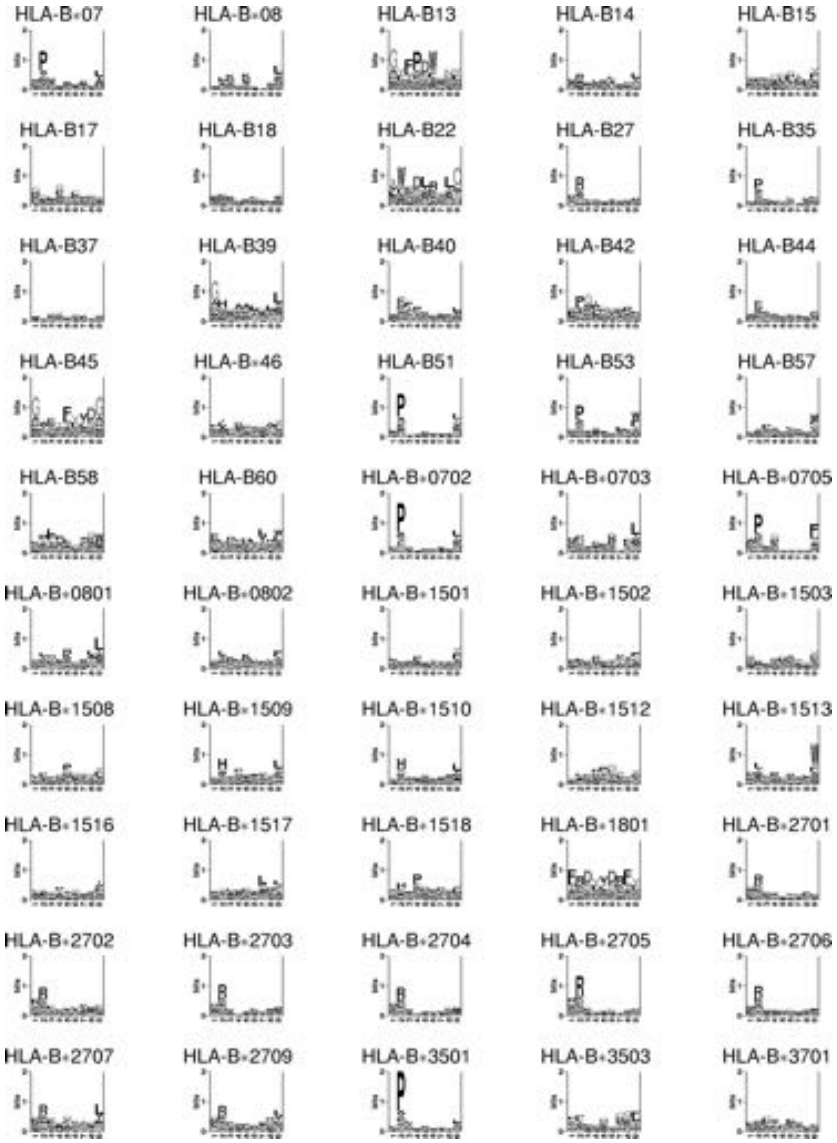


Figure 13.1: Logos displaying the binding motifs for HLA-A molecules. The height of each column of letters is equal to the information content (in bits) at the given positions in the binding motif. The relative height of each letter within each column is proportional to the frequency of the corresponding amino acid at that position. Figure reprinted from Lund et al. [2004]. See plate 22 for color version.


HLA-B*07
HLA-B*08
HLA-B13
HLA-B14
HLA-B15
HLA-B17
HLA-B18
HLA-B22
HLA-B27
HLA-B35
HLA-B37
HLA-B39
HLA-B40
HLA-B42
HLA-B44
HLA-B45
HLA-B*46
HLA-B51
HLA-B53
HLA-B57
HLA-B58
HLA-B60
HLA-B*0702
HLA-B*0703
HLA-B*0705
HLA-B*0801
HLA-B*0802
HLA-B*1501
HLA-B*1502
HLA-B*1503
HLA-B*1508
HLA-B*1509
HLA-B*1510
HLA-B*1512
HLA-B*1513
HLA-B*1516
HLA-B*1517
HLA-B*1518
HLA-B*1801
HLA-B*2701
HLA-B*2702
HLA-B*2703
HLA-B*2704
HLA-B*2705
HLA-B*2706
HLA-B*2707
HLA-B*2709
HLA-B*3501
HLA-B*3503
HLA-B*3701

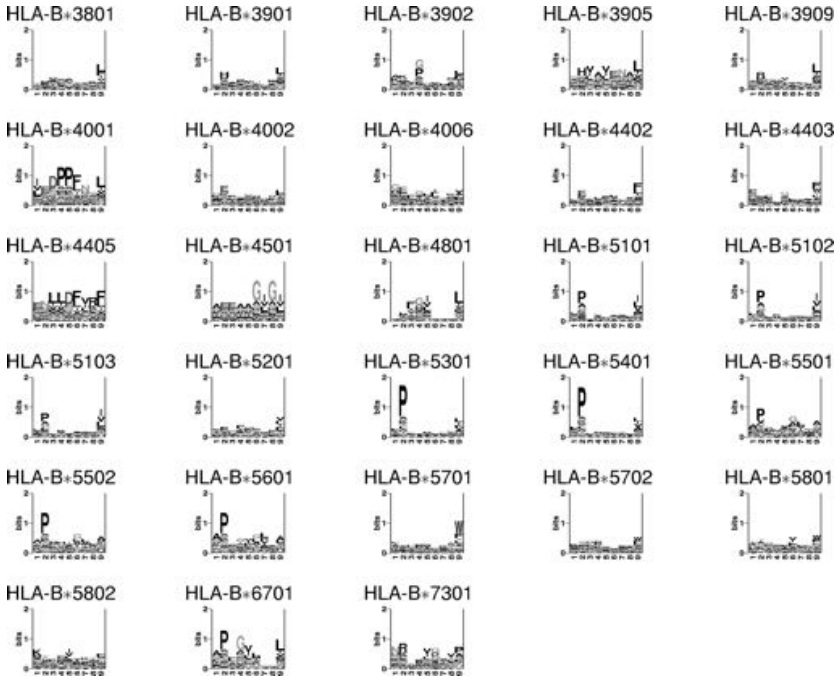


Figure 13.2: Logos displaying the binding motifs for HLA-B molecules. For details on the logo representation, see figure 13.1. Figure reprinted from Lund et al. [2004]. See plates 23 and 24 for color versions.

HLA-A1	A1	.[TS][DE].....Y	HLA-A*0101[a]	A1t	.[–].....[–]
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	A2o	.[LA].....[V-]	HLA-A*0214	A2o	.[QV].....[LV]
HLA-A*0217	A2o	.[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	A24t	.[YF].....[LF]	HLA-A*2402	A24	.[YF].....[LF]
HLA-A*2403[a]	A24t	.[–].....[–]	HLA-A*2404[a]	A24t	.[–].....[–]
HLA-A25	A1t	.[–].....[–]	HLA-A*2501[a]	A1t	.[–].....[–]
HLA-A26	A1t→A26	.[–].....[–]	HLA-A*2601	A1t→A26	E[TI].....[FY]
HLA-A*2602	A1t→A26	[DE][I.....[FY]	HLA-A*2603	A26o	E[VL].....[ML]
HLA-A*2604[a]	A1t→A26	.[–].....[–]	HLA-A28	A1→A26	.[–].....[–]
HLA-A29	outlier	.[FN].....[YC]	HLA-A*2902	outlier	K[E-].....[YL]
HLA-A30[a]	A24t	.[–].....[–]	HLA-A*3001	A24→A1	K[TF].....[FL]
HLA-A*3002	A24t→A1	R[YV].....[YK]	HLA-A*3003	A24t→A1	R[YL].....[Y-]
HLA-A*3004	A1o	K[YT].....[YL]	HLA-A31	.	.[–].....[–]
HLA-A*3101	A3	[RK]QL.....[R-]	HLA-A32	.	.[–].....[–]
HLA-A*3201[a]	A1t	.[–].....[–]	HLA-A*3301	A3	.[LV].....[RK]
HLA-A*3303	A3o	[DE]I.....[R-]	HLA-A*3402	A3o	.[V-].....[R-]
HLA-A*3601[a]	A1	.[–].....[–]	HLA-A*4301[a]	A1	.[–].....[–]
HLA-A*6601	A3o	[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-B07X	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]	B62t	.[–].....[–]
HLA-B*1302[a]	B62t	.[–].....[–]	HLA-B14	outlier	.[R-].....[LV]
HLA-B*1401[a]	B27	.[–].....[–]	HLA-B*1402[a]	B27	.[–].....[–]
HLA-B39	B62o	.[FM].....[YF]	HLA-B*1501	B62	.[QL].....[YV]
HLA-B*1502	B62	.[LQ].....[YF]	HLA-B*1503	B27t	.[QK].....[YV]
HLA-B*1508	B7	.[PV].....[YS]	HLA-B*1506[a]	B62t	.[–].....[–]
HLA-B*1509	B27→B39	.[H-].....[LF]	HLA-B*1510	B27t→B39	.[H-].....[LF]
HLA-B*1512	B62t	.[QL].....[YS]	HLA-B*1513	B62	.[IL].....[W-]
HLA-B*1514[a]	B62t	.[–].....[–]	HLA-B*1516	B58	.[TS].....[IV]
HLA-B*1517	B58	.[–].....[–]	HLA-B*1518	B27t	.[–].....[–]
HLA-B*1519[a]	B62t	.[–].....[–]	HLA-B*1521[a]	B62t	.[–].....[–]
HLA-B17	.	.[–].....[–]	HLA-B18	B44h	.[E-].....[–]
HLA-B*1801	.	.[–].....[–]	HLA-B22	.	.[–].....[–]
HLA-B27	B27o	.[R-].....[–]	HLA-B*2701	B27t	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]	B27t	.[–].....[–]	HLA-B*2709	B27o	[GR]R.....[–]
HLA-B35	B7o	.[P-].....[Y-]	HLA-B*3501X	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	B27o	.[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	B27o→B39	.[RH].....[L-]	HLA-B40	B44o	E[F-].....[L-]
HLA-B*4001	B44	.[E-].....[L-]	HLA-B*4002	B44o	.[E-].....[L-]
HLA-B*4006	B44	.[E-].....[VA]	HLA-B*4101[a]	B44h	.[–].....[–]
HLA-B42	.	.[PL].....[–]	HLA-B44	B44	.[E-].....[–]
HLA-B*4402	B44	.[E-].....[FL]	HLA-B*4403	B44	E[E-].....[FW]
HLA-B*4405	B44o	.[E-].....[R-]	HLA-B45	.	.[–].....[–]
HLA-B*4501	B44h	.[E-].....[L-]	HLA-B*4601	B62	.[MI].....[YF]
HLA-B*4801	B27t	.[QK].....[L-]	HLA-B*4802[a]	B27t	.[–].....[–]
HLA-B*4901[a]	B44h	.[–].....[–]	HLA-B*5001[a]	B44h	.[QK].....[-L]
HLA-B51	B7	.[AP].....[IL]	HLA-B*5101	B7	.[AP].....[LY]
HLA-B*5102	B7o	.[PA].....[IV]	HLA-B*5103	.	.[FG].....[YI]
HLA-B52[a]	B62	.[–].....[–]	HLA-B*5201	B62o	.[QF].....[VF]
HLA-B53	B7o	.[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	B58o	.[TS].....[WF]
HLA-B*5802	B58o	.[ST].....[FM]	HLA-B*6701	B7	.[P-].....[L-]
HLA-B*7301	B27	.[R-].....[P-]	HLA-B*7801	B7	.[GP].....[S-]

Table 13.1: HLA type (column 1,4), supertype (column 2,5) and amino acid motif (column 3,6) for all alleles described by Sette and Sidney [1999] and Rammensee et al. [1999]. Letters in square parenthesis correspond to the same position. X: X-ray structure exists. Table adopted from Lund et al. [2004]. ^{a}: Allele is not in SYFPEITHI. h/t: hypothetical/tentative supertype assignment according to Sette and Sidney [1999]. o: the supertype assignment presented here. →: assignment changed by Lund et al. [2004].

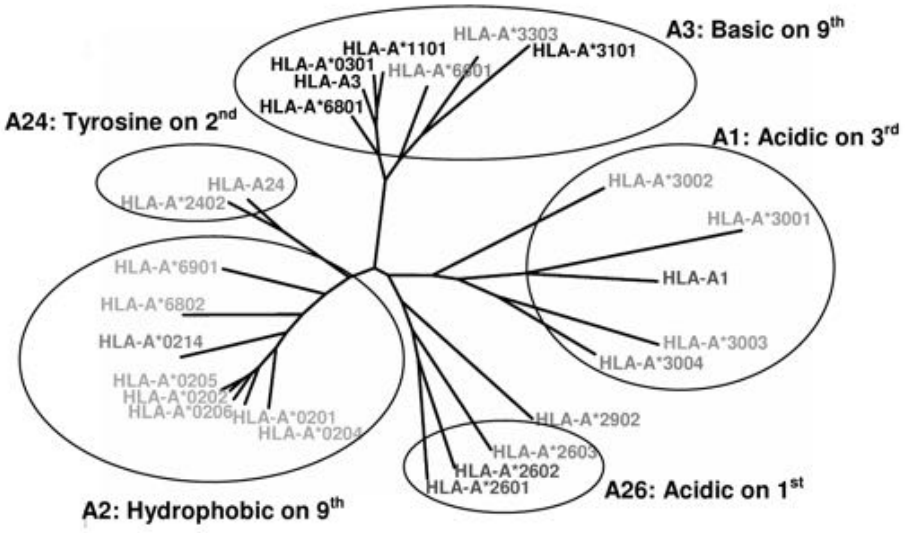


Figure 13.3: Tree showing clustering of HLA-A specificities. The alleles are colored according to the supertype classification by Sette and Sidney [1999]: A1: red, A2: orange, A3: black, A24: green, A29 and nonclassified alleles: gray. Figure reprinted from Lund et al. [2004]. See plate 25 for color version.

13.2.3 HLA-A Supertypes

The tree describing the HLA-A alleles is characterized by five clusters: A1, A2, A3, A24, and A26. 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: The resulting definition of this supertype largely overlaps with the definition by Sette and Sidney [1999]. The unassigned HLA-A*0214 and HLA-A*0217 is added to the A2 supertype.

A3 supertype — basic amino acids in position 9: A*3303 and A*6601 are assigned 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 [1999].

A1 supertype — acidic amino acids in position 3: Here the clustering shows a large difference from the A1 supertype defined by Sette and Sidney [1999]. The clustering suggests splitting the A1 supertype into two clusters. One cluster is the A1 cluster that contains the A1 and A*3001-4 alleles based on their common preference for acidic amino acids in position 3, and Y or F at posi-

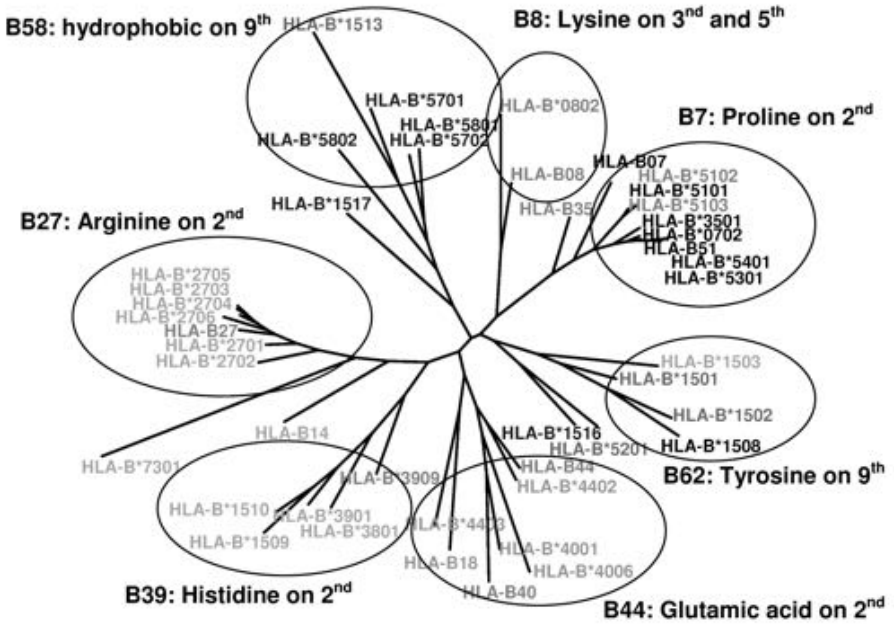


Figure 13.4: Tree showing clustering of HLA-B specificities. The alleles are colored according to the supertype classification by Sette and Sidney [1999]: B7: black, B27: orange, B44: green, B58: blue, B62: violet, and nonclassified alleles and outliers (B8 and B46): gray. Figure reprinted from Lund et al. [2004]. See plate 26 for color version.

tion 9. The other cluster is a proposed new A26 supertype. Sette and Sidney 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 stand 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 at position 3 in HLA-A1. This difference is consistent with the motif descriptions by Marsh et al. [2000]. These alleles therefore 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 this allele is left as an outlier.

A24-supertype — tyrosine or hydrophobic in position 2: A24 and A*2402 is assigned to the A24 supertype. These alleles have a bootstrap value of 0.98.

Based on the background of what is described above, a redefinition of the A1 supertype is made: The HLA-A*2601/2 alleles may form a new separate A26 supertype. The following alleles remain unclassified: A23, A31, A32, A*7401.

13.2.4 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, the HLA-B clustering is limited to the alleles where at least 15 peptide sequences are available in either the SYFPEITHI 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 figure 13.4, but with one important exception. Sette and Sidney place the HLA-B*1508 in the B7 supertype. However, the bootstrap branch value for the Sette and Sidney B7 cluster is 0.042, whereas the corresponding value for the 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 figure 13.4.

B62 supertype — tyrosine in position 9: The B62 cluster shown in figure 13.4 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. Including the alleles HLA-B*1516 and HLA-B*5201 make the bootstrap value drop to 0.06. These two alleles are thus left out 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 figure 13.4 has a branch value of 0.22. The low branch value for the Sette and Sidney B27 supertype is due to a misplacement of the HLA-B*1503 allele. As described above, this allele is placed in the B62 cluster. Splitting up the B27 cluster into two subclusters leaving out the HLA-B*7301 and the HLA-B*14 alleles as outliers, leads to a bootstrap branch value for the remaining 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 [1999]. The alleles of HLA-B*40 and HLA-B*44 are included in the supertype, and the bootstrap branch value for the cluster is then 0.36.

B58 supertype — hydrophobic at position 9: The branch bootstrap value for the B58 cluster defined in figure 13.4 is found to be 0.42. Including the HLA-B*1517 allele this value drops to 0.18, thus this allele is left out as an outlier. The bootstrap value for the Sette and Sidney [1999] 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 lead to the B58 cluster defined in figure 13.4.

There is generally good consistency between the supertypes defined by Sette and Sidney [1999] and the HLA-B tree. In addition, B8 is a novel 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, 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), and B*5301. Only one or two sequences were found in SYFPEITHI for these alleles, except for B17, where five sequences were found.

13.2.5 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. Making a tree of all the HLA-A and HLA-B alleles included in the analysis described above, no mixing of the HLA-A and HLA-B clusters is found. 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 Ebstein-Barr virus-transformed cell line SWEIG which coexpresses B*4402, and the apparent similarity may be an experimental artifact caused by cross-reactivity 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

obtain 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.

13.2.6 HLA-DR

For most class II molecules relatively few binding peptides are known. To compensate for that the similarities between different alleles are calculated, based on other published specificity matrices. Specificity matrices for HLA class II molecules can be downloaded from, e.g., the ProPred website (http://www.imtech.res.in/raghava/propred/page4.html). The list of alleles is given in table 13.2. These 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 are similar to those in the TEPITOPE program, test sequences can be submitted to both programs as well as to a program using the matrices from ProPred. The matrix scores are 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 position p in a binding peptide relative to the probability of observing that amino acid in proteins in general. Thus,

$$p_{pa} = \frac{\exp(s_{pa})q_a}{\sum_i \exp(s_{pi})q_i}, \tag{13.2}$$

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

Sequence logos were constructed to visualize the specificities. By visual inspection of different HLA class II molecules (figure 13.5) it is clear that some of these are quite similar. In order to quantify the similarities, the distance between all pairs of matrices was calculated. These distances were then used to construct a tree visualizing the similarities between the peptides that each allele binds (figure 13.6). Based on this tree, the HLA-DR molecules are divided 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 correspond to the supertype name assigned to each cluster in figure 13.5, and the cluster bootstrap branch value, respectively. The alleles in figure 13.5 are colored according to the serotype.

Allele	Sero type	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

Table 13.2: A list of the HLA class II alleles used. The list contains the allele, serotype (Type), pocket profile, and our supertype assignment. The pocket profiles used in assembly of virtual DR matrices are from Sturniolo et al. [1999]. For each allele the list of numbers in square parenthesis 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) when the work was done. Table adopted from Lund et al. [2004].

The clustering roughly corresponds to the serotype classification, but with some important exceptions. Note, e.g., 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 are, e.g., < 0.001 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 13.2 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 pocket specificities.

13.2.7 Experimental Verification of Supertypes

To verify the clustering suggested above, weight matrices for all the class I alleles in this study were constructed as earlier described. These weight matrices can then be used to *predict* the binding affinity for sets of peptides, where the binding affinity to a specific HLA allele had been *measured* experimentally. Alleles for which experimental binding information is available are, e.g., HLA-A*0101 (A1), HLA-A0202 (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) [Sylvester-Hvid et al., 2004]. Here the name written in parentheses refers to the supertype classification. The linear correlation coefficient, also known as Pearson's r [Press et al., 1992], is calculated between the prediction score and the log of the measured binding affinity. It is now expected that alleles with similar specificity to that of the allele used in the experiments will obtain a positive correlation, and that other alleles will get a correlation close to zero. This calculation actually supports most of the results obtained from the clustering analysis [Lund et al., 2004].

One of the advantages with this kind of clustering is that it 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, and several large-scale epitope discovery projects funded by the NIH have been started. Additional material is available at:
http://www.cbs.dtu.dk/researchgroups/immunology/supertypes.html
and will be updated whenever new data become available.

Figure 13.5: Logos displaying the binding motifs for 50 different HLA class II molecules. For details of the logo representation, see figure 13.1. Figure reprinted from Lund et al. [2004]. See plate 27 for color version.

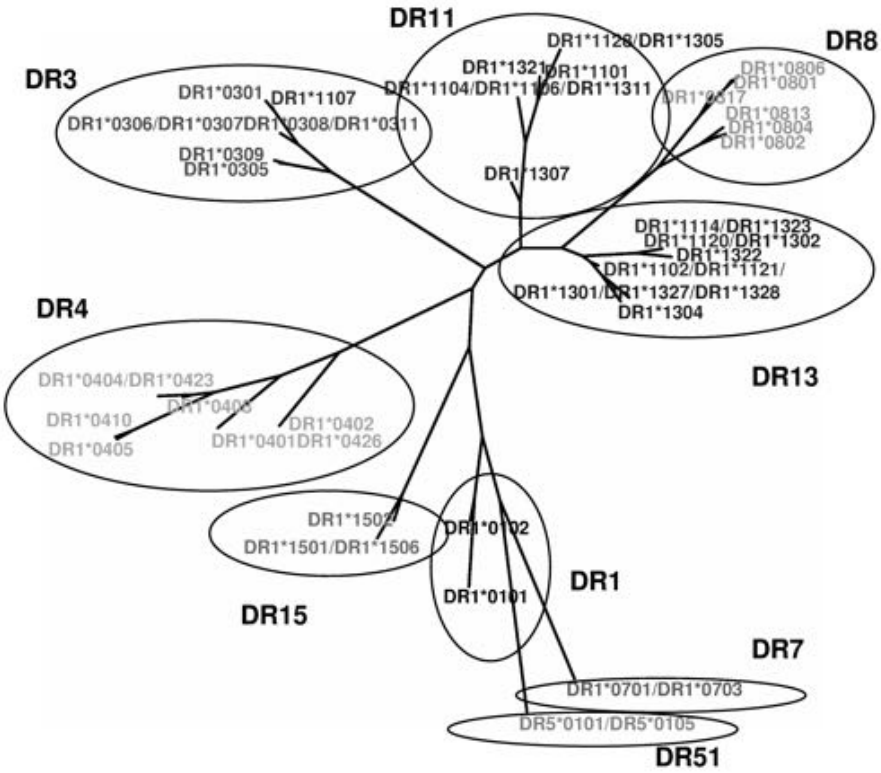


Figure 13.6: Tree showing the clustering of 50 different HLA class II molecules based on their peptide-binding specificity. The proposed clusters are encircled and labeled. Figure reprinted from Lund et al. [2004]. See plate 28 for color version.

The clusters define groups of alleles with similar binding specificities. In order to get a broad coverage of the human population with an epitope-based vaccine, it must be ensured that most people from all ethnic groups have an 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.