Generation of tissue-specific and promiscuous HLA ligand databases using DNA microarrays and virtual HLA class II matrices

Tiziana Sturniolo[1], Elisa Bono[1], Jiayi Ding[2], Laura Raddrizzani[2], Oezlem Tuereci[3], Ugur Sahin[3], Michael Braxenthaler[2], Fabio Gallazzi[1], Maria Pia Protti[4], Francesco Sinigaglia[1], and Juergen Hammer[1,2]*

[1]Roche Milano Ricerche, 20132 Milan, Italy. [2]Department of Genomics and Information Sciences, Hoffman-La Roche, Nutley, NJ 07110. [3]Department of Internal Medicine, University of Saarland, 66421 Homburg, Germany. [4]Laboratory of Tumor Immunology, Scientific Institute H. San Raffaele, 20132 Milan, Italy.
**Corresponding author (e-mail: juergen.hammer@roche.com).*

Received 26 January 1999; accepted 15 April 1999

Most pockets in the human leukocyte antigen–group DR (HLA-DR) groove are shaped by clusters of polymorphic residues and, thus, have distinct chemical and size characteristics in different HLA-DR alleles. Each HLA-DR pocket can be characterized by "pocket profiles," a quantitative representation of the interaction of all natural amino acid residues with a given pocket. In this report we demonstrate that pocket profiles are nearly independent of the remaining HLA-DR cleft. A small database of profiles was sufficient to generate a large number of HLA-DR matrices, representing the majority of human HLA-DR peptide-binding specificity. These virtual matrices were incorporated in software (TEPITOPE) capable of predicting promiscuous HLA class II ligands. This software, in combination with DNA microarray technology, has provided a new tool for the generation of comprehensive databases of candidate promiscuous T-cell epitopes in human disease tissues. First, DNA microarrays are used to reveal genes that are specifically expressed or upregulated in disease tissues. Second, the prediction software enables the scanning of these genes for promiscuous HLA-DR binding sites. In an example, we demonstrate that starting from nearly 20,000 genes, a database of candidate colon cancer–specific and promiscuous T-cell epitopes could be fully populated within a matter of days. Our approach has implications for the development of epitope-based vaccines.

Keywords: HLA matrix, epitope prediction, DNA microarrays, tumor immunology, genomics

Helper T-cell activation is essential for the initiation of a protective immune response to pathogens and tumors[1,2]. Human leukocyte antigen–group DR (HLA-DR), the predominant isotype of the human class II major histocompatibility complex (MHC), plays a central role in helper T-cell selection and activation. Proteins of HLA-DR bind peptide fragments derived from protein antigens and display them on the surface of antigen-presenting cells for interaction with antigen-specific receptors of T lymphocytes[1].

X-ray crystallographic studies demonstrated that the HLA-DR ligand binding groove consists of pockets, resulting in strong preferences for interaction with particular amino acid side chains of the ligands[3–6]. Molecules of HLA-DR are extremely polymorphic. Polymorphic residues are often involved in forming HLA-DR pockets; consequently, pockets of different HLA-DR alleles can be of distinct chemical and size characteristics. Some of the ligand side chains interact with the pockets and increase the overall binding affinity and specificity of ligands, whereas others interfere with pocket residues and reduce binding[7]. Therefore, the pocket specificity can be characterized either topographically (i.e., by differences in the amino acid residues forming the pockets) or functionally (i.e., by substituting the corresponding peptide ligand position with all natural amino acid residues and by quantifying their effects on binding ["pocket profiles"]). The sum of all pocket profiles of a given HLA-DR allele is defined as a "quantitative matrix"[8].

We and others have demonstrated that matrices are powerful tools to predict HLA class II ligands[8,9]. In contrast to previous all-or-nothing rules and approaches that are based on artificial neural networks[10,11], matrix-based predictions rely on mathematical processing of individual peptide side chain effects (see ref. 12 for a detailed comparison of bioinformatic tools used for HLA class II ligand prediction). A typical matrix-based algorithm first extracts all possible peptide frames from a given protein sequence. Subsequently, the corresponding position- and amino acid–specific matrix values are assigned to each residue of these peptide frames. Finally, the sum of these matrix values is determined for each frame. It has been shown that the resulting numerical values ("peptide scores") correlate with the binding affinity of HLA-DR ligands, thus making matrices important tools for the prediction of candidate T-cell epitopes[13,14].

HLA-DR molecules account for more than 90% of the HLA class II isotypes expressed on antigen-presenting cells. Although the HLA-DRA locus is monomorphic, more than 100 alleles have been described for the HLA-DRB1 locus[15]. Matrices have so far been determined by measuring all possible pocket profiles on a given HLA-DR allele. Hence, the determination of a single HLA-DR matrix required hundreds of individual peptides and thousands of peptide binding assays[13]; a global coverage of HLA class II binding specificity seemed, therefore, unlikely in the near future. In this report, we demonstrate that pocket profiles are nearly independent

of the remaining HLA-DR groove. Thus, once a pocket profile has been determined in vitro, it can be shared among other HLA-DR alleles as long as their amino acid residues contributing to the pocket are identical. Consequently, a relatively small number of pocket profiles can be assigned to a large number of HLA-DR alleles via sequence comparison. The resulting virtual matrices cover the majority of human HLA-DR specificity.

A comprehensive database of candidate promiscuous T-cell epitopes in tumors or pathogens would be of great value for vaccine strategies. Major bottlenecks so far have included not only the need to determine quantitative matrices for each polymorphic HLA-DR allele, but also the lack of gene expression data enabling, for example, a comprehensive selection of genes expressed in disease but not in normal tissue. The latter has become feasible by the recent development of DNA microarray technology[16,17]: DNA microarrays are used to monitor and compare the expression of thousands of genes simultaneously and are thus capable of identifying large pools of differentially expressed candidate antigens. The former is resolved in this study by applying the above concept of virtual HLA matrices.

Results and discussion

Allele independence of pocket profiles. The value of matrix-based computational algorithms for the prediction of helper T-cell epitopes has been demonstrated beyond doubt, as exemplified by the recent discovery of a human leukocyte function-associated antigen-1 (LFA-1) peptide as the candidate autoantigen in Lyme arthritis[18] or by a recent x-ray crystal structure of a DRB1*0401–collagen II peptide complex[6,19]. Both the LFA-1 and the collagen peptides were identified using our previously described DRB1*0401 matrix[8]. In this report we propose a new strategy (Fig. 1A) that leads to both a broad coverage of human HLA-DR binding specificity and the possibility of creating genomic databases of candidate T-cell epitopes.

We have previously demonstrated that pocket specificity profiles are mostly independent of neighboring ligand side chains[8]. Moreover, initial analyses on DRB1*04 subtypes have also suggested that pocket profiles might be independent of the remaining HLA-DR groove[19]. Obviously, the latter would have important implications in that pocket profiles are only determined once and can subsequently be shared among alleles as long as they are predicted to have similar

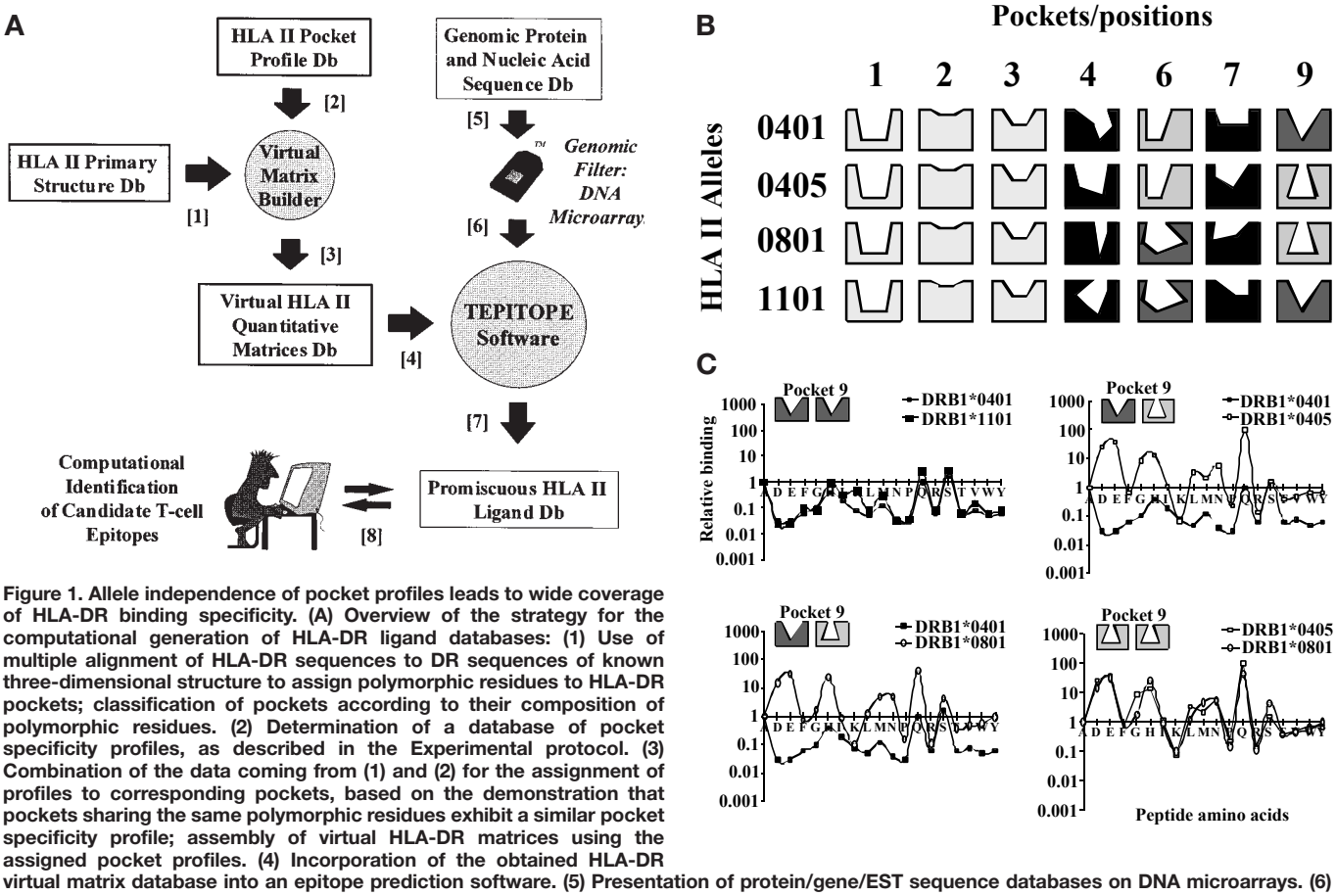

Figure 1. Allele independence of pocket profiles leads to wide coverage of HLA-DR binding specificity. (A) Overview of the strategy for the computational generation of HLA-DR ligand databases: (1) Use of multiple alignment of HLA-DR sequences to DR sequences of known three-dimensional structure to assign polymorphic residues to HLA-DR pockets; classification of pockets according to their composition of polymorphic residues. (2) Determination of a database of pocket specificity profiles, as described in the Experimental protocol. (3) Combination of the data coming from (1) and (2) for the assignment of profiles to corresponding pockets, based on the demonstration that pockets sharing the same polymorphic residues exhibit a similar pocket specificity profile; assembly of virtual HLA-DR matrices using the assigned pocket profiles. (4) Incorporation of the obtained HLA-DR virtual matrix database into an epitope prediction software. (5) Presentation of protein/gene/EST sequence databases on DNA microarrays. (6) Identification of specifically expressed or upregulated genes in disease tissues by DNA microarray expression mapping. (7) Scanning of the identified sequences using the prediction software, allowing the identification of candidate promiscuous HLA-DR ligands. (8) Use of the generated HLA class II ligand database for the identification of candidate promiscuous helper T-cell epitopes. (B) Schematic representation of the modular structure of the HLA-DR binding groove. The binding clefts of four HLA-DR allotypes are compared. The cleft regions 1–3 are constituted by monomorphic residues mostly coming from the DR α chain (except for one dimorphic residue from the DR β chain (Gly/Val86), composing pocket 1); positions 5 and 8 were excluded because peptide side chains at these positions are oriented away from the DR binding cleft, as shown in crystal structure analyses[3,5,6]; pockets 4, 6, 7, and 9 are mainly formed by DR β chain polymorphic residues and are responsible for the allele specificity of HLA-DR–ligand interaction. The modular structure of the HLA-DR binding groove enables the free exchange of functional pocket profiles, as long as the polymorphic residues forming the pockets are the same. (C) Pockets on different alleles sharing the same polymorphic residues exhibit similar pocket specificity profiles. Comparison of pocket specificity profiles obtained for pocket 9 from HLA-DR alleles, which are formed either by identical polymorphic residues (top left and bottom right panels) or by different ones (top right and bottom left panels), demonstrating that HLA-DR primary structure homology can be sufficient to assign defined binding specificity profiles to given pockets. Pocket specificity profiles were determined in HLA-DR competitive binding assays by quantifying the effects of all amino acid substitutions at a given position of affinity optimized, alanine-based designer peptides, as described in the Experimental Protocol. Relative binding values were calculated by normalizing experimental IC_{50} data with the IC_{50} value obtained for alanine at the same peptide position (IC_{50} Ala/IC_{50} substitution).

Table 1. Pocket profile database.

DR source	Pocket	Profile ID#	Polymorphic pocket residues (β-chain)	Amino acid residue																	
				A	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
B1*0101	4	1	13F;70Q;71R;74A;78Y	0	-2.4	-0.4	0.08	-0.7	-0.7	0.5	-2.1	0.9	0.8	0.04	-1.9	0.1	-2.1	-0.7	-1	-0.05	-1.8	-1.1
B1*1501	4	2	13R;70Q;71A;74A;78Y	0	-0.4	-0.6	2.4	0	1.1	0.6	-0.7	0.5	1	-0.2	-0.3	-0.8	0.2	-0.3	-0.3	0.2	0.4	2.5
B1*0301	4	3	13S;70Q;71K;74R;78Y	0	2.3	-1	-1	0.5	0	0.5	-1	0	0	0.2	-1	0	-1	0.7	-1	0	-1.0	-1
B1*0401	4	4	13H;70Q;71K;74A;78Y	0	1.4	1.5	-0.9	-1.6	1.1	0.8	-1.7	0.8	0.9	0.9	-1.6	0.8	-1.9	0.8	0.7	-0.9	-1.2	-1.6
B1*0402	4	5	13H;70D;71E;74A;78Y	0	-2.3	-2.3	0.3	-0.7	1.2	0.08	0.1	-0.6	0.6	-0.4	-1.3	-0.4	1	-1	-0.5	-0.7	1.6	-0.4
B1*0404	4	6	13H;70Q;71R;74A;78Y	0	-1.1	-1.1	1	-2.4	-1	1.1	-1.5	1	1.8	-0.7	-1.3	0	-2.4	-0.7	-0.9	0.5	-0.05	-0.4
B1*1101	4	7	13S;70D;71R;74A;78Y	0	-1.7	-1.7	0.4	-1.7	-0.6	0.9	-0.5	1.1	1	0	-1.7	-0.4	-0.7	-0.7	-0.6	0.4	-0.1	-0.7
B1*0701	4	8	13Y;70D;71R;74Q;78V	0	-1.6	-1.4	0.2	-1.1	0.1	1.1	-1.3	-0.8	-0.4	-1.1	-1.2	-1.5	-1.1	1.5	1.4	0.9	-1.1	-0.9
B1*0801	4	9	13G;70D;71R;74L;78Y	0	-1	-1	0.5	-1	-1	0.3	2.3	0.7	1.4	0	-1	-1	2.3	-1	-1	0.3	0	2.2
B5*0101	4	10	13Y;70D;71R;74A;78Y	0	-1.9	-1.3	-0.6	-1.6	-1.4	1.3	-1.7	0.6	1.7	-1.7	-1.5	-0.7	-1.7	-0.5	0.3	1.1	-1.4	-0.6
B1*1302	4	11	13S;70D;71E;74A;78Y	0	-1.4	-1.1	0.8	-1.5	1.5	-0.6	0.8	0.4	0.8	0.1	-1.5	0.6	0.2	-0.6	-1.1	-0.9	0.7	0.4
B1*0101	6	1	11L	0	-2.7	-2.4	-2.1	-0.3	-2.2	-1.9	-2	-2	-1.8	-1.1	-0.2	-1.8	-1.8	-0.6	-1.2	-1.1	-2.4	-2
B1*1501	6	2	11P	0	-0.4	-1	-0.3	0.5	-0.5	0.05	-0.3	0.2	0.1	0.7	-0.2	-0.8	1	0.6	-0.04	-0.3	-0.4	0.4
B1*0801	6	3	11S	0	-2.4	-1.4	-1.4	-0.7	-0.1	0.7	1.3	0.2	-0.9	-0.6	0.5	-0.3	1	-0.1	0.8	1.2	-1.4	-1.4
B1*0404	6	4	11V	0	-1.1	-2.4	-1.1	-1.5	-1.4	-0.1	-2.4	-1.1	-1.1	1.3	0	-1.5	-2.4	1	1.9	0.9	-1	-1.5
B1*0701	6	5	11G	0	-2.5	-2.5	-0.8	-0.6	-0.8	-0.5	-1.1	-0.9	-0.8	-0.6	-0.5	-1.1	-1.1	0.6	-0.08	0.1	-0.9	-1
B5*0101	6	6	11D	0	-2	-2	-1.7	-0.3	-1.2	-1.4	-1.5	-1	-1.5	-1.3	0.2	-1.4	-1.3	-0.5	-0.8	-1.3	-1.7	-1
B1*0101	7	1	28E;30C;47Y;61W;67L;71R	0	-2	-0.6	0.3	-1.1	0.1	0.6	-0.2	0.3	0.09	0.1	0.07	0.2	0.09	-0.2	0.09	0.7	-0.08	0.5
B1*1501	7	2	28D;30Y;47F;61W;67I;71A	0	-0.7	-0.7	1.4	0	0.6	1.5	-0.3	1.9	1.7	0.7	0.3	-0.3	-0.5	0.3	0.2	0.3	0.6	0.7
B1*0301	7	3	28D;30Y;47F;61W;67L;71K	0	-0.6	-0.2	0.5	0.1	-0.8	0.4	-0.9	0.2	1.1	-0.09	0.7	-0.1	-0.9	0.07	-0.1	0.2	-0.6	-0.05
B1*0401	7	4	28D;30Y;47Y;61W;67L;71K	0	-0.3	0.2	-1	-1.3	0	0.08	-0.3	0.7	0.8	0.6	-0.7	0	-1.2	-0.2	-0.1	0.08	-1.4	-1.2
B1*0402	7	5	28D;30Y;47Y;61W;67I;71E	0	-2.1	-1.2	0.5	-2.1	0.5	0.5	0	1	0.8	0.6	-1	1.1	1.7	-0.4	0.1	0.2	1.4	0.9
B1*0404	7	6	28D;30Y;47Y;61W;67L;71R	0	-1.2	-0.7	-0.05	-1.2	-0.4	0.08	-1.3	0.3	0.7	0.7	-1	-0.2	-0.9	0.5	0.4	-0.1	-0.7	-0.2
B1*1101	7	7	28D;30Y;47F;61W;67F;71R	0	-2.7	-1.3	-0.4	-0.4	-0.2	0.8	-0.2	1.5	1.3	-1	-0.05	-1.1	-0.4	-1.3	-1.3	-0.6	-0.4	-0.4
B1*0701	7	8	28E;30L;47Y;61W;67I;71R	0	-1.3	0.9	2.1	0	0.9	2.4	0.5	2.2	1.8	1.4	-0.2	1.1	0.7	0.4	0.9	1.6	1.4	1.7
B1*0801	7	9	28D;30Y;47Y;61W;67F;71R	0	-2.4	-2.4	-0.9	-0.5	-0.9	-0.8	-0.8	-0.3	-0.3	-1.3	-1.2	-1.2	-0.6	-1.3	-2.4	-1.1	-1.2	-1.1
B5*0101	7	10	28H;30D;47Y;61W;67F;71R	0	-1.5	-0.9	1.5	0.6	1.2	1.2	0.9	0.6	0.4	0.5	-0.6	0.7	1.3	-0.2	0.3	-0.3	0.4	1.2
B1*1302	7	11	28D;30Y;47F;61W;67I;71E	0	-1.5	-1	0.6	-1.5	0.3	-0.5	0	0.4	0	0.1	-0.5	-0.4	1.2	-0.9	-0.9	-0.1	0.4	-0.2
B1*0101	9	1	9W;37S;57D;60Y;61W	0	-1.9	-1.9	-0.4	-0.8	-1.1	0.7	-1.7	0.5	0.08	-1.2	-1.1	-1.6	-1	-0.3	-0.2	0.3	-1.4	-0.9
B1*0301	9	2	9E;37N;57D;60Y;61W	0	-0.6	-0.3	0.9	0.4	-0.5	0.6	-0.2	-0.04	1.1	-0.6	-0.3	-0.2	0.5	1.1	-0.5	0.3	-1	0.3
B1*1101	9	3	9E;37Y;57D;60Y;61W	0	-1.7	-1.7	-1	-1	0.08	-0.3	-0.3	-1	-0.4	-1.4	-1.3	0.5	-1	0.7	-1.2	-0.7	-1	-1
B1*0701	9	4	9W;37F;57V;60S;61W	0	-1.2	-0.3	2.1	-0.6	-0.2	3.4	-1.1	3.4	2	-0.5	-0.6	-0.9	-0.8	-0.3	0.4	2	0.8	1.1
B1*0801	9	5	9E;37Y;57S;60Y;61W	0	1	1.3	-0.1	0.3	1.3	-0.1	-1	0	0.7	0.6	-0.9	1.3	-1	0.7	-0.3	-0.4	-0.1	0.1
B5*0101	9	6	9Q;37D;57D;60Y;61W	0	-1.5	-0.6	1.2	0.4	1	1.2	2.7	1.3	0.5	0	-0.8	0.7	2.5	0.7	-0.2	-0.2	-0.7	1.3

This table shows the HLA-DR alleles used to generate the different profiles for the polymorphic pockets 4, 6, 7, and 9. Different profiles for each pocket are indicated with identification numbers. For each specific profile the DR β-chain polymorphic residues composing each pocket have been specified and the relative values of the effects of all natural amino acid side residues at each position are reported. Data are expressed as the logarithm of the alanine-normalized relative binding data calculated as in Figure 1C.

pocket topographies (Fig. 1B). To test this hypothesis, we determined pocket specificity profiles for several HLA pockets and compared them with each other (Fig. 1C). The alignment of HLA-DR sequences with DR sequences of known three-dimensional structures[3–6] indicated that the polymorphic residues constituting pocket 9 in DRB1*0401 and DRB1*1101 allotypes are identical. This finding is consistent with the observation that the pocket profiles for both alleles are similar (Fig. 1C). In contrast, the alignment to HLA-DR sequences with known three-dimensional structure revealed differences in the amino acid composition of pocket 9 between DRB1*0401 and DRB1*0405 subtypes, and between DRB1*0401 and DRB1*0801 allotypes. Once again, this is consistent with the resulting pocket profiles (Fig. 1C). Furthermore, a comparison of DRB1*0405 and DRB1*0801 sequences via alignment to three-dimensional structures indicated identical pocket 9 topographies and, consequently, predicted similar pocket profiles. The profiles shown in Figure 1C demonstrate that this was indeed the case. We performed a similar set of experiments for the polymorphic pocket 6 (data not shown) and were able to further confirm the approximation that profiles are mainly independent of the remaining HLA-DR groove and that primary HLA-DRB structures are sufficient to assign profiles to given HLA-DR pockets.

Table 2. Assembled DR virtual matrices.

DRB1*0101 [1;1;1;1;1]	DRB1*0102 [2;1;1;1;1]	DRB1*1501 [2;2;2;2;1]
DRB1*1502 [1;2;2;2;1]	DRB1*1506 [2;2;2;2;1]	DRB1*0301 [2;3;3;3;2]
DRB1*0305 [1;3;3;3;3]	DRB1*0306 [2;3;3;4;3]	DRB1*0307 [2;3;3;4;3]
DRB1*0308 [2;3;3;4;3]	DRB1*0309 [1;3;3;3;2]	DRB1*0311 [2;3;3;4;3]
DRB1*0401 [1;4;4;4;3]	DRB1*0402 [2;5;4;5;3]	DRB1*0404 [2;6;4;6;3]
DRB1*0405 [1;6;4;6;5]	DRB1*0408 [1;6;4;6;3]	DRB1*0410 [2;6;4;6;5]
DRB1*0421 [1;4;4;4;2]	DRB1*0423 [2;6;4;6;3]	DRB1*0426 [1;4;4;4;3]
DRB1*1101 [1;7;3;7;3]	DRB1*1102 [2;11;3;11;3]	DRB1*1104 [2;7;3;7;3]
DRB1*1106 [2;7;3;7;3]	DRB1*1107 [2;3;3;3;3]	DRB1*1114 [1;11;3;11;3]
DRB1*1120 [1;11;3;11;2]	DRB1*1121 [2;11;3;11;3]	DRB1*1128 [1;7;3;7;2]
DRB1*1301 [2;11;3;11;2]	DRB1*1302 [1;11;3;11;2]	DRB1*1304 [2;11;3;11;5]
DRB1*1305 [1;7;3;7;2]	DRB1*1307 [1;7;3;9;3]	DRB1*1311 [2;7;3;7;3]
DRB1*1321 [1;7;3;7;5]	DRB1*1322 [2;11;3;11;3]	DRB1*1323 [1;11;3;11;3]
DRB1*1327 [2;11;3;11;2]	DRB1*1328 [2;11;3;11;2]	DRB1*0701 [1;8;5;8;4]
DRB1*0703 [1;8;5;8;4]	DRB1*0801 [1;9;3;9;5]	DRB1*0802 [1;9;3;9;3]
DRB1*0804 [2;9;3;9;3]	DRB1*0806 [2;9;3;9;5]	DRB1*0813 [1;9;3;6;3]
DRB1*0817 [1;9;3;7;5]	DRB5*0101 [1;10;6;10;6]	DRB5*0105 [1;10;6;10;6]

Virtual DR matrices were assembled according to the modular structure of the HLA-DR groove as indicated in Figure 1B. Profiles for pockets 4, 6, 7, and 9 were derived from the database shown in Table 1. Profiles for the relative peptide positions 2 and 3 were derived from the DRB1*0401 matrix[8] (not shown). For relative peptide position 1, only aliphatic (Ile, Leu, Met, Val) and aromatic (Phe, Trp, Tyr) amino acid residues were considered; more specifically, for HLA-DR alleles with a β86 Gly composing pocket 1, values of 0 were assigned to aromatic and -1 to aliphatic residues at relative P1, while the reverse was done for DR alleles with a β86 Val composing pocket 1 (ref. 8). The virtual matrix values for each allele are encoded by a set of five numbers listed after each allele: The first number indicates whether the allele has a Gly (= 1) or a Val (= 2) at β86 (see above). The second number represents the identification number of the pocket 4 profile (see Table 1). The third, fourth, and fifth number indicates the identification number of the pocket 6, 7, and 9 profiles, respectively.

The approximation that pocket profiles show allele independence enables the generation of virtual matrices; that is, profiles for identical pockets are recycled from a pool rather than determined repeatedly for each allele. The important consequence is that a relatively small number of profiles can be used to build a large number of HLA-DR matrices. The synthesis of approximately 1,000 synthetic designer peptides and the accomplishment of 10,000 HLA-ligand binding assays allowed us to create a database of 35 independent pocket profiles (Table 1). These 35 profiles were used to build 51 virtual HLA-DR matrices (Table 2), which represent the majority of the human HLA-DR peptide binding specificity[20].

HLA-DR ligand prediction with virtual matrices. Are virtual matrices suitable for the prediction of HLA class II ligands and candidate T-cell epitopes? To answer this question, we created a new software package named TEPITOPE, in which the pocket profiles and the resulting virtual matrix data were incorporated (Fig. 2A). The basic ligand prediction algorithm works, in principle, like earlier quantitative matrix-based algorithms (see above). However, instead of calculating only peptide scores for every peptide frame in a given protein sequence, it enables the calculation of score distribution curves for each HLA-DR allotype using natural protein sequence databases as a source (Fig. 2B). Thus, peptides are predicted based on a user-selected threshold defined as the percentage of best scoring natural peptides (Fig. 2B). This compensates in part for the allelic differences of absolute peptide scores caused by variations in the sensitivity of HLA-DR peptide binding assays (data not shown).

The predictive power of virtual matrices was tested on both individual T-cell epitopes and large peptide repertoires. Gross et al.[18] have recently used our previously described quantitative DRB1*0401 matrix to identify a candidate autoantigenic peptide for Lyme arthritis. Figure 2A shows that the virtual matrices incorporated into our software would have predicted this peptide too, using a stringent threshold setting of "1% best scoring natural peptides." Figure 2C shows that our softward can also be used to determine 'threshold profiles' for peptides. For example, the threshold profile of MAGE-3 281–295, a peptide originally identified with TEPITOPE[21], revealed that it is predicted to bind to many HLA-DR allotypes, even when stringent threshold settings are used (Fig. 2C). Notably, we confirmed the promiscuity of MAGE-3 281–295 by in vitro binding studies, and we also demonstrated that MAGE-3 281–295 was indeed presented by melanoma cells[21]. In contrast to MAGE-3 281–295, the melanoma-specific helper T-cell epitope tyrosinase 448–462 (ref. 22) was described as being a DRB1*0401-restricted low-affinity ligand. This again is consistent with the threshold profile for this peptide (Fig. 2C).

Obviously, larger ligand repertoires are required for a better estimation of the predictive power of virtual matrices. Therefore, we tested several molecular repertoires. The first repertoire consisted of both HLA-DR–selected and nonselected peptides originally generated by the bacteriophage M13 display technology[23,24]. We then tested whether we could computer-simulate the screening of M13 display libraries. We combined both the selected and nonselected peptide repertoires and examined whether the virtual matrices could "separate" them again computationally (Fig. 3A). Up to 80% of the HLA-DR selected peptides could be predicted using a stringent threshold setting of 1–3%, whereas <5% of the nonselected peptides were predicted under the same conditions (Fig. 3B). These results clearly demonstrated the ability of TEPITOPE to computationally separate HLA selected and nonselected peptide repertoires. To further assess the predictive power of virtual matrices, we performed peptide binding assays with hundreds of randomly selected natural peptide sequences, generating yet another repertoire of HLA-DR binding and nonbinding peptides. We demonstrated that stringent threshold settings were sufficient for the preferential prediction of HLA-DR ligands (data not shown). Finally, we tested natural ligands and T-cell epitopes using the natural ligand database generated by Rammensee's group[25]. More than half of all natural ligands could be predicted using a 1–3% threshold setting and more than 75% with a 1–6% threshold setting (Fig. 3C). In conclusion, the use of large data sets that were either derived experimentally in our laboratory (Fig. 3B and data not shown) or from the literature (Fig. 3C) demonstrated the utility of the threshold setting for the prediction of HLA-DR ligands. In addition, it allowed us to estimate the potential false-positive and false-negative rate at different threshold stringencies (Fig. 3B and data not shown).

Generation of promiscuous HLA-DR ligand databases. The computational prediction of candidate T-cell epitopes by virtual

matrices is not limited to well-defined protein sequences. Various genome projects are generating huge amounts of new sequence information[26–29], and high-throughput sequencing of cDNA libraries has led to the discovery of several millions of expressed sequence tags (ESTs)[30,31]. The availability of these sequences makes it possible to quantify mRNA levels for tens of thousands of genes simultaneously by using high-density oligonucleotide arrays[16,17]. Moreover, comparative transcript profiling studies with DNA microarrays enable the discovery of large repertoires of genes that are either specifically expressed or upregulated in disease tissues (data not shown).

We propose to employ TEPITOPE on a genome-wide level for the generation of comprehensive HLA-DR ligand databases. For example, helper T cells have been shown to play a crucial role for the optimal induction of protective immunity against certain types of tumors[32]. A database of promiscuous candidate T-cell epitopes of genes upregulated or specifically expressed in tumor tissues could be a valuable tool for the design of epitope-based vaccines. To demon-

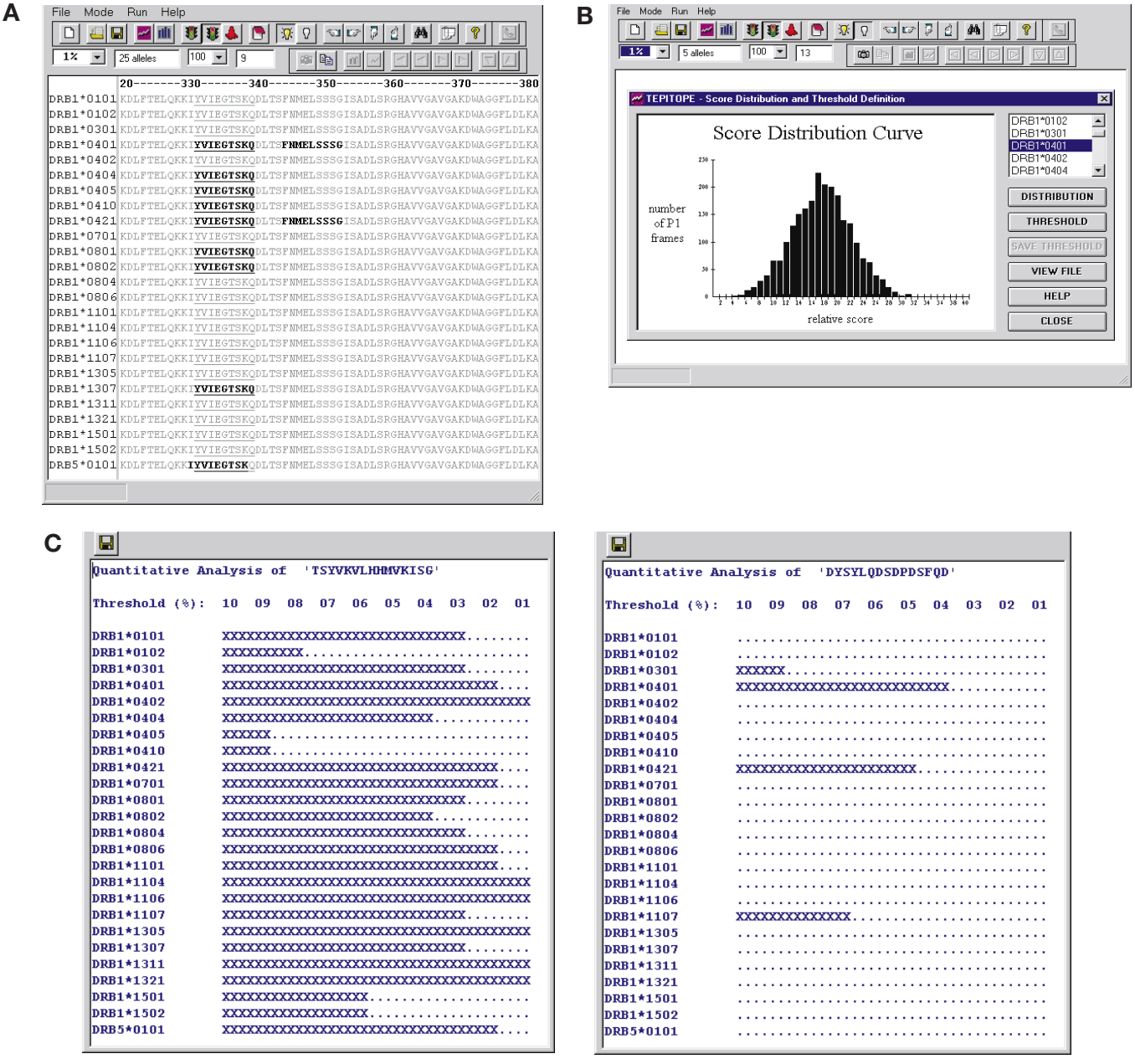

Figure 2. Function of the TEPITOPE software. (A) User interface and prediction of a selective peptide in human leukocyte function–associated antigen-1. The predicted region (bold) corresponds to a recently described candidate autoantigenic peptide (underlined) for Lyme arthritis (human leukocyte function–associated antigen-1, hLFA-1α_L 332–340, [ref. 18]). The prediction threshold was set to 1% (Fig. 2B). (B) Calculation and display of score distribution curves. TEPITOPE allows the calculation of score distribution curves for each HLA-DR allele based on any natural protein database. The Figure shows the DRB1*0401 score distribution of all possible peptide frames in a database of natural protein sequences (8,000 peptide frames). This database is used to normalize the prediction for each HLA-DR allele. Prediction thresholds (chosen by the operator) are expressed as percentage of the best scoring peptides in natural peptide frames. (C) Quantitative evaluation of threshold profiles for given peptides. For any submitted peptide sequence, a histogram displays the predictability for each DR allele according to the threshold stringency: bars indicate the threshold setting at which the peptide is predicted as a ligand for each listed DR allele. Examples of quantitative evaluations are shown for DR promiscuous MAGE-3 281–295[21] (left) and allele-specific DRB1*0401 restricted (tyrosinase 448–462 [ref. 22]) (right) peptides derived from tumor-associated antigens.

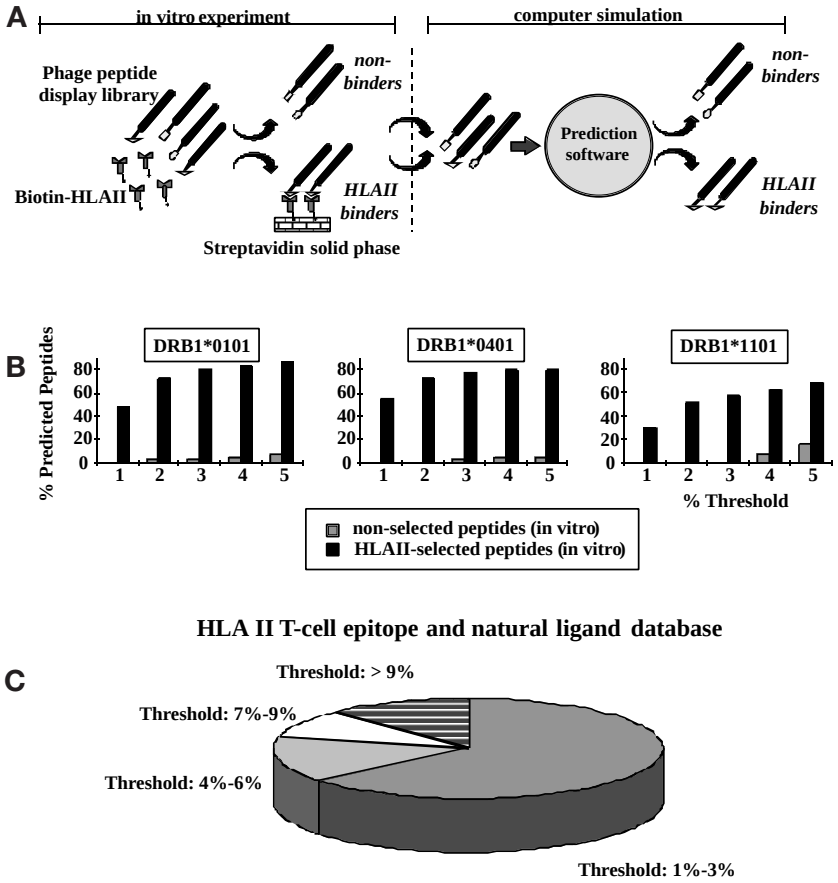

Figure 3. Validation of TEPITOPE using large peptide repertoires. (A and B) Simulation of a bacteriophage peptide display library screening with HLA-DR molecules. (A) Schematic representation of the validation experiment: (left) bacteriophage-displayed peptides binding to biotinylated DR molecules are separated by solid phase-bound streptavidin capture, and isolated by low pH elution; (right) the peptide pools generated with the M13 bacteriophage display technology are submitted to analysis for the computational "isolation" of DR-selected peptides. (B) Prediction of in vitro generated DR repertoires consisting of 60 DRB*0101-selected, 52 DRB*0401-selected, 52 DRB*1101-selected, and 60 nonselected bacteriophage displayed peptide sequences. Results for each HLA-DR allele are expressed as the percentage of predicted peptides out of the total of the respective HLA-DR-selected (black bars) and nonselected (gray bars) peptide repertoires at the indicated threshold settings. (C) Prediction of HLA-DR binders from the natural HLA-DR peptide ligands and T-cell epitopes database[25]. Natural HLA-DR peptide ligands and T-cell epitopes (*n* = 223) were selected for prediction based on the following two criteria: (1) as many peptides are represented as sets of truncated peptides, only one sequence per set was selected, and (2) as the current alpha version of TEPITOPE represents only 25 alleles, only available ligands for these alleles were selected (Fig. 2A). The peptides were analyzed at the indicated different threshold settings. Results are expressed for all HLA-DR alleles as the percentage of predicted peptides out of the total of the peptide repertoire. Predicted peptides are only counted once for a given HLA-DR allele; that is, a peptide predicted at a stringent threshold (e.g., 1–3%) would not be counted again at a less stringent threshold setting (e.g., 4–6%).

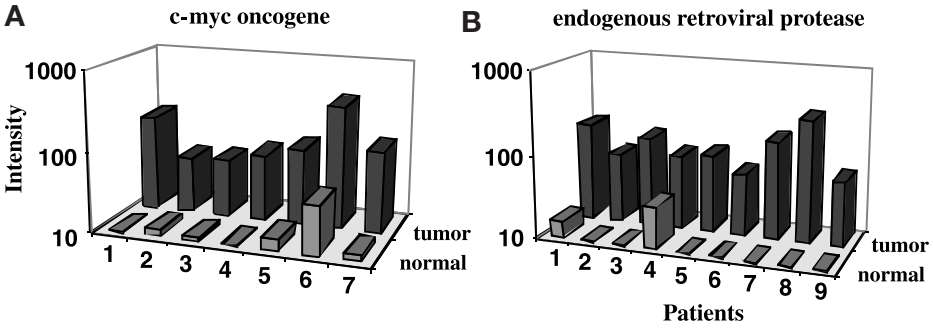

Figure 4. Examples for tumor antigen identification by DNA microarray technology. *C*-myc and an endogenous retroviral protease were upregulated in 7/20 and 9/20 colon cancer patients, respectively. Antibodies have been described in the serum of cancer patients for both antigens[36] (data not shown). Quantification for any mRNA is given by the sum of all perfect match intensities subtracted from mismatch intensities divided by the total number of probe pairs (= 'intensity')[16].

strate that such a comprehensive database can easily be generated by combining DNA microarray technology with epitope prediction software, we performed a simple pilot study: Using both a commercially available Affymetrix (Santa Clara, CA) DNA microarray set (~7,000 genes) and two of our own microarray designs (~12,000 genes; Fig. 4A and B), we have recently profiled 20 primary colon cancer tissues together with the corresponding adjacent normal tissues (data not shown). Although more than 1,000 independent genes were found to be differentially expressed in a population of 20 patients, only 34 genes were upregulated or specifically expressed in ≥50% of all patients (data not shown). These 34 genes gave rise to approximately 19,000 peptide frames. Of these 19,000 peptide frames, 130 candidate promiscuous T-cell epitopes were predicted by TEPITOPE using the following criteria: First, threshold (1–3% best scoring natural peptides); second, promiscuity (predicted to bind to 5/7 HLA-DR allotypes); and third, peptide length (15 amino acid residues). This example demonstrates both the relative ease of generating such a database and the manageable data output. Moreover, the fact that antibodies have been described in serum of cancer patients for some of the microarray-selected candidate antigens (Fig. 4) further supports the feasibility of such an approach.

Databases of candidate HLA-DR ligands and helper T-cell epitopes could ultimately be determined for every gene in a genome. However, the combination of epitope prediction software with other "filters," as demonstrated in this report, will obviously be more practical. DNA microarray/prediction software–based approaches to generate databases of promiscuous candidate T-cell epitopes could be widely applicable in other areas. For example, the current genome project for the malaria-causing pathogen *Plasmodium falciparum* should soon make it possible to generate similar databases (e.g., for life cycle–specific candidate antigens). Similarly, approaches that use epitope prediction software in combination with serological identification of antigens by recombinant expression cloning (SEREX) technology[33] might also prove very useful. SEREX allows the systematic identification of antigens in human cancers and has led to the definition of a wealth of new tumor antigens in many different tumor entities.

Experimental protocol

Determination of a pocket profile database. Pocket profiles were derived from side-chain scanning data obtained by substituting allele-specific peptide ligands (basis peptides) in position 4, 6, 7, and 9, with all natural amino acid residues. Peptide interactions with detergent-solubilized HLA-DR molecules were measured using an ELISA-based high-throughput competitive binding assay as described[13,34]. For each HLA-DR molecule analyzed a specific basis peptide was selected after several optimization experiments to guarantee a highly sensitive analysis of the effects of each peptide side chain on HLA-DR binding[13]. The following basis peptides were used in this study: Gly-Phe-Lys-Ala-Ala-Ala-Ala-Ala-Ala-Ala for DRB1*0101, DRB5*0101, and DRB1*0701; Ile-Ala-Tyr-Asp-Ala-Ala-Ala-Ala-Ala for DRB1*0301; Tyr-Arg-Ser-Met-Ala-Ala-Ala-Ala-Ala for DR1*0401, DRB1*0801, and DRB1*1101; Gly-Ile-Arg-Ala-Ala-Tyr-Ala-Ala-Ala-Ala for DRB1*1501. Competition assays were conducted to measure the ability of substituted basis peptides to compete with a biotinylated indicator peptides for binding to purified DR molecules. At least five dilutions were determined for each competitor peptide. The resulting data points were plotted[34] and the shape of the curves were used for quality control; that is, data sets that did not display a sigmoid shape were repeated. The following biotinylated indicator peptides were used: Gly-Phe-Lys-Ala-Ala-Ala-Ala-Ala-Ala-Ala for DRB1*0101 and DRB1*0701, Gly-Ile-Arg-Ala-Ala-Tyr-Ala-Ala-Ala-Ala for DRB1*1501, myelin-based protein 85–99 for DRB5*0101, Ile-Ala-Tyr-Asp-Ala-Ala-Ala-Ala-Ala for DRB1*0301, Tyr-Pro-Lys-Phe-Val-Lys-Gln-Asn-Thr-Leu-Lys-Ala-Ala for DRB1*0401 (ref. 19), tetanus toxoid$_{830–843}$ for DRB1*1101 (ref. 35), and Gly-Tyr-Arg-Ala-Ala-Ala-Ala-Ala-Ala-Leu for DRB1*0801. The relative binding data of the competitor peptides were expressed as the concentration of competitor peptide required to inhibit 50% of binding of the biotinylated indicator peptide (IC_{50}).

Assembly of virtual matrices and software. Virtual matrices were assembled as follows. First, multiple alignments of HLA-DR sequences to DR sequences of known three-dimensional structures were performed to link polymorphic DR residues to given DR pockets. Second, pockets were classified according to their composition of polymorphic residues; that is, pockets from different alleles constituted by identical residues were considered identical. Third, a pocket profiles database was determined in vitro on 11 HLA-DR alleles (Table 1). Fourth, pocket profiles were assigned to pockets of all HLA-DR alleles according to their classification. And fifth, 51 fully assembled virtual DR matrices were generated by combining the assigned profiles of pockets 4, 6, 7, and 9, and cleft region 2 and 3. Expert rules were used for pocket 1, as previously described[19] (Table 2, legend). Profiles for peptide positions 5 and 8 were not considered due to their minimal effect on binding[3] (data not shown). TEPITOPE is a Windows 98/NT application. The visual user interface allows the identification of promiscuous HLA-DR ligands independent of whether identical and/or shifted HLA-DR binding frames constitute promiscuity. Twenty-five virtual HLA-DR matrices were incorporated into the current alpha version of the application. Requests to use the alpha version of TEPITOPE should be addressed via e-mail to juergen.hammer@roche.com.

Microarray design, RNA sample preparation, hybridization, and analysis. Three microarray designs were used for transcript mapping of primary colon cancer tissue. Microarray 1 is commercially available (6.8k Human Chip; Affymetrix, Santa Clara, CA) and covers 7,071 genes. Microarrays 2 and 3 are custom designs, each covering 6,088 genes. The commercially available microarray 1 represents the currently known set of functionally characterized genes, which are all available in the public domain. The design of microarrays 2 and 3 will be described elsewhere (data not shown). In brief, microarray 2 consists of transcripts for which high-quality consensus sequences could be generated from public and proprietary EST databases. “High quality” means that the consensus is based on an EST sequence redundancy of at least five to correct for the vast majority of EST sequencing errors. The minimum length requirement of 500 nucleotides applied in this design is significantly exceeded by most of the sequences. Microarray 3 contains 3′ sequences with a minimum of five contributing ESTs from public and proprietary sources. In addition, sequences were selected to exclude any significant sequence homology between genes represented on the microarray set.

RNA was extracted from primary colon cancer and adjacent normal human tissue using the Ultraspec method (Biotecx, Houston, TX). RNA was converted into cDNA by reverse transcription and then into cRNA with an in vitro transcription reaction that contained biotin-labeled CTP and UTP nucleotides[16]. Hybridization of cRNA to the microarrays and quantification of RNA expression was performed as described in ref. 16.

1. Germain, R.N. MHC-dependent antigen processing and peptide presentation: providing ligands for T lymphocyte activation. *Cell* **76,** 287–299 (1994).
2. Topalian, S.L. MHC class II restricted tumor antigens and the role of CD4 T cells in cancer immunotherapy. *Curr. Opin. Immunol.* **6,** 741–745 (1994).
3. Stern, L.J. et al. Crystal structure of the human class II MHC protein HLA-DR1 complexed with an Influenza virus peptide. *Nature* **368,** 215–221 (1994).
4. Brown, J.H. et al. The three-dimensional structure of the human class II histocompatibility antigen HLA-DR1. *Nature* **364,** 33–39 (1993).
5. Ghosh, P., Amaya, M., Mellins, E. & Wiley, D.C. The structure of an intermediate in class II MHC maturation: CLIP bound to HLA-DR3. *Nature* **378,** 457–462 (1995).
6. Dessen, A., Lawrence, M.C., Cupo, S., Zaller, D.M. & Wiley, D.C. X-ray crystal structure of HLA-DR4 (DRA*0101, DRB1*0401) complexed with a peptide from human collagene II. *Immunity* **7,** 473–481 (1997).
7. Sinigaglia, F. & Hammer, J. Defining rules for the peptide-MHC class II interaction. *Curr. Opin. Immunol.* **6,** 52–56 (1994).
8. Hammer, J. et al. Precise prediction of major histocompatibility complex class II peptide interaction based on peptide side chain scanning. *J. Exp. Med.* **180,** 2353–2358 (1994).
9. Marshall, K.W. et al. Prediction of peptide affinity to HLA DRB1*0401. *J. Immunol.* **154,** 5927–5933 (1995).
10. Brusic, V., Rudy, G., Honeyman, G., Hammer, J. & Harrison, L. Prediction of MHC class II-binding peptides using an evolutionary algorithm and artificial neural network. *Bioinformatics* **14,** 121–30 (1998).
11. Sette, A. et al. Structural characteristics of an antigen required for its interaction with Ia and recognition by T cells. *Nature* **328,** 395–399 (1987).
12. Hammer, J., Sturniolo, T. & Sinigaglia, F. HLA class II peptide binding specificity and autoimmunity. *Adv Immunol* **66**, 67-100 (1997).
13. Hammer, J. & Sinigaglia, F. in *MHC* Vol. 2, Vol. 181 (eds Fernandez, N. & Butcher, G.) 197–228 (Oxford Univ. Press, New York, 1998).
14. Hammer, J. New methods to predict MHC-binding sequences within protein antigens. *Curr. Opin. Immunol.* **7,** 263–269 (1995).
15. Marsh, S.G.E. Nomenclature for factors of the HLA system, update January 1998. *Tissue Antigens* **51,** 582–583 (1998).
16. Lockhart, D.J. et al. Expression monitoring by hybridization to high-density oligonucleotide arrays [see comments]. *Nat. Biotechnol.* **14,** 1675–1680 (1996).
17. Wodicka, L., Dong, H., Mittmann, M., Ho, M.H. & Lockhart, D.J. Genome-wide expression monitoring in Saccharomyces cerevisiae. *Nat. Biotechnol.* **15,** 1359–1367 (1997).
18. Gross, D.M. et al. Identification of LFA-1 as a candidate autoantigen in treatment-resistant Lyme arthritis. *Science* **281,** 703–706 (1998).
19. Hammer, J. et al. Peptide binding specificity of HLA-DR4 molecules: correlation with rheumatoid arthritis association. *J. Exp. Med.* **181,** 1847–1855 (1995).
20. Tsuji, K., Aizawa, M. & Sasazuki, T. *HLA 1991; Proceedings of the eleventh international histocompatibility workshop and conference* (Oxford Univ. Press, New York, 1992).
21. Manici, S. et al. Melanoma cells present a MAGE-3 epitope to CD4(+) cytotoxic T cells in association with histocompatibility leukocyte antigen DR11. *J. Exp. Med.* **189,** 871–876 (1999).
22. Topalian, S.L. et al. Melanoma-specific CD4 T cells recognize nonmutated HLA-DR-restricted tyrosinase epitopes. *J. Exp. Med.* **183,** 1965–1971 (1996).
23. Hammer, J., Takacs, B. & Sinigaglia, F. Identification of a motif for HLA-DR1 binding peptides using M13 display libraries. *J. Exp. Med.* **176,** 1007–1013 (1992).
24. Hammer, J. et al. Promiscuous and allele-specific anchors in HLA-DR binding peptides. *Cell* **74,** 197–203 (1993).
25. Rammensee, H.G., Bachmann, J., Emmerich, N. & Stevanovic, S. SYFPEITHI: an internet database for MHC ligands and peptide motifs (access via http:/www.uni-tuebingen.de/uni/kxi/). (data extracted in 4Q, 1998).
26. Blackwell, J.M. Parasite genome analysis. Progress in the Leishmania genome project. *Trans. R. Soc. Trop. Med. Hyg.* **91,** 107–110 (1997).
27. Collins, F.S. et al. New goals for the U.S. Human Genome Project: 1998-2003. *Science* **282,** 682–689 (1998).
28. Harwood, C.R. & Wipat, A. Sequencing and functional analysis of the genome of Bacillus subtilis strain 168. *FEBS Lett.* **389,** 84–87 (1996).
29. Degrave, W., Levin, M.J., da Silveira, J.F. & Morel, C.M. Parasite genome projects and the Trypanosoma cruzi genome initiative. *Mem. Inst. Oswaldo Cruz* **92**, 859–862 (1997).
30. Adams, M.D. et al. Sequence identification of 2,375 human brain genes [see comments]. *Nature* **355,** 632–634 (1992).
31. Adams, M.D. et al. Complementary DNA sequencing: expressed sequence tags and human genome project. *Science* **252,** 1651–1656 (1991).
32. Ossendorp, F., Mengede, E., Camps, M., Filius, R. & Melief, C.J. Specific T helper cell requirement for optimal induction of cytotoxic T lymphocytes against major histocompatibility complex class II negative tumors. *J. Exp. Med.* **187,** 693–702 (1998).
33. Sahin, U., Türeci, Ö. & Pfreundschuh, M. Serological identification of human tumor antigens. *Curr. Opin. Immunol.* **9,** 709–716 (1997).
34. Raddrizzani, L. et al. Different modes of peptide interaction enable HLA-DQ and HLA-DR molecules to bind diverse peptide repertoires. *J. Immunol.* **159,** 703–711 (1997).
35. Panina-Bordignon, P. et al. Universally immunogenic T cell epitopes: promiscuous binding to human MHC class II and promiscuous recognition by T cells. *Eur. J. Immunol.* **19,** 2237–2242 (1989).
36. Ben-Mahrez, K., Thierry, D., Sorokine, I., Danna-Muller, A. & Kohiyama, M. Detection of circulating antibodies against c-myc protein in cancer patient sera. *Br. J. Cancer* **57,** 529–534 (1988).