11

Definition of MHC Supertypes Through Clustering of MHC Peptide-Binding Repertoires

Pedro A. Reche* **and Ellis L. Reinherz**

Summary

Identification of peptides that can bind to major histocompatibility complex (MHC) molecules is important for anticipation of T-cell epitopes and for the design of epitope-based vaccines. Population coverage of epitope vaccines is, however, compromised by the extreme polymorphism of MHC molecules, which is in fact the basis for their differential peptide binding. Therefore, grouping of MHC molecules into supertypes according to peptide-binding specificity is relevant for optimizing the composition of epitope-based vaccines. Despite the fact that the peptide-binding specificity of MHC molecules is linked to their specific amino acid sequences, it is unclear how amino sequence differences correlate with peptide-binding specificities. In this chapter, we detail a method for defining MHC supertypes based on the analysis and subsequent clustering of their peptide-binding repertoires

Key Words: MHC; supertypes; clustering; peptide-binding repertoire

1. Introduction

Major histocompatibility complex (MHC) molecules play a key role in the immune system by capturing peptide antigens for display on cell surfaces. Subsequently, these peptide–MHC (pMHC) complexes are recognized by T cells through their T-cell receptors (TCRs). MHC molecules fall into two

* Address for correspondence: Department of Immunology, Faculated de Medicina, Universidad Complutense de Madrid, Ave Complutense S/N Madrid, 28040, SPAIN. TL: +34 91 394 7299; FX: +34 91 394 1641; Email: parecheg@med.ucm.es

From: *Methods in Molecular Biology, vol. 409: Immunoinformatics: Predicting Immunogenicity In Silico*
Edited by: D. R. Flower © Humana Press Inc., Totowa, NJ

major classes, MHC class I (MHCI) and MHC class II (MHCII). Antigens presented by MHCI and MHCII are recognized by two distinct sets of T cells, $CD8^+$ T and $CD4^+$ T cells, respectively ***(1)***. Because T-cell recognition is limited to those peptides presented by MHC molecules, prediction of peptides that can bind to MHC molecules is important for anticipating T-cell epitopes and designing epitope-based vaccines ***(2–4)***. Furthermore, the availability of computational methods that can readily identify potential epitopes from primary protein sequences has fueled a new epitope discovery-driven paradigm in vaccine development.

A major complication to the development of epitope-based vaccines is the extreme polymorphism of the MHC molecules. In the human, MHC molecules are known as human leukocyte antigens (HLAs), and there are hundreds of allelic variants of class I (HLA I) and class II (HLA II) molecules. These HLA allelic variants bind distinct sets of peptides ***(5)*** and are expressed at vastly variable frequencies in different ethnic groups ***(6)***. Consequently, the potential population coverage of epitope-based vaccines is greatly compromised. Interestingly, it has been noted that some HLA molecules can bind largely overlapping sets of peptides ***(7,8)***. Therefore, grouping of MHC molecules into supertypes according to peptide-binding specificity is of relevance for the formulation of epitope vaccines providing a wide population coverage.

The first supertypes were defined by Sidney, Sette, and co-workers ***(7,8)*** (hereafter Sidney–Sette et al.) upon inspection of the reported peptide-binding motifs of individual HLA alleles. However, the relationships between peptide-binding specificities of HLA molecules may be too subtle to be defined by visual inspection of these peptide-binding motifs. Furthermore, such sequence patterns have proven to be too simple to describe the binding ability of a peptide to a given MHC molecule ***(9,10)***. In view of these limitations, we developed an alternative method to define MHC supertypes by clustering the peptide-binding repertoire of MHC molecules. The core of the method consists of the generation of a distance matrix whose coefficients are inversely proportional to the peptide binders shared by any two MHC molecules. Subsequently, this distance matrix is fed to a phylogenic clustering algorithm to establish the kinship among the distinct MHC peptide-binding repertoires. The peptide-binding repertoire of any given MHC molecule is unknown, and thereby, defining supertypes through this method requires the estimation of the peptide-binding repertoire of MHC molecules. In this chapter, we will use position-specific scoring matrices (PSSMs) as the predictor of peptide–MHC binding ***(11,12)*** and describe in detail the generation of supertypes using, for example, a selection of HLA class I (HLA I) molecules for which PSSMs are readily available.

2. Materials

2.1. Prediction of Peptide–MHC Binding Repertoires

We consider the peptide-binding repertoire of any MHC molecule as the subset of peptides predicted to bind from a reference set consisting of a random protein of 1,000 amino acids. A selection of public online resources that can be used for the prediction of peptide–MHC binding is summarized in Table 1. In our study PSSMs derived from aligned MHC ligands as the predictors of peptide–MHC binding ***(11,12)***. In this approach, the binding potential of any peptide sequence (query) to the MHC molecule is determined by its similarity to a set of known peptide–MHC binders and can be obtained by comparing the query to the PSSM. Prediction of peptide–MHC binding is threshold-dependent, and here we use the same threshold for all MHC molecules. Thus, the size of the peptide-binding repertoire of all MHC molecules is considered to be same (same number of peptides).

2.2. Supertype Construction

MHC supertypes are derived following the general scheme illustrated in Fig. 1 . First, the overlap between the predicted peptide-binding repertoires (see Section 2.1) of any two MHC molecules, pMHC_i and pMHC_j, is computed as the number of peptide binders shared by the two molecules. Let that number be n_{ij}. Subsequently, a distance coefficient (d_{ij}) is defined as follows:

$$d_{ij} = N - n_{ij}, \tag{1}$$

where N is the size of the peptide-binding repertoire of the MHC molecule. Thus, if the peptide-binding repertoire between two MHC molecules is identical, then $d_{ij} = 0$. Alternatively, if they share no peptides in common, d_{ij} will match the size of the binding repertoire, N. Through the repetition of this process over all distinct pairs of MHC molecules, a quadratic distance matrix is derived containing the d_{ij} coefficients for all distinct pairs of MHC molecules. Once the distance matrix is obtained, we use the Phylogeny Inference Package (PHYLIP; http:// evolution.genetics.washington.edu/phylip.html) ***(13)*** to generate a phylogenic tree where the MHC molecules appear clustered according to their peptide-binding specificity. Specifically, within the PHYLIP package one must use applications such as *kitsch* and *neighbor* that take distance matrices as input. The *kitsch* application uses a Fitch–Margoliash criterion and assumes an evolutionary clock ***(14)***. On the other hand, the *neighbor* application uses the popular neighbor-joining method to derive an unrooted tree without the assumption of a

Table 1
Public online resources for the prediction of peptide–major histocompatibility complex (MHC) binding

Name	URL	Method	Class	Reference
HLABIND	http://www.bimas.dcrt.nih.gov/molbio/hla_bind/	QM	I	*(22)*
MHCpred	http://www.wehih.wehi.edu.au/mhcpep	QM	I and II	*(23)*
PROPRED	http://www.imtech.res.in/raghava/propred/	QM	II	*(24)*
SVMHC	http://www.sbc.su.se/svmhc/information.html	SVM	I	*(25)*
SYFPEITHI	http://www.syfpeithi.de/Scripts/MHCServer.dll/EpitopePrediction.htm	Motif matrix	I and II	*(26)*
RANKPEP	http://www.mif.dfci.harvard.edu/Tools/rankpep.html	Motif PSSM	I and II	*(11)*
NetMHC	http://www.cbs.dtu.dk/services/NetMHC/	ANN	I	*(27)*
PREDEP	http://www.margalit.huji.ac.il/	Structure based	I	*(28)*

ANN, Artificial Neural Network; PSSM, position-specific scoring matrix; QM, quantitative matrix.

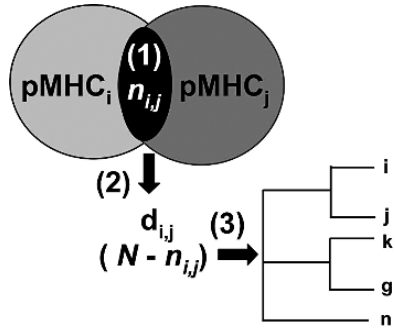

Fig. 1. Strategy to define major histocompatibility complex (MHC) supertypes. MHC supertypes are identified as follows: (1) estimate number of common peptides, n_{ij}, between the binding repertoires of any two MHC molecules, pMHC$_i$ and pMHC$_j$; (2) obtain a distance matrix whose coefficients, d_{ij}, are inversely proportional to the peptide-binding overlap between any pair of MHC molecules; and (3) derive a dendrogram using a phylogenic clustering algorithm to visualize MHC supertypes (groups of MHC molecules with similar peptide-binding specificity). N is the size of the peptide-binding repertoire of the MHC molecule.

clock ***(15)***. For instance, to generate a tree using the neighbor-joining algorithm method one can use the command:

$$echo\ Y\ \ |\ neighbor > /dev/null.$$

This command will generate a tree from a distance matrix that must be named as *infile* using the default options of the *neighbor* application. Likewise, one may use similar commands to generate trees using other applications. In any case, these applications will generate two files, one named *outfile* displaying the tree and another named *treefile* describing the tree in NEWICK format, which can be used to visualize and manipulate the tree using third party applications such as TREEVIEW (http:// taxonomy.zoology.gla.ac.uk/rod/treeview.html).

3. Methods

3.1. HLA I Supertypes

Definition of MHC supertypes using the method described here requires the estimation of the peptide-binding repertoire of the MHC molecules using predictors of peptide–MHC binding. The prediction of peptide–MHCII binding is generally less reliable than that of peptide–MHCI binding ***(12)***. Therefore, to illustrate the definition of MHC supertypes, we focused on 55 HLA I molecules

(human MHCI) for which we can readily predict their peptide-binding repertoires using PSSMs (see Section 2). Given that MHCI ligands are usually nine residues in length, we selected PSSMs for the prediction of binders of that same size (nine residues). In previous studies we have shown that depending on the specific MHCI molecule, the accuracy of peptide–MHCI binding predictions is optimal by considering as binders the top 2–5% scoring peptides (2–5% threshold) within a protein query ***(12)***. Here we have estimated the peptide-binding repertoire of the selected HLA I molecules using a 2% threshold. Thus, following the method described above with a Fitch and Margoliash clustering algorithm ***(14)*** (Section 2.2; *kitsch* application), we generated the phylogenic tree, which is shown in Fig. 2. In this tree, HLA I molecules with similar peptide-binding specificity (large overlap in their peptide-binding repertoires) branch together in groups or supertypes. The relationship between the peptide-binding specificities of HLA I molecules is extensive, and although affinities are mostly confined to alleles belonging to the same gene, they also reach to alleles belonging to different genes (Fig. 2, B15 cluster; B*4002 and A*2902; and A*2402 and B*3801). We clearly identified the classic A2, A3, B7, B27, and B44 supertypes previously defined by Sidney–Sette et al., as well as three new potential supertypes, BX, AB, and B57 (Fig. 2). Furthermore, this analysis indicates that classic HLA I supertypes may be larger than that previously thought. For instance, the A2 supertype would also include the A0207, A0209, and A0214, and the A3 supertype will also include A*6601.

3.2. Combined Phenotypic Frequency of HLA I Supertypes

HLA I-restricted peptides are the targets of $CD8^+$ cytotoxic T lymphocytes (CTLs). The population protection coverage (PPC) of a vaccine composed of CTL epitopes is given by the combined phenotypic frequency (CPF) of the HLA I molecules restricting the epitopes, and it can be computed from the gene and haplotype frequencies ***(16)***. Using the allelic and haplotype frequencies reported by Cao et al. ***(17)*** corresponding to five major American ethnic groups (Black, Caucasian, Hispanic, Native American, and Asian), we have computed the CPF for the HLA I supertypes defined in the previous section (Section 3.1), and the values are tabulated in Table 2 . Targeting HLA I supertypes for the prediction of promiscuous peptide binders allows to minimize the total number of predicted epitopes without compromising the population coverage required in the design of multi-epitope vaccines. However, including many distantly related HLA I molecules in the supertypes may result in too few or no epitopes predicted to bind to all the alleles included in the supertype. Therefore, for the CPF calculations, we have limited the composition of HLA I supertypes to

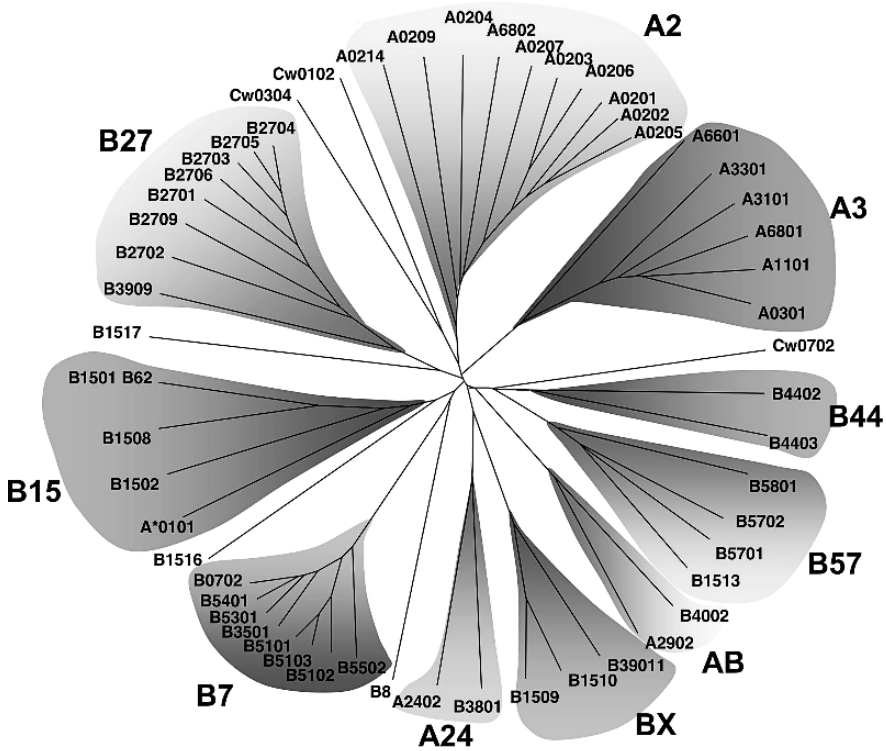

Fig. 2. Human leukocyte antigen (HLA) I supertypes. This figure shows an unroot dendrogram reflecting the relationships between the peptide-binding specificities of HLA I molecules. The closer the HLA I alleles branch, the larger the overlap between their peptide-binding repertoires. Groups of HLA I alleles with similar peptide-binding specificities branch together defining supertypes (shaded groups).

include only those HLA I alleles with $\geq$20% peptide-binding overlap (pairwise between any pair of alleles).

The A2, A3, and B7 supertypes have the largest CPF in the five studied ethnic groups, providing a CPF close to 90%, regardless of ethnicity. To increase the CPF to 95% in all ethnicities, it is necessary to include at least two more supertypes. Specifically, the supertypes A2, A3, B7, B15, and A24 or B44 represent the minimal supertypic combination providing a CPF $\geq$95%. These results indicate that as few as five epitopes restricted by the mentioned HLA I supertypes may be enough to develop a vaccine eliciting CTL responses in the whole population, regardless of ethnicity.

Table 2
Cumulative phenotype frequency of defined supertypes

Supertype	Alleles	Blacks (%)	Caucasians (%)	Hispanics (%)	N.A. Natives (%)	Asians (%)
A2	A*0201–7, A*6802	43.7	49.9	51.8	52.4	44.7
A3	A*0301, A*1101, A*3101, A*3301, A*6801, A*6601	35.4	46.9	41.5	40.7	47.9
B7	B*0702, B*3501, B*5101–02, B*5301, B*5401	45.9	42.2	40.5	52.0	31.3
B15	A*0101, B*1501_B62, B1502	13.06	37.80	16.75	27.26	21.04
A24	A*2402, B*3801	15.5	17.28	25.85	41.94	35.0
B44	B*4402, B*4403	10.4	27.7	17.15	14.4	10.1
B57	B5701–02, B5801, B*1503	19.2	10.3	5.9	5.8	16.5
ABX	A*2902, B*4002	7.4	11.3	19.1	16.3	16.3
B27	B*2701–06, B*2709, B*3909	2.3	4.8	5.1	16.9	4.7
BX	B*1509, B*1510, B*39011	3.1	0.7	4.2	7.8	4.1
AB	A*2902, B4002	7.4	0.11	0.19	0.16	0.07

N.A., North American.

4. Conclusions

HLA molecules are represented by hundreds of allelic variants displaying distinct peptide-binding specificities, and grouping them into supertypes is relevant for developing epitope-based vaccines with a wide PPC. The peptide-binding specificity of HLA molecules stems from the specific amino acids lining their binding groove, and consequently, supertypes may be defined from structural analysis ***(18–20)***. However, it is not always clear how amino acid sequence differences among HLA molecules translate into distinct peptide-binding specificities. Indeed, structure-based methods for the prediction of peptide–MHC binding are still in their infancy. Therefore, in thischapter, we described a method for defining HLA supertypes based on the analysis and subsequent clustering of their predicted peptide-binding repertoires. Furthermore, we have shown that the method can identify experimentally defined HLA I supertypes, suggesting in addition new potential relationships between the peptide-binding specificity of HLA I molecules. When the predictor of peptide–MHC binding is a specificity matrix such as a PSSM, clustering of the HLA molecules according to peptide-binding specificity may alternatively be achieved by comparison of the matrix coefficients ***(21)***. However, it is important to stress that the clustering method described here to derive supertypes can be applied in combination with any predictor of peptide–MHC binding. Although, not indicated in this chapter, minor differences in the defined supertypes appear depending on the phylogenic algorithm used to cluster the HLA I molecules. There are also two other limitations to the method described here. First, the method is limited by both the quality and availability of the peptide–MHC binding predictors. Thus, we do not discard the possibility that the fine structure of the supertypes may suffer some changes as new and better predictors of peptide–MHC binding develop. The second limitation is that we have considered the size of the peptide-binding repertoire of all MHC molecules to be the same. However, that might not always be the case. Indeed, it has been noted that, for instance, the A*0201 appears to be quite promiscuous, binding larger sets of peptides than the other HLA I molecules (Azouz, Reinhold, and Reinherz, unpublished results).

References

1. Margulies, D.H. 1997. Interactions of TCRs with MHC-peptide complexes: a quantitative basis for mechanistic models. *Curr Opin Immunol* 9:390–395.
2. Yu, K., Petrovsky, N., Schonbach, C., Koh, J.Y., and Brusic, V. 2002. Methods for prediction of peptide binding to MHC molecules: a comparative study. *Mol Med* 8:137–148.

3. Flower, D. 2003. Towards in silico prediction of immunogenic epitopes. *Trends Immunol* 24:667–674.
4. Flower, D., and Doytchinova, I.A. 2002. Immunoinformatics and the prediction of immunogenicity. *Appl Bioinformatics* 1:167–176.
5. Reche, P.A., and Reinherz, E.L. 2003. Sequence variability analysis of human class I and class II MHC molecules: functional and structural correlates of amino acid polymorphisms. *J Mol Biol* 331:623–641.
6. David W. Gjertson, and Paul I. Terasaki, E. (Eds) 1998. *HLA 1998*. American Society for Histocompatibility and Immunogenetics, Lenexa.
7. Sette, A., and Sidney, J. 1999. Nine major HLA class I supertypes account for the vast preponderance of HLA-A and -B polymorphism. *Immunogenetics* 50:201–212.
8. Sette, A., and Sidney, J. 1998. HLA supertypes and supermotifs: a functional perspective on HLA polymorphism. *Curr Opin Immunol* 10:478–482.
9. Bouvier, M., and Wiley, D.C. 1994. Importance of peptide amino acid and carboxyl termini to the stability of MHC class I molecules. *Science* 265:398–402.
10. Ruppert, J., Sidney, J., Celis, E., Kubo, T., Grey, H.M., and Sette, A. 1993. Prominent role of secondary anchor residues in peptide binding to HLA-A2.1 molecules. *Cell* 74:929–937.
11. Reche, P.A., Glutting, J.-P., and Reinherz, E.L. 2002. Prediction of MHC class I binding peptides using profile motifs. *Hum Immunol* 63:701–709.
12. Reche, P.A., Glutting, J.-P, Zhang, H., and Reinherz, E.L. 2004. Enhancement to the RANKPEP resource for the prediction of peptide binding to MHC molecules using profiles. *Immunogenetics* 56:405–419
13. Retief, J.D. 2000. Phylogenetic analysis using PHYLIP. *Methods Mol Biol* 132:243–258.
14. Fitch, W.M., and Margoliash, E. 1967. Construction of phylogenetic trees. *Science* 155:279–284.
15. Saitou, N., and Nei, M. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol* 4:406–425.
16. Dawson, D.V., Ozgur, M., Sari, K., Ghanayem, M., and Kostyu, D.D. 2001. Ramifications of HLA class I polymorphism and population genetics for vaccine development. *Genet Epidemiol* 20:87–106.
17. Cao, K., Hollenbach, J., Shi, X., Shi, W., Chopek, M., and Fernandez-Vina, M.A. 2001. Analysis of the frequencies of HLA-A, B, and C alleles and haplotypes in the five major ethnic groups of the United States reveals high levels of diversity in these loci and contrasting distribution patterns in these populations. *Hum Immunol* 62:1009–1030.
18. Doytchinova, I.A., Guan, P., and Flower, D.R. 2004. Quantitative structure-activity relationships and the prediction of MHC supermotifs. *Methods* 34:444–453.
19. Doytchinova, I.A., and Flower, D.R. 2005. In silico identification of supertypes for class II MHCs. *J Immunol* 174:7085–7095.

20. Doytchinova, I.A., Guan, P., and Flower, D.R. 2004. Identifying human MHC supertypes using bioinformatic methods. *J Immunol* 172:4314–4323.
21. Lund, O., Nielsen, M., Kesmir, C., Petersen, A.G., Lundegaard, C., Worning, P., Sylvester-Hvid, C., Lamberth, K., Roder, G., Justesen, S., Buus, S., and Brunak, S. 2004. Definition of supertypes for HLA molecules using clustering of specificity-matrices. *Immunogenetics* 55:797–810.
22. Parker, K.C., Bednarek, M.A., and Coligan, J.E. 1994. Scheme for ranking potential HLA-A2 binding peptides based on independent binding of individual peptide side chains. *J Immunol* 152:163–175.
23. Guan, P., Doytchinova, I.A., Zygouri, C., and Flower, D. 2003. MHCPred: a server for quantitative prediction of peptide-MHC binding. *Nucleic Acids Res* 31:3621–3624.
24. Singh, H., and Raghava, G.P. 2001. ProPred: prediction of HLA-DR binding sites. *Bioinformatics* 17:1236–1237.
25. Donnes, P., and Elofsson, A. 2002. Prediction of MHC class I binding peptides, using SVMHC. *BMC Bioinformatics* 3:25.
26. Rammensee, H.G., Bachmann, J., Emmerich, N.P.N., Bacho, O.A., and Stevanovic, S. 1999. SYFPEITHI: database for MHC ligands and peptide motifs. *Immunogenetics* 50:213–219.
27. Buus, S., Lauemoller, S.L., Worning, P., Kesmir, C., Frimurer, T., Corbet, S., Fomsgaard, A., Hilden, J., Holm, A., and Brunak, S. 2003. Sensitive quantitative predictions of peptide-MHC binding by a 'Query by Committee' artificial neural network approach. *Tissue Antigens* 62:378–384.
28. Altuvia, Y., Sette, A., Sidney, J., Southwood, S., and Margalit, H. 1997. A structure based algorithm to predict potential binding peptides to MHC molecules with hydrophobic binding pockets. *Hum Immunol* 58:1–11.