13

Prediction of Peptide–MHC Binding Using Profiles

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

Summary

Prediction of peptide binding to major histocompatibility complex (MHC) molecules is a basis for anticipating T-cell epitopes. Peptides that bind to a given MHC molecule are related by sequence similarity. Therefore, a position-specific scoring matrix (PSSM)—also known as profile—derived from a set of aligned peptides known to bind to a given MHC molecule can be used as a predictor of both peptide–MHC binding and T-cell epitopes. 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. Following structural considerations of the peptide–MHC interaction, we will describe here how to derive alignments and PSSMs that are suitable for the prediction of peptide–MHC binding.

Key Words: PSSM; MHC; binding; epitopes; profile; prediction

1. Introduction

T-cell immune responses are triggered by the recognition of foreign peptide antigens bound to cell membrane-expressed major histocompatibility complex (MHC) molecules ***(1–3)***. Because T-cell recognition is limited to those peptides presented by MHC molecules, prediction of peptides that can bind to MHC molecules is the basis for the anticipation of T-cell epitopes ***(4–6)***. Peptides binding to MHC molecules must fit into a specific chemical and physical

*Address for correspondence: Department of Immunology, Facultad de Medicina, Universidad Complutense de Madrid, Ave Complutense S/N, Madrid, 28040, SPAIN. TL: +34 91 394 7229; 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

environment conditioned by polymorphic residues in the MHC molecule ***(7–9)***. Consequently, distinct MHC molecules have distinct peptide-binding specificities ***(9)***. In addition, the peptides that bind to the same MHC molecule are related by sequence similarity. Sequence patterns reflecting amino acid preferences in peptide–MHC binders (anchor residues) are routinely used for defining peptide–MHC binding motifs and prediction of peptide–MHC binding ***(10,11)***. For example, the binding motif of the human MHC class I (MHCI) molecule A*0201 may be described by the following sequence pattern:

$$\text{X-[AILMVT]-X}_6\text{-[AILMVT]}.$$

This motif indicates that A*0201 will preferentially bind peptides of nine residues having an Ala, Ile, Leu, Met, Val, or Thr residue at positions 3 and 9, which act as anchor positions. However, the binding ability of a peptide to a given MHC molecule cannot be explained by the presence of a few anchor residues, and indeed, non-anchor residues contribute to peptide–MHC binding ***(12,13)***. Instead, a position-specific scoring matrix (PSSM) or profile created from a set of aligned sequences of peptide–MHC binders provides a better alternative for capturing the complexity of peptide–MHC binding motifs. These PSSMs can be also used to quantify the relatedness of any peptide to the known peptide–MHC binders, thus serving as predictors of peptide–MHC binding.

PSSMs were first introduced by Gribskov et al. ***(14)*** for the detection of distantly related proteins and are now widely used for the representation and identification of sequence motifs ***(15,16)***. In essence, a PSSM consists of a table containing a form of frequency count of each one of the 20 amino acids observed in every column of an alignment divided by the corresponding expected frequency of that amino acid in the background (usually the frequency of the amino acid in a reference database). In addition, methods for the derivation of profiles also provide corrections for missing data and sequence redundancy in the alignments, which are essential to increase the detection limits of PSSMs ***(17,18)***. Missing and/or low counts in the alignments are corrected using pseudo-counts estimated from substitution matrices ***(17)***, whereas sequence redundancy is corrected by applying sequence weights before the estimation of the amino acid counts.

A PSSM is a good descriptor of the peptide–MHC binding motif, only if the peptide–MHC binders are aligned by structural and/or sequence similarity. There are two types of MHC molecules, class I (MHCI) and class II (MHCII), which actually present peptide antigens for recognition by two distinct sets of T cells, $CD8^+$ and $CD4^+$, respectively ***(7)***. MHCI and MHCII molecules bind peptides in a different mode, and thus, for aligning MHCI and MHCII ligands,

we devised two distinct procedures that are compatible with the structural and molecular basis of the peptide–MHCI and peptide–MHCII interactions. In thischapter, we will describe these two procedures, and we will illustrate the prediction of peptide–MHC binding through the use of PSSMs.

2. Materials

2.1. Databases

Prediction of peptide–MHC binding using profiles require the availability of the sequence of peptides known to bind to MHC molecules. These sequences can be retrieved from any of the available public databases of MHC ligands (Table 1). However, in this study, we used the EPIMHC database ***(19)*** as the only source of MHC ligands (Table 1). All peptides in EPIMHC are MHC binders, and their binding strength is reported as unknown, low, moderate, or high. Importantly, the EPIMHC database (http://bio.dfci.harvard.edu/epimhc/) has been designed to facilitate the query, extraction, and analysis of data by third parties. To illustrate the prediction of peptide–MHC binding using PSSMs, we selected from the EPIMHC the sequences of 178 and 80 peptides annotated to bind with high affinity to A*0201 (human MHCI molecule) and DRB1*0401 (human MHCII molecule), respectively. The protein sources of the peptides were also retrieved from the EPIMHC database. All A*0201 peptide binders had a length of nine residues (9 mers), whereas the DRB1*0401 peptide binders were variable in length with at least nine residues. These sets of peptides are available as supplemental data from the site http://bio.med.ucm.es/methods/.

2.2. Software

The applications used in this study are indicated in Table 2. All these packages are freely available for academia users and were compiled and/or under the LINUX operating system. The core applications used for deriving alignments and profiles from MHC ligands are PROFILEWEIGHT ***(18)***, BLIMPS ***(20)***, and MEME ***(21)***. In addition to these applications, we used a set of Perl scripts to format data and/or handle the applications described above. These scripts are summarized in Table 2, and their use will be described elsewhere in Methods.

2.3. Leave-One Out Cross-Validation

Performance of PSSMs predicting peptide–MHC binding was evaluated using a leave-one out cross-validation (LOOCV). Briefly, for a set of peptides n known to bind to a given MHC molecule, a PSSM is generated from $n - 1$ peptides and used to test the binding of the remaining peptide (target peptide). This process is repeated n times until the binding of each peptide is tested.

Table 1
Selected public databases of MHC ligands

	Availability	Description	Reference
MHCPEP	http://wehih.wehi.edu.au/mhcpep	Database of MHC–binding peptides	***(44)***
MHCBN	http://www.imtech.res.in/raghava/mhcbn	Database of MHC-binding and nonbinding peptides	***(45)***
ANTIGEN	http://www.jenner.ac.uk/antijen/	Database of quantitative functional peptide data for immunology	***(46)***
FIMM	http://research.i2r.a-star.edu.sg/fimm/	A database of functional molecular immunology	***(47)***
EPIMHC	http://bio.dfci.harvard.edu/epimhc/	Curated database of MHC Ligands	***(19)***

Table 2
Computer applications used in this study

	Download	Description
BLIMPS	ftp://ftp.ncbi.nih.gov/repository/blocks/unix/	Motif discovery program
MEME	ftp://ftp.sdsc.edu/pub/sdsc/biology/meme/	Motif discovery program
PROFILEWEIGHT	ftp://ftp.ebi.ac.uk/pub/software/unix/profile.tar.Z	Program to create a PSSM from a GCG/MSF alignment
READSEQ	ftp://iubio.bio.indiana.edu/molbio/redseq/classic/	Program for sequence format conversion
epimhc.pl	http://bio.med.ucm.es/software/	Perl script to get peptides into a FASTA file from EPIMHC
meme2fasta.pl	http://bio.med.ucm.es/software/	Perl script to format block alignment in MEME output in FASTA
Mkmatrix.pl	http://bio.med.ucm.es/software/	Perl script to generate PSSMs using BLIMPS and PROFILEWEIGHT
rankpep.pl	http://bio.med.ucm.es/software/	Perl script to score and rank peptides using PSSMs

PSSM, position-specific scoring matrix.

3. Methods

3.1. Structural Alignments of MHCI and MHCII ligands

Capturing the complexity of the peptide–MHC binding motif in the form of a PSSM that can be used for the prediction of peptide–MHC binding requires the alignment of known MHC ligands by structural and/or sequence similarity. Peptides bound to MHCI molecules are in an extended conformation with several side chains accommodated in the binding pockets of the MHCI binding groove, and the N-terminal and C-terminal pinned into the groove ***(7,8)*** (Fig. 1A). As a consequence, MHCI ligands are of short length (8–11 residues), and proper structural alignment can be best accomplished by piling up peptides that have the same length ***(22)***. In contrast, the peptide binding groove of MHCII molecules is open, allowing both the N-terminal and C-terminal of a peptide to extend beyond the binding groove ***(7,8)*** (Fig. 1B). Consequently, peptides bound to MHCII molecules display a great variability in length (9–22 residues). Nevertheless, only a peptide core of nine residues fits into the MHCII binding groove per se and is responsible for anchoring the peptide to the MHCII molecule ***(3)***. This peptide core of nine residues binds in a conserved mode across the different peptide–MHCII complexes, sitting in the groove in an extended conformation connected through a network of hydrogen bonds between its backbone and conserved residues in the MHCII molecule ***(3,7,8,23)***. As a result, the peptide-binding repertoire of MHCII molecules is broader than that of MHCI molecules, and MHCII ligands share less sequence similarity than MHCI ligands. Poor amino acid sequence similarity between MHCII ligands together with their great variability in sequence length makes their alignment difficult, hampering the use of global alignment algorithms such as CLUSTALW ***(24)***. Because alignment of the MHCII ligands requires the identification of their binding core, we use the motif discovery program MEME ***(21)*** for aligning them. MEME uses an expectation-maximization algorithm in combination with a priori information to identify sequence motifs. The a priori information we use for aligning MHCI ligands is consistent with the interaction of peptides and MHCII molecules, namely, the existence of a single peptide-binding register per se MHCII ligand stretching nine residues.

3.2. Generation of Alignments and PSSMs from MHCI and MHCII Ligands

The strategy to derive alignments and profiles from known MHCI and MHCII ligands for the prediction of peptide–MHC binding consists of three basic steps: (i) peptide collection and subsequent subsetting by their MHC-binding

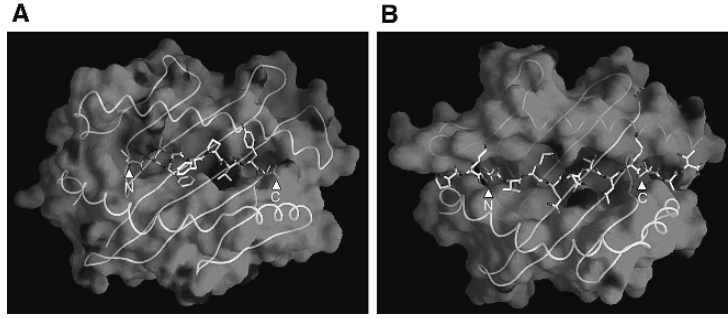

Fig. 1. Binding of peptide ligands to major histocompatibility complex class I (MHCI) and MHCII molecules. The figure shows the top of the molecular surface of the antigen-presenting platform of representative human MHCI (**A**) and MHCII (**B**) molecules as viewed by the T-cell receptor. The MHCI molecule corresponds to HLA-A*0201 in complex with a peptide LLFGYPVYV from HTLV-1 TAX protein [PDB: 1HHK;***(41)***]. The MHCII molecule corresponds to HLA-DR1 in complex with peptide PKYVKQNTLKLAT from influenza hemagglutinin protein [PDB:1FYT ***(42)***]. Peptides bound to these molecules are represented by sticks to highlight the contours of the binding groove. Note how the peptide binding groove of the MHCI molecule is closed, and peptides bind in a manner such that both the N-terminal and C-terminal ends of the peptide (indicated by arrows) are nested into the MHCI binding groove, restricting their lengths to 8–11 residues. In contrast, the peptide binding groove of the MHCII molecule is open, thereby imposing no limitation to the size of ligands, whose N-terminal and C-terminal ends can extend beyond the binding grove. The side chains of N-terminal and C-terminal ends of the 9-mer peptide core fitting into the MHCII binding groove are indicated. The figure was prepared using GRASP ***(43)***.

specificity and length in the case of MHCI ligands; (ii) generation of ungapped alignments; and (iii) generation of PSSMs from alignments. An outline of this strategy is shown in Fig. 2 and the detailed description is as follows.

1. Peptide collection and subsetting: In the case of MHCI ligands, the sequences must be subgrouped into files according to their MHCI-binding specificity and subsequently by sequence length. Peptides with 12 or more amino acids bind to MHCI molecules only exceptionally, and therefore, alignments and profiles should only be made from subsets of peptides of length 8, 9, 10, and 11. Furthermore, given that most of the known MHCI-restricted peptides are 9 mers (~90%) (data not shown), we suggest to preferentially make/use profiles from peptides of nine residues (9 mers). In the case of MHCII ligands, sequences must be subgrouped into distinct files only by their MHCII-binding specificity, and peptides with less than nine residues must be discarded. MHC

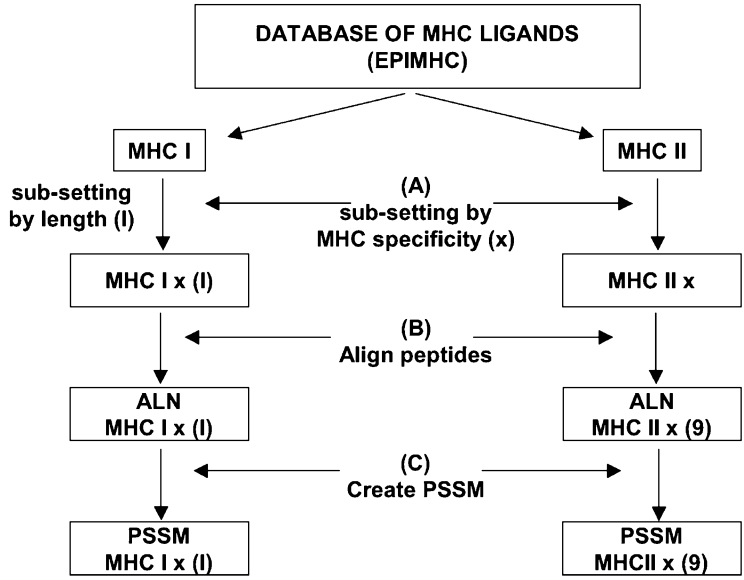

Fig. 2. Overview of the strategy for defining position-specific scoring matrices (PSSMs) from major histocompatibility complex class I (MHCI) and MHCII ligands. The basic steps for defining PSSMs are (A) peptide collection and subsetting of peptides by their MHC-binding specificity (x) and length (l) in the case of MHCI ligands; (B) generation of ungapped alignments; and (C) generation of PSSMs from alignments.

ligands meeting the above criteria can be obtained using the web interface of EPIMHC. Peptides should be saved as plain TEXT and in FASTA format. Alternatively, the Perl script *epimhc.pl* can be used to retrieve peptides from EPIMHC on the command line. For example, to create a FASTA file with all peptides in EPIMHC binding with high affinity to HLA-A*0201, one can use the following command:

*epimhc.pl -m 'HLA-A*0201' -s 9 -b high*

Likewise, the command:

*epimhc.pl -m 'HLA-DRB1*0401' -b high*

will generate a FASTA file with all peptides in EPIMCH binding to DRB1*0401 with high affinity. Peptides with less that seven residues will not be included in this file.

2. Generation of ungapped motif alignments: MHCI ligands of the same length in the FASTA files generated in the step above are already aligned. For aligning MHCII ligands, we use MEME with the following command:

```
meme mhcii_lig.fasta -protein -mod oops -nmotifs 1 -minsites 4\
  -maxsites 300 -minw 9 -maxw 9 -evt 10000 > mhcii_lig.meme
```

where *mhcii_lig.fasta -protein* corresponds to each of the MHCII-specific subsets of peptide sequences in FASTA format; *-mod oops* indicates that each sequence has a binding site; *-minsites 4 -maxsites 500* indicates that the motif should contain between 4 and 500 sequences; *-min 9 -maxw 9* indicates that the size of the motif is exactly 9; and finally *-evt 10000* is the expected threshold value for a sequence to be included in the motif. The output of MEME (*mhcii_lig.meme*) will contain a log-odd and a probability PSSM of the MHCII ligands' binding core which can readily be used for the prediction of peptide–MHCII binding. However, for consistency with the profiles derived from MHCI ligands, we obtain instead the motif alignment in the MEME output using the Perl script *meme2fasta.pl* and built the PSSM in the next step. The use of the script *meme2fasta.pl* is as follows:

```
meme2fasta.pl -i mhcii_lig.meme
```

This command will format the motif alignment in the output of MEME into FASTA format, discarding repeated sequences. The alignment obtained with MEME encompasses the binding core of the MHCII ligands.

3. Generation of PSSMs from alignments of MHC ligands: There are many methods to derive profiles from alignments that differ in the sequence weighting and in the computation of amino acid counts and pseudocounts. Here, we will describe the generation of profiles using PROFILEWEIGHT ***(18)*** and the applications included in the BLIMPS package ***(17,25)***. In both cases, pseudocounts are estimated using the BLOMUS62 substitution matrix-derived protein blocks ***(26)***. To learn about the actual equations used in these packages see Thompson et al. ***(18)*** and Henikoff and Henikoff ***(27)***. PROFILEWEIGHT uses a branch-proportional weighting method and requires an alignment in GCG/MSF format as input. BLIMPS PSSMs are generated through the sequential use of the following three applications: *mablock*, to translate alignments from FASTA format to BLOCK format; *blweight*, to apply weights to the sequences in the alignment, and *blk2pssm*, to generate the actual PSSM. The application *blweight* supports four distinct weighting methods: P, position-based method ***(28)***; A, pairwise distance method ***(29)***; V, Voroni method ***(30)***, and Cn,

clustering method ***(26)***. The generation of matrices with PROFILEWEIGHT and BLIMPS can be facilitated using the Perl script *mkmatrix.pl*. For example, the command:

mkmatrix.pl -i peptides.tfa -w pw

will convert the alignment *peptides.tfa* into GSF/MSF format and create a PSSM using PROFILEWEIGHT. The PSSM will be saved under the name *peptides.pw.mtx*. Likewise the command:

mkmatrix.pl -i peptides.tfa -w p

will generate a PSSM under the name *peptides.p.mtx* using BLIMPS and position-based weights.

3.3. Scoring Peptide–MHC Binding Using PSSMs

PSSMs can be used to provide scores indicating the similarity (and hence binding potential) of any peptide to the set of aligned peptides known to bind to a given MHC molecule. These scores are computed by aligning the PSSM with the protein segments with the same length than the width of the PSSM (length of the alignment) and adding up the appropriate profile coefficients matching the residue type and position in the protein segment. Scoring all peptides in an entire protein sequence requires a dynamic algorithm that starts scoring at the beginning of the sequence and then moves the PSSM over the entire sequence one residue at a time to score the remaining peptides. Here, we provide the Perl script *rankpep.pl* as an example of dynamic scoring algorithm. The use of the script is as follows:

rankpep.pl -i sequence.fasta -m file.mtx

where *sequence.fasta* is the sequence query in FASTA format and *file.mtx* is the PSSM. The output of the program is a list of all peptides in the input sequence ranked by their score. Rank per se may, however, be insufficient to assess whether a peptide is a potential binder. Consequently, to better address whether a peptide might bind or not to a given MHC molecule, one should consider scoring all the peptides in the alignment from which the PSSM was obtained. Then, any given peptide can be considered a binder if it has a score within the range of scores of the peptides known to bind to the relevant MHC molecule.

3.4. Performance of PSSMs Predicting Peptide Binding to MHCI and MHCII Molecules

Only peptides that bind to MHC with an affinity above a necessary threshold are able to elicit a T-cell response. Therefore, determining whether known peptide–MHC binders can be identified among the high-scoring peptides within their protein sources is the best way to check whether prediction of peptide–MHC binding using PSSMs is of practical utility. Here, we have tested this notion for two sets of peptide ligands, one consisting of high-affinity binders to the human MHCI molecule A*0201 (Fig. 3A) and another of high-affinity binders to the human MHCII molecule DRB1*0401 (DR4) (Fig. 3B). These MHC ligands were aligned as indicated in section 3.2, and the binding of each of the peptides in the resulting alignments to the relevant MHC molecule was tested at different thresholds (0.5%, 1%, 2%, 3%, 4%, 5%, 10%, and 20%) under a LOOCV (see Section 2.3). At a given threshold, a peptide is computed as "to bind" if it is among the top scoring peptides from its protein source at that threshold. It is known that sequence weighting increases the sensitivity of profiles. Therefore, we carried out these prediction tests using PSSMs generated with PROFILEWEIGHT which applies branch-proportional weights (empty bars) and BLIMPS with position-based weights (black bars). The results

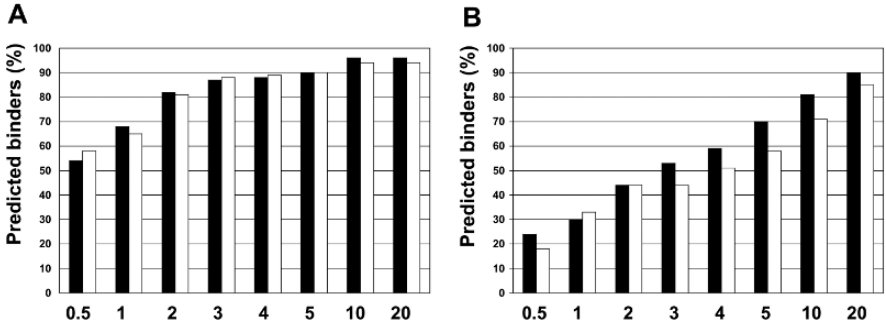

Fig. 3. Performance of position-specific scoring matrices (PSSMs) predicting peptide–major histocompatibility complex (MHC) binding. Performance of PSSMs predicting the binding of 178 peptides to A*0201 (**A**) and 66 peptides to DR4 (**B**) was evaluated by testing whether the peptides are predicted from their protein sources under a leave-one out cross-validation (LOOCV). Predictions were carried out at different thresholds (abscissa), and the percentage of correctly predicted peptides is plotted in the figure (ordinate). PSSMs were derived using PROFILEWEIGHT (empty bars) and BLIMPS with position-based weights (black bars).

indicated that ≥80% of the A*0201 peptide binders are predicted at a 2% threshold (Fig. 3A), whereas ~45% of the DR4 peptide binders are predicted at this threshold (Fig. 3B). As previously reported ***(22,31)***, PSSMs derived with BLIMPS and PROFILEWEIGHT are comparable predicting peptide binding to A*0201. However, for the prediction of peptide binding to DR4, PSSMs obtained with BLIMPS applying position-based weights were significantly better than those obtained using PROFILEWEIGHT (Fig. 3B).

4. Concluding Remarks

PSSMs are powerful tools to detect new and diverse sequences that are functionally related to those included in the original alignment (peptides binding to MHC molecules) and can be used to identify those peptides that can bind to MHC molecules. Prediction of peptide–MHC binding using PSSMs appears to be more accurate for MHCI molecules. (Fig. 3). This observation does not necessarily indicate that the MHCII-specific PSSMs were derived from incorrect alignments but rather could reflect the greater structurally inherent peptide-binding promiscuity of MHCII molecules (see Section 3.1).

Prediction of peptide–MHC binding has been approached by a large array of methods including quantitative matrices ***(32–34)***, machine learning algorithms (MLAs) ***(35,36)***, and peptide threading ***(37,38)***. Despite the fact that direct comparison between the various methods is not straightforward, some reports have indicated that MLAs such as artificial neural networks yield the best predictors of peptide–MHC binding, and it has been linked to the fact that MLAs can model binding interferences between peptide side chains, whereas the remaining methods, including PSSMs, assume independent binding of each side chain. Nevertheless, independent binding is generally supported by experimental evidence ***(32,39)***, and furthermore, considering side chain pair interactions only results in marginal improvement peptide–MHC binding predictions ***(40)***. Likewise, in a recent study, we have shown that PSSMs give similar or better results than those reported for MLAs ***(31)***. Thus, there may be more disadvantages than benefits when applying MLAs to the prediction of peptide–MHC binding. Thus, unlike PSSMs, MLAs are very prone to overfit data and are very sensitive to "dirty data." Consequently, much care and time has to be taken in preprocessing the data before training. Also, MLAs, as well as most data-driven methods used for predicting peptide–MHC binding, do not account for unseen data, instead only fitting the data they are provided with. Not surprisingly, it has also been shown that simple motif matrices outperform MLAs predicting peptide–MHC binding when the training sets are composed of a reduced set of samples (≤100 peptides) which is by large the most frequent scenario ***(4)***.

Prediction of peptide–MHC binding using PSSMs is also available at the RANKPEP web site (http://bio.dfci.harvard.edu/Tools/rankpep.html). Currently, 88 and 50 different MHCI and MHCII molecules, respectively, can be targeted for peptide-binding predictions in RANKPEP. This server is very versatile providing a framework for the prediction of MHC–peptide binding using profiles provided by the user.

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. Garcia, K.C., Teyton, L., and Wilson, I.A. 1999. Structural basis of T cell recognition. *Annu Rev Immunol* 17:369–397.
3. Wang, J.-H., and Reinherz, E.L. 2001. Structural basis of T cell recognition of peptides bound to MHC molecules. *Mol Immunol* 38:1039–1049.
4. 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.
5. Flower, D. 2003. Towards in silico prediction of immunogenic epitopes. *Trends Immunol* 24:667–674.
6. Flower, D., and Doytchinova, I.A. 2002. Immunoinformatics and the prediction of immunogenicity. *Appl Bioinformatics* 1(4):167–176.
7. Stern, L.J., and Wiley, D.C. 1994. Antigen peptide binding by class I and class II histocompatibility proteins. *Structure* 2:245–251.
8. Madden, D. 1995. The three-dimensional structure of peptide-MHC complexes. *Annu Rev Immunol* 13:587–622.
9. 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.
10. D'Amaro, J., Houbiers, J.G., Drijfhout, J.W., Brandt, R.M., Schipper, R., Bavinck, J.N., Melief, C.J., and Kast, W.M. 1995. A computer program for predicting possible cytotoxic T lymphocyte epitopes based on HLA class I peptide binding motifs. *Hum Immunol* 43:13–18.
11. 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.
12. 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.
13. 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.
14. Gribskov, M., McLachlan, A.D., and Eisenberg, D. 1987. Profile analysis: detection of distantly related proteins. *Proc Natl Acad Sci USA* 84:4355–4358.

15. Gribskov, M., and Veretnik, S. 1996. Identification of sequence pattern with profile analysis. *Methods Enzymol* 266:198–212.
16. Pearson, W. 1997. Identifying distantly related protein sequences. *Comput Appl Biosci* 13:325–332.
17. Henikoff, J.G., and Henikoff, S. 1996. Using substitution probabilities to improve position-specific scoring matrices. *Comput Appl Biosci* 12:135–143.
18. Thompson, J.D., Higgins, D.G., and Gibson, T.J. 1994. Improved sensitivity of profile searches through the use of sequence weights and gap excision. *Comput Appl Biosci* 10:19–29.
19. Reche, P.A., Zhang, H., Glutting, J.-P., and Reinherz, E.L. 2005. EPIMHC: a curated database of MHC-binding peptides for customized computational vaccinology. *Bioinformatics* 21:2140–2141.
20. Henikoff, S., Henikoff, J.G., Alford, W.J., and Pietrokovski, S. 1995. Automated construction and graphical presentation of protein blocks from unaligned sequences. *Gene* 163:17–26.
21. Bailey, T.L., and Elkan, C. 1995. The value of prior knowledge in discovering motifs with MEME. *Proc Int Conf Intell Syst Mol Biol* 3:21–29.
22. 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.
23. Barber, L.D., and Parham, P. 1993. Peptide binding to major histocompatibility complex molecules. *Annu Rev Cell Biol* 9:163–206.
24. Thompson, J.D., Higgins, D.G., and Gibson, T.J. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weigh matrix choice. *Nucleic Acids Res* 2:4673–4680.
25. Henikoff, S., Henikoff, J.G., and Pietrokovski, S. 1999. Blocks+: a non-redundant database of protein alignment blocks derived from multiple compilations. *Bioinformatics* 15:471–479.
26. Henikoff, S., and Henikoff, J.G. 1992. Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci USA* 89:10915–10919.
27. Henikoff, J.G., and Henikoff, S. 1996. Substitution probabilities to improve position-specific scoring matrices. *Comput Appl Biosci* 12:135–143.
28. Henikoff, S., and Henikoff, J.G. 1994. Position-based sequence weights. *J Mol Biol* 243:574–578.
29. Vingron, M., and Sibbald, P. 1993. Weighting in sequence space: a comparison of methods in terms of generalized sequences. *Proc Natl Acad Sci USA* 90: 8777–8781.
30. Sibbald, P., and Argos, P. 1990. Weighting aligned protein or nucleic acid sequences to correct for unequal representation. *J Mol Biol* 216:813–818.
31. Reche, P.A., Glutting, J.-P., 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.

32. 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.
33. Stryhn, A., Pederson, L.O., Romme, T., Holm, A., and Buus, S. 1996. Peptide binding specificity of major histocompatibility complex class I resolved into an array of apparently independent subspecificities: quantitation by peptide libraries and improved prediction of binding. *Eur J Immunol* 26: 1911–1918.
34. Udaka, K., Wiesmuller, K.H., Kienle, S., Jung, G., Tamamura, H., Yamigishi, H., Okumura, K., Walden, P., Suto, T., and Kawasaki, T. 2000. An automated prediction of MHC class I-binding peptides based on positional scanning with peptide libraries. *Immunogenetics* 51:816–828.
35. Adams, H.P., and Koziol, J.A. 1995. Prediction of binding to MHC class I molecules. *J Immunol Methods* 185:181–190.
36. Gulukota, K., Sidney, J., Sette, A., and DeLisi, C. 1997. Two complementary methods for predicting peptides binding major histocompatibility complex molecules. *J Mol Biol* 267:1258–1267.
37. 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.
38. Schueler-Furman, O., Altuvia, Y., Sette, A., and Margalit, H. 2000. Structure-based prediction of binding peptides to MHC class I molecules: application to a broad range of MHC alleles. *Protein Sci* 9:1838–1846.
39. Sturniolo, T., Bono, E., Ding, J., Raddrizzani, L., Tuereci, O., Sahin, U., Sinigaglia, F., and Hammer, J. 1999. Generation of tissue-specific and promiscuous HLA ligand databases using DNA microarrays and virtual HLA class II matrices. *Nat Biotechnol* 17:555–561.
40. Peters, B., Tong, W., Sidney, J., Sette, A., and Weng, Z. 2003. Examining the independent binding assumption for binding of peptide epitopes to MHC-I molecules. *Bioinformatics* 19:1765–1772.
41. Madden, D., Garboczi, D.N., and Wiley, D.C. 1993. The antigenic identity of peptide-MHC complexes: a comparison of the conformations of five viral peptides presented by HLA-A2. *Cell* 75:693–708.
42. Hennecke, J., Carfi, A., and Wiley, D.C. 2000. Structure of a covalently stabilized complex of a human alpha beta T-cell receptor, influenza HA peptide and MHC class II molecule, HLA-DR1. *EMBO J* 19:5611–5624.
43. Nicholls, A., Sharp, K., and Honig, B. 1991. Protein folding and association insights from the interfacial and thermodynamic properties of hydrocarbons. *Proteins* 11:281–296.
44. Brusic, V., Rudy, G., Kyne, A.P., and Harrison, L.C. 1998. MHCPEP, a database of MHC-binding peptides: update 1997. *Nucleic Acids Res* 26:368–371.

45. Bhasin, M., Singh, H., and Raghava, G.P. 2003. MHCBN: a comprehensive database of MHC binding and non-binding peptides. *Bioinformatics* 19:665–666.
46. Blythe, M.J., Doytchinova, I.A., and Flower, D. 2002. JenPep: a database of quantitative functional peptide data for immunology. *Bioinformatics* 18: 434–439.
47. Schonbach, C., Koh, J.L., Flower, D., Wong, L., and Brusic, V. 2002. FIMM, a database of functional molecular immunology: update 2002. *Nucleic Acids Res* 30: 226–229.