ELSEVIER

Prediction of MHC Class I Binding Peptides Using Profile Motifs

Pedro A. Reche, John-Paul Glutting, and Ellis L. Reinherz

ABSTRACT: Peptides that bind to a given major histocompatibility complex (MHC) molecule share sequence similarity. Therefore, a position specific scoring matrix (PSSM) or profile derived from a set of peptides known to bind to a specific MHC molecule would be a suitable predictor of whether other peptides might bind, thus anticipating possible T-cell epitopes within a protein. In this approach, the binding potential of any peptide sequence (query) to a given MHC molecule is linked to its similarity to a group of aligned peptides known to bind to that MHC, and can be obtained by comparing the query to the PSSM. This article describes the derivation of alignments and profiles from a collection of peptides known to bind a specific MHC, compatible with the structural and molecular basis of the peptide-MHC class I (MHCI) interaction. Moreover, in order to apply these profiles to the prediction of peptide-MHCI binding, we have developed a new search algorithm (RANKPEP) that ranks all possible peptides from an input protein using the PSSM coefficients. The predictive power of the method was evaluated by running RANKPEP on proteins known to bear MHCI K^b- and D^b-restricted T-cell epitopes. Analysis of the results indicates that $>$ 80% of these epitopes are among the top 2% of scoring peptides. Prediction of peptide-MHC binding using a variety of MHCI-specific PSSMs is available on line at our RANKPEP web server (www.mifoundation.org/Tools/rankpep.html). In addition, the RANKPEP server also allows the user to enter additional profiles, making the server a powerful and versatile computational biology benchmark for the prediction of peptide-MHC binding. *Human Immunology 63, 701–709 (2002).* © American Society for Histocompatibility and Immunogenetics, 2002. Published by Elsevier Science Inc.

KEYWORDS: PSSM; profile; epitopes; MHC; prediction

INTRODUCTION

Major histocompatibility complex (MHC) molecules play a key role in the immune system by capturing peptide antigens for display on cell surfaces. Different MHC molecules bind distinct sets of peptides. Subsequently, these peptide-MHC complexes (pMHC) are recognized by T cells via their T-cell receptors (TCR) (reviewed in references [1–4]). T-cell recognition is thus restricted to those peptides that the MHC molecules can present. Therefore, prediction of peptides that can bind to MHC molecules is important for identification of peptides capable of eliciting a T-cell response.

There are two major classes of MHC molecules, class I and class II (MHCI and MHCII, respectively) that, despite their structural similarity, differ in many ways [5]. MHCI are recognized by CD8 cytotoxic T lymphocytes (CTL), whereas MHCII are recognized by CD4 helper T cells. The type of peptide that MHCI and MHCII bind is also different. MHCI molecules bind short peptides, usually between 8 and 10 residues, with their N- and C-terminal ends pinned in the peptide binding groove [1]. In contrast, peptides bound to MHCII are longer, more variable in length (9 to 22 residues), and both the N- and C-terminal ends of the peptide can extend beyond the peptide binding groove [1, 3]. The binding motifs for MHCII are less well defined than those for MHC class I [6]. In this article, we will focus on prediction of peptide binding to MHCI molecules.

MHCI binding peptides are related by sequence similarity, and therefore prediction of pMHCI binding has traditionally been accomplished using sequence motif patterns as predictors [7]. These sequence patterns are usually extracted from large numbers of existing known peptides, or from pool sequencing experiments [6, 8]. The specific amino acids present in the pattern are called anchor residues, and the positions where they occur are

Laboratory of Immunobiology (P.A.R., J.-P.G., E.L.R.), Dana-Farber Cancer Institute, and the Department of Medicine (P.A.R., E.L.R.), Harvard Medical School, Boston, MA, USA.

Address reprint requests to: Dr. Ellis L. Reinherz, Department of Medicine, Harvard Medical School, 44 Binney Street, Boston, MA 02115, USA; Tel: +1 (617) 632-3412; Fax: +1 (617) 632-3351; E-mail: ellis_reinherz@dfci.harvard.edu.

Received March 29, 2002; revised June 12, 2002; accepted June 17, 2002.

0198-8859/02/$–see front matter
PII S0198-8859(02)00432-9

termed anchor positions [8]. For example, the sequence patterns described [8] for K^b octamers and D^b nanomers are the following:

K^b X-X-Y-X-[YF]-X-X-[LMIV]
D^b X-X-X-X-N-X-X-X-[LMIV].

Such sequence patterns, however, have proven to be too simple, as the binding ability of a peptide to a given MHC molecule cannot be explained exclusively in terms of the presence or absence of a few anchor residues [9, 10]. In response to these limitations, motif matrices have also been developed to account for the preference of every amino acid type at every position in the peptide [6, 11]. Coefficients in these matrices relate to the strength of the amino acid signals in a pool sequence of peptides eluted from a given MHCI molecule, or to the occurrence of an amino acid in a set of binding peptides. However, the precise way in which the coefficients are derived is not clear.

The above matrices offer two good efforts at representing the complexity of MHCI binding motifs [6, 11]. Nevertheless, it is well-established that position specific scoring matrices (PSSM) or profiles created from a set of aligned sequences provides a better way for defining and recognizing sequence motifs [12]. There are several methods to generate PSSM from aligned sequences, usually including distinct sequence weighting methods [13, 14]. In all cases, profile coefficients relate to the observed frequency of every amino acid at the position column of the alignment, corrected by the expected frequency of that amino acid in the background using a reference database. Thus, in this approach the binding potential of any peptide (query) to a given MHC molecule can be obtained by comparing the query to a PSSM created from a set of aligned MHCI-specific peptides. In this article we describe a new search algorithm, RANKPEP, that ranks all possible peptides from a test protein using PSSM coefficients. In addition, this study describes, for K^b and D^b molecules, that profiles created from aligned peptides are very sensitive in identifying MHCI-restricted epitopes. These profiles are guided by recent structural data indicating differences in binding residues involving peptides of distinct length. Peptide-MHC binding prediction using PSSMs are available at our RANKPEP web server (www.mifoundation.org/Tools/rankpep.html), where users can select the provided PSSMs or enter their own.

MATERIALS AND METHODS

Peptide and Protein Sequences

Sequences of peptides that bind to MHC molecules were collected from the MHCPEP database [15], which is available for downloading from the worldwide web (http://wehih.wehi.edu.au/mhcpep/). The MHC database contains 13,423 peptide entries distributed between 281 MHC specificities. All peptides in the MHCPEP database are binders, but their binding strength for specific MHC molecule is reported as unknown, low, moderate or high. This work has excluded MHC class I ligands that were ranked as low binders. Sequences of K^b (8 mers) and D^b (9 mers) restricted T-cell epitopes were collected from the literature, and from the SYFPEITHI database [6], and their protein sources were collected from the Genbank database following a blast search [16] against Genbank using these peptides as queries. In total, 37 K^b- and 34 D^b-restricted epitopes were identified.

Block Alignments and PSSM of MHCI-Specific Ligands

Peptides binding specific MHC molecules were isolated from the MHCPEP database in fasta format and curated from sequences that are closely related using the purge utility of the Gibbs sampler [17], choosing an exhaustive method and a maximum blosum62 relatedness score of 30. Typically, this purge protocol guarantees that in a set of 8 mers any peptide differs in at least four residues from any other peptide. Peptide sequences were then parsed by size in five sets of 8, 9, 10, 11, and 12+ mers to give ungapped block alignments of peptides, and profiles [12] were built for those individual sets containing at least five sequences. Profiles basically consist of a table containing the sequence-weighted frequency of each one of 20 amino acids observed in every column of the alignment divided by the corresponding expected frequency of that amino acid in the background (usually the frequency of the amino acid in the SWISSPROT database). There are, however, various protocols to make the profiles that usually vary in the weighting method used to reduce sequence redundancy. In this study we tested the profiles generated by PROFILEWEIGHT [14] and the BLK2PSSM utility included in the BLIMPS package [13, 18].

PROFILEWEIGHT uses a branch proportional weighing method, whereas BLK2PSSM can be used with the following weighting methods: P = position-based method [19]; A = pairwise distance method [20]; V = Voroni method [21]; and Cn = clustering method [22]. BLK2PSSM and PROFILEWEIGHT differ not only in the weighting method they apply, but also in the actual formula by which amino acids counts are translated into profile coefficients (to learn about the actual equations see Henikoff *et al.* [13] and Thompson *et al.* [14]).

Searching Sequences With MHCI-Specific PSSMs of MHC-Binding Peptides

To prospect protein sequences for MHC ligands using PSSM, we have written a dynamic algorithm in Python

that scores all protein segments (peptides) with the length of the PSSM width, and sorts them accordingly. Scores are obtained by aligning the PSSM with the protein segments, and adding up the profile scores that match the residue type and position in the profile. Scoring starts at the beginning of each sequence, and the PSSM is slid over the sequence one residue at a time until the end of the sequence.

Prediction of K^b- and D^b-Restricted Peptides Using PSSM

The power of profiles to correctly predict peptide-MHCI binding was assessed by checking whether empirically determined K^b- and D^b-restricted epitopes were among the top ranking peptides when their protein sources were scored using profiles derived from the relevant alignment of peptides binding K^b and D^b. For each alignment of binders, we built five different PSSMs following the methods described above. To investigate the effect of the number of MHC binding peptides in the matrix on the sensitivity of predictions, block alignments containing a decreasing number of peptides binding K^b and D^b were obtained by randomly removing ten peptides at a time from the original alignment ($35 \rightarrow 25 \rightarrow 15 \rightarrow 5$ for K^b, for example). No grouping with fewer than five alignments was utilized. PSSMs were then created for the different alignments and used to predict known K^b- and D^b-restricted epitopes. The process of randomly removing sequences from the original alignment, creating the profiles, and scoring the protein sources of K^b- and D^b-restricted epitopes was repeated 100 times, and the mean and standard deviation of the number of known epitopes found among the top 5 and 10 predicted peptides was obtained.

RESULTS AND DISCUSSION

MHCI Molecules: Correlation Between Structure and Specificity

Peptides bound to an MHCI molecule are in an extended conformation with several side chains accommodated in the binding pockets of the MHCI binding groove (Figure 1), and the N- and C-terminal pinned into the groove, connected by a network of hydrogen bonds with conserved residues of the MHCI molecule [1, 23, 24]. In turn, the binding pockets of the MHCI are delineated by polymorphic side chains, providing the molecular basis for the peptide specificity of the different MHC molecules and the correlation between peptide-binding sequence patterns (motifs) and various MHC alleles. Thus, anchor residues present in peptide-binding sequence patterns have side chains that have been selected for the geometry and chemical environment of the MHCI binding pockets. Moreover, the constraints that MHCI molecules impose on the specific residue type and overall length of the peptides to achieve binding have indeed facilitated the identification of many of these sequence patterns, which in turn have been used to predict pMHC binding [7, 8]. Nevertheless, the interaction of the peptide with the MHCI molecule is not only restricted to the primary anchor residues, and, indeed, the importance of secondary anchors and deleterious residues at nonconserved positions [9, 10] places a limitation on the usefulness of these simple patterns. Therefore, better descriptors than sequence patterns alone are required to represent the complexity of peptide-MHC binding motifs.

Defining MHCI-Specific Binding Motifs Using PSSM

Peptides binding to a specific MHC molecule are functionally related, and, therefore, a PSSM or profile derived from them should capture the complexity of the binding motif. However, for a profile to be a good descriptor of the binding motif, binding peptides must be aligned by structural and/or sequence similarity. MHCI molecules can bind peptides that differ in length by one or two amino acids, and when all peptides are aligned independently of length, gapped alignments will result. Unfortunately, the sequence similarity of peptide ligands can be very low, making the generation of such gapped alignments difficult. Moreover, peptides of different sizes can frequently bind to the same MHCI molecule in two different modes [25], through utilization of alternative binding pockets. Hence, the structure and sequence relationship between these peptides is unclear (Figure 2A). For example, Figure 2A illustrates how p2I, p3I, p5F, and p8L anchor residues of OVA octapeptide (1VAC) interact with K^b, whereas p2R, p3D, p6R, and p9M anchor residues of YGS (2VAD) interact with the same MHCI molecule differently. In particular, the third anchor residues (p5F vs. p6R in OVA versus YGS, respectively) insert into separate pockets. As a consequence, the YGS peptide mainchain arches upward and exposes more atomic contacts to the TCR. In contrast, peptides of the same size that bind to a given MHCI typically share the same binding mode, superimpose well in three-dimensional space, and possess side chains accommodated by the same binding pockets of the MHCI binding groove (Figure 2B). In view of these considerations, we have separated the peptides bound to a given MHCI molecule into subsets containing only peptides of the same length, thus creating separate, ungapped block alignments and profiles.

Prediction of Peptide-MHCI Binding Using PSSMs

Once the PSSM has been created, the binding potential of any peptide sequence (query) to the MHC molecule is

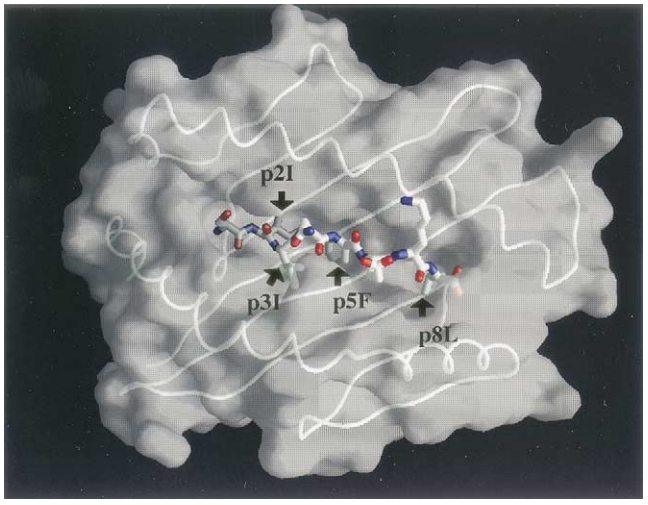

FIGURE 1 Peptide binding groove of class I MHC molecules. The figure illustrates a view of the molecular surface of the peptide binding α1-α2 antigen-presenting platform of the mouse K^b class I major histocompatibility complex (MHCI) as seen by the T-cell receptor (TCR). The K^b molecule is in complex with a peptide derived from chicken ovalbumin (SIINFEKL) represented by sticks to highlight the contours of the binding groove. The binding domain of MHCI molecules is composed of two antiparallel α-helices sitting over a base of eight antiparallel β-strands (a worm representation of these secondary structures is depicted under the molecular surface). Peptide side chains that are facing the MHCI binding groove (anchor residues p2I, p3I, p5F, and p8L) are indicated. Anchor residues make a major contribution to binding. Nevertheless, the peptide is deeply buried in the binding groove and the interaction between peptide and MHCI molecule is not restricted to the anchor residues, explaining why sequence patterns are not adequate to describe the complexity of the MHCI binding motif. The figure was derived from pdb 1VAC [39] and was prepared using the GRASP program [40]. Note that the unlabeled p4N, p6E, and p7K are exposed to the TCR.

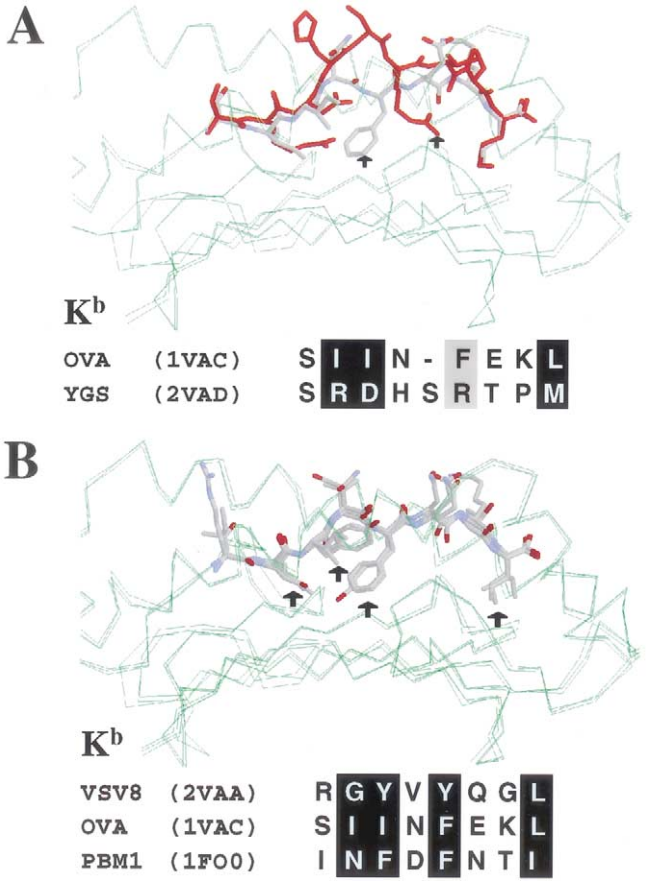

FIGURE 2 Binding mode of peptides bound to major histocompatability complex class I (MHCI) molecules. The illustration reveals the superimposition of the K^b molecule in complex with peptides of different length (panel A), and in complex with peptides of the same length (panel B). The structure-based sequence alignment of the peptides bound to the K^b molecules are depicted under the drawing. Peptide side chains that face the MHCI molecule (anchor residues) are shadowed in the alignment. Panel A: K^b molecules are in complex with peptides derived from chicken ovalbumin (OVA) and yeast α-glucosidase (YGS). YGS is indicated in red in the structure. Note the different binding modes of the two peptides with an alternative use of binding pockets resulting in poor structural superimposition. Peptide side chains using the alternative binding pocket are shadowed in gray in the sequences, and indicated with arrows in the structural drawing. Panel B: K^b molecules are in complex with eight residue peptides derived from vesicular stomatitis virus nucleoprotein (VSV8), OVA, and a naturally processed mouse octapeptide (PBM1). Note the same binding mode used by the three peptides, with their anchor residues superimposed very well in three-dimensional space, and occupying the same binding pockets. PDB names of the structures used for this analysis are given in parentheses. Structures were superimposed using TOP [41], and the figure was prepared using RASMOL [42]. For K^b only the Cα trace is given in green.

linked to its similarity to the group of aligned binding peptides and can be obtained by comparing the query to the PSSM using dynamic programming algorithms. Thus, in order to prospect a query protein for potential peptide-MHCI binders we developed a new search algorithm (RANKPEP), which uses the profile coefficients that score all possible fragments the width of the PSSM, and ranks them by score (see the Methods section for details). However, rank *per se* is insufficient to assess whether a peptide is a potential binder. Consequently, to more specifically identify potential binders, we score all the peptide sequences included in the alignment from which a profile is derived and define a binding threshold as the score value that includes 90% of the peptides within the PSSM. This binding threshold is built into each of our matrices, delineating the range of putative binders among the top scoring peptides. For example, given a random protein of 1000 amino acids, around 2% (18) of the peptides are in the binding range for the K^b molecule (Figure 3). This number may vary from profile

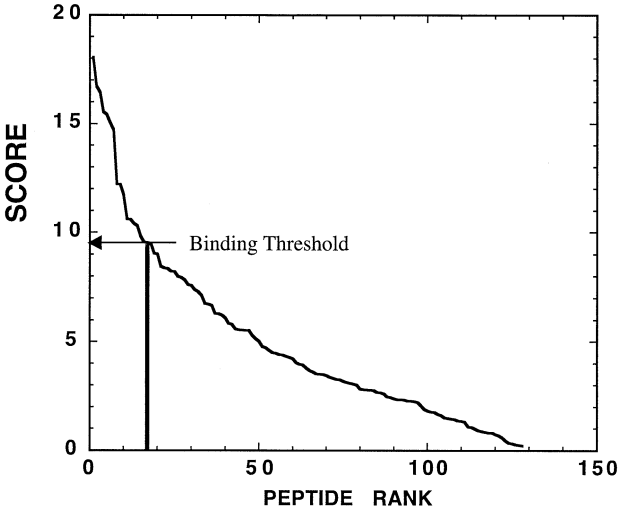

FIGURE 3 Score distribution of a random protein using a K^b-specific profile. The figure illustrates a graph of the scores of the peptides from a random protein of 1000 amino acids plotted against the ranking of the peptides. Only positive scores have been represented. Scoring was carried out using a K^b-specific profile generated using BLK2PSSM [13, 18]. The binding threshold for this specific profile had a value of 9.5. This means that 90% of the peptides from which the position specific scoring matrice (PSSM) was derived had a score ≥ 9.5 (indicated in the figure). Thus, peptides with a score equal or above the binding threshold will likely bind to the K^b molecule.

to profile, but is in accord with the fact that a given MHCI molecule binds only a subset of potential peptides derived from one protein. False-positives cannot be excluded among those selected, but in view of the small number of peptides selected by RANKPEP, this is of little practical consequence. On the other hand, the possibility of a false-negative result is more uncertain, but would imply the lack of complete descriptor of binding. At this time we can only describe the sensitivity of the predictions yielded through various PSSMs, because a false-negative or false-positive peptide assignment requires additional empirical data.

Sensitivity of PSSMs in the Prediction of K^b- and D^b-Restricted Epitopes: Comparison of Various Sequence Weighting Methods and Alignment Sizes

MHCI-specific T-cell epitopes should be expected among the high scoring peptides from within their protein sources, if PSSMs are good predictors of pMHCI binding. We checked the validity of this notion with a practical example regarding murine K^b and D^b MHCI molecules. Specifically, we identified 37 K^b and 34 D^b T-cell epitopes and their protein sources, and scored all peptides fragments from their respective proteins using relevant profiles derived from K^b- and D^b-binding peptides. Subsequently, we determined whether the naturally restricted peptides were among the top scoring peptides according to RANKPEP. It is known that sequence weighting increases the sensitivity of the profiles. However, in the absence of a general consensus about the optimal sequence weighting methodology, five different types of profile predictors were tested: one generated using PROFILEWEIGHT [14], which uses a branch-proportional sequence weighting method; and four generated with BLK2PSSM [13, 18], in combination with different weighting methods (see the Materials and Methods section for more details). The correctly predicted K^b- and D^b-restricted epitopes among the top 1, top 3, top 5, and top 10 scoring peptides are listed in Table 1. Overall, the predictions are quite robust given that alignments and profiles were derived in an automated way. Thus, over 80% of the known K^b- and D^b-restricted epitopes were found within the top 10 scoring peptides, regardless of the specific profile used.

TABLE 1 Prediction of K^b- and D^b-restricted peptides from their protein sources

	MHC	Top 1	Top 3	Top 5	Top 10
PW	K^b	11 (29.73%)	22 (59.46%)	24 (64.86%)	30 (81.08%)
	D^b	15 (44.12%)	21 (61.76%)	23 (67.65%)	28 (82.35%)
P	K^b	14 (37.84%)	24 (64.86%)	26 (70.27%)	30 (81.08%)
	D^b	18 (52.94%)	24 (70.59%)	28 (82.35%)	29 (85.29%)
A	K^b	14 (37.84%)	23 (62.16%)	28 (75.68%)	31 (83.78%)
	D^b	16 (47.06%)	24 (70.59%)	27 (79.41%)	29 (85.29%)
V	K^b	14 (37.84%)	22 (59.46%)	26 (70.27%)	31 (83.78%)
	D^b	15 (44.12%)	25 (73.53%)	28 (82.35%)	29 (85.29%)
Cn	K^b	14 (37.84%)	24 (64.86%)	28 (75.68%)	31 (83.78%)
	D^b	16 (47.06%)	24 (70.59%)	27 (79.41%)	28 (82.35%)

For each alignment of K^b and D^b binders, five profiles were built. PW profiles were built using PROFILEWEIGHT, which uses a branch proportional weighting method [14]. P, A, V, and Cn matrices were built using BLK2PSSM with the following weighting methods: P = position-based method [19]; A = pairwise distance method [20]; V = Voroni method [21]; and Cn = clustering method [22]. K^b profiles were then used to score 37 protein sequences, each bearing one of the 37 K^b-restricted epitopes. D^b profiles were used to score 34 proteins, each bearing one of the 34 identified D^b-restricted epitopes. The table illustrates the number and percentage of K^b- and D^b-restricted epitopes that were included among the top 1, top 3, top 5, and top 10 scoring peptides.

Also, taking into account that the average sequence length of the protein sources of K^b- and D^b-restricted peptides were 502 aa and 597 aa, respectively, it follows that over 80% of the known restricted peptides appeared in ~2% of the top scoring peptides. It is also noteworthy that only 8 of K^b-restricted peptides, and 11 of D^b-restricted peptides were actually included in the alignment from which the profiles were derived. We see no clear differences between the five sets of prediction results obtained from profiles derived using BLK2PSSM and PROFILEWEIGHT.

In order to address the question of how many peptides are required for a proper representation of the MHC-peptide binding motif that would yield appropriate predictions, we have also used profiles derived from alignments of different size (see the Materials and Methods section). The results illustrated in Figure 4 indicate that ~60% of the restricted epitopes are found within the top 2% of scoring peptides when using profiles derived from alignments that contained only five sequences. Interestingly, for the two smallest alignments (15 or less sequences), the trend was for a marginally better prediction if profiles were made using PROFILEWEIGHT. Given the results, we have chosen PROFILEWEIGHT for generation of all matrices.

Availability

Predictions of peptide-MHCI binding using PSSMs are available online from the Molecular Immunology Foundation web server hosted by the Dana-Farber Cancer Institute (http://www.mifoundation.org/Tools/rankpep.html). Currently, the site contains profiles from 57 different MHCI molecules that, by default, are made using PROFILEWEIGHT. The average number of peptides in the block alignments from which we derived the profiles is 29, and profiles were built only if the alignment contained a minimum of five sequences. The largest alignment contains 162 peptide sequences, corresponding to those binding to the HLA-A2*0201 allele. Generally, for each MHCI specificity, individual alignments are derived corresponding to sets of peptides of each different length accommodated by that MHC molecule (see the Materials and Methods section). Every profile yields an optimum sequence (consensus) that gives the highest score, and thus for every sorted peptide, the server outputs its score and the percentage of the optimum score. The number of scored peptides returned by the server is selected by the user. In addition, those peptides whose scores are equal or greater than the binding threshold score will be highlighted. Finally, the server also returns the amino acid position of the peptide in the original sequence as well as its molecular weight. We are confident that the protocol we have followed to build the PSSMs is appropriate to represent the binding motif of that set of MHCI binding peptides. Nevertheless, our matrices are limited by the quality of the

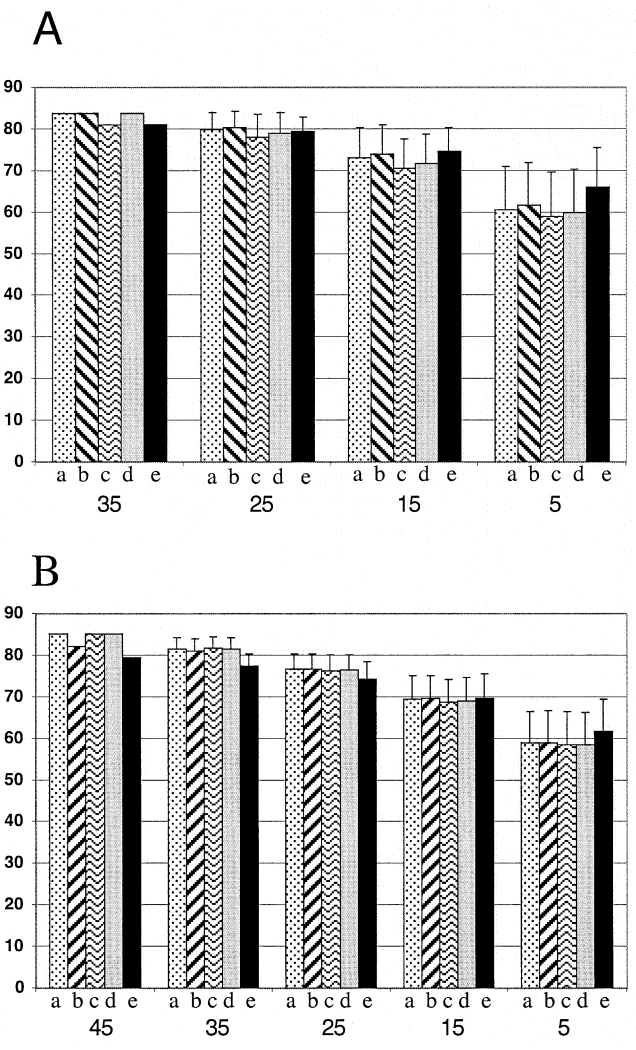

FIGURE 4 Prediction of K^b- and D^b-restricted peptides using position specific scoring matrice (PSSM) derived from alignments of different size. Proteins known to contain K^b- and D^b-restricted epitopes (37 K^b and 34 D^b ligands) were scored using various PSSMs derived from alignments containing a variable number of peptides (indicated in the figure), and the percentage of restricted peptides found in the top ten scoring peptides are represented in the figure. Alignments were generated by removing ten peptides at random from the previous alignment (see the Materials and Methods section), and five profiles were derived for each alignment: a = BLK2PSSM with a pairwise distance sequence weighting method [20]; b = BLK2PSSM with a clustering sequence weighting method [22]; c = BLK2PSSM with position-based sequence weighting method [19]; d = BLK2PSSM with Voroni distance sequence weighting method [21]; and e = PROFILEWEIGHT [14]. The process of creating the alignments, profiles, and running the predictions was repeated 100 times, and thus the values represent the mean of the percentage with standard deviations noted. Panel A = K^b, Panel B = D^b.

sequences we started with, and, therefore, we have also given the user the possibility of entering their own matrices. In fact, the server can input most profile formats as long as a header line with the amino acid types in the profile columns is included.

CONCLUSIONS

CTL responses rely on the recognition of peptides that must be presented on the target cell surface by MHC class I molecules. Therefore, determination of peptides that bind to MHCI molecules is important and has been approached by several methods, including quantitative matrices [26–28], neural networks [29, 30], and peptide threading [31, 32]. Although a direct comparison between the various methods is not straightforward due to the different criteria followed by authors to assess the power of their predictors, it seems that overall quantitative matrices and neural networks yield similarly good results, whereas peptide threading remains under development. Quantitative matrices are generated from actual binding measurements of peptide interactions with a given MHC molecule, and those generated by Parker *et al.* [26] are indeed publicly available. Prediction matrices of Parker *et al.* [26] were generated from a limited set of peptides, perhaps explaining a recent report [33] describing poor correspondence between the predicted MHC binding peptides and those determined experimentally. Quantitative matrices have also been derived from positional scanning combinatorial peptide libraries (PSCPL) [27, 28], where all possible peptides of a given length are represented by sets of sublibraries and in each sublibrary, one amino acid is kept fixed whereas the remaining positions contain mixtures of all amino acids. Unfortunately, to date, prediction of pMHCI binding using these PSCPL-derived matrices is not freely accessible. Moreover, the generation of those matrices requires substantial investment of money and time. Prediction of pMHCI binding through neural network algorithms is also unavailable to the public, and although predictors could be derived by training on existing collections of peptide binding data, the technique itself is not readily within reach of the average researcher.

For the above reasons, sequence motifs remain one of the most popular methods applied to the prediction of pMHCI binding. In this regard, peptides are useful for representing sequence motifs, and indeed popular databases such as the BLOCK [18] and PROSITE databases [33] contain sets of motif profiles derived from protein families that are used for the functional classification of new sequences via their similarity to these profiles. In this study, we applied the concept to the prediction of peptide-MHCI binding. Thus, we have generated a collection of profile motifs from MHCI-specific binding peptides that are available online for the detection of pMHCI binding sequences using a dynamic search algorithm (RANKPEP). Our profiles of MHCI-specific binding peptides have been generated following a protocol to minimize redundancy of the initial data, and, importantly, taking into account recent structural insights into the basis of the interaction between the peptide and the MHCI molecule [5]. A similar approach to profiles using a hidden Markov motif (HMM) was previously applied to the prediction of HLA-A2 binding peptides [34]. However, prediction of pMHCI binding using HMMs is not broadly available, and moreover, HMMs were built without taking into account the fact that peptides of different lengths can utilize different binding modes.

Our RANKPEP server provides a framework for the prediction of MHC-peptide binding using custom profiles, such as those obtained from the automated motif discovering programs MEME (http://meme.sdsc.edu/meme/website/) [35], the MOTIF SAMPLER [17], and PROTOMAT (http://www.blocks.fhcrc.org/) [36], making our server also useful for the prediction of class II MHC peptides. Class II MHC motifs are harder to define than class I MHC motifs for two reasons: (1) MHC class II molecules bind peptides that range from 9 to 22 residues in length, yet only nine residues fit in the binding groove [3, 37]; and (2) MHC class II molecules impose fewer restrictions than MHCI on the type of side chains that can be accommodated into their binding pockets [3, 37]. Thus, most sequence alignment tools, such as CLUSTALW [38] that perform global alignments will fail to correctly align a set of MHC class II binding peptides. In contrast, motif discovering programs are optimized to find short ungapped sequence motif patterns from within a set of related sequences of variable length, and therefore are more likely to succeed in finding the sequence binding motif from a collection of peptides binding to a given MHC molecule. We believe that RANKPEP constitutes a powerful as well as a flexible benchmark for the prediction of peptide-MHCI binding at a level readily accessible to most researchers investigating immune recognition-based diseases.

ACKNOWLEDGMENTS

This work was supported by NIH grant AI50900. Doctor Reche is supported by funds from the Molecular Immunology Foundation. We thank Drs. Linda Clayton, Masha Fridkis-Hareli, Rob Meijers, Esther Lafuente, Bruce Reinhold, and Jia-huai Wang for reading and comments.

REFERENCES

1. Madden DR: The three-dimensional structure of peptide-MHC complexes. Annu Rev Immunol 13:587, 1995.
2. Margulies DH: Interactions of TCRs with MHC-peptide

complexes: a quantitative basis for mechanistic models. Curr Opin Immunol 9:390, 1997.

3. Stern LJ, Wiley DC: Antigen peptide binding by class I and class II histocompatibility proteins. Structure 2:245, 1994.
4. Wang J-H, Reinherz EL: Structural basis of cell-cell interactions in the immune system. Curr Opin Structural Biol 10:656, 2000.
5. Marsh SGE, Parham P, Barber LD: The HLA Facts Book. New York: Academic Press, 2000.
6. Rammensee HG, Bachmann J, Emmerich NPN, Bacho OA, Stevanovic S: SYFPEITHI: database for MHC ligands and peptide motifs. Immunogenetics 50:213, 1999.
7. D'Amaro J, Houbiers JG, Drijfhout JW, Brandt RM, Schipper R, Bavinck JN, Melief CJ, Kast WM: A computer program for predicting possible cytotoxic T lymphocyte epitopes based on HLA class I peptide binding motifs. Hum Immunol 43:13, 1995.
8. Falk K, Rotzschke O, Stevanovic S, Jung G, Rammensee HG: Allele-specific motifs revealed by sequencing of self-peptides eluted from MHC molecules. Nature 351:290, 1991.
9. Bouvier M, Wiley DC: Importance of peptide amino acid and carboxyl termini to the stability of MHC class I molecules. Science 265:398, 1994.
10. Ruppert J, Sidney J, Celis E, Kubo RT, Grey HM, Sette A: Prominent role of secondary anchor residues in peptide binding to HLA-A2: 1 molecules. Cell 74:929, 1993.
11. De Groot AS, Jesdale BM, Szu E, Schafer JR, Chicz RM, Deocampo G: An interactive web site providing major histocompatibility ligand predictions: application to HIV research and AIDS. AIDS Res Hum Retroviruses 13:529, 1997.
12. Gribskov M, McLachlan AD, Eisenberg D: Profile analysis: detection of distantly related proteins. Proc Natl Acad Sci USA 84:4355, 1987.
13. Henikoff JG, Henikoff S: Substitution probabilities to improve position-specific scoring matrices. Comput Appl Biosci 12:135, 1996.
14. Thompson JD, Higgins DG, Gibson TJ: Improved sensitivity of profile searches through the use of sequence weights and gap excision. Comput Appl Biosci 10:19, 1994.
15. Brusic V, Rudy G, Kyne AP, Harrison LC: MHCPEP, a database of MHC-binding peptides: update 1997. Nucl Acids Res 26:368, 1998.
16. Altschul SF, Madden TL, Schaffer AA, Zhang Z, Miller W, Lipman DJ: Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucl Acids Res 25:3389, 1997.
17. Neuwald AF, Liu JS, Lawrence CE: Gibbs motif sampling detection of bacterial outer membrane protein repeats. Protein Sci 4:1618, 1995.
18. Henikoff S, Henikoff JG, Pietrokovski S: Blocks+: a non-redundant database of protein alignment blocks derived from multiple compilations. Bioinformatics 15:471, 1999.
19. Henikoff S, Henikoff JG: Position-based sequence weights. J Mol Biol 243:574, 1994.
20. Vingron M, Sibbald PR: Weighting in sequence space: a comparison of methods in terms of generalized sequences. Proc Natl Acad Sci USA 90:8777, 1993.
21. Sibbald PR, Argos P: Weighting aligned protein or nucleic acid sequences to correct for unequal representation. J Mol Biol 216:813, 1990.
22. Henikoff S, Henikoff JG: Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci USA 89: 10915, 1992.
23. Matsumura M, Fremont D, Peterson PA, Wilson IA: Emerging principles for the recognition of peptide antigens by MHC class I molecules. Science 257:927, 1992.
24. Zhang C, Anderson A, DeLisi C: Structural principles that govern the peptide-binding motifs of class I MHC molecules. J Mol Biol 281:929, 1998.
25. Madden DR, Garboczi DN, Wiley DC: The antigenic identity of peptide-MHC complexes: a comparison of the conformations of five viral peptides presented by HLA-A2. Cell 75:693, 1993.
26. Parker KC, Bednarek MA, Coligan JE: Scheme for ranking potential HLA-A2 binding peptides based on independent binding of individual peptide side chains. J Immunol 152:163, 1994.
27. Stryhn A, Pederson LO, Romme T, Holm A, Buus S: 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, 1996.
28. Udaka K, Wiesmuller KH, Kienle S, Jung G, Tamamura H, Yamigishi H, Okumura K, Walden P, Suto T, Kawasaki T: An automated prediction of MHC class I-binding peptides based on positional scanning with peptide libraries. Immunogenetics 51:816, 2000.
29. Adams HP, Koziol JA: Prediction of binding to MHC class I molecules. J Immunol Methods 185:181, 1995.
30. Gulukota K, Sidney J, Sette A, DeLisi C: Two complementary methods for predicting peptides binding major histocompatibility complex molecules. J Mol Biol 267: 1258, 1997.
31. Altuvia Y, Sette A, Sidney J, Southwood S, Margalit H: A structure-based algorithm to predict potential binding peptides to MHC molecules with hydrophobic binding pockets. Hum Immunol 58:1, 1997.
32. Schueler-Furman O, Altuvia Y, Sette A, Margalit H: Structure-based prediction of binding peptides to MHC class I molecules: application to a broad range of MHC alleles. Protein Sci 9:1838, 2000.
33. Hofmann K, Bucher P, Falquet L, Bairoch A: The PROS-

ITE database, its status in 1999. Nucl Acids Res 27:215, 1999.

34. Mamitsuka H: Predicting peptides that bind to MHC molecules using supervised learning of hidden Markov models. Proteins 33:460, 1998.
35. Bailey TL, Elkan C: The value of prior knowledge of discovering motifs with MEME. Proc Intern Conf Intell Syst Mol Biol 3:21, 1995.
36. Henikoff S, Henikoff JG, Alford WJ, Pietrokovski S: Automated construction and graphical presentation of protein blocks from unaligned sequences. Gene 163:17, 1995.
37. Hammer J, Bono E, Gallazzi F, Belunis C, Nagy Z, Sinigaglia F: Precise prediction of major histocompatibility complex class II-peptide interaction based on peptide side chain scanning. J Exp Med 267:1258, 1994.
38. Thompson JD, Higgins DG, Gibson TJ: CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weigh matrix choice. Nucl Acids Res 2:4673, 1994.
39. Fremont DH, Stura EA, Matsumura M, Peterson PA, Wilson IA: Crystal structure of an H-2K ovalbumin peptide complex reveals the interplay of primary and secondary anchor positions in the major histocompatibility complex binding groove. Proc Natl Acad Sci USA 92:2479, 1995.
40. Nicholls A, Sharp K, Honig B: Protein folding and association insights from the interfacial and thermodynamic properties of hydrocarbons. Proteins 11:281, 1991.
41. Lu G: TOP: A new method for protein structure comparison and similarity searches. J Appl Cryst 33:176, 2000.
42. Sayle RA, Milner-White EJ: RASMOL: biomolecular graphics for all. Trends Biochem Sci 20:374, 1995.