PROTEINS: Structure, Function, and Bioinformatics 63:512–526 (2006)

Ab Initio Prediction of Peptide-MHC Binding Geometry for Diverse Class I MHC Allotypes

Andrew J. Bordner[1,2*] **and Ruben Abagyan**[1]
[1]*Department of Molecular Biology, The Scripps Research Institute, San Diego, California*
[2]*Computer Science and Mathematics Division, Oak Ridge National Laboratory, Oak Ridge, Tennessee*

ABSTRACT **Since determining the crystallographic structure of all peptide-MHC complexes is infeasible, an accurate prediction of the conformation is a critical computational problem. These models can be useful for determining binding energetics, predicting the structures of specific ternary complexes with T-cell receptors, and designing new molecules interacting with these complexes. The main difficulties are (1) adequate sampling of the large number of conformational degrees of freedom for the flexible peptide, (2) predicting subtle changes in the MHC interface geometry upon binding, and (3) building models for numerous MHC allotypes without known structures. Whereas previous studies have approached the sampling problem by dividing the conformational variables into different sets and predicting them separately, we have refined the Biased-Probability Monte Carlo docking protocol in internal coordinates to optimize a physical energy function for all peptide variables simultaneously. We also imitated the induced fit by docking into a more permissive smooth grid representation of the MHC followed by refinement and reranking using an all-atom MHC model. Our method was tested by a comparison of the results of cross-docking 14 peptides into HLA-A*0201 and 9 peptides into H-2*K*^{b} as well as docking peptides into homology models for five different HLA allotypes with a comprehensive set of experimental structures. The surprisingly accurate prediction (0.75 Å backbone RMSD) for cross-docking of a highly flexible decapeptide, dissimilar to the original bound peptide, as well as docking predictions using homology models for two allotypes with low average backbone RMSDs of less than 1.0 Å illustrate the method's effectiveness. Finally, energy terms calculated using the predicted structures were combined with supervised learning on a large data set to classify peptides as either HLA-A*0201 binders or nonbinders. In contrast with sequence-based prediction methods, this model was also able to predict the binding affinity for peptides to a different MHC allotype (H-2*K*^{b}), not used for training, with comparable prediction accuracy. Proteins 2006;63:512–526.** © 2006 Wiley-Liss, Inc.

Key words: **peptide docking; major histocompatibility complex (MHC); Monte Carlo optimization; homology models; potential grid; peptide binding prediction**

INTRODUCTION

The binding of short peptide fragments of endogenous and foreign proteins to class I major histocompatibility complex (MHC) glycoproteins is a necessary first step in the immune surveillance by circulating cytotoxic T-cells. Peptides resulting from proteosomal processing of cytosolic proteins are transported to the endoplasmic reticulum by the transporter associated with antigen processing (TAP) where they bind to newly synthesized MHC molecules. The resulting complex is then transported to the cell surface where the MHC is inserted into the membrane. These complexes are then recognized by $CD8^+$ T-cells through peptide and MHC allele specific interactions with the T-cell receptor (TCR) as well as conserved interactions with the CD8 coreceptor.

MHC molecules are polymorphic with most variable residues in the peptide binding pocket so that each allotype preferentially binds a distinct subset of peptides. Since, for example, an individual human can have cells expressing up to six different allotypes, this diversity presumably prevents potential antigens from escaping recognition by the cellular immune system. Also, a particular MHC allotype can strongly bind a large number of 8–11 residue peptides. Although most have preferred residue types in primary or secondary anchor positions, this is neither necessary nor sufficient for strong binding.[1,2] This extreme variability in both components of the peptide-MHC complex together with the limited number of available X-ray structures make computational prediction of the complex an important goal in molecular biology.

Accurate models of peptides bound to MHC are essential for structure-based prediction of peptide binding affinities. Position-specific scoring matrices[3–5] and machine learning methods[6–9] can predict peptide-MHC binding affinity reasonably accurately when a large amount of experimental

The Supplementary Material referred to in this article can be found at http://www.interscience.wiley.com/jpages/0887-3585/suppmat/

*Correspondence to: Andrew J. Bordner, Computer Science and Mathematics Division, Oak Ridge National Laboratory, P.O. Box 2008, MS 6173, Oak Ridge, TN 37831. E-mail: bordner@ornl.gov

Grant sponsor: National Institute of Health; Grant number: 1R01GM071872-01. Grant sponsor: U.S. Department of Energy Genomics: GTL and Biopilot grants; Grant number: ORNL is operated under DOE contract number DE-AC05-00OR22725.

Received 3 May 2005; Revised 12 September 2005; Accepted 11 October 2005

Published online 7 February 2006 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.20831

binding data for related equal-length peptides binding to a particular MHC allotype is available. However, their use is limited to a small number of MHC allotypes with sufficient quantities of such data. In contrast, physics-based scoring functions combined with accurate structural models should provide general binding affinity predictions, applicable to different peptide lengths and MHC allotypes, for use in vaccine design.

Peptide-MHC docking methods also provide a starting point for future computational studies with important biomedical applications. These include computational screening of peptide analogues that bind MHC molecules, predicting specific structures of ternary complexes with T-cell receptors (TCR), and designing small molecules that interact with the complex. Peptide analogues have a similar size, number of rotatable bonds, and structure so that adapting peptide docking methods is expected to be straightforward. Peptide analogues designed to either stimulate[10, 11] or block[12] an immune response have previously been shown to bind to class I MHC. Another possible future application, the structural prediction of TCR binding to the peptide-MHC complex, should give a detailed understanding of the atomic interactions responsible for binding affinity and specificity. Because comparison of some X-ray crystal structures show little change in peptide conformation upon TCR binding and others show larger but local changes[13–17] no general conclusion of peptide conformational changes upon TCR binding can be drawn and thus some local flexibility in the peptide may be necessary for successful TCR docking. It also opens up the possibility of designing therapeutic compounds that modulate this interaction and consequently the immune system response.

The peptide-class I MHC complex is particularly conducive to computational docking approaches to structure prediction because all X-ray structures, with one exception,[18] show little deviation in the positions of the peptide N- and C-terminii.[19] The experimental structures also indicate a conserved network of hydrogen bonds between MHC side chain atoms and peptide main chain atoms near the terminii that contribute to the common peptide main chain structure in those regions. Thus, the peptide's conformational freedom may be viewed as being limited to the center and docking becomes more like protein loop modeling. However, the conformational prediction of the peptide-MHC complex remains a formidable problem both because of the exponential dependence of the size of the conformational space to be searched on the large number of flexible degrees of freedom in the interface and errors in the physical energy function that make it difficult to distinguish the correct conformation from those with similar energies.

All previous methods for calculating the peptide-binding conformation divide the structure prediction into separate parts, either (1) the peptide terminii and central portion, (2) the peptide main chain and side chains, or (3) individual peptide residues. One study[20] used a multiple copy approximation, in which multiple peptide conformations are sampled simultaneously with each peptide independent of the other but with the receptor residues moving in the mean potential of all peptides, to predict the conformations of peptide-MHC complexes. The terminal residues were first docked, followed by sampling of the remainder of the peptide using a bond-scaling relaxation algorithm to insure loop closure. Another study,[21] whose main focus was the prediction of peptide binding affinity, also predicted the peptide terminii and center separately, using other peptide-MHC structures to assign the terminal residue conformation and a search for loops in the Protein Data Bank (PDB) with similar sequences and stem orientations to model the center, followed by local optimization. A recent study by Tong et al.[22] used rigid docking of the peptide end residues followed by homology-based prediction of the central portion of the peptide using MODELLER,[23] and finally refinement using energy minimization with the peptide C_α atoms restrained to their original positions.

Three peptide–MHC docking methods divided the prediction into subproblems for the peptide main chain and side chains. One method[24] modeled the peptide using a threading approach based on X-ray structures to rank the binding affinity of the peptide. Another method[25] first predicted the conformations of independent peptide side chain conformations by optimizing precomputed free energy maps for individual residue side chains in each receptor pocket followed by the assignment of the backbone conformation and local energy optimization. A third method[26] used dead end elimination (DEE) to select peptide side chain rotamers modeled on a fixed backbone from available X-ray structures.

Finally, one study[27] described a docking method using DEE with a combinatorial buildup algorithm, in which consecutive residues are added to peptide fragments of increasing length, to sample the conformational space of the peptide and nearby receptor side chains. This method was evaluated using the structures of two H-2K^b complexes as well as a peptide binding calmodulin.

In contrast to previous studies, we tested and optimized fully converged peptide-MHC docking simulations using a flexible all-atom model of the complete peptide. A biased-probability Monte Carlo minimization method[28] implemented in the ICM[29] program, combined with grid potentials to represent interactions with the MHC, allows a computationally efficient sampling of the peptide degrees of freedom. The lowest energy conformations from the grid docking simulations were then reranked using the energy of an all-atom model of the complex after local minimization. The ICM Monte Carlo method has previously been applied to two other peptide docking problems, phosphotyrosine peptides binding to SH2 and PTB domains[30] and peptides epitopes binding to an IgG1 monoclonal antibody.[31]

All docking simulations were started with the peptide in an extended conformation and the length of the simulations was shown to be sufficient for convergence, as measured using three independent runs. Three different types of docking calculations were performed: (1) redocking of peptides into the MHC structure from the correspond-

ing complex, (2) cross-docking of peptides into the MHC structure from the complex with another peptide, and (3) docking of peptides into homology models of various MHC allotypes. While docking calculation (1) provides a check of the adequacy of the energy function and conformational sampling, docking calculations (2) and (3) solve the realistic problems of predicting the peptide geometry when the structure of the MHC in complex with the peptide is unknown or even when the structure of the particular MHC allotype is unknown. The latter problem is particularly acute since crystal structures are currently available for only 10 class-I HLA allotypes even though there are more than 1,000 allotypes that have been catalogued so far.[32,33]

Since examination of the HLA-A*0201 side chain conformations for all available X-ray crystal structures showed that they clustered into two groups, one representative structure from each group, PDB entries 1JF1 and 1I7U, was used for docking. Peptides from 16 HLA-A*0201 structures were docked and then compared with the structure of the corresponding complex in order to validate the procedure. The peptides were docked into both HLA models and the lowest energy conformation from both simulations was selected as the final prediction. Next, the peptides for 10 H-2K^b murine MHC structures were docked into an MHC model based on PDB entry 1KPU and compared with the respective structures of the complex. Only a single MHC model was used since MHC side chain conformations in the binding interface had less variability and did not fall into well-separated clusters. In addition, peptides were docked into homology models of five different allotypes, HLA-B*0801, B*2705, B*3501, B*5101, and B*5301, using an HLA-A*0201 structure as a template and compared with all available X-ray crystal structures in order to assess the accuracy of the prediction results.

Finally, we demonstrated the utility of our peptide–MHC docking method by predicting the binding affinity of peptides, even to a different MHC allotype than that used for training the model. A Support Vector Machine (SVM) trained on the binding energy components calculated from the predicted geometry of the complex combined with peptide residue composition was used to discriminate binders from nonbinders in a large data set comprised of 304 HLA-A*0201 binding peptides and an equal number of random nonbinding peptides. Furthermore, the universality of the energy-based binding prediction was shown by using the SVM trained on HLA-A*0201 peptides to predict the binding affinities of 54 H-2K^b peptides. This is the first time, to the best of our knowledge, that a peptide–MHC binding prediction model trained for a particular MHC allotype has been used for an accurate prediction on a significantly different MHC, with different anchor residues and even from a different organism (mouse vs. human). This illustrates that a structure-based peptide–MHC binding affinity prediction method, although slower than sequence-based methods, is generalizable to other MHC allotypes, unlike the sequence-based approach.

MATERIALS AND METHODS

ICM docking

Grid potentials

An all-atom model of the peptide was docked into grid potentials derived from an X-ray structure of the MHC molecule using a stochastic global optimization in internal coordinates with pseudo-Brownian and collective "probability-biased" random moves[28] as implemented in the ICM 3.0 program.[29] Five types of potentials[34,35] for the peptide–MHC interaction energy

$$E_{peptide-MHC} = E_{Cvw} + E_{Hvw} + 0.87E_{hb} + 3.68E_{el} + 1.58E_{hp} \quad (1)$$

were precomputed on a rectilinear grid with 0.5 Å spacing that fills a 34 Å × 34 Å × 25 Å box containing the peptide binding domain of the MHC (residues 1–180). The determination of the weights multiplying E_{hb}, E_{el}, and E_{hp}, is described below. The van der Waals grid potentials E_{Cvw} and E_{Hvw}, for nonhydrogen and hydrogen atoms, respectively, were calculated from a van der Waals (vdW) energy, which is smoothed by introducing a cutoff value E_{vw}^{max} = 3.0 kcal/mol [refer to Eq. (3) in Fernández-Recio *et al.*[36]]. The energy cutoff reduces the extreme sensitivity of the vdW potential to small conformational changes and speeds convergence of local minimization of the energy function. The hydrogen bonding (E_{hb}) and hydrophobic (E_{hp}) potentials were calculated as described previously[36] and the electrostatic energy (E_{el}) was calculated using a distance-dependent dielectric constant $\varepsilon = 4r$. The peptide-MHC intermolecular energy calculated using these grid potentials was added to the peptide intramolecular energy, $E_{peptide}$, which was calculated using the truncated vdW energy with cutoff E_{vw}^0 = 7.0 kcal/mol, the distance-dependent dielectric electrostatic term, ECEPP/3[37–39] hydrogen-bonding and torsional potentials, and a side chain entropic term proportional to the fractional SASA.[28]

The weights multiplying the grid potentials E_{hb}, E_{el}, and E_{hp} in Eq. 1 were determined by simulated annealing minimization[40] of the average rank of the near-native conformation in the stack. Simulation results with all weights set to 1 for the HLA-A*0201 and H-2K^b peptides in Tables I and II docked to 1JF1 and 1KPU MHC structures, respectively, were used to calculate the objective function. The optimal weights in Eq. 1 were then used in all docking simulations. As expected, these weights differ from those previously derived for protein-protein docking.[36]

Grid docking protocol

An all-atom model of the peptide with charged N- and C-terminii and idealized covalent geometry was first generated in an extended conformation. Comparison of all available X-ray structures reveals conserved hydrogen bonds between the N- and C-terminii of the peptide and particular MHC residues, which cause the peptide backbone to adopt similar conformations in these regions.[19] Based on this observation, a quadratic restraint energy $E_{restraint} = kR_{ij}^2$ with strength k = 10 kcal/(molÅ^2) was imposed between corresponding atoms on the peptide to be

TABLE I. Comparison of the Results of Docking Peptides Into the Dual Grid HLA-A *0201 MHC Model With the Corresponding X-Ray Crystal Structures of the Peptide–MHC Complexes[†]

				RMSD (Å)			
PDB entry (Ref.)	Resolution (Å)	Peptide sequence	Best MHC	Backbone	Central backbone	Buried non-H	MHC
Cross-docking to the dual HLA-A *0201 model							
1B0G ([58])	2.50	ALWGFFPVL	B	0.54	0.59	0.66	0.34
1EEY ([51])	2.25	ILSALVGIV	A	0.54	0.56	0.88	0.36
1EEZ ([51])	2.30	ILSALVGIL	A	0.76	0.81	1.28	0.41
1HHG ([59])	2.60	TLTSCNTSV	A	1.21	1.33	2.10	0.48
1HHH ([59])	3.00	FLPSDFFPSV	A	1.56	1.67	2.42	0.51
1HHI ([59])	2.50	GILGFVFTL	B	0.81	0.88	1.57	0.36
1HHJ ([59])	2.50	ILKEPVHGV	A	1.74	1.92	2.53	0.43
1HHK ([59])	2.50	LLFGYPVYV	A	1.26	1.39	2.50	0.42
1I1F ([52])	2.80	FLKEPVHGV	A	1.51	1.66	2.55	0.39
1I1Y ([52])	2.20	YLKEPVHGV	A	1.42	1.56	2.28	0.38
1I7R ([60])	2.20	FAPGFFPYL	A	0.87	0.92	1.41	0.40
1I7T ([60])	2.80	ALWGVFPVL	B	0.40	0.41	0.86	0.36
1I4F ([61])	1.40	GVYDGREHTV	A	0.75	0.79	1.62	0.52
1JHT ([62])	2.15	ALGIGILTV	A	1.88	2.08	2.26	0.27
Average RMSD (Å)				1.09	1.18	1.78	0.40
Self-docking to the dual HLA-A *0201 model							
1JF1 ([62])	1.85	ELAGIGILTV	A	0.22	0.23	0.56	0.00
1I7U ([60])	1.80	ALWGFVPVL	B	0.23	0.25	0.21	0.00

[†]All peptides were docked into both the 1JF1 and 1I7U MHC structures and the lowest energy conformation from both calculations chosen as the prediction. The MHC structure used is denoted as either A or B for 1JF1 or 1I7U, respectively. The RMSD between the docked peptide and the X-ray structure was calculated after first aligning the backbone atoms of the MHC peptide binding domain (residues 1–180). The resulting RMSDs for the backbone atoms, central backbone atoms (residues 3-M-1), all nonhydrogen atoms in buried residues, and the MHC backbone atoms are shown.

TABLE II. Comparison of Results of Docking Peptides Into an H-2K^b MHC Model Based on the 1KPU Structure With the Corresponding X-Ray Crystal Structures of the Peptide–MHC Complexes[†]

			RMSD (Å)			
PDB entry (Ref.)	Resolution (Å)	Peptide sequence	Backbone	Central Backbone	Buried non-H	MHC
Cross-docking to the single H-2K^b model						
1G7P ([63])	1.50	SRDHSRTPM	1.40	1.54	2.09	0.25
1G7Q ([64])	1.50	SAPDTRPA	0.73	0.78	1.73	0.21
1KJ3 ([16])	2.30	KVITFIDL	0.57	0.55	1.17	0.76
1KPV ([65])	1.71	FAPGNYPAL	0.76	0.79	0.97	0.29
1LEG ([66])	1.75	EQYKFYSV	0.46	0.47	2.10	0.23
1N59 ([67])	2.95	AVYNFATM	0.29	0.27	1.16	0.39
1NAN ([17])	2.30	INFDFNTI	0.71	0.71	1.38	0.97
1OSZ ([68])	2.10	RGYLYQGL	0.52	0.51	1.00	0.26
1VAC ([69])	2.50	SIINFEKL	0.53	0.56	0.98	0.55
Average RMSD (Å)			0.66	0.69	1.40	0.43
Self-docking to the single H-2K^b model						
1KPU ([65])	1.50	RGYVYQGL	0.76	0.86	1.06	0.00

[†]See Table I for an explanation of the values.

docked and atoms on the peptide in the original MHC structure: N in residues 1,2, and M and carbonyl C and O in residues 1,2, M-1, and M, where M is the length of the peptide. The restraint energy was first minimized to position the peptide in the binding site before the docking simulation. The sum of the energy terms $E_{peptide-MHC} + E_{peptide} + E_{restraint}$ was then optimized by the ICM Biased-Probability Monte Carlo sampling[28] of side chain torsion angles, ϕ and ψ angles for residues 3 to M-1, and 6 orientational variables of the peptide. Local deformations that approximately preserve loop closure were used to sample the backbone torsion angles.[41] Conformations were sampled according to a Metropolis criterion[42] with temperature 700K followed by up to 2,000 steps of conjugate gradient minimization after each stochastic move. A set of 200 of the lowest energy accepted conformations within 30° RMSD in torsion angle coordinates was used to prevent oversampling of nearby points in coordinate space.[43] The simulation was terminated after 5×10^7 function calls. This value was chosen after examining the convergence of the simulations, as discussed in the following section. The simulations required an average CPU

time of about 21 h on a single 1.3 GHz Athlon processor. The results of several docking simulations without constraints for HLA-A*0201 showed both lower prediction accuracy and longer computational times (data not shown), thus demonstrating that constraining the positions of the peptide terminii backbone atoms improves the docking method's performance.

All-Atom Reranking of the Conformational Stack

An all-atom model of the MHC and peptide was used to rerank the 50 lowest energy conformations from the grid docking simulations using a more realistic physical energy function. The energy function, which uses ECEPP/3 force field parameters,[37–39] is

$$E = E_{vw} + E_{to} + E_{hb} + E_{el} + E_{hp} + E_{cn} \qquad (2)$$

in which E_{to} and E_{hb} are the original ECEPP/3 energy functions. E_{vw} is the smoothed van der Waals term described above with a cutoff energy $E_{vw}^{max} = 7.0$ kcal/mol. The electrostatics term E_{el} was calculated using the boundary element method[44] with ECEPP/3 atomic charges and an internal dielectric constant of 4.0. The hydrophobic term E_{hp} was calculated as the product of a surface tension parameter 12 cal/(mol Å^2) and the molecule's solvent accessible surface area (SASA). The side chain entropy term $E_{en} = TS_{max} * A/A_{max}$ is proportional to the total SASA of the side chain atoms A, in which $T = 300K$, A_{max} is the total SASA of the atoms with the residue between two glycine residues in an extended peptide conformation, and S_{max} was calculated using approximate rotamer distributions.[28] Starting from the grid docking conformations, the conformation of the peptide and nearby MHC residues (non-H atoms within 4 Å) were locally optimized using the sum of the energy in Eq. 2 and the quadratic restraint energy described in Grid Docking Protocol. This resulted in relatively small conformational changes that primarily reduced steric clashes caused by the steeper all-atom van der Waal's potential as compared to the smoother grid version of the potential. Next, the binding energy was calculated as the difference between the energy of the peptide-MHC complex and the energy of the isolated components, without the restraint potential. The 50 conformations were then ranked according to this binding energy and the lowest energy conformation chosen as the final docking solution. Although it is expected that the conformations of the isolated peptide and the corresponding MHC binding cleft are different than in the peptide-MHC complex, optimizing the conformations of the isolated components before calculating their energies yielded larger errors in the predicted geometry (data not shown). This is possibly due to the noise introduced from relaxing these degrees of freedom because of inaccuracies in the force field energy as well as the lack of conformational averaging.

Treating HLA-A*0201 Through Multiple Grid Models

The conformations of side chains nearby the peptides for all HLA-A*0201 structures listed in Table I were compared after aligning the MHC backbone atoms. The result is shown in Figure 1. It is evident from Figure 1 that, although most interacting side chain conformations differ little between the MHC molecules binding different bound peptides, the conformations of residues R97 and Y116 group into two distinct clusters. The conformations of residue H114, which is hydrogen bonded to R97 also cluster into two more closely separated clusters. Based on this observation, we attempted to incorporate receptor flexibility for HLA-A*0201 by docking the peptides into potential maps calculated for a representative structure belonging to each cluster, namely PDB entries 1JF1 and 1I7U. All peptides in Table I were docked independently into maps calculated using these two structures. The lowest energy conformation in the combined conformational stack for both docking simulations was then selected as the docking solution.

Homology Model Generation and Peptide Docking

The homology models of different MHC allotypes were generated using the ICM method.[45] Briefly, the method consists of the following steps for models without loops insertions or deletions, as considered in this study: (1) Calculate a global alignment with zero end gap penalties between the target and template sequences, (2) build an extended polypeptide chain with idealized covalent geometry from the target sequence, (3) assign torsion angles for the aligned portions of the backbone and identical aligned residues to be those in the template structure, (4) assign the most likely rotamer to nonidentical aligned residues, and finally (5) iteratively minimize a sum of the physical energy and quadratic restraints between corresponding atoms in the model and template structures, reducing the strength of the restraint potential with each iteration. Homology models were made for the following HLA allotypes using the 1JF1 HLA-A*0201 structure as a template: HLA-B*0801, B*2705, B*3501, B*5101, and B*5301.

Grid potentials were then calculated using these models. Peptides from all X-ray crystal structures for each of these allotypes were then docked into the grid potentials for the corresponding model and reranked using all-atom models using the same procedure as for HLA-A*0201 and H-2K^b.

Support Vector Machine Peptide-MHC Binding Prediction

*HLA-A*0201 and H-2*K^b *peptide data sets*

First, a set of 304 peptides that bind HLA-A*0201 strongly, with IC50 $<$ 500 nM, was extracted from the reports of Doytchinova and Flower[46, 47] and references therein. This set contained 3 octamers, 242 nonamers, 53 decamers, and 6 11-mers. Because little published data is available for nonbinding peptides, we created an equal number of nonbinding peptides by concatenating the binding peptide sequences, randomly shuffling this sequence, and then partitioning it so that the distribution of peptide lengths is unchanged. A data set with equal numbers of binding and nonbinding peptides for H-2K^b was created in a similar manner using 27 peptides classified as high affinity in the MHCPEP database.[48] All peptide data sets are provided as Supplementary Material.

Support Vector Machine training

The *SVM*^{light} [49] (http://svmlight.joachims.org/) program was used for SVM training and prediction. The input data for each peptide included the differences of all-atom energy terms ΔE_{vw} + ΔE_{to}, ΔE_{hb}, ΔE_{el}, ΔE_{hp}, and ΔE_{en}, described in Materials and Methods, with the energy difference calculated as $\Delta E_X = E_X^{peptide-MHC} - E_X^{peptide} - E_X^{MHC}$ in which the energies of the isolated peptide and MHC, $E_X^{peptide}$ and E_X^{MHC}, respectively, are calculated with the bound conformation. A pairwise empirical energy term[50] and the number of peptide residues of each type were also included. Contacting residues, used to calculate the empirical energy, were defined as those with at least one side chain atom from each residue separated less than 4 Å. The six energy terms were each normalized to the interval [0, 1]. It is important to note that no explicit position-dependent peptide residue information is included in the data vectors used for machine learning. Because many peptides that strongly bind to a particular MHC allotype have specific position-dependent anchor residues, including this information may increase prediction accuracy for a particular allotype, but at the expense of accuracy for peptides without standard anchor residues or for predicting peptides binding to a different MHC allotype.

A Gaussian kernel function $k(\mathbf{x},\mathbf{y}) = \exp(-\gamma \parallel \mathbf{x} - \mathbf{y} \parallel^2)$ with $\gamma = 0.1$ and regularization constant $C = 1$ were used for all SVM calculations. The residue count values were also scaled by a factor of 0.5. These parameter values were determined to give the best performance on the HLA-A*0201 peptide set.

RESULTS

Grid Docking Energy Optimization Convergence

It is important to verify that the ICM stochastic energy minimization has converged since otherwise the global minimum may be missed and the results will not be reproducible from independent simulations. The convergence of the grid-docking simulations was studied by examining the lowest values of the energy function in Eq. 1 achieved after a given number of function calls for three independent Monte Carlo docking runs. The maximum difference between the lowest energies attained in each simulation and the lowest energy attained by all simulations was used as a measure of convergence. The convergence was evaluated for docking all HLA-A*0201 peptides in Table I into the potential maps derived from the 1JF1 and 1I7U structures and docking all H-2K^b peptides in Table II into maps from the 1KPU structure. The maximum energy difference after 5×10^7 function calls was only 0.33 kcal/mol, indicating that all simulations had converged within the characteristic energy scale $kT = 0.6$ kcal/mol for $T = 300$ K.

All three independent simulations converged within 0.6 kcal/mol of the lowest energy in considerably less iterations for most peptide-MHC complexes. Simulations for all but four of the HLA-A*0201 complexes and one of the H-2K^b complexes were converged to this degree in only half of the number of function calls, or 2.5×10^7. The slowest converging simulations were for the HLA-A*0201 complex with the 1I4F peptide, which did not converge, according to the criteria stated above, until 4.5×10^7 iterations. This is likely due to the fact that this is one of the longest peptides, with 10 residues. Likewise, the slowest converging simulations with H-2K^b were for the 1G7P peptide, which was one of the longest peptides for this MHC. The redocking of the original peptides for 1JF1 and 1KPU was among the fastest converging simulations, as expected since the MHC interface side chains are all correctly oriented.

Essentially, the same lowest energy conformations, with all non-hydrogen atom root mean square deviation (RMSD) < 0.02 Å, for the three independent runs were reached for most peptide-MHC combinations. However for 6 out of the 32 HLA-A*0201 complexes and one H-2K^b complex, the lowest energy conformations differed between two independent simulations, even though the energy differences between the three independent simulations were less than 0.33 kcal/mol. The same conformations were present within the four lowest energy stack conformations indicating a few peptide conformations that are approximately degenerate in energy. Also, in all but one case, the differences between the lowest energy conformations from independent runs were limited to at most two side chain torsion angles with the remainder of the peptide structure essentially identical. Most of the conformational differences were the relative rotation of a serine, isoleucine, or valine by 120° so as to be approximately isosteric. This is due to the smoothness of the van der Waal's grid potential term, which yields about the same energy for the two conformations, as compared with the corresponding energy term in an all-atom MHC model. This emphasizes the importance of retaining a number of low-energy conformations from each grid-docking simulation and reranking them according to a more accurate energy function with more restrictive steric constraints.

Self-Docking Selects the Correct Structure in the HLA-A*0201 Dual Grid Model

First of all, it should be noted that the all-atom energy function used to select the final conformation after refinement successfully chose the correct corresponding MHC structure for the docking of 1JF1 and 1I7U peptides i.e., the self-docked conformation had lower energy than the cross-docked conformation. The all-atom RMSD of the docking solutions as very low. Furthermore, the lowest energy conformations were also the ones closest to the correct structure in both cases. This is a nontrivial result since the near-native conformation from the 1JF1 docking moved from the sixth lowest energy conformation for the grid potential docking to the lowest energy conformation after reranking using an all-atom model of the MHC.

Cross-Docking Into the Dual Grid HLA-A*0201 Model

A comparison of the predicted peptide conformation with the corresponding X-ray crystal structure is shown in Table I. The RMSD of the peptide atoms was calculated

TABLE III. Comparison of Results of Docking Peptides Into HLA-B*0801, B*2705, B*3501, B*5101, B*5301 MHC Homology Models With the Corresponding X-Ray Crystal Structures of the Peptide–MHC Complexes[†]

			RMSD (Å)			
PDB entry (Ref.)	Resolution (Å)	Peptide sequence	Backbone	Central Backbone	Buried non-H	MHC
HLA-B*0801 homology model						
1AGB ([54])	2.20	GGRKKYKL	0.74	0.70	1.43	0.69
1AGC ([54])	2.10	GGKKKYQL	0.93	0.94	1.70	0.70
1AGD ([54])	2.05	GGKKKYKL	0.98	1.01	1.79	0.70
1AGE ([54])	2.30	GGKKKYRL	0.95	0.98	1.61	0.68
1AGF ([54])	2.20	GGKKRYKL	0.74	0.67	1.79	0.69
Average RMSD (Å)			0.87	0.86	1.66	0.69
HLA-B*2705 homology model						
1HSA ([70])	2.10	ARAAAAAAA	0.70	0.76	0.92	0.76
1JGE ([71])	2.10	GRFAAAIAK	1.10	1.22	1.44	0.51
Average RMSD (Å)			0.90	0.99	1.18	0.64
HLA-B*3501 homology model						
1A9E ([72])	2.50	LPPLDITPY	1.07	1.13	1.59	0.87
HLA-B*5101 homology model						
1E27 ([73])	2.20	LPPVVAKEI	2.21	2.45	3.44	0.64
1E28 ([73])	3.00	TAFTIPSI	1.12	1.21	1.77	0.66
Average RMSD (Å)			1.67	1.83	2.61	0.65
HLA-B*5301 homology model						
1A1M ([74])	1.40	TPYDINQML	1.39	1.54	2.81	0.59
1A1O ([74])	2.30	KPIVQYDNF	1.20	1.29	2.23	0.75
Average RMSD (Å)			1.29	1.42	2.52	0.67

[†]All homology models used the 1JFI structure as a template. See Table I for an explanation of the values.

after structural alignment of the backbone atoms of the peptide backbone atoms (residues 1–180) for the predicted structure and the crystal structure for these results as well as all other docking results in Tables II and III. Although this procedure leads to a small additional contribution to the peptide RMSD values due to the imperfect alignment of the MHC backbone atoms (e.g., an average of 0.4 Å for HLA-A*0201 structures in Table I), it is consistent for the comparison of multiple structures since the the peptide conformations do not affect the relative orientation of the MHC molecules. All peptides were docked into both the 1JF1 and 1I7U structures and the lowest energy conformation selected as the predicted conformation, as discussed in Materials and Methods.

The 1I7U MHC model gave the lowest energy conformation for only three peptides in the cross-docking results. Two of these peptides, those for 1B0G and 1I7T, differ by only one and two residues, respectively, from the native 1I7U peptide. This likely explains the exceptionally accurate predictions for these peptides but, like the self-docking results, it also reflects the accuracy of the all-atom energy function in differentiating the correct MHC model.

The highest accuracy predictions for peptides with sequences that are significantly different from the one present in the complex used for the MHC structure are 1EEY, 1EEZ, 1HHI, and 1I4F. The actual and predicted conformations for the 1EEY peptide are shown in Figure 2. The deviation of the peptide main chain atoms is quite small, only 0.54 Å, with only a slight difference at the center. The predicted conformation for the 1EEZ peptide, which only differs by one C-terminal residue, was predicted with similar regions and degree of deviation. The prediction for the 1HHI peptide had the main chain atoms close to those in the X-ray crystal structure and the largest error due to a 90° rotation of residue F5. Finally, the docking results for the 1I4F peptide had only a small backbone deviation of 0.75 Å even though it is a longer 10-residue peptide with a larger central bulge and, consequently, greater flexibility. The interactions of R6 with a crystal symmetry-related MHC molecule probably affects the peptide conformation, making its prediction difficult without the the inclusion of these additional interactions.

Computational and Physical Factors Affecting Docking Accuracy

Difficulties in predicting the docked peptide geometry resulted both from inaccuracies in the energy function used to rank the conformations and from physical factors that made computational prediction difficult, such as interactions with nearby symmetry-related molecules and conformational disorder evident from high crystallographic B-factors. These factors affecting prediction accuracy are apparent from examining the four least accurate docking results, as measured by the all-atom RMSD in Table I, namely the peptides corresponding to PDB entries 1HHH, 1HHJ, 1HHK, and 1I1F. The all-atom RMSD of the near-native conformation from the grid potential docking results (1.55, 1.24, 0.81, and 1.44 Å, respectively) also is significantly lower for all of these peptides, indicating that the all-atom energy incorrectly ranked the docking conformations. The results for another better predicted peptide, 1JHT,

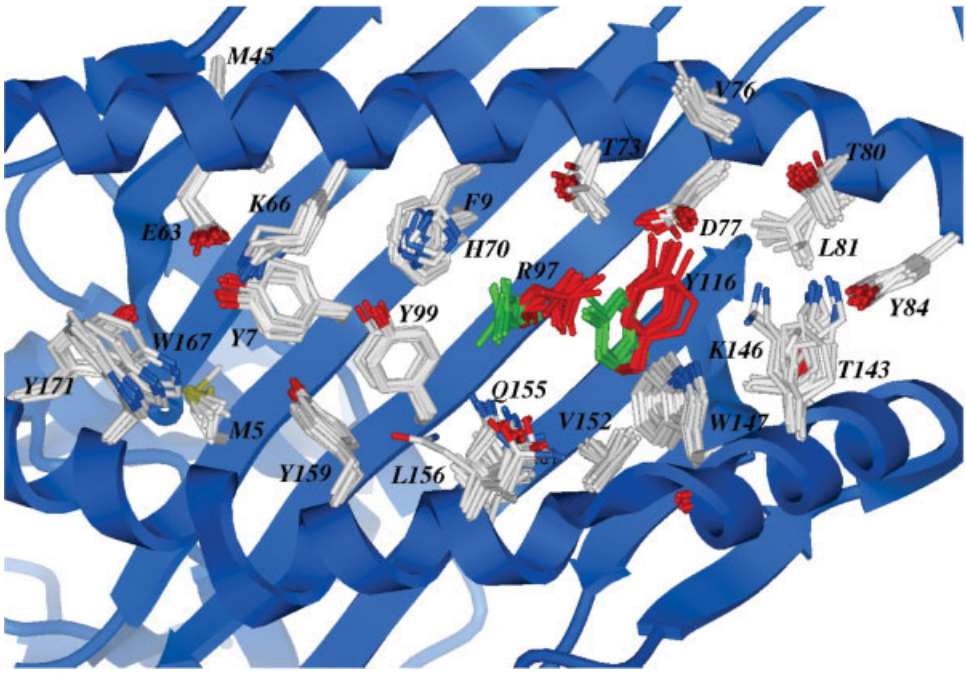

Fig. 1. HLA-A *0201 peptide binding pocket showing residues nearby the peptide for all structures listed in Table I, after alignment of the backbone atoms. The HLA backbone is shown in ribbon representation and the interacting residues are shown in stick representation. The residues whose side chain conformations cluster into two groups, R97 and Y116, are colored red and green according to the cluster to which they belong. Most of the remaining interacting residues have similar conformations for all structures. The peptides, which bind in the center of the groove, are not shown.

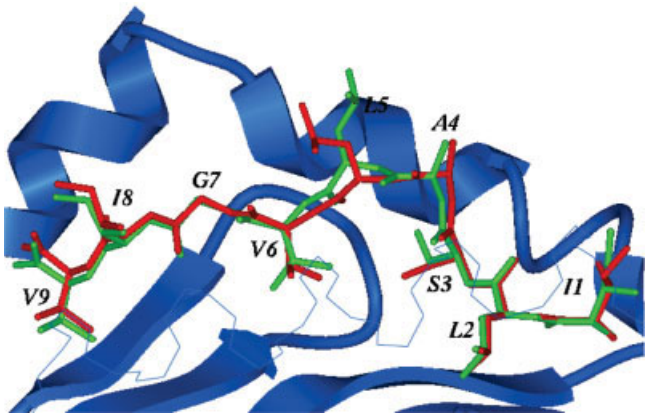

Fig. 2. Cross-docking result for the 1EEY peptide docked into the 1JF1 MHC structure. The docked conformation is shown in red and the peptide conformation from the 1EEY X-ray crystal structure, after aligning the MHC backbone atoms, is shown in green. The MHC molecule is shown in blue with the foreground helix in wire representation. The RMSD is only 0.54 Å for the backbone atoms and 0.88 Å for all buried residue nonhydrogen atoms, even though the docked peptide (ILSALVGIV) and the original bound peptide (ELAGIGILTV) are dissimilar; except for the P_2 and P_{M-1} anchor residues, and have different lengths.

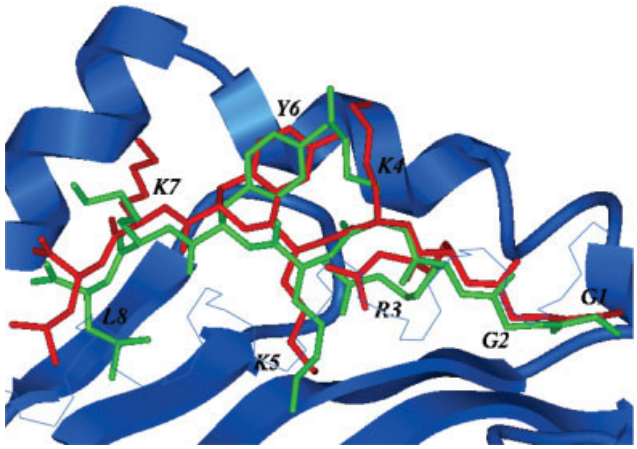

Fig. 3. Docking result for the 1AGB peptide (GGRKKYKL) docked into the HLA-B*0801 homology model. The docked conformation is shown in red and the peptide conformation from the 1AGB X-ray crystal structure, after aligning the MHC backbone atoms, is shown in green. The MHC is shown in blue with the foreground helix in wire representation. The RMSD is 0.74 Å for the backbone atoms and 1.43 Å for all nonhydrogen atoms in buried residues. The systematic shift in the predicted backbone structure may be due to a shift in portions of the flanking MHC helices,[54] which affects the alignment of the two structures.

also had a grid potential docking solution that was considerably closer to the correct conformation but was not the lowest all-atom energy conformation. The all-atom RMSDs for the near-native conformations were only 0.73 Å, even though the lowest energy conformations had an RMSD of 2.26 Å, respectively. The near-native solution for the 1JHT peptide has only the T8 side chain conformation significantly differing from the X-ray crystal structure. However, this residue is hydrogen-bonded to the C-terminal oxygen atom in a symmetry-related MHC molecule in the crystal structure, possibly perturbing its conformation. The central 1HHK peptide residue Y5 has the largest conformational errors. One

reason for the difficulty in predicting its conformation is that it has atomic B-factors > 40 Å^2, even though the average B-factor for this structure is only 21 Å^2.

It is also instructive to compare the HLA-A*0201 cross-docking predictions for pairs of peptides that differ by only a few residues. One such pair is the 1EEY and 1EEZ peptides, which differ only in the C-terminal residue and are both different from the original peptides bound to the MHC structures used for docking (1JF1 and 1I7U). Residue L5 had the largest conformational error for both the 1EEY and 1EEZ peptides. This central residue adopts significantly different conformations in the 1EEY and 1EEZ structures. In fact, this residue has two different conformations in noncrystallographic symmetry-related molecules and low electron density for 1EEZ and has an orientation that is not well defined due to low electron density for 1EEY.[51] This suggests that the difficulty in predicting its conformation is due to a lack of significant energetic constraints and the consequent disorder in the X-ray crystal structure. A similar pattern appears for two other peptides that differ by only a single N-terminal residue, 1I1F and 1I1Y. The predicted conformations are also quite similar but E4, a residue with different side chain conformations in the 1I1F and 1I1Y X-ray crystal structures, has a large conformational error, even in the near-native conformation. The study describing the experimental structures[52] speculated that the different conformations for this residue in the two structures is due to a water-mediated hydrogen bond with residue Y1 in the 1I1Y structure, which is not present in the 1I1F structure. However, low density for the water oxygen suggests that it has low occupancy. In any case, a more computationally expensive docking method that explicitly accounts for bound water molecules may be necessary to accurately predict the bound peptide conformations in these structures.

Cross-Docking Into a Single Grid H-2K^b Model

The results for docking the peptides for all H-2K^b X-ray crystal structures in the PDB into the 1KPU MHC structure are shown in Table II. The prediction accuracy is even higher than that for HLA-A*0201, with an average all-atom RMSD of only 1.38 Å. This may be partially due to the generally shorter length of the peptides, with all having 8 residues except 1G7P and 1KPV. This means that the peptide is in a more extended conformation because of the conserved hydrogen bonds at the peptide ends, which keep them effectively fixed. Although the docking result for one nonapeptide, 1G7P, had the lowest accuracy, the result for the other nonapeptide, 1KPV, was better than average. Another related factor that may have contributed to the improved accuracy, as compared with HLA-A*0201, is that the H-2K^b allotype has a central anchor residue, at position 5, which constrains the conformation of the peptide center to a greater degree. It is interesting to note that, unlike the HLA-A*0201 results, many of the H-2K^b cross-docking results were more accurate than the self-docking of the 1KPU peptide. We speculate that this is because of errors in ranking the conformation using the energy function since the difference in the all-atom RMSDs between the lowest energy and the near-native conformations is highest for the native peptide (1KPU), in contrast to the HLA-A*0201 docking results.

Accuracy of Homology Model Structures

The structures for the HLA homology models were first compared with all X-ray crystal structures for the corresponding allotypes in order to assess the accuracy of the model geometry. Only residues that contact the flexible portions of the peptide i.e., all side chain atoms and the central backbone atoms for residues 3 to M-2, with M the peptide length, were compared since these are the most relevant for peptide docking. An MHC residue was considering contacting the peptide if at least one non-hydrogen atom was within 4 Å of a peptide non-hydrogen atom.

*HLA-B*0801*

The interface residue conformations differed little between X-ray structures so only the 1AGB structure was compared with the homology model. Only 6 of the 27 interface residues types are different from those in HLA-A*0201. The assignment of the most prevalent rotameric conformations for nonidentical residues in the homology-modeling procedure was quite successful for this allele since only one residue, E76, out of the six nonidentical ones had a side chain conformation that was significantly different from the 1AGB structure. This residue forms a salt bridge with the lysine at P7. Residue E76 assumes a conformation in all other HLA-B*0801 structures and likewise interacts with the basic or polar residues at P7. This is not one of the anchor residues for HLA-B*0801, which are instead P3 and P5.[53]

*HLA-B*2705*

Comparison of the HLA-B*2705 homology model to the X-ray crystal structures shows that again only one interface residue, D116, has a significantly different conformation. This residue interacts with a lysine side chain at P9, which is a preferred residue at this position but not a primary anchor.

*HLA-B*3501*

A total of 12 out of 26 interface residue types differ between HLA-B*3501 and the template HLA-A*0201. This is considerably more than for B*0801 and B*2705. Four of these residues have significantly different conformations in the homology model as compared with the two HLA-B*3501 X-ray crystal structures, Y9, R62, F67, and S116. However, the conformational difference for R62 is likely due to its interaction with E161 in a nearby crystal symmetry partner. Only one conserved residue, R97, has a different conformation, which is slightly shifted relative to the A*0201 structure, possibly due to its interaction with a non-anchor C-terminal tyrosine in the peptide cocrystalized in both X-ray structures.

*HLA-B*5101*

HLA-B*5101 also has about half of the interface residue types, 12 out of 23, differing with HLA-A*0201. Interest-

ingly, the conformation of Y74 agreed with that in the 1E27 structure and Y99 agreed with that in the 1E28 structure even though each of these residues have different conformations in the two structures. An additional five residues, W95, R62, E76, F67, and E152, all of which differ in type from their HLA-A*0201 counterparts, have different conformations in the homology model. However, two of these, R62 and E76, are on the side of the adjoining alpha helix facing the solvent and, with the nearest peptide atom almost 4 Å away, likely have weak interactions with the peptide.

*HLA-B*5301*

A total of 10 out of the 25 interface residue types differ between HLA-B*5301 and the HLA-A*0201 template. As in the case of HLA-B*5101, two interface residues, N70 and Y74 have different conformations between the two X-ray crystal structures but the conformations of these residues in the homology model agree with one of the crystal structures. Three other residues have different conformations in the homology model, R97, Y90, and R62. The latter residue, as in HLA-B*5101, likely has weak interactions with the peptide.

Homology Model Docking Results

The results for docking the peptides for all PDB structures into homology models for HLA-B*0801, B*2705, B*3501, B*5101, and B*5301 are given in Table III. All homology models used the 1JF1 HLA-A*0201 structure as a template (see Materials and Methods for details).

The accuracies of each prediction for the HLA-B*0801 peptides are mutually comparable in accordance with the peptides' sequence similarity. The peptides are all variant peptides from the HIV-1 Gag protein p17. The docking conformations are quite close to the experimentally determined conformations with an average backbone RMSD of 0.87 Å and an average all-atom RMSD of 1.76 Å. This is not dramatically higher than the average deviation for the H-2K^b peptides, most of which are also octamers, docked into the 1KPU MHC X-ray crystal structure. This demonstrates that the accuracy for docking into a homology model of an MHC may be comparable to that for docking into a crystal structure. The shift in the main chain of the 1AGB peptide near R3 relative to the reference 1AGD peptide, which may be partially responsible for eliminating the activity of almost all T-cell clones tested in one study,[54] was reproduced by the docking result. The predicted and experimental conformations for the 1AGB peptide are shown in Figure 3. The uniform shift of the entire predicted peptide structure relative to the 1AGB X-ray crystal structure is probably due to a change in the MHC structural alignment used to compare the conformations because of a shift of the flanking MHC alpha helices.[54] The largest contribution to the all-atom RMSD for the HLA-B*0801 peptide docking results was from the large solvent-exposed lysine residues, some of which are indirectly hydrogen bonded to HLA atoms through water molecules. The lack of explicit water molecules that indirectly participate in peptide-MHC binding is a general problem with implicit solvent models.

The docking results for the HLA-A*2705 peptides were also reasonably close to the experimental structures, with accuracy similar to the HLA-A*0801 results. The predicted backbone structure for the 1HSA peptide was close to the crystal structure and the only extended side chain, R2, had a small deviation beyond C_γ. The predicted 1JGE peptide geometry had a localized backbone deviation only near A4, with the remaining backbone close. Most side chain conformations had little deviation, except A4 and R4, which was rotated approximately 90° in the last torsion angle.

As mentioned above, the structure of the complex of HLA-A*3501 with an octamer peptide (PDB entry 1A1N) has nonstandard conformations of the N- and C-terminii and so was not included in this study. This degree of structural variability in the N- and C-terminii of bound peptides has not been observed in other octamer-MHC complexes, such as HLA-B*0801, HLA-B*5101, and H-2K^b so this is not likely to commonly occur. However, only further experimental structures of complexes can resolve this issue.

Except for a localized twist of the peptide main chain that displaces only residue T7, the conformation for the 1A9E peptide bound to HLA-3501 is close to the experimental structure. The near-native structure from the docking calculation is similar, except that it does not have the main chain twist, making it even closer, with an all-atom RMSD of only 1.05 Å. Presumably, it is difficult to distinguish the energy difference between the conformation with T7 buried and with a hydrogen bond interaction and its native conformation with this residue solvent exposed.

The conformations of the N- and C-terminal residues for the 1E27 peptide bound to HLA-A*5101, L1, P2, E8, and I9, are predicted correctly, probably because of the restraints imposed on the terminii main chain atoms, but the central portion of the peptide deviates considerably from the correct conformation. If the conformation of the peptide in the X-ray crystal structure is superimposed on the HLA-A*5101 homology model, it is evident that the steric clash of P2 with the MHC Y99 side chain, which is in the incorrect conformation for 1E28 but the correct conformation for 1E27, appears to be the principal cause for this large deviation. Likewise, for the 1E28 peptide, the steric clash of I5 in the correct conformation with MHC residue Y74, which is in the incorrect conformation for 1E28 but the correct conformation for 1E27, causes a localized main chain deviation. Unlike the prediction for 1E27, this deviation is not large enough to disrupt the remainder of the peptide so the overall RMSD, 1.77 Å, is not too large. It is interesting that the near-native conformation from the 1E28 prediction has only an all-atom RMSD of 1.32 Å since the peptide is accommodated in the modeled MHC by only a slight shift of the I5 side chain to avoid clashing with the MHC tyrosine.

A number of factors[18] contribute to the lower accuracy of the docking results for HLA-A*5301. First, the MHC main chain residues 66–75 in the α_1 helix delimiting one side of the peptide binding groove have a larger shift, compared to HLA-A*0201, than the other allotypes considered and contribute to a small B pocket. Second, there are three

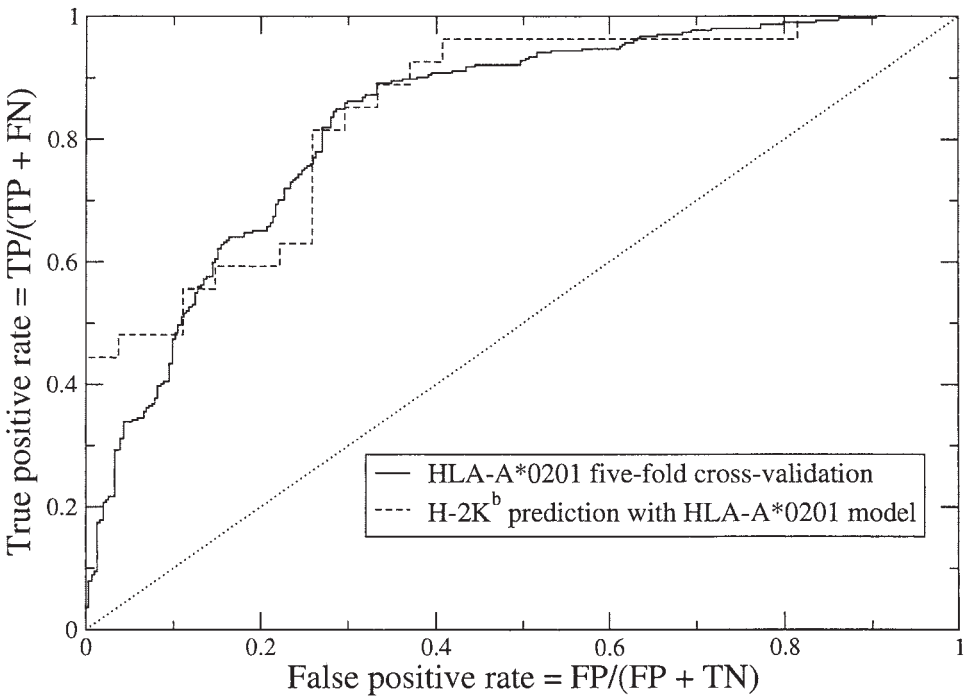

Fig. 4. Receiver Operating Characteristic (ROC) curves for fivefold cross-validation of the SVM binding affinity prediction for the HLA-A*0201 peptide data set (solid line) and for the prediction of peptide binding affinity to H-2K^b using the SVM trained on the HLA-A*0201 data set (dashed line). The areas under the curves are 0.83 and 0.85, respectively.

peptide intramolecular hydrogen bonds in the 1A1M structure that couple the conformations of the central peptide side chains. Third, crystallographically resolved water molecules mediate hydrogen-bonding interactions between MHC and central peptide side chains. This appears to be particularly important for the 1A1O peptide. The specific directional interactions of the fixed water molecules are not accounted for by the implicit solvation model used in this study. Finally, MHC residue R97 adopts significantly different side chain conformations in the two HLA-A*5301 complex structures, both of which differ from the conformation in the homology model. This same residue adopts alternate conformations upon binding different peptides along with a concerted shift of the Y116 side chain in HLA-A*0201. This is the motivation for using two MHC models for docking peptides to HLA-A*0201, as described in Materials and Methods. In fact, the R97 conformation in the HLA-A*5301 1A1O structure is essentially the same as in the HLA-A*0201 1I7U structure used for the other MHC model in docking. This implies that the use of multiple MHC template structures with different conformations for R97 could possibly improve model accuracy. The all-atom energy function used to rank the grid-docking solutions does not appear to be a limiting factor for the HLA-A*5301 complexes since the near native RMSDs are either equal or close to those for the lowest energy conformation chosen as the docking solution.

Peptide-HLA-A*0201 Binding Affinity Prediction

The performance of the SVM binding affinity prediction on the set of 304 binders and 304 nonbinders was assessed by fivefold cross-validation. The overall accuracy, recall, and precision were 78, 85, and 75%, respectively. This indicates the significant discrimination between binding and nonbinding peptides since the corresponding random expected values were only 50% for the accuracy and precision and 57% for the recall. The Receiver Operating Characteristic (ROC) curve, which plots the trade-off between sensitivity and selectivity at different classifier cutoff values, is shown in Figure 4. The total area under the ROC curve, which was calculated from the Mann-Whitney-Wilcoxon rank sum statistic, was 0.83.

Peptide-H-2K^b Binding Affinity Prediction Using the HLA-A*0201 Trained SVM

Next, the SVM trained on peptide-binding affinity to human HLA-A*0201 was used to predict the peptide-binding affinity to murine H-2K^b, in order to test the general applicability of the prediction model to other MHC allotypes for which no data were included in the training data set. A total of 40/54, or 74% of the peptides were correctly classified as either binders or nonbinders. The prediction recall and precision were 96 and 67%. Although the overall performance is not quite as good as for HLA-A*0201 peptides, these statistics represent good discrimination between binding and nonbinding peptides since the random expected values were only 50% for the accuracy and precision and 72% for the recall. The higher recall, both for the observed and random expected values, for this prediction than for the HLA-A*0201 cross-validation results, given above, is due to a larger fraction of peptides predicted as binders. The difference between the observed

and random recall values, 24%, however, is only slightly lower than the difference for HLA-A*0201 cross-validation results, 28%, demonstrating that our prediction method, which uses energy values calculated from the predicted geometry of the peptide–MHC complex, can successfully predict peptide-binding affinities for other MHC allotypes not used for training the model. The ROC curve is shown in Figure 4 and the area under the curve of 0.85 is close to that for the HLA-A*0201 cross-validation results, further demonstrating the comparable prediction performance.

DISCUSSION

Comparison With Previous Studies

Peptide-MHC binding geometry prediction

While peptide–MHC docking results from previous studies give a useful qualitative indication of the accuracy of the respective docking methods, a number of differences with our study make a quantitative comparison difficult. These differences include one or more of the following: (1) the results published previously are for self-docking (redocking the peptide into the MHC structure from the corresponding complex) rather than cross-docking, (2) the peptide backbone structure is employed in the prediction method, (3) the structural alignment method used for calculating the RMSD is unspecified, and (4) the prediction method is tested on only a small number of peptides. Difference (1) probably has a large effect on prediction accuracy because, according to our results, the self-docking accuracy is typically higher than for cross-docking. In particular, the two self-docking results for HLA-A*0201 had a higher accuracy than any cross-docking result for this MHC. This is likely due to small rearrangements of MHC side chains in order to accommodate the bound peptide and the consequent binding pocket surface complementarity with the correct peptide conformation.

Two early studies[20,55] demonstrated that a multiple copy algorithm gave a reasonably accurate prediction for cross-docking the 1HHI peptide into the 3HLA MHC structure (1.4 Å backbone RMSD) but poor accuracy for self-docking the 1KPV peptide in H-2K^b (2.7 Å backbone RMSD). An improvement of this method,[25] by incorporating a translational search, gave excellent results (1.0 Å backbone RMSD) for self-docking the nonameric 1HHI and 1HHJ peptides. A dead-end elimination algorithm that includes sampling of nearby MHC side chains was used in another study[27] to dock the 1KPU and 1KPV peptides (0.79 Å and 1.33 Å backbone RMSD, respectively). The accuracy of our docking result was comparable for the 1KPU peptide but better for 1KPV peptide. The knowledge-based structure prediction of 23 peptide-MHC complexes in another report[26] made use of both the MHC and bound peptide backbone conformations from the complex to predict the peptide side chain conformations and, thus, is not comparable to our results. Finally, a predominantly knowledge-based method[21] was applied to self-docking five HLA-A*0201 peptides in another study. The resulting backbone RMSDs were 1.27, 1.82, 0.46, 0.87, and 1.44 Å for 1HHG-1HHK, respectively. These values are comparable to the accuracy of our corresponding *cross-docking* results, which is a more difficult prediction than self-docking.

There are several advantages and disadvantages of our method, compared with the previous ones described in the Introduction. One advantage is that, unlike threading methods that model the peptide backbone conformation using X-ray crystal structures, our prediction method can be used for any length peptide, even if there is no available structure for the same length peptide bound to the same MHC. Furthermore, even if such structures are present, there may not be enough to fully define the variability in the peptide conformations or to validate the predictions, since structures containing the same peptides must be eliminated before applying the prediction procedure in order to fairly evaluate its performance. Also, as discussed in the Introduction, other methods separately predict different portions of the peptide, which may make it difficult for them to recover from conformational errors in the early steps of the procedure. In addition, since we optimize a physical energy function, our method may be used, in principle, to dock any comparably sized molecule to MHC, such as peptide analogues. Although we provided evidence that the Monte Carlo simulations were converged, the main disadvantage of our method is its speed. Also, while our use of potential maps certainly made energy evaluation faster than using an all-atom MHC model, it also prevented the explicit flexibility of the MHC interface that has been included in other methods. It is not yet clear, however, whether sampling the MHC side chain conformations improves accuracy since additional degrees of freedom introduce more error in the energy function and possibly more false low-energy minima. Clearly, a detailed large-scale comparison of the predicted peptide conformations using a uniform criterion is needed in order to determine the relative strengths and weaknesses of each docking method.

In summary, the results from testing our ab initio docking method with a comprehensive set of peptide–MHC complexes with known structures indicate that (1) multiple MHC models are useful to approximately incorporate receptor flexibility, (2) accurate prediction results may be obtained using our docking method, even for highly flexible decameric peptides, and (3) docking peptides into homology models of MHC allotypes without too many nonconserved residues in the binding pocket yields accurate conformation predictions. Peptide docking to MHC is a difficult problem but it has many important applications that will drive future efforts to develop improved methods.

Peptide–MHC binding affinity prediction

We have attempted to classify peptides as binders/nonbinders rather than predicting the binding free energy of peptides binding to MHC, as was done in several previous peptide–MHC prediction methods.[21,46,47] Although the latter class of methods may give insight into the molecular interactions contributing to peptide-MHC binding, most practical applications, such as epitope prediction, involve identifying the relatively small fraction of peptides of suitable lengths that bind to a particular MHC

allotype. Also a supervised learning method, SVM, trained on both positive (binder) and negative (nonbinder) examples, is expected to perform better at this task than a method that is trained only on positive examples and that predicts nonbinders by extrapolating the predicted binding free energy to higher values than are present in the training data. Two previous studies by Schueler-Furman et al.[56] and Logean and Rognan[57] used peptide–MHC complex structures predicted by threading to classify peptides as binders or nonbinders using empirical energy scoring functions. Although the threading methods used for geometry prediction in these methods is faster than the global optimization of the energy function employed in our method, the accuracy of the geometry prediction was not examined. Furthermore, the performance statistics for data sets with a higher fraction of nonbinders given in these studies are not comparable with our cross-validation results.

Future Directions

A number of extensions and applications of the peptide–MHC docking method presented here are possible. First, MHC homology models may be improved through the use of multiple template structures and all-atom refinement in the presence of a strongly binding peptide. In addition, it may be possible to include limited flexibility of the MHC interface through a hybrid representation that uses an all-atom model for residues that contact the central region of the peptide and uses the grid potentials for the remainder of the interface. This would introduce flexibility without a prohibitively large increase in the number of variables in the global energy optimization. Also, it would be interesting to apply a similar docking method to predict the geometry of peptides bound to class-II MHC. This is a more difficult problem, however, as the peptide-binding groove is open at the ends, allowing multiple registrations of the peptide within the cleft. In addition, the SVM-binding prediction method described above could be applied to docking results for homology models in order extend the binding-affinity prediction to relatively uncharacterized MHC allotypes. Finally, because the peptide ends are effectively fixed for docking to class-I MHC molecules, the geometry prediction method presented here could be applied with little modification to a different class of problems: the prediction of external loop conformations in homology models. This is a critical problem for comparative modeling since loops often do not have sufficient sequence similarity to existing structures and, therefore, must be predicted using energy-based methods.

ACKNOWLEDGMENTS

We thank Molsoft LLC for providing the ICM molecular modeling software.

REFERENCES

1. Townsend AR, Rothbard J, Gotch FM, Bahadur G, Wraith D, McMichael AJ. The epitopes of influenza nucleoprotein recognized by cytotoxic T lymphocytes can be defined with short synthetic peptides. Cell 1986;44:959–968.
2. 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 1993;74:929–937.
3. Parker KC, Bednarek MA, Coligan JE. Scheme for ranking potential HLA-A2 bind-ing peptides based on independent binding of individual peptide side-chains. J Immunol 1994;152:163–175.
4. Brusic V, Schonbach C, Takiguchi M, Ciesielshi V, Harrison LC. Application of genetic search in derivation of matrix models of peptide binding to mhc molecules. Proc Int Conf Intell Syst Mol Biol 1997;5:75–83.
5. Reche PA, Glutting JP, Reinherz EL. Prediction of MHC class I binding peptides using profile motifs. Hum Immunol 2002;63:701–709.
6. Gulukota K, Sidney J, Sette A, DeLisi C. Two complementary methods for pre-dicting peptides binding major histocompatibility complex molecules. J Mol Biol 1997; 267:1258–1267.
7. Mamitsuka H. Predicting peptides that bind to MHC molecules using supervised learn-ing of hidden Markov models. Proteins 1998;33:460–474.
8. Milik M, Sauer D, Brunmark AP, Yuan L, Vitiello A, Jackson MR, Peterson PA, Skolnick J, Glass CA. Application of an artificial neural network to predict specific class I MHC binding peptide sequences. Nat Biotechnol 1998;16:753–756.
9. Nielsen M, Lundegaard C, Worning P, Lauemoller SL, Lamberth K, Buus S, Brunak S, Lund O. Reliable prediction of T-cell epitopes using neural networks with novel sequence representations. Protein Sci 2003;12:1007–1017.
10. Baratin M, Kayibanda M, Ziol M, Romieu R, Briand JP, Guiller JG, Viguier M. Amino acid modifications in the wild type sequence p53 232–240 overcome the poor immunogenicity of this self tumour epitope. J Pept Sci 2002;8:327–334.
11. Stemmer C, Quesnel A, Prevost-Blondel A, Zimmermann C, Muller S, Briand JP, Pircher H. Protection against lymphocytic choriomeningitis virus infection induced by a reduced peptide bond analogue of the H-2Db-restricted CD8(+) T cell epitope GP33. J Biol Chem 1999;274:5550–5556.
12. Krebs S, Folkers G, Rognan D. Binding of rationally designed non-natural peptides to the human leukocyte antigen HLA-B*2705. J Pept Sci 1998;4:378–388.
13. Ding YH, Smith KJ, Garboczi DN, Utz U, Biddison WE, Wiley DC. Two human T cell receptors bind in a similar diagonal mode to the HLA-A2/Tax peptide complex using different TCR amino acids. Immunity 1998;8:403–411.
14. Garcia KC, Degano M, Pease LR, Huang M, Peterson PA, Teyton L, Wilson IA. Structural basis of plasticity in t cell receptor recognition of a self peptide-mhc antigen. Science 1998;279:1166–1172.
15. Reiser JB, Darnault C, Guimezanes A, Gregoire C, Mosser T, Schmitt-Verhulst AM, Fontecilla-Camps JC, Malissen B, Housset D, Mazza G. Crystal structure of a T cell receptor bound to an allogeneic MHC molecule. Nature Immunol 2000;1:291–297.
16. Reiser JB, Gregoire C, Darnault C, Mosser T, Guimezanes A, Schmitt-Verhulst AM, Fontecilla-Camps JC, Mazza G, Malissen B, Housset D. A T cell receptor CDR3β loop undergoes conformational changes of unprecedented magnitude upon bind-ing to a peptide/MHC class I complex. Immunity 2002;16:345–354.
17. Reiser JB, Darnault C, Gregoire C, Mosser T, Mazza G, Kearney A, van der Merwe PA, Fontecilla-Camps JC, Housset D, Malissen B. CDR3 loop flexibility contributes to the degeneracy of TCR recognition. Nat Immunol 2003;4:241–247.
18. Smith KJ, Reid SW, Stuart DI, McMichael AJ, Jones EY, Bell JI. An altered position of the alpha 2 helix of MHC class I is revealed by the crystal structure of HLA-B*3501. Immunity 1996;4:203–213.
19. Batalia MA, Collins EJ. Peptide binding by class I and class II MHC molecules. Biopolymers 1997;43:281–302.
20. Rosenfeld R, Zheng Q, Vajda S, DeLisi C. Computing the structure of bound peptides. application to antigen recognition by class I major histocompatibility complex receptors. J Mol Biol 1993;234:515–521.
21. Rognan D, Lauemoller SL, Holm A, Buus S, Tschinke VV. Predicting binding affini-ties of protein ligands from three-dimensional models: application to peptide binding to class I major histocompatibility proteins. J Med Chem 1999;42:4650–4658.
22. Tong JC, Tan TW, Ranganathan S. Modeling the structure of bound peptide ligands to major histocompatibility complex. Protein Sci 2004;13:2523–2532.

23. Sali A, Blundell TL. Comparative protein modelling by satisfaction of spatial re-straints. J Mol Biol 193;234:779–815.
24. Altuvia Y, Schueler O, Margalit H. Ranking potential binding peptides to mhc molecules by a computational threading approach. J Mol Biol 1995;249:244–250.
25. Sezerman U, Vajda S, DeLisi C. Free energy mapping of class I MHC molecules and structural determination of bound peptides. Protein Sci 1996;5:1272–1281.
26. Schueler-Furman O, Elber R, Margalit H. Knowledge-based structure prediction of MHC class I bound peptides: a study of 23 complexes. Fold Des 1998;3:549–564.
27. Desmet J, Wilson IA, Joniau M, Maeyer M De, Lasters I. Computation of the binding of fully flexible peptides to proteins with flexible side chains. FASEB J 1997; 11:164–172.
28. Abagyan RA, Totrov MM. Biased probability Monte Carlo conformational searches and electrostatic calculations for peptides and proteins. J Mol Biol 1994;235:983–1002.
29. Molsoft, LLC. ICM software manual. Version 3.0. 2004.
30. Zhou Y, Abagyan R. How and why phosphotyrosine-containing peptides bind to the SH2 and PTB domains. Fold Des 1998;3.
31. Stigler RD, Hoffmann B, Abagyan R, Schneider-Mergener J. Soft docking an L and a D peptide to an anticholera toxin antibody using internal coordinate mechanics. Struct Fold Des 1999;7:663–670.
32. Marsh SGE, Parham P, Barber LD. The HLA factsbook. San Diego: Academic Press. 2000.
33. Marsh SG, Albert ED, Bodmer WF, Bontrop RE, Dupont B, Erlich HA, Ger-aghty DE, Hansen JA, Mach B, Mayr WR, Parham P, Petersdorf EW, Sasazuki T, Schreuder GM, Strominger JL, Svejgaard A, Terasaki PI. Nomenclature for factors of the HLA system. Tissue Antigens 2002;60:407–464.
34. Goodford PJ. A computational-procedure for determining energetically favorable binding-sites on biologically important macromolecules. J Med Chem 1985;28:849–857.
35. Totrov M, Abagyan R. Flexible protein-ligand docking by global energy optimization in internal coordinates. Proteins 1997; (Suppl 1):215–220.
36. Fernández-Recio J, Totrov M, Abagyan RA. Soft protein-protein docking in internal coordinates. Protein Sci 2002;11:280–91.
37. Momany FA, McGuire RF, Burgess AW, Scheraga HA. Energy parameters in polypeptides. VII. geometric parameters, partial atomic charges, nonbonded interac-tions, hydrogen bond interactions, and intrinsic torsional potentials for the naturally occurring amino acids. J Phys Chem 1975;79:2361–2381.
38. Némethy George N, Pottle Marcia S, Scheraga HA. Energy parameters in polypep-tides. 9. updating of geometrical parameters, nonbonded interactions and hydrogen bond interactions for the naturally occurring amino acids. J Phys Chem 1983;87:1883–1887.
39. Némethy GN, Gibson KD, Palmer KA, Yoon CN, Paterlini G, Zagari A, Rumsey S, Scheraga HA. Energy parameters in polypeptides. 10. improved geometrical parameters and nonbonded interactions for use in the ECEPP/3 algorithm, with application to proline-containing peptides. J Phys Chem 1992;96:6472–6484.
40. Corana A, Marchesi M, Martini C, Ridella S. Minimizing multimodal functions of continuous-variables with the simulated annealing algorithm. ACM Trans Math Soft 1987;13:262–280.
41. Abagyan RA, Mazur AK. New methodology for computer-aided modelling of biomolecular structure and dynamics. 2. local deformations and cycles. J Biomol Struct Dyn 1989;6:833–845.
42. Metropolis NA, Rosenbluth AW, Rosenbluth NM, Teller AH, Teller E. Equation of state calculations by fast computing machines. J Chem Phys 1953;21:1087–1092.
43. Abagyan RA, Argos P. Optimal protocol and trajectory visualization for conforma-tional searches of peptides and proteins. J Mol Biol 1992;225:519–532.
44. Totrov M, Abagyan R. Rapid boundary element solvation electrostatics calculations in folding simulations: successful folding of a 23-residue peptide. Biopolymers 2001;60:124–133.
45. Cardozo T, Totrov M, Abagyan R. Homology modeling by the ICM method. Proteins 1995;23:403–414.
46. Doytchinova IA, Flower DR. Toward the quantitative prediction of T-cell epitopes: CoMFA and CoMSIA studies of peptides with affinity for the class I MHC molecule HLA-A*0201. J Med Chem 2001;44:3572–3581.
47. Doytchinova IA, Flower DR. Physicochemical explanation of peptide binding to HLA-A*0201 major histocompatibility complex: A three-dimensional quantitative structure-activity relationship study. Proteins 2002;48:505– 518.
48. Brusic V, Rudy G, Kyne AP, Harrison LC. MHCPEP, a database of MHC-binding peptides: update 1997. Nucleic Acids Res 1998;26:368–371.
49. Joachims T. Learning to classify text using support vector machines. New York: Springer. 2002.
50. Betancourt MR, Thirumalai D. Pair potentials for protein folding: Choice of reference states and sensitivity of predicted native states to variations in the interaction schemes. Prot Sci 1999;8:361–369.
51. Sharma AK, Kuhns JJ, Yan S, Friedline RH, Long B, Tisch R, Collins EJ. Class I major histocompatibility complex anchor substitutions alter the conformation of T cell receptor contact. J Biol Chem 2001;276:21443–21449.
52. Kirksey TJ, Pogue-Caley RR, Frelinger JA, Collins EJ. The structural basis for the increased immunogenicity of two HIV-reverse transcriptase peptide variant/class I major histocompatibility complexes. J Biol Chem 1999;274:37259–37264.
53. Arnett KL, Huang W, Valiante NM, Barber LD, Parham P. The Bw4/Bw6 differ-ence between HLA-B*0802 and HLA-B*0801 changes the peptides endogenously bound and the stimulation of alloreactive T cells. 1998 1998;48:56–61.
54. Reid SW, McAdam S, Smith KJ, Klenerman P, O'Callaghan CA, Harlos K, Jakob-sen BK, McMichael AJ, Bell JI, Stuart DI, Jones EY. Antagonist HIV-1 Gag peptides induce structural changes in HLA B8. J Exp Med 1996;184:2279–2286.
55. Rosenfeld R, Zheng Q, Vajda S, DeLisi C. Flexible docking of peptides to class I major-histocompatibility-complex receptors. Genet Anal 1995;12:1–21.
56. 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. Prot Sci 2000;9:1838–1846.
57. Logean A, Rognan D. Recovery of known T-cell epitopes by computational scanning of a viral genome. J Comput Aided Mol Des 2002;16:229–243.
58. Zhao R, Loftus DJ, Appella E, Collins EJ. Structural evidence of T cell xeno-reactivity in the absence of molecular mimicry. J Exp Med 1999;189:359–370.
59. Madden DR, Garboczi DN, Wiley DC. The antigenic identity of peptide-MHC com-plexes: a comparison of the conformations of five viral peptides presented by HLA-A. Cell 1993;75:693–708.
60. Busslep J, Zhao R, Dinnini D, Loftus D, Saad M, Appella E, Collins EJ. T cell activity correlates with oligomeric peptide-major histocompatibility complex binding on T cell surfac. J Biol Chem 2001;276:47320–47328.
61. Hillig RC, Coulie PG, Stroobant V, Saenger W, Ziegler A, Hulsmeyer M. High-resolution structure of HLA-A*0201 in complex with a tumour-specific antigenic peptide encoded by the MAGE-A4 gene. J Mol Biol 2001;310:1167–1176.
62. Sliz P, Michielin O, Cerottini JC, Luescher I, Romero P, Karplus M, Wiley DC. Crystal structures of two closely related but antigenically distinct HLA-A2/melanocyte-melanoma tumor-antigen peptide complexes. J Immunol 2001;167:3276–3284.
63. Apostolopoulos V, Yu M, Corper AL, Li W, McKenzie IF, Teyton L, Wilson IA, Plebanski M. Crystal structure of a non-canonical high affinity peptide complexed with MHC class I: a novel use of alternative anchors. J Mol Biol 2002;318:1307–1316.
64. Apostolopoulos V, Yu M, Corper AL, Teyton L, Pietersz GA, McKenzie IF, Wilson IA, Plebanski M. Crystal structure of a non-canonical low-affinity peptide complexed with MHC class I: a new approach for vaccine design. J Mol Biol 2002;318:1293–1305.
65. Fremont DH, Matsumura M, Stura EA, Peterson PA, Wilson IA. Crystal structures of two viral peptides in complex with murine MHC class I H-2Kb. Science 1992;257:919–927.
66. Luz JG, Huang M, Garcia KC, Rudolph MG, Apostolopoulos V, Teyton L, Wilson IA. Structural comparison of allogeneic and syngeneic T cell receptor-peptide-Major Histocompatibility Complex complexes: A buried alloreactive mutation subtly alters peptide presentation substantially increasing V(beta) interactions. J Exp Med 2002; 195:1175–1186.
67. Achour A, Michaelsson J, Harris RA, Odeberg J, Grufman P, Sandberg JK, Levitsky V, Karre K, Sandalova T, Schneider G. A structural basis for LCMV immune evasion: subversion of H-2D(b) and H-2K(b) presentation of gp33 revealed by compar-ative crystal structure. Anal Immun 2002;17:757–768.

68. Ghendler Y, Teng MK, Liu JH, Witte T, Liu J, Kim KS, Kern P, Chang HC, Wang JH, Reinherz EL. Differential thymic selection outcomes stimulated by focal structural alteration in peptide/major histocompatibility complex ligands. Proc Natl Acad Sci USA 1998;95:10061–10066.
69. Fremont DH, Stura EA, Matsumura M, Peterson PA, Wilson IA. Crystal structure of an H-2Kb-ovalbumin peptide complex reveals the interplay of primary and secondary anchor positions in the major histocompatibility complex binding groove. Proc Natl Acad Sci USA 1995;92:2479–2483.
70. Madden DR, Gorga JC, Strominger JL, Wiley DC. The three-dimensional structure of HLA-B27 at 2.1 A resolution suggests a general mechanism for tight peptide binding to MHC. Cell 1992;70:1035–1048.
71. Hulsmeyer M, Hillig RC, Volz A, Ruhl M, Schroder W, Saenger W, Ziegler A, Uchanska-Ziegler B. HLA-B27 subtypes differentially associated with disease exhibit subtle structural alterations. J Biol Chem 2002;277:47844–47853.
72. Menssen R, Orth P, Ziegler A, Saenger W. Decamer-like conformation of a nonapeptide bound to HLA-B*3501 due to non-standard positioning of the C terminus. J Mol Biol 1999;285:645–653.
73. Maenaka K, Maenaka T, Tomiyama H, Takiguchi M, Stuart DI, Jones EY. Non-standard peptide binding revealed by crystal structures of HLA-B*5101 complexed with HIV immunodominant epitopes. J Immunol 2000;165:3260–3267.
74. Smith KJ, Reid SW, Harlos K, McMichael AJ, Stuart DI, Bell JI, Jones EY. Bound water structure and polymorphic amino acids act together to allow the binding of different peptides to MHC class I HLA-B53. Immunity 1996;4:215–228.