Journal of Computer-Aided Molecular Design, **15:** 573–586, 2001.
KLUWER/ESCOM

Predicting sequences and structures of MHC-binding peptides: a computational combinatorial approach

Jun Zeng[a,d,*], Herbert R. Treutlein[a,b,d] & George B. Rudy[c,e]
[a]*Molecular Modelling Laboratory, Ludwig Institute for Cancer Research, P.O. Box 2008, Royal Melbourne Hospital, Parkville, VIC 3050, Australia;* [b]*Cooperative Research Centre for Cellular Growth Factors, P.O. Royal Melbourne Hospital, Parkville, VIC 3050, Australia;* [c]*Genetics and Bioinformatics Division, Walter and Eliza Hall Institute of Medical Research, P.O. Royal Melbourne Hospital, Parkville, VIC 3050, Australia;* [d]*Current Address: Cytopia Pty Ltd, 7th Floor, Daly Wing, St Vincent's Hospital, 41 Victoria Parade, Fitzroy, VIC 3065, Australia;* [e]*Current address: GeneType Pty Ltd, P.O. Box 115, Fitzroy, VIC 3065, Australia*

Received 20 September 2000; accepted 18 April 2001

Key words: computational combinatorial chemistry, docking, major histocompatibility complex, MCSS, peptide design

Summary

Peptides bound to MHC molecules on the surface of cells convey critical information about the cellular milieu to immune system T cells. Predicting which peptides can bind an MHC molecule, and understanding their modes of binding, are important in order to design better diagnostic and therapeutic agents for infectious and autoimmune diseases. Due to the difficulty of obtaining sufficient experimental binding data for each human MHC molecule, computational modeling of MHC peptide-binding properties is necessary. This paper describes a computational combinatorial design approach to the prediction of peptides that bind an MHC molecule of known X-ray crystallographic or NMR-determined structure. The procedure uses chemical fragments as models for amino acid residues and produces a set of sequences for peptides predicted to bind in the MHC peptide-binding groove. The probabilities for specific amino acids occurring at each position of the peptide are calculated based on these sequences, and these probabilities show a good agreement with amino acid distributions derived from a MHC-binding peptide database. The method also enables prediction of the three-dimensional structure of MHC-peptide complexes. Docking, linking, and optimization procedures were performed with the XPLOR program [1].

Abbreviations: MCSS – Multiple Copy Simultaneous Search; MHC – Major Histocompatibility Complex; CCLD – Computational Combinatorial Ligand Design; RMSD – Root-mean-square deviation; PBG – Peptide Binding Groove; RPP – Representative Polyglycine Peptides; PSSM – Position-Specific Scoring Matrix; ANN – Artificial Neural Network; HMM – Hidden Markov Model.

Introduction

Predicting sequences of MHC-binding peptides

Major Histocompatibility Complex (MHC) molecules bind short protein fragments (peptides) derived from foreign (e.g., viral) proteins and present them at the cell surface to T cells, thereby initiating specific immune responses [2]. The mechanism of MHC-peptide binding is thus of central importance to immune defence. Peptides bind MHC molecules in a specific manner; only certain peptides can bind the MHC molecule (allomorph) encoded by a specific MHC allele. Being able to predict which peptides bind which MHC allomorphs would greatly advance our understanding of the immune response and facilitate the design of diagnostic and therapeutic agents targeting T-cell immunity.

*To whom correspondence should be addressed. E-mail: Jun.Zeng@ludwig.edu.au

Over the past decade, many attempts have been made to characterize the peptide-binding properties of particular MHC allomorphs. Most have employed statistical methods to discover features common to the amino acid sequences of peptides found experimentally to bind the MHC molecule in question. Approaches using artificial neural networks (ANNs) [3–7], position-specific scoring matrices (PSSMs) [8–11], multiple regression [12], stepwise discriminant analysis [13], and hidden Markov models (HMMs) [14] have been published. By comparison, relatively few attempts have been made to exploit the structural information from X-ray crystallographic studies of MHC-peptide complexes. 'Threading' [15, 16], binding energy [17, 18], and molecular dynamics [19] – based approaches have been investigated for a limited number of MHC allomorphs, yielding only modest success despite the considerable computational costs entailed. Recently, three groups [20–22] have explored the possibility of deriving "pocket profiles" for MHC molecules that as yet lack solved crystal structures and experimental binding data. While this approach appears promising, it makes only limited use of existing structural information. Moreover, it is not clear how easily it can be generalized to all MHC molecules, nor how readily it can be refined to reflect new structural and sequence data that becomes available.

This paper describes a different approach to the study of MHC-peptide binding. Beginning with published X-ray crystal coordinates for an MHC molecule, a large set of potentially binding peptides is generated by a novel computational combinatorial ligand design (CCLD) method. Well-established multiple alignment and PSSM or HMM algorithms can then operate on these 'constructed' peptides, supplemented with experimental data when available, in order to create a useful model of the peptide-binding properties of the MHC allomorph under consideration. Joint treatment of structural and sequence-based information in this manner should lead to a more powerful analysis than would be possible with either alone. Moreover, additional experimentally-derived data can be easily integrated into an updated model.

MHC-peptide interactions

Two sorts of MHC molecules participate in immune responses – class I and class II. Here, a human MHC class I allomorph has been used for analysis; however, the approach outlined can be applied to either class. Peptides that bind MHC class I molecules are usually eight to eleven amino acids in length. X-ray crystallographic studies of complexes formed by ten different MHC class I molecules and various peptides have been reported: HLA-A*0201 [23, 24], HLA-Aw68 [25], HLA-B*2705 [26], HLA-B*5301 [27], HLA-B*3501 [28], HLA-B*0801 [25, 29, 30], H-2L^d [31], H-2D^d [32], H-2K^b [34], and H-2D^b [33]. These structures reveal that the peptides bind in a groove formed by two alpha helices and a seven-stranded beta sheet (Figure 1a). This peptide-binding groove (PBG) contains six distinct binding sites, A to F, some of which accommodate peptide sidechains [35]. In particular, the peptide C-terminal residue sidechain, which binds to site F near one end of the groove, is an important determinant of binding specificity [33, 36]. Sidechains at the peptide second and third positions generally target binding sites B and D, respectively, although in a few complexes sites C and D in the middle of the groove may interact with residues five or six of the peptide. Peptide residues that bind sites B-F are called 'anchor residues'; the positions they occupy in the peptide are termed 'anchor positions'. In general, anchor positions in MHC class I-binding peptides include positions 2 (P2), 3 (P3), 5 or 6 (P5/P6), and the C-terminus (PΩ).

Methods

This paper presents a CCLD-based method for predicting sequences and structures of MHC-binding peptides. The method is an extension of a previously described computational combinatorial approach for the design of peptides or ligands of a target protein with known three-dimensional structure [37–41]. In this study, the target protein is a human MHC class I molecule, HLA-A*0201, referred to subsequently as 'A2.'

The method consists of five steps (Figure 2): in the first, an exhaustive multiple copy simultaneous search (MCSS) is used to identify the positions of chemical functional group fragments on the target protein region of interest, the A2 PBG in this case. A combinatorial library is assembled from those fragments docking within a specified distance of the protein surface. In the second step, a polyglycine peptide mainchain is assembled from formamide (FORM) group fragments in the combinatorial library. Other functional group fragments from the library are then selected and linked to the mainchain C_α atoms (step three, 'Sidechain Selection and Linking'). The 'constructed' peptides are then 'completed' by energy minimization and exclusion

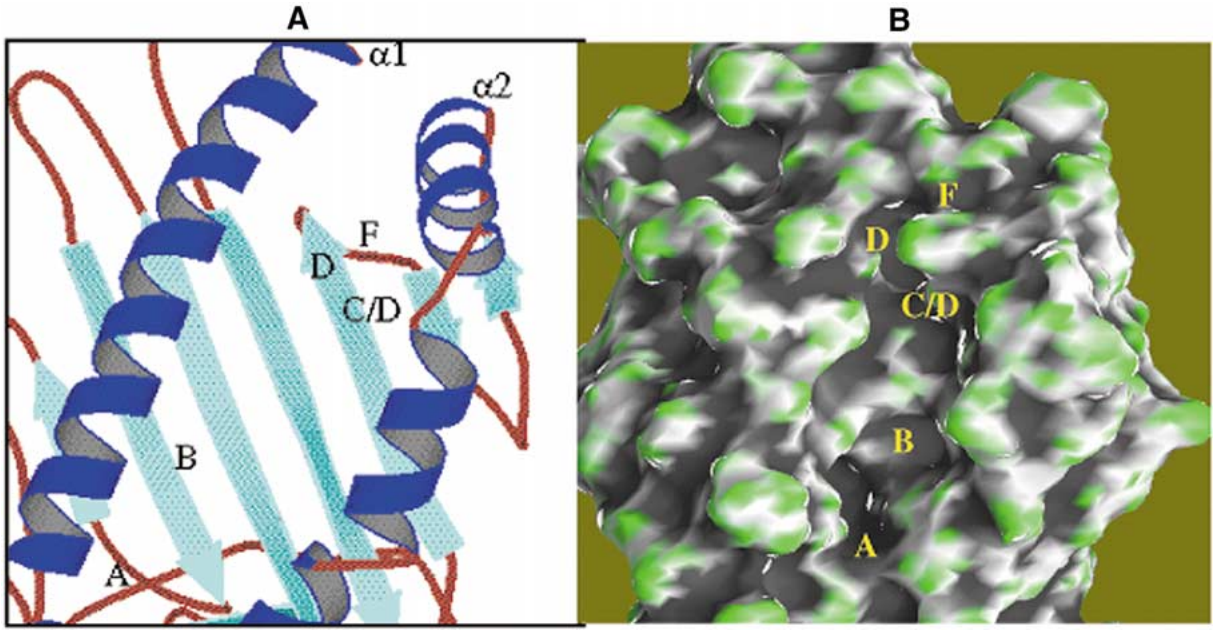

Figure 1. (A) The peptide binding groove (PBG) with binding sites A-F in the MHC class I molecule, HLA-A*0201 [35]. Helices are blue, beta strands cyan. The binding site E is not shown. (B) Accessible surface of the peptide binding groove (PBG) with binding sites A-F in the MHC class I molecule.

of those with unfavorable composition or conformation (step four). Lastly, a reduced-alphabet PSSM is created. Final peptide ligand sequences are predicted based on the relative frequencies of amino acid residues at each position in an alignment of the peptide sequences remaining after the fourth step.

Besides predicting the sequences of peptides that might bind an MHC molecule, the CCLD approach can also be used to determine the structure of MHC-peptide complexes. Using the MCSS-derived combinatorial library of functional group minima, the structure of an Influenza A matrix peptide (residues 58–67) bound in the A2 PBG is predicted and subsequently compared to the published crystal structure (PDB [42] entry code 1hhi [23]).

Functional groups (amino acid models)

Ten functional groups were chosen as amino acid models: formamide (FORM), benzene (BENZ), methyl (METY), propane (PROP), ethane (ETHA), isobutane (IBUT), phenol (PHEN), methanol (MEOH), acetate ion (ACET) and methylammonium (MAMM) (Table 1). FORM, representing glycine residues, was used for construction of the mainchain. Other groups correspond to apolar, polar, negatively charged, and lysine residues. TIP3P water potential [43] was used to model the explicit solvent molecules attached to the charged ACET and MAMM groups (see below). Parameters for functional groups, proteins and TIP3P water were taken from the CHARMM22 all-hydrogen atom force field [44].

Multiple Copy Simultaneous Search (MCSS)

The MCSS protocol used in this study is similar to the one described previously [37]. In brief, the structure of the MHC class I molecule was taken from the crystal structure of the complex between A2 and an Influenza A matrix peptide (PDB entry 1hhi, [23]). Replicas of each functional group were randomly distributed onto the A2 surface within a 20 Å radius around Tyr99-C, located at the centre of the PBG. With MHC atoms held fixed, positions of all replicas were minimized simultaneously for 1000 steps. For charged functional groups, the solvent molecules can play an important role in the protein-peptide interaction. Two water molecules were therefore attached to the oxygen atoms of ACETs, and three to the hydrogen atoms of MAMMs, so that each replica consisted of a functional group and its associated water molecules.

The interaction energy for each replica in the MCSS calculation was defined as

$$U_{\mathrm{MCSS}} = U_{\mathrm{target\_protein-replica}} + U_{\mathrm{replica}} \qquad (1)$$

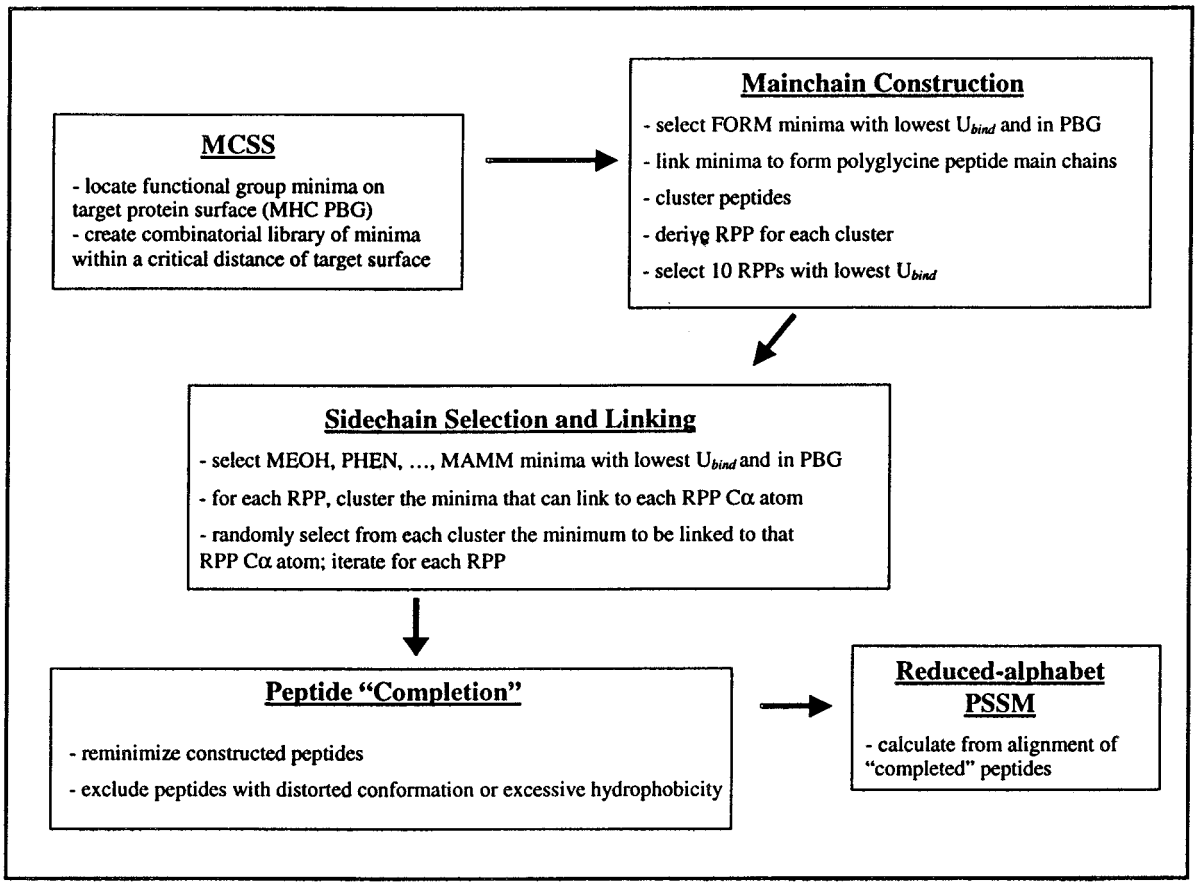

Figure 2. Overall scheme of the five-stage prediction method.

Table 1. Functional groups, their amino acid correspondences, and maximum distances for linking replicas to mainchain atoms.

Functional group	Abbreviation	Copies of group initially placed on surface	Docked minima	Minima included in library	Range of U_{bind} for minima (kcal/mol)	Amino acid correspondences	Maximum distance for linking minima to C_α atoms (Å)[b]
Formamide	FORM	300	190	56	−8.41,−0.90	Gly (Mainchain)	–
Methanol	MEOH	300	180	52	−5.42,−1.18	Ser,Thr	4.5
Phenol	PHEN	300	182	67	−6.66,−0.66	Tyr	4.5
Benzene	BENZ	300	187	62	−7.16,−0.75	Phe	4.5
Propane	PROP	300	187	53	−11.92,−0.41	Ile,Leu,Val	4.5
Isobutane	IBUT	300	189	50	−6.42,−0.50	Leu,Val	4.5
Ethane	ETHA	300	190	59	−5.68,−0.65	Ile,Leu,Val	4.5
Methyl	METY	300	221	61	−4.53,−0.25	Ala	3.0
Acetate ion	ACET	1000[a]	803	103	−37.30,−28.32	Asp,Glu	4.5
Methylammonium	MAMM	1000[a]	391	150	−22.42,−5.34	Lys	7.0

[a]200 acetate and methylammonium replicas (a functional group and two or three associated water molecules) were initially placed on the surface. 20 ps molecular dynamics simulation results were then obtained for 1000 replicas, as described in the text.
[b]Maximum distances are specified from a particular carbon atom to the C_α atom of the RPP. These atoms are C_β, C_γ, C_δ, and C_ε for MEOH, PHEN, ACET and MAMM, respectively. For apolar functional groups BENZ, PROP, IBUT and ETHA, any carbon atom may be used. To allow for mainchain conformational flexibility during sidechain construction, the distance is chosen to be 1.5 Å greater than that for inserting a CH_2 group with standard bond length.

where $U_{\text{target\_protein-replica}}$ and U_{replica} represent the nonbonded (van der Waals (vdW) and electrostatic) interactions between the MHC molecule and the replica, and the internal energy (bonded and nonbonded) of the replica itself, respectively. The binding energy for a functional group in each minimised replica obtained from the MCSS calculations was defined as

$$U_{\text{bind}} = U_{\text{target\_protein-functional\_group}} + U_{\text{functional\_group}} - U^{0}_{\text{functional\_group}} \quad (2)$$

where $U_{\text{target\_protein-functional\_group}}$, $U_{\text{functional\_group}}$, and $U^{0}_{\text{functional\_group}}$ represent the nonbonded interactions between the MHC molecule and the functional group, the internal energy of the functional group within the complex, and the internal energy of the isolated functional group, respectively. Each replica only interacts with MHC atoms within a 20 Å radius around Tyr99-Cζ, but not with the other replicas. Nonbonded interactions were truncated at 20 Å. A dielectric constant of 10 was used to mimic the screening effect of the solvent [45].

After several hundred charged replicas (i.e., ACET and MAMM) were generated randomly, 20 ps multiple copy molecular dynamics simulations [46] were performed at a temperature of 400 K with interaction energies defined as above. The coordinates of all the charged replicas were saved every 5 ps, followed by a 400-step MCSS minimization. This resulted in 1000 charged replicas being saved. For apolar and polar groups, 300 replicas were generated and docked onto the MHC molecule in the MCSS calculation. All the charged, apolar, and polar functional group minima less than 4.0 Å from A2 atoms and within a 20.0 Å radius around Tyr99-Cζ were selected, and those with U_{bind} less than -0.25 kcal/mol and within the PBG were included in the combinatorial library. All calculations were performed using the program XPLOR [1]. The number of replicas initially placed on the MHC molecule (300 each for apolar and polar groups and 1000 each for charged groups) was sufficient for mapping the important PBG binding sites, given the well-defined structure of the PBG. This was confirmed by a calculation in which additional BENZ replicas were added; no significant difference was found in the distribution of minima inside the PBG.

Peptide construction

Based on the combinatorial library of selected functional group minima, peptides were constructed in a three-step strategy: (i) mainchain construction, (ii) sidechain selection and linking, and (iii) peptide 'completion'. The peptide mainchain was defined by the formamide (FORM) minima identified in the MCSS calculations. The algorithm for linking the FORM replicas to make polyglycine mainchains was similar to that described previously [37], except that FORM minima were used to make an NH-CO frame rather than NMA minima to make a C_α trace; C_α atoms were subsequently added onto the NH-CO frame. Fifty polyglycine 9-mers were constructed, and a clustering procedure then employed to group the carbonyl and amide atomic positions [37]. Two polyglycine peptides were placed in the same cluster if the RMS deviation (RMSD) between their heavy atoms was less than 1 Å. A representative peptide for each cluster was derived by averaging over all member-peptide structures. These representative polyglycine peptides (RPPs) were then subjected to an 800-step conjugate gradient energy minimization. The ten RPPs with lowest interaction energy with the A2 molecule were selected for subsequent analysis and sidechain construction.

Peptide sidechains were constructed next by linking combinatorial library functional group minima to the C_α atoms of RPP mainchains [37] when the distance between them was less than a specified maximum (Table 1). Maximal distances were defined as being those required for inserting a CH_2 group with standard bond length, plus 1.5 Å. This results in increased flexibility of the peptide mainchain to accommodate MHC-peptide sidechain contacts, thereby reducing the conformational bias of the RPP, which was selected based on its strong interaction with the MHC molecule. The amino acid type for each peptide position was then randomly chosen from among those represented by replicas docked within the specified distance, and sidechain atoms grown to connect the functional group to the mainchain Cα atom. Based on their distance from the RPP, some minima could be equally well linked to either of two positions along the mainchain. In such cases, if the same minimum was randomly chosen more than once it was linked to the N-terminal-most candidate C_α atom. In this way, all constructed peptides were constrained to be 9-mers.

Finally, each constructed peptide was energy minimized to obtain its optimal position and orientation within the A2 groove. All linking and optimization procedures were performed using XPLOR [1]. Linking each of the 10 RPPs with the selected functional groups resulted in 500 constructed peptides predicted to bind with high affinity in the A2 PBG. Those

with significant geometric distortion (i.e., large internal energy) were excluded, as were those likely to be insoluble (i.e., more than 7 hydrophobic residues). The remaining 364 9-mer peptide sequences were then aligned and examined to determine the relative frequency of each amino acid at each position, thereby creating a PSSM reflecting predicted peptide-binding properties of the A2 molecule.

Results

Mainchain construction

Formamide minima. Potential A2-binding peptide mainchains were constructed from FORM minima identified in the MCSS calculation. In total, 300 FORM replicas were placed on the MHC molecule; 51 were retained for mainchain construction based on location in the PBG and binding energy (*Peptide Construction*), and are highlighted by thick lines (Figure 3A). These minima span residues (66–70, 73–77) in the α_1 helix and (150, 155–159) in the α_2 helix of A2; three interact with binding site B or F.

Mainchain conformers. Fifty polyglycine peptides were constructed from the retained FORM minima and clustered into subsets based on their heavy atom positions (RMSD between heavy atoms of less than 1 Å). The ten RPP with lowest U_{bind} were retained for subsequent construction of 'completed' peptides. Their structures are shown, along with those of four viral 9-mer peptide mainchains from published A2-peptide complex crystal structures (Figure 3B). Two sets of mainchain conformations can be distinguished (black and yellow in the figure). Mainchain conformations for nine of the RPPs are reasonably close throughout their length to those of at least one of the four viral peptides (RMSD less than 2.74 Å) (Table 2). All RPP structures showed excellent agreement with the viral mainchains in their N- and C-termini, apart from a discrepancy centered around non-anchor position P4.

Sidechain selection and linking

In addition to FORM, replicas of nine other functional groups were placed onto the A2 surface in the region of the PBG. In each case, over half the replicas docked to within 4 Å of A2 atoms. The approximately 200 minima with lowest U_{bind} from each group were retained for further consideration; only about one third of these were located within the PBG and so included into combinatorial libraries (Table 1). Their distributions on the A2 surface are shown (Figure 4, Panels A–I); those selected for subsequent sidechain attachment are highlighted by thick lines.

Table 2. RMSD (Å) between mainchain atoms of RPPs and crystal structures of four viral peptides[a].

RPP	I	II	III	IV
1	2.77	2.85	2.62	3.04
2	1.98	2.37	2.39	2.44
3	2.58	2.74	2.75	2.72
4	3.34	3.02	2.74	2.98
5	2.95	2.78	2.69	2.86
6	3.04	2.98	2.70	2.76
7	2.79	2.69	2.35	2.80
8	3.43	3.03	2.73	3.03
9	2.16	3.13	2.73	2.95
10	3.35	3.20	3.07	3.08

[a]Peptides forming complexes with the HLA-A*0201 molecule: I – HIV-1 gp120 (197-205, TLTSCNTSV); II – Influenza A matrix (58-66, GILGFVFTL); III – HIV-1 RT (476-484, ILKEPVHGV); IV – HTLV-1 Tax (11-19, LLFGYPVYV).

Apolar minima. The A2 PBG comprises 19 hydrophobic and 15 polar residues and, as expected, is primarily occupied by apolar functional groups (Figure 4A–E). Sixty-two BENZ minima are found inside the PBG, mostly concentrated in the region around site D between A2 residues 152–159 and 65–73 (Figure 4A). Fifty-three PROP minima dock within the PBG (Figure 4B). The 50 IBUT minima within the PBG (Figure 4C) lie over the β-sheet floor and the α_1 helix. The majority of the 59 ETHA minima concentrate in pockets B–D (Figure 4D). On the other hand, the 61 METY minima (Figure 4E) are scattered over the PBG.

Polar minima. Fifty-two MEOH minima are dispersed around the A2 PBG, but located primarily close to residues Arg65, Glu63, Lys66, Tyr159, Tyr99 and Tyr7 in pockets B and C (Figure 4F). The orientation of these minima is controlled by hydrogen bonds to polar and charged A2 residues. Sixty-seven PHEN minima are distributed in a similar manner to BENZ (Figure 4G). Overall, the polar minima are con-

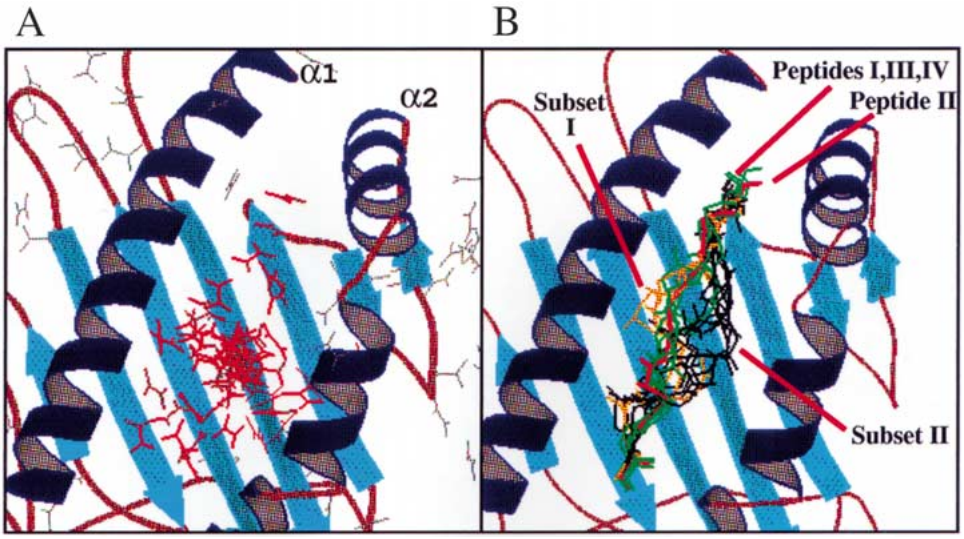

Figure 3. Stereo views of polyglycine peptides built with FORM minima obtained in the MCSS calculations. (A) Distribution of FORM minima used for the construction of polyglycine peptides. Those minima used to form the NH-CO frames of polyglycine peptides are highlighted by thick lines (see text for details). (B) Structures of the ten polyglycine peptides inside the PBG of the MHC HLA-A*0201 molecule compared to the crystal structures of four viral peptide backbones [23]. The two conformational subsets of the 10 polyglycine peptides are in yellow and black, respectively. The peptides (I, III, IV) from the X-ray structures are colored green, and the Influenza A matrix peptide (peptide II) is in red (see text for details).

centrated around pockets B and C due to significant hydrophobicity in pockets D and F.

Charged minima. Multiple negatively charged residues are situated near the A2 PBG pocket A and attract significant numbers of MAMM minima, resulting in a well-defined distribution for the 150 MAMM (Figure 4H). Similarly, positively charged residues in the vicinity of pocket D invite aggregation of 103 ACET minima (Figure 4I).

For each RPP, all the apolar, polar, and charged minima in the combinatorial library were clustered into subsets that could be linked to each RPP C_α atom. Fifty sequences were generated by randomly selecting minima according to their distributions in the cluster associated with each C_α atom in each RPP, resulting in a total of 500 constructed peptides (*Peptide Construction*). However, after connecting the sidechains, 136 (27%) of the constructed peptides showed large conformational distortions or unrealistically high hydrophobicity and were therefore excluded.

The position specific scoring matrix (PSSM)

The remaining 364 completed 9-mer peptides were used to build a reduced-alphabet PSSM representing peptides predicted to bind in the A2 PBG. Figure 5 shows graphically the relative frequencies of amino acids in each position of the aligned 9-mers. For comparison, the relative frequencies of amino acids in each position of 69 9-mer sequences reported in the literature to bind A2 with high affinity [47] are also plotted. Overall, the patterns of apolar, polar and charged amino acids in the constructed peptides show good agreement with those of the known high-binding-affinity peptides. Major differences were found at positions P1-P3, where frequencies of polar residues were systematically overestimated in the constructed sequences, and at position P8 where no Ser/Thr (MEOH) were predicted. The frequent occurrence of polar residues at positions P1-P3 was possibly due to hydrogen bond formation between the MEOH minima and residues Tyr7, Tyr99, and Glu63 in pockets A and B. The discrepancy at P8, where no MEOH minima were predicted, is probably attributable to the hydrophobic nature of pockets D and F. In fact, Ser/Thr at position P8 is exposed to solvent in the crystal structures of A2 complexed with the HIV gp120 and Influenza A matrix peptides [23].

Predicting MHC-bound peptide structure from sequence

Given a polyglycine mainchain conformation, the three-dimensional structure of a 'completed' peptide (i.e., with a specific amino acid sequence) docked to an

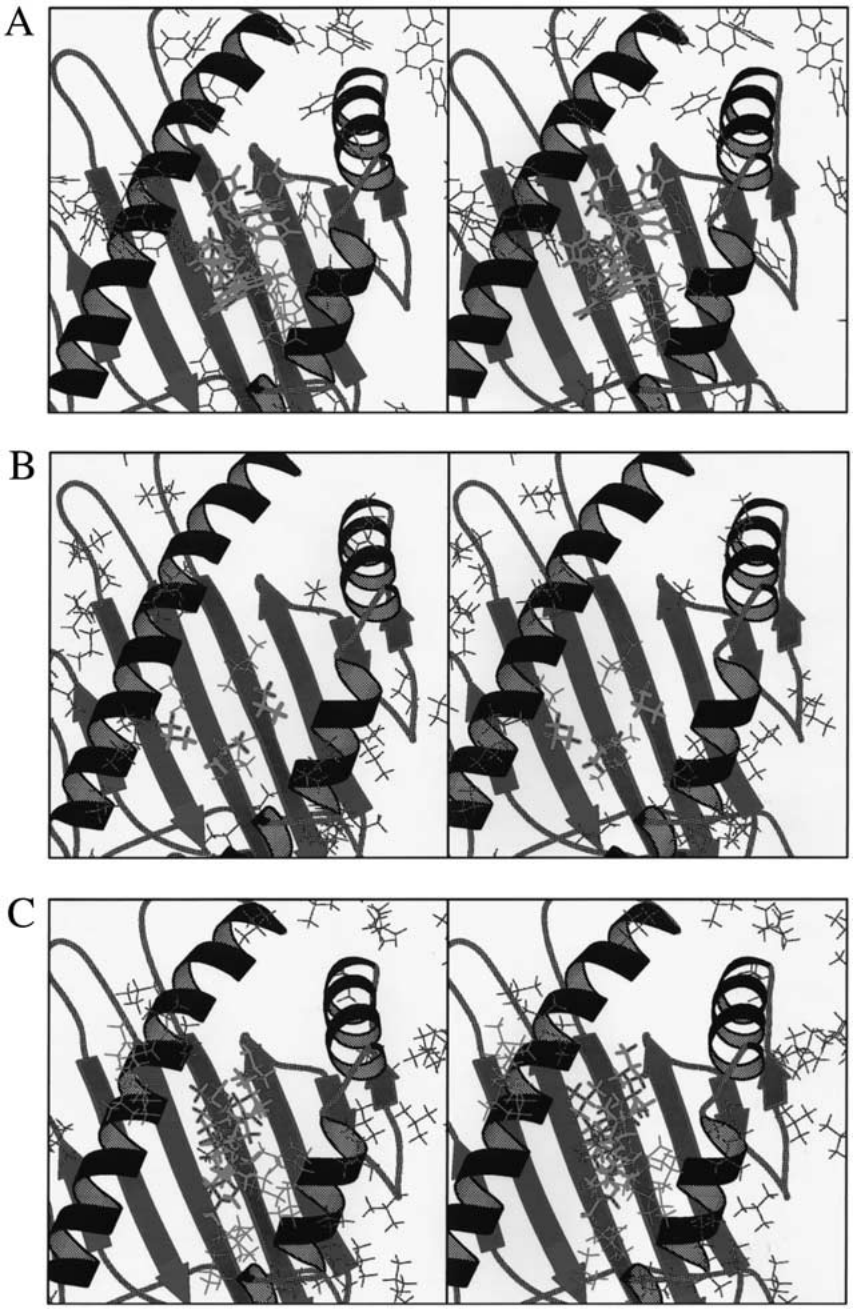

Figure 4. Stereo views of distributions of functional group minima identified in the MCSS calculations. Minima selected for linking to RPP C_α atoms are highlighted by thick lines. (A) Benzene (BENZ). (B) Propane (PROP). (C) Isobutane (IBUT). (D) Ethane (ETHA). (E) Methyl (METY). (F) Methanol (MEOH). (G) Phenol (PHEN). (H) Methylammonium (MAMM). (I) Acetate ion (ACET).

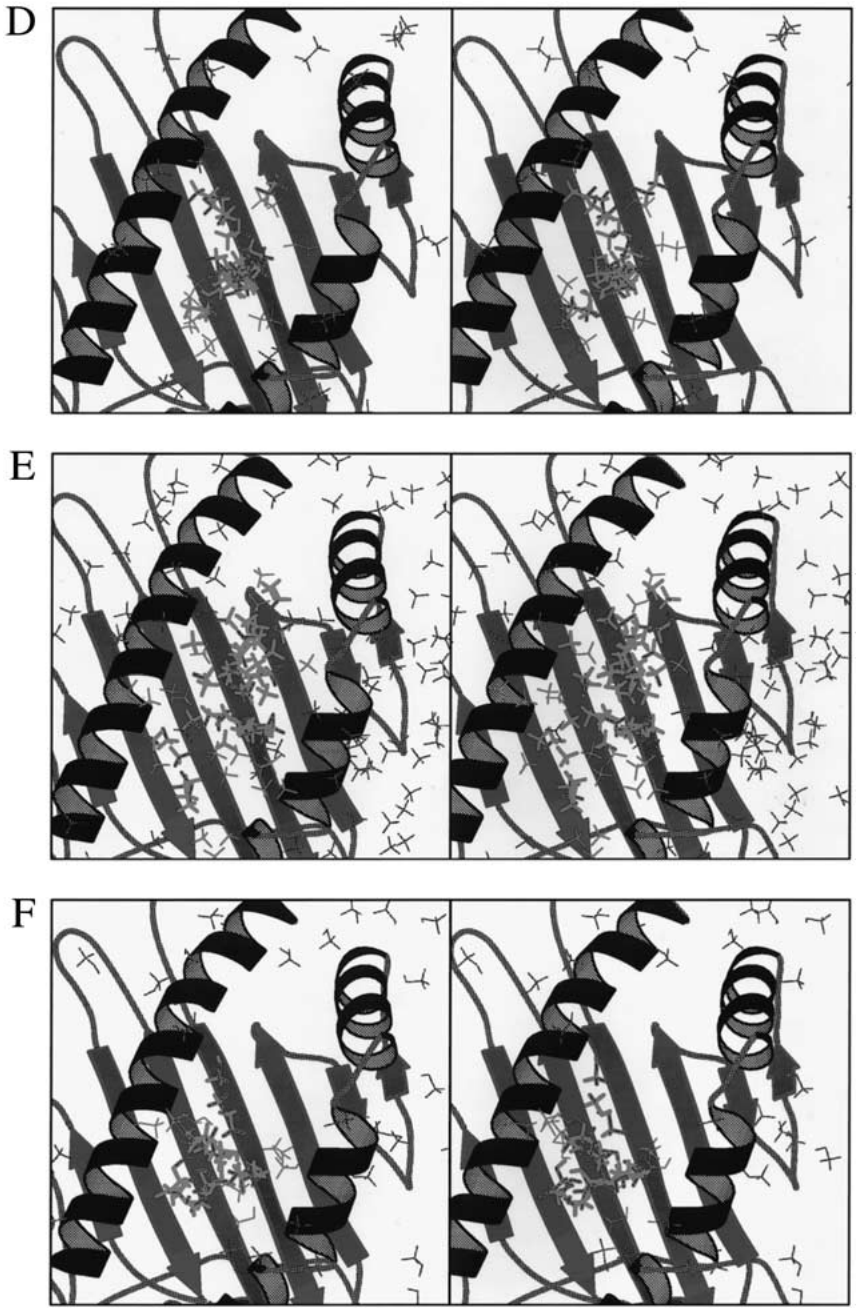

Figure 4. Continued.

MHC molecule can be estimated. Unlike the method described above, which focuses on the generation of peptide sequences, here a specific functional group minimum is selected for each peptide mainchain C_α atom. These specific functional groups correspond to the amino acid sequence specified for the peptide. The Influenza A matrix peptide complexed with A2 [23] was examined here as a test case, as the sequence of this peptide (GILGFVFTL) could in principle be obtained from the functional groups used in this study.

The second RPP mainchain (Table 2) was selected as the template for construction of the GILGFVFTL peptide for two reasons: first, all amino acids in the GILGFVFTL sequence can in principle be represented

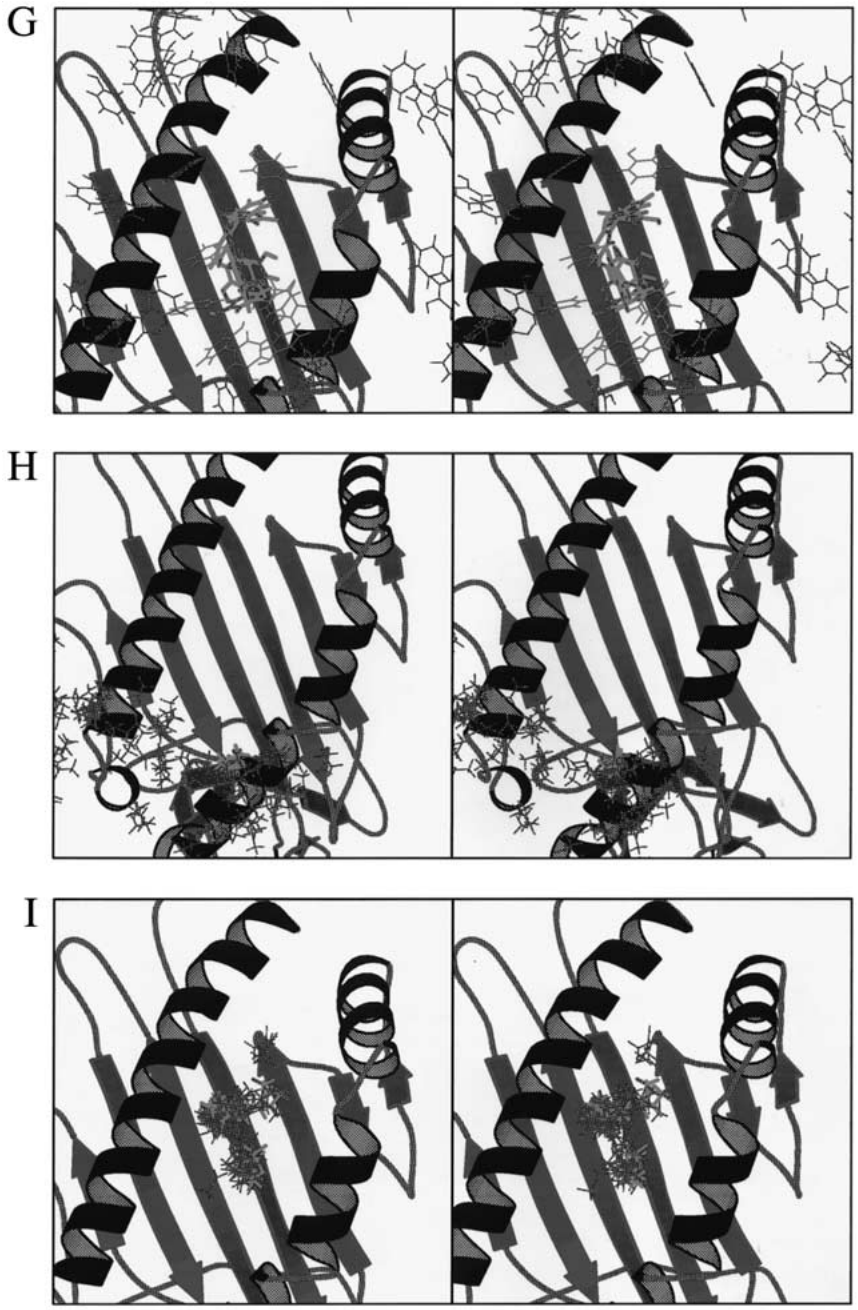

Figure 4. Continued.

by functional group minima included in the clusters corresponding to the C_α atoms of this RPP; second, the RMSD for mainchain atoms of this RPP from those of the GILGFVFTL crystal structure (2.37 Å) is the smallest for any of the ten RPPs in the generated set (Table 2, Column II). Residues Thr-P8 and Leu-P9 were replaced by Gly and Ala, respectively, as no Ser/Thr were predicted at position P8 (see above), and only small, apolar METY minima docked into pocket F. Using the protocol described (*Peptide Construction*), 50 conformers of the peptide were generated. In this case, however, mainchain atoms were allowed to move freely during sidechain construction so as to

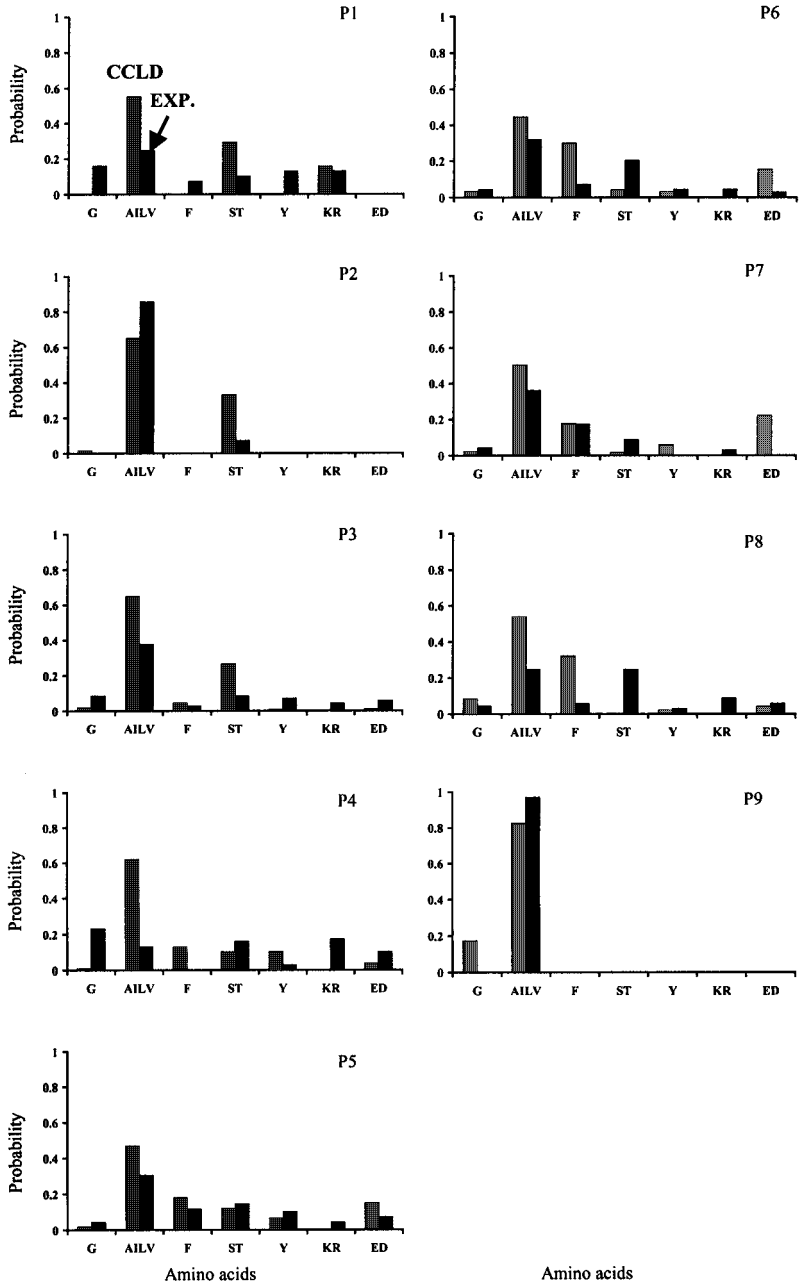

Figure 5. Graphical representation of PSSMs – relative frequencies of amino acid residues at positions 1 through 9 of peptides. Frequencies for peptides predicted by CCLD are in gray; those for peptides reported to have high experimental binding affinity [47] are in black.

maintain optimal orientations of the functional group minima.

Table 3 lists the result of structural analyses of the 10 conformers with amino acid sequence (GILGFVFGA) closest to the target sequence (GILGFVFTL) and with lowest interaction energies with the A2 PBG surface. Interaction energy is dominated by vdW forces, as expected from the hydrophobicity of the peptide, resulting in a ranking of the conformers based primarily on vdW contributions. Structures of the two best-predicted (i.e., minimum RMSD) conformers – first and fourth in Table 3, with vdW interaction energies with A2 of −74.83 kcal/mol and −71.07 kcal/mol – and the crystal structure of the

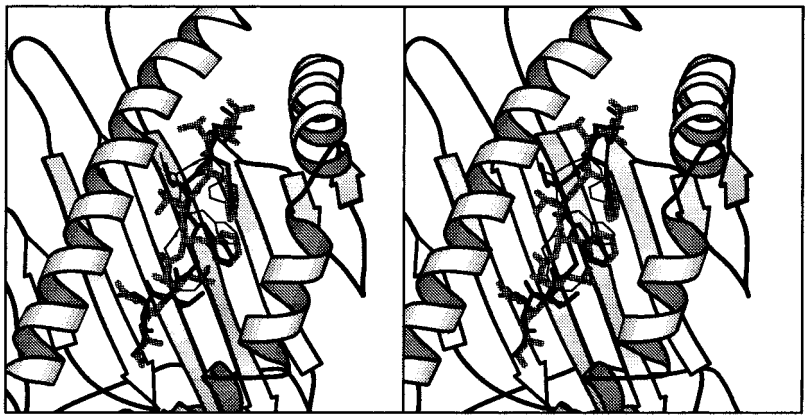

Figure 6. Comparison of the two best structures predicted by the CCLD method with the crystal structure of the Influenza A matrix peptide (residues 58–67) in the PBG. The predicted conformers are drawn in thin and thick black lines, and the experimentally-determined structure is in gray.

Table 3. RMSD (Å) and energy (kcal/mol) of the ten lowest energy complexes between CCLD-predicted peptide conformers[a] and HLA-A*0201

Conforme	RMSD				U_{bind}	
	Heavy atoms	Main chain	Side chain	Total	vdW	Electrostatic
1[b]	2.42	2.43	2.40	−71.77	−74.83	2.30
2	2.89	2.70	3.11	−72.47	−72.67	0.20
3	3.57	2.56	4.51	−74.58	−71.81	−2.71
4[b]	2.52	2.38	2.69	−71.07	−71.07	0.0
5	2.65	2.77	2.50	−68.84	−69.91	1.07
6	3.41	2.14	4.50	−62.40	−67.79	5.39
7	3.27	3.00	3.58	−66.47	−66.87	0.44
8	3.09	2.31	3.83	−66.07	−66.78	0.71
9	2.92	2.59	3.29	−62.98	−65.57	2.59
10	3.11	2.38	3.83	−63.64	−62.35	−1.29

[a]The sequence of the predicted peptide is not identical to that of the Influenza A matrix peptide from the crystal structure (GILGFVFGA vs. GILGFVFTL). Therefore, RMSD were calculated using only backbone atoms of Thr8 and matching atoms at position P9 (Ala vs. Leu). See text for details.
[b]Two best conformers are shown in Figure 6.

native GILGFVFTL peptide [23] are compared (Figure 6). The two conformers lie quite close to the crystal structure, except again at position P4 where considerable variation of conformer mainchains is present (*Mainchain Construction*). The RMSDs at Gly4 were calculated to be 3.88 Å and 4.65 Å. As a consequence of this P4 variation, the sidechains deviate significantly from those of the crystal structure at residues Leu3 and Val6. While Leu3 in the two conformers has RMSDs of 1.32 Å and 1.55 Å for mainchains, 1.62 Å and 3.22 Å for sidechains, and 1.52 Å and 3.86 Å for heavy atoms, Val6 has RMSDs of 2.64 Å and 1.71 Å for mainchains, 4.38 Å and 3.41 Å for sidechains, and 3.49 Å and 2.58 Å for heavy atoms. The RMSD of the predicted structures was calculated with only the backbone atoms of Thr8 and those atoms matching at position P9 (Leu replaced by the similar Ala). Overall, these two conformers are in good agreement with the A2-Influenza A matrix peptide crystal structure (Table 3), with RMSDs of 2.43 Å and 2.38 Å for mainchains, 2.40 Å and 2.69 Å for sidechains, and 2.42 Å and 2.52 Å for heavy atoms.

Discussion

This paper introduces a novel method for generating both sequences and structures of peptides predicted to bind an MHC molecule of known three-dimensional structure. The results presented demonstrate that the CCLD approach is sensitive enough to capture important features of both sequence and structural experimental data. Specifically, the distributions in experimental data of hydrophobic, polar and charged residues at A2 anchor positions P2, P3, P5/P6, and PΩ were reproduced in the constructed peptides. Predicted binding motifs involving the P1 and P7 secondary anchor positions were also in generally good agreement with experiment. There is likely bias in the experimental data used, due to the relatively small number of high-affinity binding sequences available, and that may be a factor in the discrepancies noted at P1-P3 and P8.

Results of the CCLD structural modeling were similarly encouraging, with RMSD of 2.43 Å for the best-predicted A2-binding conformer and the crystal structure, i.e. comparable to the result (2.1 Å) obtained for this same Influenza A matrix peptide using a general sidechain rotamer library [48]. Of note, RMSDs among crystal structures of A2-binding nonamer mainchains themselves range from 0.7 to 1.4 Å, and up to 2.1 Å for decamers [48]. The overall quality of the predicted conformer structures correlated well with their vdW binding energies (Table 3), as to be expected. Although atomic position Gly4 in the two best conformers deviates significantly from that in the crystal structure (Figure 6), the local structures of Phe5, Val6, and Phe7 are well-predicted.

In spirit, the CCLD method lies somewhere between structure and sequence-based approaches, combining at least some of the advantages of each. On the one hand, it considers peptide sidechain and mainchain atom interactions with all residues in the MHC PBG, not just with those residues associated with binding pockets A–F. On the other hand, by reducing the model to a PSSM with the attendant assumptions of independent and additive contributions to binding of residues at each peptide position, it sacrifices some of the structural information in exchange for the convenience of sequence manipulation. That this approach performs well in the case of modeling MHC-peptide binding may be due in part to the particular circumstances. First, the PBG is a well-defined region of the MHC surface and has been extremely well characterized in the multiple crystal structures solved to date. This permits a drastic reduction in the size of the conformational space searched during the sidechain construction step of the protocol. Second, comparison of A2-peptide crystal structures [23, 24] has revealed that conformational changes induced in the MHC molecule on binding of peptide are limited to only slight changes in the sidechain conformations of only three residues. This provides a significant theoretical advantage, as it allows the structure of the target protein (A2) to be fixed to the known crystal structure for the purposes of CCLD. Finally, given that observed binding modes for MHC class I-peptide interactions are primarily hydrophobic [45], certain current assumptions of the CCLD approach, e.g., neglect of explicit solvation, might be expected to have only minor effects on the results.

Currently, the technique is limited by a restricted alphabet of functional groups used for modeling amino acids. However, this alphabet is being expanded to effectively cover all naturally occurring amino acids. Other improvements, in particular to the force field parameters, are also envisaged. Inclusion of solvation effects, either by implicit continuum via solution of the Poisson–Boltzmann equation or by explicit solvent molecules in the MCSS calculation, and consideration of target protein flexibility should result in better binding affinity estimates. Finally, for those cases where experimental binding data are available, it may be desirable to incorporate a scoring function to calculate the binding affinities of designed peptides to their MHC targets, with parameter values estimated from the empirical data.

The CCLD method has several apparent strengths: first, the approach taken – MCSS followed by peptide construction and minimization – does not require the prior choice of rotamer libraries [48] or the lengthy conformational space search involved in the explicit free-energy mapping [18] of other structure-based techniques. Second, unlike techniques based purely on experimental binding data, it is not allomorph-specific; that is, homology modeling should permit extension to other MHC structures, as has been previously demonstrated for another structure-based approach [18]. Finally, and perhaps most useful, the method presented here permits easy integration of experimentally-determined binding sequence data with that generated via CCLD. It is trivial to recalculate the PSSM associated with a given MHC allomorph whenever new experimental data become available, compared to the effort required to retrain an artificial neural network (ANN) or hidden Markov

model (HMM). And as the amount of experimental evidence increases, the contributions of experimentally-determined and "constructed" sequences can be differentially weighted [49, 50]. In this manner, it should be feasible to create and maintain a database of PSSMs, one for each human MHC allomorph. Predicting MHC-binding subsegments of a protein sequence will then reduce to conducting a search against a PSSM database [51].

Acknowledgements

J.Z. gratefully acknowledges a C.J. Martin fellowship awarded by the Australian National Health and Medical Research Council (No. 967362). The authors would like to thank R. Flegg for initially suggesting the investigation of MHC-peptide binding properties.

References

1. Brunger, A.T., X-PLOR version 3.1, Yale University (1992).
2. Rothbard, J.B. and Gefter, M.L., Annu. Rev. Immunol., 9 (1991) 527.
3. Bisset, L.R. and Fierz, W., J. Mol. Recognit., 6 (1993) 41.
4. Adams, H.P. and Koziol, J.A., J. Immunol. Methods, 185 (1995) 181.
5. Gulukota, K., Sidney, J., Sette, A. and DeLisi, C., J. Mol. Biol., 267 (1997) 1258.
6. Brusic, V., Rudy, G., Honeyman, G., Hammer, J. and Harrison, L., Bioinformatics, 14 (1998) 121.
7. Milik, M., Sauer, D., Brunmark, A.P., Yuan, L., Vitiello, A., Jackson, M.R., Peterson, P.A., Skolnick, J. and Glass, C.A., Nat. Biotechnol., 16 (1998) 753.
8. Hammer, J., Bono, E., Gallazzi, F., Belunis, C., Nagy, Z. and Sinigaglia, F., J. Exp. Med., 180 (1994) 2353.
9. Parker, K.C., Bednarek, M.A. and Coligan, J.E., J. Immunol., 152 (1994) 163.
10. Davenport, M.P., Ho Shon, I.A. and Hill, A.V., Immunogenetics, 42 (1995) 392.
11. Marshall, K.W., Wilson, K.J., Liang, J., Woods, A., Zaller, D. and Rothbard, J.B., J. Immunol., 154 (1995) 5927.
12. Mallios, R.R., J. Theor. Biol., 166 (1994) 167.
13. Mallios, R.R., Bioinformatics, 15 (1999) 432.
14. Mamitsuka, H., Proteins, 33 (1998) 460.
15. Altuvia, Y., Sette, A., Sidney, J., Southwood, S. and Margalit, H., Hum. Immunol., 58 (1997) 1.
16. Rognan, D., Lauemoler, S.L., Holm, A., Buus, S. and Tschnike, V., J. Med. Chem., 42 (1999) 4650.
17. Froloff, N., Windemuth, A. and Honig, B., Protein Sci, 6 (1997) 1293.
18. Sezerman, U., Vajda, S. and DeLisi, C., Protein Sci, 5 (1996) 1272.
19. Lim, J.S., Kim, S., Lee, H.G., Lee, K.Y., Kwon, T.J. and Kim, K., Mol. Immunol., 33 (1996) 221.
20. Chelvanayagam, G., Hum. Immunol., 58 (1997) 61.
21. Sturniolo, T., Bono, E., Ding, J., Raddrizzani, L., Tuereci, O., Sahin, U., Braxenthaler, M., Gallazzi, F., Protti, M.P., Sinigaglia, F. and Hammer, J., Nature Biotechnology, 17 (1999) 555.
22. Seeger, F.H., Schirle, M., Gatfield, J., Arnold, D., Keilholz, W., Nickolaus, P., Rammensee, H.G. and Stevanovic, S., Immunogenetics, 49 (1999) 571.
23. Madden, D.R., Garboczi, D.N. and Wiley, D.C., Cell, 75 (1993) 693.
24. Collins, E.J., Garboczi, D.N. and Wiley, D.C., Nature, 371 (1994) 626.
25. Silver, M.L., Guo, H.C., Strominger, J.L. and Wiley, D.C., Nature, 360 (1992) 367.
26. Madden, D.R., Annu. Rev. Immunol., 13 (1995) 587.
27. Smith, K.J., Reid, S.W., Harlos, K., McMichael, A.J., Stuart, D.I., Bell, J.I. and Jones, E.Y., Immunity, 4 (1996) 215.
28. Smith, K.J., Reid, S.W., Stuart, D.I., McMichael, A.J., Jones, E.Y. and Bell, J.I., Immunity, 4 (1996) 203.
29. Guo, H.C., Jardetzky, T.S., Garrett, T.P., Lane, W.S., Strominger, J.L. and Wiley, D.C., Nature, 360 (1992) 364.
30. Collins, E.J., Garboczi, D.N., Karpusas, M.N. and Wiley, D.C., Proc. Natl. Acad. Sci. USA, 92 (1995) 1218.
31. Speir, J.A., Garcia, K.C., Brunmark, A., Degano, M., Peterson, P.A., Teyton, L. and Wilson, I.A., Immunity, 8 (1998) 553.
32. Achour, A., Persson, K., Harris, R.A., Sundback, J., Sentman, C.L., Lindqvist, Y., Schneider, G. and Karre, K., Immunity, 9 (1998) 199.
33. Fremont, D.H., Matsumura, M., Stura, E.A., Peterson, P.A. and Wilson, I.A., Science, 257 (1992) 919.
34. Young, A.C., Zhang, W., Sacchettini, J.C. and Nathenson, S.G., Cell, 76 (1994) 39.
35. Zhang, C., Anderson, A. and DeLisi, C., J. Mol. Biol., 281 (1998) 929.
36. Parker, K.C., Biddison, W.E. and Coligan, J.E., Biochemistry, 33 (1994) 7736.
37. Zeng, J. and Treutlein, H.R., Protein Eng., 12 (1999) 457.
38. Caflisch, A., Miranker, A. and Karplus, M., J. Med. Chem., 36 (1993) 2142.
39. Caflisch, A. and Karplus, M., Perspect. Drug Discov. Des., 3 (1996) 51.
40. Caflisch, A., J. Comput. Aided Mol. Des., 10 (1996) 372.
41. Joseph-McCarthy, D., Hogle, J.M. and Karplus, M., Proteins, 29 (1997) 32.
42. Bernstein, F.C., Koetzle, T.F., Williams, G.J., Meyer, E.E., Jr., Brice, M.D., Rodgers, J.R., Kennard, O., Shimanouchi, T. and Tasumi, M., J. Mol. Biol., 112 (1977) 535.
43. Neria, E., Fischer, S. and Karplus, M., J. Chem. Phys., 105 (1996) 1902.
44. MacKerell, A.D., J. Phys. Chem., B102 (1998) 3586.
45. Simonson, T. and Brunger, A.T., Biochemistry, 31 (1992) 8661.
46. Pearlman, D.A. and Murcko, M.A., J. Med. Chem., 39 (1996) 1651.
47. Brusic, V., Rudy, G. and Harrison, L.C., Nucleic Acids Res., 26 (1998) 368.
48. Schueler-Furman, O., Elber, R. and Margalit, H., Folding & Design, 3 (1998) 549.
49. Thompson, J.D., Higgins, D.G. and Gibson, T.J., Comput. Appl. Biosci., 10 (1994) 19.
50. Henikoff, S. and Henikoff, J.G., J. Mol. Biol., 243 (1994) 574.
51. Schaffer, A.A., Wolf, Y.I., Ponting, C.P., Koonin, E.V., Aravind, L. and Altschul, S.F., Bioinformatics, 15 (1999) 1000.