Modeling the Peptide−T Cell Receptor Interaction by the Comparative Molecular Similarity Indices Analysis−Soft Independent Modeling of Class Analogy Technique

Irini A. Doytchinova[†] and Darren R. Flower*

The Jenner Institute, University of Oxford, Compton, Berkshire RG20 7NN, United Kingdom

Received September 5, 2005

A set of 38 epitopes and 183 non-epitopes, which bind to alleles of the HLA-A3 supertype, was subjected to a combination of comparative molecular similarity indices analysis (CoMSIA) and soft independent modeling of class analogy (SIMCA). During the process of T cell recognition, T cell receptors (TCR) interact with the central section of the bound nonamer peptide; thus only positions 4−8 were considered in the study. The derived model distinguished 82% of the epitopes and 73% of the non-epitopes after cross-validation in five groups. The overall preference from the model is for polar amino acids with high electron density and the ability to form hydrogen bonds. These so-called "aggressive" amino acids are flanked by small-sized residues, which enable such residues to protrude from the binding cleft and take an active role in TCR-mediated T cell recognition. Combinations of "aggressive" and "passive" amino acids in the middle part of epitopes constitute a putative TCR binding motif.

Introduction

T lymphocytes are an essential component of the adaptive immune response and have been documented in almost all chordates.[1] They recognize degraded intracellular protein fragments with lengths of 8−12 amino acids bound to major histocompatibility complex (MHC) class I proteins. Intracellular peptide fragments are from two sources: self-proteins and antigenic proteins.[2] Self-proteins are degraded at a fast rate, including some newly synthesized proteins, producing large quantities of short peptides. Antigenic proteins are derived from external agents, such as viruses and bacteria, which are degraded by the host in a similar way to self-proteins. Intracellular protein degradation is undertaken by a supramolecular complex known as the proteasome. After peptides are generated, they are translocated into the endoplasmic reticulum (ER) lumen by the transporter associated with antigen processing (TAP). In the ER, peptides associate with MHC class I molecules and the resulting peptide−MHC complexes are transferred to the cell surface, where they are recognized by T cells via T cell receptors (TCR). However, not all presented peptides are recognized by T cells. Those that are recognized are commonly referred to as epitopes.

The TCR−peptide−MHC complex is shown in Figure 1. TCR molecules are membrane-bound glycoproteins. Most TCR molecules consist of two polypeptide chains, α and β.[3] TCRs are associated with the CD3 complex, which helps to transport TCRs to the cell surface and send activating intracellular signals to the T cell when peptide−MHC complexes are recognized.[4] The TCR proteins are produced by gene rearrangement, as are immunoglobulins.[5] The α chain is formed by the rearrangement of the variable (V) to the joining (J) segment, and the β chain is produced by the rearrangement of the variable (V), diversity (D), and joining (J) genes.[6] This rearrangement creates a potential repertoire of ∼10^{13} different T cells.[7] The rearranged genes are attached to the constant (C) gene to form the complete α and β chains. There are four hypervariable complementarity-determining regions (CDRs) on α and β chains, three of which (CDR1, CDR2, and CDR3) resemble the CDRs of immunoglobulins.[8] These hypervariable regions form the contact site between the TCR and the peptide−MHC complex. CDR3 is the most variable and is considered to be principally responsible for TCR specificity. CDR3 contacts position 4 to position 8 of the peptide.[9] Mutations on the CDR3 loops are known to abolish peptide−MHC recognition.[10] The crystallized structures of a TCR complex with class I MHC[8−18] shows that the TCR−MHC binding surface is not parallel to the major axis of the MHC peptide binding groove. Instead, it varies between 20° [9] and

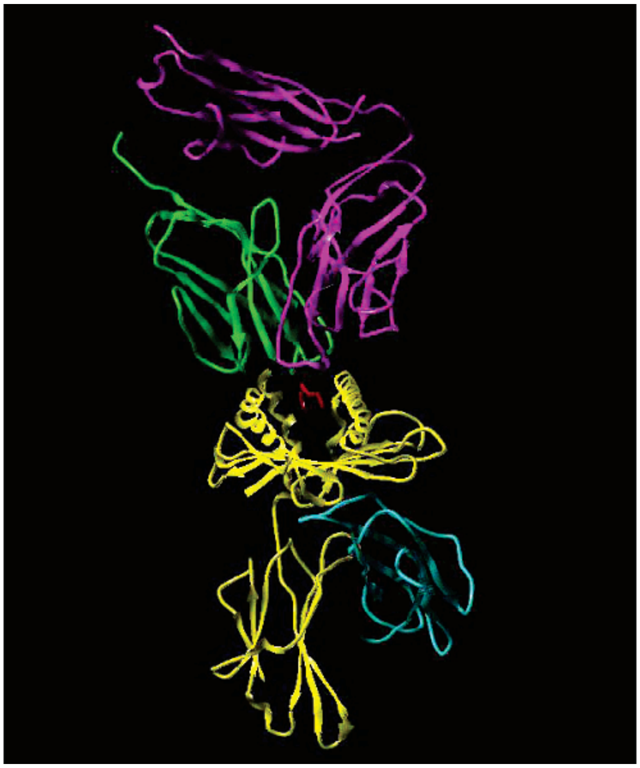

Figure 1. Crystal structure of TCR complexed with HLA-A*0201 and viral peptide Tax (PDB code 1AO7).[8] The TCR α chain is in green, and the β chain is in purple. The HLA α chain is in yellow, and β2-microglobulin is in cyan. The peptide is in red.

* Address correspondence to this author: tel +44 1635 577954; fax +44 1635 577908; e-mail darren.flower@jenner.ac.uk.

[†] Present address: Faculty of Pharmacy, Medical University, 1000 Sofia, Bulgaria.

10.1021/jm050876m CCC: $33.50
Published on Web 03/09/2006

70° [16] toward the diagonal. A hydrophobic pocket was formed above the binding site between residues 93−104 of the α chain and 95−107 of the β chain, which could accommodate a peptide side chain.[9] The great structural variability and large conformational changes induced in the TCR upon binding allows the receptor to interact with many different peptide−MHC interfaces.[13]

Recognition of peptide−MHC by the T cell is a multistep process.[7] Initially, the T cell probes cells presenting peptide−MHC complexes with pseudopodial extensions (scanning phase). This contact with the MHC orients the TCR so that it can quickly determine whether the peptide occupying the binding groove is appropriate. This is followed by folding of the CDR3 loops around the peptide to achieve a stable state. The T cell then becomes activated with a sustained release of calcium from internal stores (early activation phase). At this point, it begins to form a synapse characterized by a discrete pattern of central TCR/CD3 accumulation surrounded by a ring of coreceptors (synapse phase). Finally, the cytotoxic T cell secretes factors to induce cell death (effector phase).

Despite the tremendous potential diversity of an individual T cell repertoire, there are common patterns underlying the recognition of epitopes and initiation of the immune response by T cells, as manifest in the interactions observed between the TCR and the peptide−MHC complex. First, the TCR fits into a surface feature common to all MHC molecules, suggesting that the diagonal mode of binding might be general.[8−18] MHC helices impose steric limitations on the orientation and depth of approach of the TCR to the bound peptide; the diagonal orientation allows for the deepest docking solution of the TCR CDRs onto the peptide−MHC surface. Thus, the diverse CDR3 α and β loops interact primarily with the most exposed middle region of the bound peptide. Second, very different TCR sequences can recognize the same antigen.[12] At the same time, however, small changes in ligand structure can induce different signals when recognized by the same TCR (cross-reactivity).[19] These signals can vary from strong agonistic to full antagonistic effects.[14] Subtle changes in the structure and conformation in the middle region of peptide may generate analogues with unexpectedly high immunogenicity, defined as “heteroclitic analogues”.[20,21]

Third, there is no clear relationship between MHC binding affinity and T cell recognition. Although most of the T cell epitopes are good MHC binders, there are numerous exceptions, such as cancer epitopes.[22] This means that other structural factors, arising from the non-MHC-buried part of the peptide structure, may account for the interaction with TCR. Moreover, no clear relationship has been established between the binding of TCR to pMHC and the contingent functional response of whole T cells. Thus, instead of attempting to quantify TCR−pMHC interactions, we have addressed functional responses by discriminating predictively epitopes, which give rise to T cell responses, from peptides that do not engender either naïve immunogenic or recall antigenic T cell responses.

It is important not to confuse the capacity of being an epitope with thermodynamic measurements that characterize the binding of pMHC and TCR. It is well-known that a peptide can either be an epitope or be inactive in terms of immunogenicity (initial response by an unprimed T cell) or antigenicity (an equivalent recall response). For class I presentation, arguably the most direct approach is to measure T cell killing. CD8+ T cells, often called cytotoxic T lymphocytes or CTL, lyse cells upon antigen activation. This can be measured by use of ^{51}Cr, or radiolabeled thymidine, taken up into target cells and released upon CTL lysis. Alternatively, for class II presentation, the proliferative response of CD4+ T cells, which act more indirectly through the activation of B cells or macrophages rather than by direct cell killing, can be measured by use of tritiated thymidine that is incoporated into T cell DNA during cell division. Alternatively, enzyme-linked immunospot (ELISpot) assays measure the ability of class I and/or class II T-cells to produce cytokines (most often interferon-γ, but also interleukins IL-2 or IL-10) or other molecules when exposed to antigen. More recently attention has turned to RT-PCR and tetramers as tools for detecting T cell responses. As there are many different ways to identify T-cell epitopes, the quantitative data produced by such assays is not consistent enough to be used outside of particular experimental conditions. Ultimately the diversity of measurements means that we are obliged to accept the judgment of experimentalists as to what are, or are not, T-cell epitopes. Although somewhat subjective, it does bring the immunologist’s intimate knowledge of particular assays to bear on this difficult equation. The only general criterion for separating one epitope from another is the property of “immunodominance”: the principal epitope, with the greatest response, is said to be immunodominant and other measurable responses, as opposed to inactive peptides, are labeled subdominant. Interestingly, though this distinction is often alluded to, it is not one which is drawn sufficiently widely, or sufficiently consistently, by experimental immunologists to be useful in the present context.

In the present study, we investigate the structural differences between T cell epitopes and non-epitopes in the middle region of the peptide molecule, which is thought to contact the TCR. The focus of our study is a set of epitopes known to bind to human MHC class I molecules comprising the HLA-A3 supertype (HLA-A*0301, HLA-A*0302, HLA-A*1101, and HLA-A*3301).[23] A set of non-epitopes was generated by a multistep algorithm from the same epitope source proteins. As the study examined the middle part of the binding peptides, only positions 4−8 were included in the analysis. CoMSIA fields were generated for each binder to describe its steric, electrostatic, hydrophobic, and hydrogen-bond donor and acceptor properties.[24] A discriminant analysis performed via SIMCA[25] was applied to CoMSIA fields in order to model the structural differences between epitopes and non-epitopes associated with the peptide region directly involved in the interaction with TCRs.

Results

A set of 38 nonapeptide T cell epitopes belonging to 25 proteins was collected from AntiJen,[26] SYFPEITHI,[27] and the HIV database.[28] These epitopes bind to four HLA-A3 supertype alleles (HLA-A*0301, HLA-A*0302, HLA-A*1101, and HLA-A*3301). Only proteins consisting of less than 1000 amino acids were considered in the study. The source proteins were processed by an algorithm, called EpiJen, which mimics the antigen processing pathway of the cell. EpiJen, as described elsewhere,[29] is based on quantitative matrices, created by the additive method,[30] which are applied as a succession of filters. Briefly, each protein is presented as a set of overlapping peptides that are processed successively through three models: proteasome cleavage, TAP binding, and MHC binding. At each step, peptides are eliminated according to predefined thresholds. In the present study, the thresholds were defined as follows: 0.1 for proteasome cleavage, 5.0 for TAP binding, and 6.3 for binding to HLA-A*0301, HLA-A*0302, HLA-A*1101, and HLA-A*3301. After the last step, a small set of good MHC binders remain. In 85% of the cases, the known epitopes are

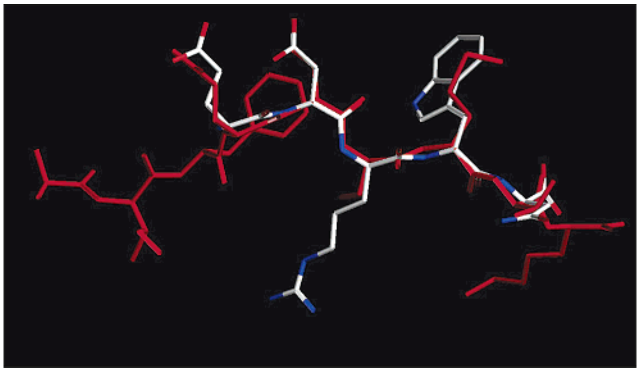

Figure 2. Alignment between the X-ray structure of the HIV RT epitope AIFQSSMTK bound to HLA-A*1101[18] and the HIV T cell epitope KLTEDRWNK bound to HLA-A*0301, shown from p4 to p8.

among the top 5% of this final set. The epitopes used in the present study corresponded to this final set of binders. This formed the class of epitopes. The remaining binders were considered as non-epitopes. As the non-epitopes were significantly more numerous than the epitopes, their number was reduced further by selecting only binders with affinities higher than those of the epitopes. The final set of non-binders consists of 183 peptides; they formed the non-epitope class. The epitopes, non-epitopes, and source proteins used in the present study are provided as Supporting Information.

The X-ray structure of the HIV RT epitope AIFQSSMTK bound to HLA-A*1101[18] was used as a starting conformation. To minimize the conformational noise, only peptides of the same length (9 amino acids) were selected for the study. The peptide structures were built with the BIOPOLYMER option in SYBYL.[31] The side chains followed the side-chain conformations of the X-ray structure. The built peptides underwent full geometry optimization with the standard Tripos molecular mechanics force field (Powell method,[32] no electrostatics, and 0.05 kcal/(mol·Å) energy gradient convergence criterion). The peptide backbone was fixed in the X-ray conformation by use of the option "Aggregates" in the Minimize Energy menu. The aggregate consists of the α-carbon atoms, the carbonyl carbon and oxygen atoms, and the amide nitrogen and hydrogen atoms. After MM optimization, the final conformations did not differ significantly from the starting ones. As the middle part of the peptide (positions 4−8) does not make significant contacts with the HLA molecule, the protein environment was not necessary for effective energy minimization. The partial atomic charges were computed by the AM1 semiempirical method[33] available in MOPAC V6, as implemented in SYBYL. Single-point calculations were performed.

The alignment of all peptides was based on the corresponding backbone atoms (the same as in the aggregate) in the conformation derived from X-ray data (Figure 2). As the study was focused on the middle region of the binding peptides, positions 1, 2, 3, and 9 were eliminated and a grid box was defined to extend 4 Å beyond the aligned pentapeptides. Five CoMSIA fields were calculated to account for the steric, electrostatic, hydrophobic, and hydrogen-bond donor and acceptor properties of the peptides. The fields underwent SIMCA discriminant analysis and the generated models were used to elucidate the structural differences between epitopes and non-epitopes in the middle region of the binding peptides. The models were internally cross-validated in five groups.

Table 1. Statistics of CoMSIA−SIMCA Models

model	PC[a]	TP[b]	TN[c]	FN[d]	FP[e]	sensitivity,[f] %	specificity,[g] %
steric field	3	26	122	12	61	68	67
electrostatic field	3	25	129	13	54	66	70
hydrophobic field	4	29	134	9	49	76	73
H-bond donor field	1	25	92	13	91	66	50
H-bond acceptor field	2	22	119	16	64	58	65
all fields	3	31	134	7	49	82	73

[a] PC, optimum number of principal components giving the highest sensitivity. [b] TP, true positives: correctly predicted epitopes. [c] TN, true negatives: correctly predicted non-epitopes. [d] FN, false negatives: incorrectly predicted epitopes. [e] FP, false positives: incorrectly predicted non-epitopes. [f] Sensitivity = true positives/(true positives + false negatives). [g] Specificity = true negatives/(true negatives + false positives).

Both single-field and combination models were developed in the study. Representative statistics for some of these are shown in Table 1. Among the single-field models, the hydrophobic-field model performed best with 76% *sensitivity* [defined as true positives/(true positives + false negatives)] and 73% *specificity* [defined as true negatives/(true negatives + false positives)], where epitopes were defined as positives and non-epitopes as negatives. The steric-field model followed with 68% sensitivity and 67% specificity. Electrostatic field and hydrogen-bond donor field models had lower sensitivity (66%), although the former model had moderate specificity (70%). In terms of sensitivity, the hydrogen-bond acceptor field model performed poorly (58%), while in terms of specificity, the hydrogen-bond donor field model was worst (50%). Two- and three-field combinations did not improve the results (data not shown). Only the all-field combination gave significant increase in the sensitivity (82%) by comparison with the single hydrophobic-field model. The best performed all-field model revealed the complexity of the TCR−peptide interaction. It indicated the importance of van der Waals, electrostatic, hydrophobic, and hydrogen-bond interactions for effective T cell recognition. This model was used to analyze each aspect of this complex interaction. The preferences at positions 4−8 (p4−p8) are summarized in Table 2. The loadings of CoMSIA fields onto principal component 1 (PC1), derived by SIMCA for the epitope class, are shown in different colors in Figure 3. The HIV T cell epitope KLTEDRWNK, which binds to HLA-A*0301, is shown within the fields.

Steric Bulk. Sterically favorable areas are shown in green in Figure 3, upper left, whereas disfavored regions are shown in yellow. For epitope activity, bulky substituents are favored at p6−p8 and disfavored at p5. In the non-epitope plot (data not shown), the preferred bulky area is along the main backbone, whereas the spatial regions around side chains show no favorable interactions.

Electron Density. Areas where the electron density is favored for T cell epitope recognition are colored red (Figure 3, upper right), and the disfavored electron density is shown in blue. Electron density is required at p6−p8 for T cell epitope recognition and disfavored at p5. In the non-epitope map, the areas of preferred and non-preferred electron density are close to the backbone, which indicates their importance for MHC binding (data not shown).

Local Hydrophobicity. Favored areas are shown in yellow, while the disfavored areas are shown in white (Figure 3, lower left). Favored areas are situated distantly at p5 and close to p6 and p7. Hydrophobic substitutents are disfavored distantly at p4 and p6−p8. For non-epitope class, the preferred hydrophobic areas are at p4 and p8 and close to the backbone at p4 and p5 (data not shown).

Hydrogen-Bond Donor and Acceptor Properties. Cyan areas depict hydrogen-bond donor preferred positions, and the purple areas show the positions where hydrogen-bond acceptor amino acids are favored (Figure 3, lower right). Amino acids

Table 2. Physicochemical Properties Preferred for Interaction with TCR

property	position 4	position 5	position 6	position 7	position 8
steric bulk		disfavored (small side chains preferred)	favored (bulky side chains preferred)	favored (bulky side chains preferred)	favored (bulky side chains preferred)
electron density		disfavored (aliphatic side chains preferred)	favored (aromatic or polar side chains preferred)	favored (aromatic or polar side chains preferred)	favored (aromatic or polar side chains preferred)
hydrophobicity	disfavored (polar side chains preferred)	favored (nonpolar side chains preferred)	disfavored (polar side chains preferred)	disfavored (polar side chains preferred)	disfavored (polar side chains preferred)
hydrogen bond	acceptor		donor	donor/acceptor	donor
suitable amino acids	Asp, Glu, Met	Ala, Leu, Ile, Pro, Val	Arg, Cys, His, Lys	Asn, Gln, His, Tyr	Arg, Cys, His, Lys

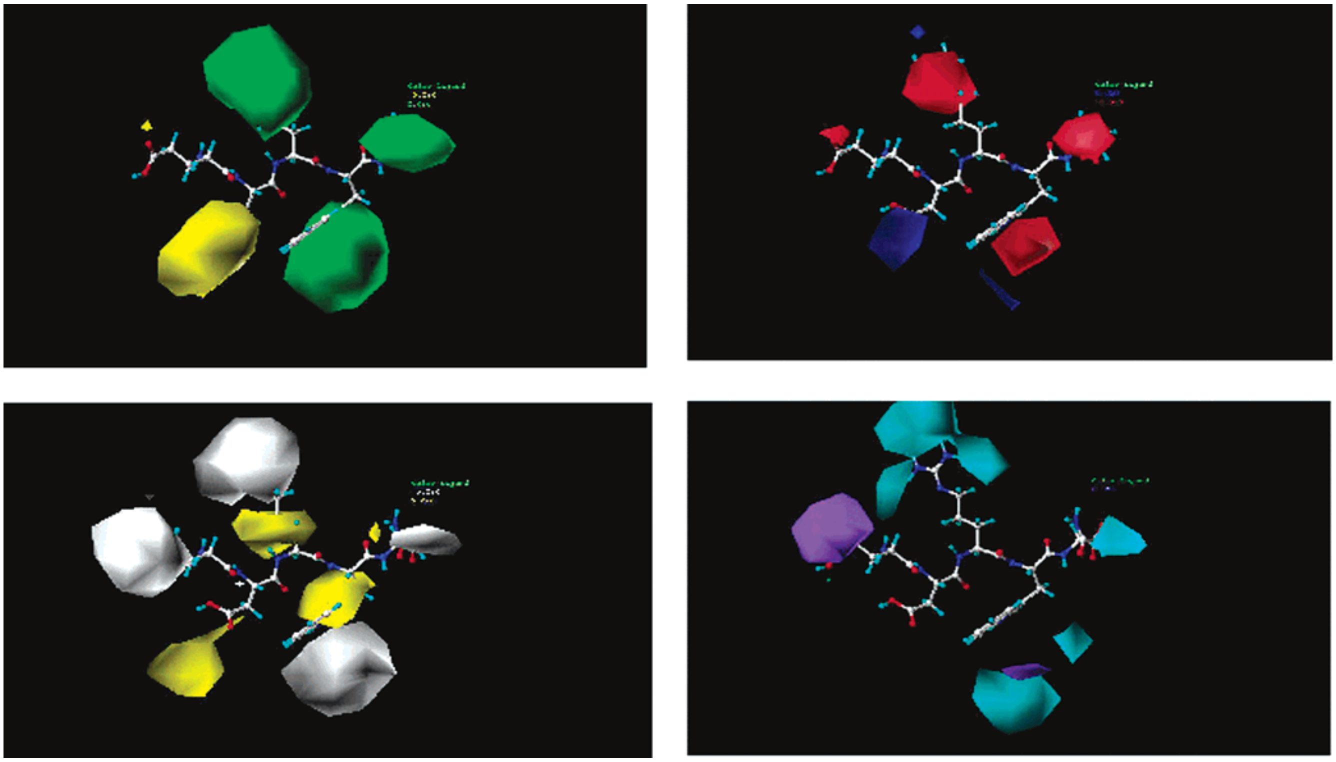

Figure 3. CoMSIA−SIMCA loading maps for PC1 of the epitope class. Positions 4−8 of the epitope KLTEDRWNK are shown inside the fields. Upper left: steric field. Contour levels are +0.02 green (steric bulk favored) and −0.02 yellow (steric bulk disfavored). Upper right: electrostatic field. Contour levels are −0.05 red (electron density favored) and +0.05 blue (electron density disfavored). Lower left: hydrophobic field. Contour levels are +0.02 yellow (hydrophobicity favored) and −0.02 white (hydrophobicity disfavored). Lower right: hydrogen bond. Contour levels are +0.05 cyan (donor favored) and +0.02 purple (acceptor favored).

at p6−p8, which could take part in hydrogen-bond formation as donors, contribute positively to TCR interaction. Hydrogen-bond acceptors contribute positively at p4 and p7. In the non-epitope map (data not shown), positively contributing donors and acceptors are located close to the backbone, indicating their significance for MHC binding.

Discriminating Power of the All-Field Model. The discriminating power is a ratio between the sum of squared residuals when fitted to a false class and the sum of squared residuals when fitted to the true class.[34] Thus, the larger the value of the discriminating power, the better the column is at differentiating between the classes. The discriminating power of the all-field model was examined for each field at the 90% contribution contour level. In the steric and electrostatic maps, the discriminating areas appeared at p5 (Figure 4, upper left and right); in hydrophobic and hydrogen-bond donor maps at p5 and p7 (Figure 4, lower left and right); and in the hydrogen-bond acceptor field map at p4 and p7 and close to NH at p6 (Figure 4, lower right).

Discussion

The models derived in this study aim to distinguish between T cell epitopes and non-epitopes among peptides that bind well to MHCs. Whereas the N- and C- termini are responsible for securing peptides within the MHC binding site, the middle region is known to bulge out of the cleft and thus interact with TCRs. This middle section of binding peptides, from p4 to p8, was examined here to identify the main physicochemical properties responsible for steric, electrostatic, hydrophobic, and hydrogen-bond interactions with the TCR. The good performance of the all-field model is indicative of the complexity of this key interaction. The model predicts 82% of epitopes and 73% of non-epitopes.

The main anchor residues for this supertype are hydrophobic (Leu, Ile, Val, Met) or hydroxyl-containing (Ser, Thr) residues at p2 and positively charged amino acids (Arg, Lys) at the C-terminal. A previously performed CoMSIA study on peptides binding to the HLA-A3 supertype[23] revealed the preferred properties at each position.[35] In the present study, the preferences for MHC binding and T cell recognition in the middle part of the peptides have been compared for each position. Despite the great diversity among TCRs and their ligands, some general conclusions could be drawn for the preferred properties of the exposed positions.

Position 4. Hydrophilic substituents with hydrogen-bond acceptor properties are favored at this position for the interaction

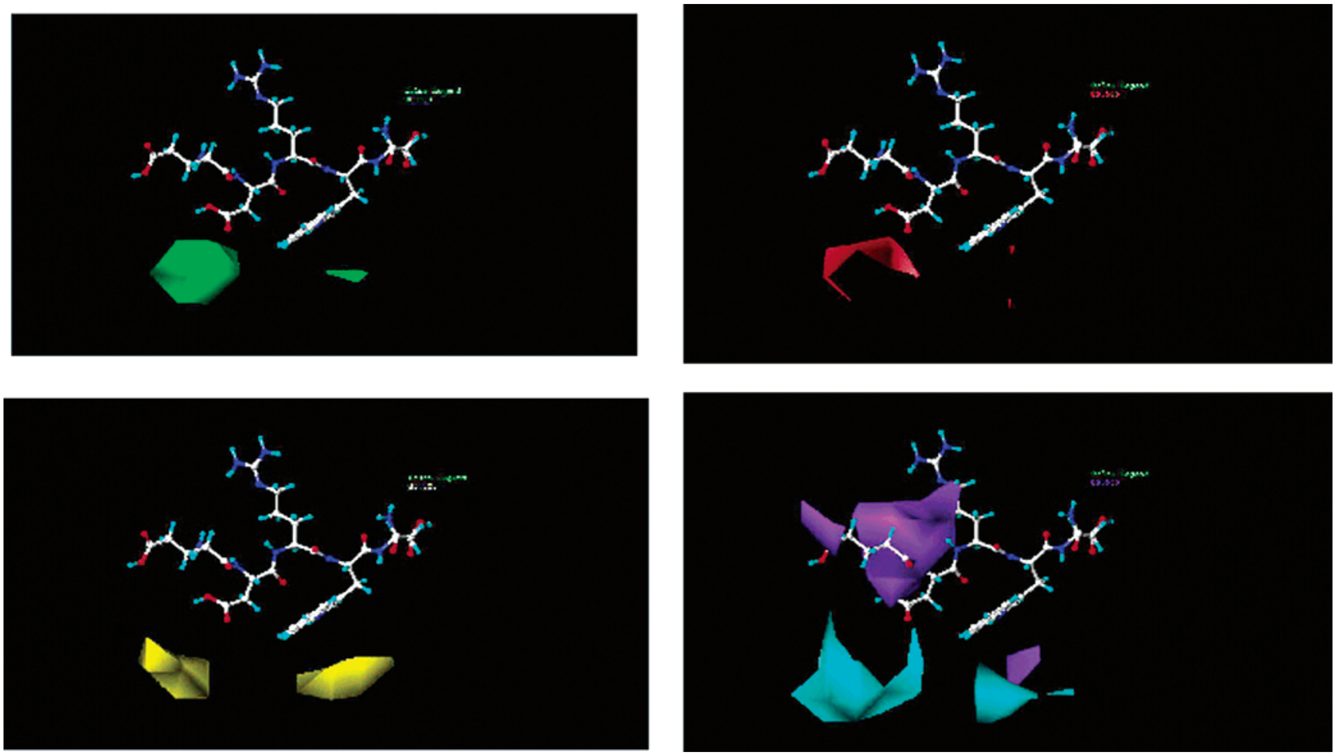

Figure 4. CoMSIA−SIMCA discriminating power maps. Positions 4−8 of the epitope KLTEDRWNK are shown inside the fields. Upper left, steric field; upper right, electrostatic field; lower left, hydrophobic field; lower right, hydrogen-bond donor (cyan) and acceptor (purple) fields.

with TCR. The last property prioritizes Asp, Glu, and Met for this position. Amino acids with mixed hydrogen-bond donor/acceptor properties such as Asn, Gln, His, Ser, Thr, and Tyr could be suitable here. However, in a previous study, hydrogen-bond donor ability was found to be a key property at this position for binding to HLA-A3 supertype molecules.[35] This is in a good agreement with the hydrogen-bond discriminating plot (Figure 4, lower right), where p4 was indicated as one of the three positions that are important for differentiating between epitopes and non-epitopes.

X-ray data has shown that Lys at p4 can interact with CDR3 α-chain residues Ser^{93}, Gly^{99}, Phe^{100}, and Ala^{101}, making van der Waals contacts with the amino acids side chains and also making hydrogen bonds with the backbone carbonyl oxygens.[13] Met at p4, together with Trp at p5, form a double side-chain “peg” about which the TCR CDR loops residues cluster, making intermolecular contacts, stabilized by hydrophobic and ring-stacking interactions.[36]

Position 5. Hydrophobic aliphatic small-sized amino acids, like Ala, Pro, Val, Ile, and Leu, are preferred at this position for TCR recognition, whereas bulky side chains with negative electrostatic potential were favored for binding to MHC proteins in the A3 supertype.[35] Thus, p5 was found to be of great importance for epitope/non-epitope discrimination. However, Garboczi et al.[8] have shown that bulky Tyr at p5 is bound in a deep pocket at the center of the TCR where the CDR3 loops converge, forming a hydrogen bond with Ser^{31} from CDR1 α chain.[14]

Position 6. Preferences here are for bulky hydrophilic amino acids with high electron density and hydrogen-bond donor properties. This corresponds to strict donors such as Arg, Lys, His, and Cys. Amino acids with mixed donor/acceptor abilities are also acceptable at this position. Similar properties are required for binding to A3 supertype molecules.[35]

The X-ray structure of HLA-A*1101 complexed with HIV-1 peptide AIFQSSMTK[18] clearly shows that p6 is an important additional anchor for binding to MHC molecule. The nitrogen atom of p6 Ser makes a hydrogen bond with a bound water molecule[18] and it is important for discriminating between epitopes and non-epitopes, as is evident from Figure 4.

Position 7. The favored amino acids for this position are bulky, hydrophilic, with high electron density and mixed hydrogen-bond donor/acceptor abilities. Asn, Gln, His, and Tyr correspond to these requirements, but amino acids with strict donor or acceptor properties are also appropriate. Along with p4 and p5, p7 is of great importance for discriminating between epitopes and non-epitopes. Accordingly, in a previous study it was found that hydrophobic amino acids at p7 will increase the A3 supertype binding affinity.[35]

Garcia et al.[13] have shown that Tyr at p6 in an octamer peptide EQYKFYSV (corresponding to p7 in a nonamer) complexed with 2C TCR and mouse MHC class I H-2K makes a hydrogen bond to Asn^{30} from the CDR1 β chain. Likewise, p7 Tyr, which resides within a bulged and exposed region of the EBV peptide, is clearly the pivotal residue for LC13 TCR recognition.[17] It protrudes deeply within the TCR pocket formed by CDR1α, CDR3α, and CDR3β chains.

Position 8. The requirements for T cell recognition at p8 are the same as at p6: bulky hydrophilic amino acids with high electron density and hydrogen-bond donor properties. Similar properties—steric bulk and negative electrostatic potential—were found to be favored for HLA-A3 supertype binding affinity.

Ser at p7 in an octamer (corresponding to p8 in a nonamer) forms a hydrogen bond with Asn^{30} from CDR1 β chain.[13] Similarly, Tyr at p8 forms a hydrogen bond to Asp^{30} of CDR1β from A6 TCR complexed with a Tax peptide and HLA-A2.[8,14]

The discriminating power analysis of the all-field model indicates that p4, p5, and p7 are the most important for distinguishing between epitopes and non-epitopes. The comparison with the requirements for binding to alleles belonging to the HLA-A3 superfamily supports this. Similar properties for MHC binding and TCR interaction are favored at p6 and p8, whereas the opposite preferences are shown at p4, p5, and p7.

Overall, the preferred amino acids in the middle section of the epitopes point to a common requirement for polar amino acids with high electron density and ability to form hydrogen bonds. These "aggressive" amino acids protrude out of the binding cleft and take an active role in the process of T cell recognition. At the same time, more than 30% of all amino acids occupying p4−p8 are small-sized, "passive" residues such as Gly, Pro, Ala, and Val. The role of these amino acids could be 2-fold. They may allow the backbone carbonyl oxygen and amide nitrogen to be reached by the extended side chains of the TCR. This is the case for the interaction of the Vβ17Vα 10.2 TCR with the influenza virus matrix protein epitope MP-(58−62) (a so-called "plain vanilla peptide"[37]) bound to HLA-A2.[16] Typically, the TCR fits over an exposed side chain of the bound peptide. In this case, however, a TCR residue (Arg^{98} from the CDR3 β loop) inserts into a notch in the peptide−MHC surface and forms a dense network of hydrogen bonds, including some to the peptide main-chain carbonyl atoms. This mode of TCR−peptide interaction is strongly facilitated by Gly residues present in the middle part of the peptide. Alternatively, the small residues flanking the long "aggressive" amino acids enable them to protrude deeply into the centrally located TCR pocket.[17] This combination of "aggressive" and "passive" amino acids in the middle part of epitopes could be thought to form a TCR binding motif.

As a three-dimensional quantitative structure−activity relationship (3D QSAR) study, the present investigation is very sensitive to any conformational ambiguity. Because of their innate flexibility, the modeling of protein-bound peptides can be a more complicated task than the modeling of small molecules. Our experience[38−40] indicates that the conformational noise could be minimized by taking into account several considerations. First, the modeled peptides should all be the same length (nine amino acids in this case). Data from numerous X-ray crystal structures shows beyond doubt that bound nine-amino-acid peptides tend toward strong isomorphism without significant differences in backbone structure. It is only as peptide length increases beyond nine that we see real differences in conformation. Second, the backbone and the side chains should be built following the conformation of the template X-ray structure. Third, MM optimization is allowed only for the side chains, while the backbone is kept fixed. Finally, the alignment is based on the corresponding backbone atoms. Given these conditions, any remaining conformational ambiguity will not preclude the development of a relevant and robust 3D QSAR analysis.

Conclusion

The SIMCA classification of 221 epitopes and non-epitopes as good binders to HLA-A3 supertype proteins, based on CoMSIA fields, defined the preferred physicochemical properties at each of the five exposed positions from the middle region of the binding peptide. Comparison with the preferred properties for MHC binding at the same positions revealed similarities at p6 and p8 and dissimilarities at p4, p5, and p7. The combination of polar amino acids with high electron density and hydrogen-bond-making abilities and small-sized residues in the middle of the peptide may form a TCR binding motif. The present study was focused on the last step of the long multistep process of antigen processing and recognition. Because the T cell repertoire is enormously diverse, the T cell recognition process has been considered as a bioinformatic task of unprecedented complexity with simply too many unknown variables to be tractable. As this process is of extreme importance for epitope-based vaccine design, the present study has attempted to look inside the "black box" that is T cell epitope recognition.

Experimental Section

Epitopes. A set of 38 nonameric peptide epitopes, which bind to four A3 supertype alleles—HLA-A*0301, HLA-A*0302, HLA-A*1101, and HLA-A*3301—was collected from AntiJen,[26] SIFPEITHI,[27] and the HIV database.[28]

CoMSIA. The set of 221 nonamer peptides (38 epitopes and 183 non-epitopes) was imported into the molecular modeling software SYBYL 6.9.[31] CoMSIA[24] fields were autofilled with the MSS option. Five physicochemical properties (steric, electrostatic, hydrophobic, and hydrogen-bond donor and acceptor) were evaluated, by use of a common probe atom with 1 Å radius, charge +1, hydrophobicity +1, and hydrogen-bond donor and acceptor properties +1. The value of the attenuation factor α was set to 0.3. Column filtering was set to 2.0 kcal/mol.

SIMCA. The table consisting of 221 rows, five CoMSIA fields, and one categorical column (epitope/non-epitope) was used by the SIMCA algorithm, as implemented within SYBYL 6.9. SIMCA (soft independent modeling of class analogy) is a technique for producing a mathematical description of the differences between rows of different categories, based on columns of explanatory properties.[25] SIMCA constructs a set of principal components (PC) for each category, relying on internally performed cross-validation in five groups to determine which components distinguish between the categories. The method is based on projecting each row into each category's factor space, reprojecting back into the original space of the explanatory columns, and measuring the difference. The category that yields the smallest differences is the category to which the compound is predicted to belong.

The correctly predicted epitopes and non-epitopes were defined as true positives (TP) and true negatives (TN), respectively, while the incorrectly predicted ones yielded false negatives (FN) and false positives (FP), respectively. On the basis of these predictions, two assessments of the models were defined as follows: *sensitivity* [true positives/(true positives + false negatives)] and *specificity* [true negatives/(true negatives + false positives)]. As the non-epitopes generated from one protein were significantly higher than the epitopes, the parameter *accuracy* [(true positives + true negatives)/total] could be misleading and has not been used in the study. For example, if 98% of the peptides in one source protein are non-epitopes, a model that simply predicts everything as non-epitope will not be very useful but will nonetheless have an overall accuracy of 98%.

To select the optimum number of components, different variants for each studied model were calculated while the number of PCs was varied from 1 to 7. An optimum number of principal components (PCs) was selected where the addition of a component decreases or does not change the number of true positives.

The loadings of different CoMSIA fields on PC1, as derived according to the best-performing all-field CoMSIA−SIMCA model, were visualized in maps contoured by actual values. A set of five maps (steric, electrostatic, hydrophobic, and hydrogen-bond donor and acceptor) were generated for both epitope and non-epitope classes. Additionally, a set of five maps visualizing the discriminating power of the all-field model contoured by a 90% contribution was generated. The HIV T cell epitope KLTEDRWNK, which binds to the HLA-A*0301 allele, is shown inside the fields of the maps.

Acknowledgment. The Jenner Institute thanks its former sponsors: GlaxoSmithKline, Medical Research Council, Biotechnology and Biological Sciences Research Council, and the U.K. Department of Health. The present study was supported in part by grants from the Royal Society, U.K., and the Ministry of Education, Bulgaria.

Supporting Information Available: A table showing the epitopes, non-epitopes, and source proteins used in the study. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Litman, G. W.; Anderson, M. K.; Rast, J. P. Evolution of Antigen Binding Receptors. *Annu. Rev. Immunol*. **1999**, *17*, 109−147.

(2) Shastri, N.; Schwab, S.; Serwold, T. Producing nature's gene-chips: The generation of peptides for display by MHC class I molecules. *Annu. Rev. Immunol*. **2002**, *20*, 463−463.

(3) de la Hera, A.; Muller, U.; Olsson, C.; Isaaz, S.; Tunnacliffe, A. Structure of the T Cell Antigen Receptor (TCR): Two CD3 Epsilon Subunits in a Functional TCR/CD3 Complex. *J. Exp. Med.* **1991**, *173*, 7−17.

(4) Gouaillard, C.; Hucheng-Champagne, A.; Arnaud, J.; Chen, C. L.; Rubin, B. Evolution of T cell Receptor (TCR) Alpha Beta Heterodimer Assembly with the CD3 Complex. *Eur. J. Immunol.* **2001**, *31*, 3798−805.

(5) Arden, B.; Clark, S. P.; Kabelitz, D.; Mak, T. W. Human T-cell Receptor Variable Gene Segment Families. *Immunogenetics* **1995**, *42*, 455−500.

(6) Krangel, M. S.; McMurry, M. T.; Hernandez-Munain, C.; Zhong, X. P.; Carabana, J. Accessibility Control of T Cell Receptor Gene Rearrangement in Developing Thymocytes. The TCR Alpha/Delta Locus. *Immunol. Res.* **2000**, *22*, 127−35.

(7) Davis, M. M.; Krogsgaard, M.; Huppa, J. B.; Sumen, C.; Purbhoo, M. A.; Irvine, D. J.; Wu, L. C.; Ehrlich, L. Dynamics of Cell Surface Molecules During T Cell Recognition. *Annu. Rev. Biochem.* **2003**, *72*, 717−742.

(8) Garboczi, D. N.; Ghosh, P.; Utz, U.; Fan, Q. R.; Biddison, W. E.; Wiley, D. C. Structure of the Complex Between Human T-cell Receptor, Viral Peptide and HLA-A2. *Nature* **1996**, *384*, 134−41.

(9) Garcia, K. C.; Degano, M.; Stanfield, R. L.; Brunmark, A.; Jackson, M. R.; Peterson, P. A.; Teyton, L.; Wilson, I. A. An Alphabeta T Cell Receptor Structure at 2.5 Å and Its Orientation in the TCR−MHC Complex. *Science* **1996**, *274*, 209−19.

(10) Engel, I.; Hedrick, S. M. Site-Directed Mutations in the VDJ Junctional Region of a T Cell Receptor Beta Chain Cause Changes in Antigenic Peptide Recognition. *Cell* **1988**, *54*, 473−84.

(11) Batalia, M. A.; Collins, E. J. Peptide Binding by Class I and Class II MHC Molecules. *Biopolymers* **1997**, *43*, 281−302.

(12) Ding, Y.-H.; Smith, K. J.; Garboczi, D. N.; Utz, U.; Biddison, W. E.; Wiley, D. C. 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.

(13) Garcia, K. C.; Degano, M.; Pease, L. R.; Huang, M.; Peterson, P. A.; Teyton, L.; Wilson, I. A. Structural Basis of Plasticity in T Cell Receptor Recognition of a Self-Peptide-MHC Antigen. *Science* **1998**, *279*, 1166−1172.

(14) Ding, Y.-H.; Baker, B. M.; Garboczi, D. N.; Biddison, W. E.; Wiley, D. C. Four A6-TCR/Peptide/HLA-A2 Structures that Genarate Very Different T Cell Signals Are Nearly Identical. *Immunity* **1999**, *11*, 45−56.

(15) Krogsgaard, M.; Prado, N.; Adams, E. J.; He, X.-I.; Chow, D.-C.; Wilson, D. B.; Garsia, K. C.; Davis, M. M. Evidence that Structural Rearrangements and/or Flexibility during TCR Binding Can Contribute to T Cell Activation. *Mol. Cell* **2003**, *12*, 1367−1378.

(16) Stewart-Jones, G. B. E.; McMichael, A. J.; Bell, J. I.; Stuart, D. I.; Jones, E. Y. A Structural Basis for Immunodominant Human T Cell Receptor Recognition. *Nat. Immunol*. **2003**, *4*, 657−663.

(17) Kjer-Nielsen, L.; Clements, C. S.; Purcell, A. W.; Brooks, A. G.; Whisstock, J. C.; Burrows, S. R.; McCluskey, J.; Rossjohn, J. A Structural Basis for the Selection of Dominant Alphabeta T Cell Receptors in Antiviral Immunity. *Immunity* **2003**, *18*, 53−64.

(18) Li, L.; Bouvier, M. Structures of HLA-A*1101 Complexed with Immunodominant Nonamer and Decamer HIV-1 Epitopes Clearly Reveal the Presence of a Middle, Secondary Anchor Residue. *J. Immunol*. **2004**, *172*, 6175−6184.

(19) Garcia, K. C.; Degano, M.; Speir, J. A.; Wilson, I. A. Emerging Principles for T Cell Receptor Recognition of Antigen in Cellular Immunity. *Rev. Immunogen*. **1999**, *1*, 75−90.

(20) Slansky, J. E.; Rattis, F. M.; Boyd, L. F.; Fahmy, T.; Jaffee, E. M.; Schneck, J. P.; Margulies, D. H.; Pardoll, D. M. Enhanced Antigen-Specific Antitumor Immunity with Altered Peptide Ligands that Stabilize the MHC−Peptide−TCR Complex. *Immunity* **2000**, *13*, 529−538.

(21) Tangri, S.; Ishioka, G. Y.; Huang, X.; Sidney, J.; Southwood, S.; Fikes, J.; Sette, A. Structural Features of Peptide Analogues of Human Histocompatibility Leukocyte Antigen Class I Epitopes that Are More Potent and Immunogenic than Wild-Type Peptide. *J. Exp. Med.* **2001**, *194*, 833−846.

(22) Tourdot, S.; Oukka, M.; Manuguerra, J. C.; Magafa, V.; Vergnon, I.; Riche, N.; Bruley-Rosset, M.; Cordopatis, P.; Kosmatopoulos, K. Chimeric Peptides: A New Approach to Enhancing the Immunogenicity of Peptides with Low MHC Class I Affinity: Application in Antiviral Vaccination. *J. Immunol*. **1997**, *159*, 2391−2398.

(23) Doytchinova, I. A.; Guan, P.; Flower, D. R. Identifying Human MHC Supertypes Using Bioinformatic Methods. *J. Immunol*. **2004**, *172*, 4314−4323.

(24) Klebe, G.; Abraham, U.; Mietzner, T. Molecular Similarity Indices in a Comparative Analysis (CoMSIA) of Drug Molecules To Correlate and Predict Their Biological Activity. *J. Med. Chem*. **1994**, *37*, 4130−4146.

(25) Wold, S.; Sjöström, M. Method for Analyzing Chemical Data in Terms of Similarity and Analogy. In *Chemometrics: Theory and Application*; Kowalski, B. R., Ed.; ACS Symposium Series; American Chemical Society: Washington, DC, 1977; Vol. 52, pp 243−282.

(26) Toseland, C. P.; Taylor, D. J.; McSparron, H.; Hemsley, S. L.; Blythe, M. J.; Paine, K.; Doytchinova, I. A.; Guan, P.; Hattotuwagama, C. K.; Flower, D. R. AntiJen: A Quantitative Immunology Database Integrating Functional, Thermodynamic, Kinetic, Biophysical, and Cellular Data. *Immunome Res.* **2005**, *1*, 4. http://www.immunome-research.com/content/1/1/4.

(27) Rammensee, H.; Bachmann, J.; Emmerich, N. P.; Bachor, O. A.; Stevanovic, S. SYFPEITHI: database for MHC ligands and peptide motifs. *Immunogenetics* **1999**, *50*, 213−219.

(28) HIV Molecular Immunology Database (last update June 8, 2005) of the Los Alamos National Laboratory (www.hiv.lanl.gov).

(29) Doytchinova, I. A.; Guan, P.; Flower, D. R. EpiJen: A Server for Multistep T Cell Epitope Prediction. *BMC Bioinf.* **2006**, *7*, in press.

(30) Doytchinova, I. A.; Blythe, M. J.; Flower, D. R. Additive Method for the Prediction of Protein−Peptide Binding Affinity. Application to the MHC Class I Molecule HLA-A*0201. *J. Proteome Res.* **2002**, *1*, 263−272.

(31) *SYBYL 6.9*; Tripos Inc.: 1699 Hanley Rd., St. Louis, MO 63144.

(32) Powell, M. J. D. Restart Procedures for the Conjugate Gradient Method. *Math. Prog.* **1977**, *12*, 241−254.

(33) Dewar, M. J. S.; Zoebisch, E. G.; Healy, E. F.; Stewart, J. J. P. AM1: A New General Purpose Quantum Mechanical Molecular Model. *J. Am. Chem. Soc.* **1985**, *107*, 3902−3909.

(34) *SYBYL Ligand-Based Design Manual*, QSAR. Version 6.9; Tripos Inc.: 1699 Hanley Rd., St. Louis, MO 63144.

(35) Guan, P.; Doytchinova, I. A.; Flower, D. R. A Comparative Molecular Similarity Indices (CoMSIA) Study of Peptide Binding to the HLA-A3 Superfamily. *Bioorg. Med. Chem*. **2003**, *11*, 2307−2311.

(36) Chen, J.-L.; Stewart-Jones, G.; Bossi, G.; Lissin, N. M.; Wooldridge, L.; Choi, E. M. L.; Held, G.; Dunbar, P. R.; Esnouf, R. M.; Sami, M.; Boulter, J. M.; Rizkallah, P.; Renner, C.; Sewell, A.; van der Merwe, P. A.; Jakobsen, B. K.; Griffiths, G.; Jones, E. Y.; Cerundolo, V. Structural and Kinetic Basis for Heightened Immunogenicity of T Cell Vaccines. *J. Exp. Med.* **2005**, *201*, 1243−1255.

(37) Davis, M. M. The Problem of Plain Vanilla Peptides. *Nat. Immunol*. **2003**, *4*, 649−650.

(38) Doytchinova, I.; Flower, D. R. Towards the Quantitative Prediction of T-cell Epitopes: CoMFA and CoMSIA Studies of Peptides with Affinity to Class I MHC Molecule HLA-A*0201. *J. Med. Chem.* **2001**, *44*, 3572−3581.

(39) Doytchinova I. A.; Flower, D. R. Physicochemical Explanation of Peptide Binding to HLA-A*0201 Major Histocompatibility Complex: A Three-Dimensional Quantitative Structure−Activity Relationship Study. *Proteins: Struct., Funct., Genet.* **2002**, *48*, 505−518.

(40) Doytchinova I. A.; Flower, D. R. A Comparative Molecular Similarity Index Analysis (CoMSIA) Study Identifies an HLA-A2 Binding Supermotif. *J. Comput.-Aided Mol. Des*. **2002**, *16*, 535−544.

JM050876M