Quantitative Structure–Activity Relationship of Peptides Binding to the Class II Major Histocompatibility Complex Molecule Aq Associated with Autoimmune Arthritis

Lotta Holm,† Kristina Frech,† Balik Dzhambazov,‡ Rikard Holmdahl,¶ Jan Kihlberg,†,§ and Anna Linusson*,†

Department of Chemistry, Umeå University, SE-901 87 Umeå, Sweden, Section of Medical Inflammation Research, Lund University, Sölvegatan 19, 111 BMC, SE-221 84 Lund, Sweden, and AstraZeneca R&D Mölndal, SE-431 83 Mölndal, Sweden

Received October 17, 2006

Introduction

The immune system’s ability to distinguish self from nonself is a crucial feature of the body’s defenses against foreign antigens. This ability is linked to a ternary complex in which T cells play a key role in the activation or nonactivation of the immune system. Foreign and autoantigens are degraded into peptides by antigen presenting cells and subsequently presented by class II major histocompatibility complex (MHCII) molecules for recognition by T-cell receptors. This recognition constitutes an important step in a series of events that forms the immune response to foreign antigens. The self-peptide MHC complexes should not lead to any response, and when this tolerance is disrupted, an immune response toward endogenous tissue(s) is elicited and an autoimmune inflammatory disease, such as rheumatoid arthritis (RA), may develop.

RA is one of the most common autoimmune inflammatory diseases and is characterized by chronic inflammation of peripheral cartilaginous joints. The symptoms include swelling, stiffness, and pain in the joints and subsequent erosion of underlying bones, which might further lead to deformity and malfunction of the joints.1 A large number of therapeutic agents against RA are commercially available, but there is no cure and most treatments do not inhibit the progression of the disease satisfactorily.2 The major difficulties hindering the development of effective treatments for RA seem to be the complexity of the system, lack of knowledge of the mechanisms involved, and the need for reliable disease models.3,4 RA has been linked to the MHC class II molecules DR1 and DR4,5–7 and the activated immune systems of severely affected RA patients include both autoreactive T cells8–10 and antibodies11–13 directed against type II collagen (CII), which is the most abundant protein in cartilage.

Injection of CII in rats14 or mice15 provokes the development of collagen-induced arthritis (CIA) with symptoms and histopathology similar to those of RA. CIA is the most commonly used animal model for RA, and susceptibility to murine CIA is linked to the mouse MHC class II molecule Aq.16,17 By use of synthetic peptides, the minimal CII peptide epitope required for both binding to the Aq molecule and inducing a T-cell response has been determined to be the octapeptide ranging from amino acid 260 to amino acid 267 (CII260–267, Figure 1).18 The isoleucine at position 260 and phenylalanine 263 have been found to be essential anchor residues for binding to the Aq molecule, according to an alanine scan.19,20 Furthermore, it has been shown that the T-cell response is often linked to specific recognition of a carbohydrate moiety resulting from post-translational modification of lysine at CII264 (β-d-galactopyranosyl modified hydroxylysine, i.e., a GalHyl moiety).21–24 In some cases, the T-cell recognition is not linked solely to the sugar moiety but also to the glutamic acid located at position CII266.25 Highly interesting results have shown that vaccination of mice with the mouse and rat galactosylated CII256–273 peptides can provide protection against development of CIA.26 In particular, vaccination by the rat glycopeptide complexed with the Aq molecule significantly retarded progression of the disease and reduced its severity in mice with ongoing chronic relapsing arthritis.27 These findings indicate that analyses with variants of the CII glycopeptide could be very useful for exploring the tricomponent Aq/glycopeptide/T-cell receptor interactions and could facilitate the development of potential drug or vaccine candidates for treating RA.

While the carbohydrate specificity has been extensively studied, less is known about the interactions between the Aq molecule and the peptide ligand in the ternary complex. Besides the proposed anchoring positions (Ile260 and Phe263), the introduction of a methylene ether amide bond mimetic between amino acids Ile260 and Ala261 in the minimal T-cell epitope CII260–267 resulted in a substantial drop (20-fold) in peptide affinity for Aq.25 Moreover, Jane-wit et al. have postulated a common autoimmune-motif (KXXS) for peptides binding to...
correlated to the biological response (e.g., binding strength).\textsuperscript{52} It has been shown that QSAR models based on properties of amino acids at each position (local models) were superior over models based on properties of whole peptides (global models) for peptides binding to the MHC class I HLA-A*0201.\textsuperscript{53} In the study reported here we designed a peptide library based on the minimal CII peptide binding to the mouse MHC class II molecule A\textsuperscript{a}, using the SMD approach. The design was performed in amino acid space, and peptides were selected using D-optimal design.\textsuperscript{50} The chosen peptides were synthesized on solid phase, and their binding strength to A\textsuperscript{a} was tested in a cell-based competition assay. Multivariate methods were used to evaluate the binding data, and a QSAR model was subsequently established. An external peptide test set was used to further verify the binding model.

**Results and Discussion**

**Peptide Scaffold.** The minimal CII peptide epitope found to bind to the A\textsuperscript{a} molecule while retaining the ability to induce T-cell responses, i.e., the octapeptide CII260–267,\textsuperscript{18} was decided to be the most suitable scaffold for this binding study (Figure 2). Although the even shorter peptide CII260–266 has been shown to bind (although very weakly),\textsuperscript{18} it was believed that modifications were likely to cause severe losses of A\textsuperscript{a} binding.

Position 264 harbors the GalHyl moiety important for T-cell recognition, and it has been shown that the moiety has minimal influence on peptide binding to the A\textsuperscript{a} molecule.\textsuperscript{15} This position was therefore kept constant and nonglycosylated in order to simplify the synthesis of the peptide library. However, it should be stressed that GalHyl plays an extremely important role in T-cell responses and thus warrants further study regarding autoimmune response. The known critical binding points, i.e., the anchor residues Ile260 and Phe263, were also left unchanged to prevent complete loss of binding among members of the peptide library. Consequently, five positions (Ala261, Gly262, Gly265, Glu266, Gln267) in CII260–267 were chosen to be systematically varied with SMD, referred to hereafter as positions 1–5 (Figure 2).

In a first attempt to synthesize the library of selected peptides considerable problems with aqueous solubility were experienced. To circumvent these problems, lysine was introduced at the N and C terminals of the peptide (N-Lys and C-Lys, respectively) and tested for two commonly used CII-specific T-cell hybridoma cell lines (HCQ.4 and HDB2, Table 1). The peptide with lysine at the C terminal was recognized by both hybridomas. Consequently, the peptide scaffold was elongated with a C-terminal lysine to improve the solubility of the peptide library.\textsuperscript{54} We also decided to N\textsuperscript{α}-acetylate the resulting peptide scaffold and to introduce a C-terminal amide to prevent loss of affinity to A\textsuperscript{a} due to the presence of charged functional groups in the peptide backbone (Figure 2).

**Building Block Selection Using Multivariate Characterization.** A representative set of peptides to synthesize and test was selected using SMD in building block space using amino acid descriptors. Several characterizations of amino acids have been presented previously,\textsuperscript{38,55–59} and QSAR models of class I and class II MHC peptides described in the literature have been based on additive descriptors.\textsuperscript{47,48,50–52} Several property descriptors,\textsuperscript{62} and the z-scale descriptors\textsuperscript{53,62} defined by Wold and Sandberg.\textsuperscript{58} The z-scale descriptors are derived on the basis of experimentally determined characterization of the amino acids followed by a subsequent principle component analysis (PCA)\textsuperscript{63} resulting in three to five principle properties (i.e., the z
Figure 2. Peptide scaffold used to study molecular property preferences for peptide binding to the A\textsuperscript{b} MHC class II molecule based on the minimal CII glycopeptide with retained T-cell response (CII260–267). Five positions in the scaffold were varied by SMD, and the effects were studied. Major anchor residues (positions 260 and 263) were left unchanged. The nonmodified lysine residue was used in position 264, and an additional lysine was added to the C-terminal for solubility purposes. The arrows indicate the amino acid sets selected for incorporation at each position.

Table 1. Amino Acid Sequences of N- and C-Terminally Extended Peptides and Resulting T-Cell Responses

<table>
<thead>
<tr>
<th>peptide</th>
<th>259</th>
<th>260</th>
<th>261</th>
<th>262</th>
<th>263</th>
<th>264</th>
<th>265</th>
<th>266</th>
<th>267</th>
<th>268</th>
<th>hybridoma&lt;sup&gt;a&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>N-Lys</td>
<td>Ac-Lys</td>
<td>Ile</td>
<td>Ala</td>
<td>Gly</td>
<td>Phe</td>
<td>Lys</td>
<td>Gly</td>
<td>Glu</td>
<td>Glu</td>
<td>Glu-NH₂</td>
<td>+</td>
</tr>
<tr>
<td>C-Lys</td>
<td>Ac-Ile</td>
<td>Ala</td>
<td>Gly</td>
<td>Phe</td>
<td>Lys</td>
<td>Gly</td>
<td>Glu</td>
<td>Gln</td>
<td>Lys-NH₂</td>
<td>+</td>
<td></td>
</tr>
<tr>
<td>CII260–267&lt;sup&gt;b&lt;/sup&gt;</td>
<td>Ac-Ile</td>
<td>Ala</td>
<td>Gly</td>
<td>Phe</td>
<td>Lys</td>
<td>Gly</td>
<td>Glu</td>
<td>Glu</td>
<td>Glu-NH₂</td>
<td>+</td>
<td></td>
</tr>
</tbody>
</table>

<sup>a</sup> + refers to an equally strong response to the corresponding longer peptide CII259–275, and − refers to no response at the tested concentrations.

<sup>b</sup> Included as reference peptide.

A comparative model of the A\textsuperscript{b} molecule<sup>20</sup> indicates that there is limited room for bulky side chains at positions 1–3 (corresponding to positions 261, 262, and 265), where small residues (glycine and alanine) occur in CII. This implies that the presence of large amino acids at any of these positions could severely reduce the peptides’ ability to bind to the A\textsuperscript{b} molecule. Therefore, the principal property design space was reduced for these three positions to cover only amino acids of small to moderate sizes. Five amino acids representing this space were selected: Thr, Ser, Val, Met, and Ala (Figures 2 and 3a,c). For positions 4 and 5 (corresponding to CII266 and CII267) there was, according to the A\textsuperscript{b} model, no need for restrictions in the chemical space. Therefore, the total property space was used to cover as large an area as possible, and six amino acids were chosen: Arg, Asn, Asp, Tyr, Ser, and Val (Figure 2). These amino acids represented a maximum spread in characteristics such as aliphatic/aromatic, hydrogen bond donors and acceptors, size, and charged/neutral, as described by the three score values in the principal property space (Figure 3a,c).

Library Selection Using D-Optimal Design. Virtual combinations yielding all possible variations of the selected amino acids for the five positions in the scaffold resulted in a peptide library consisting of 4500 peptides (5\(^3\) × 6\(^2\)). D-optimal design<sup>66</sup> was applied to this library to reduce it further. Each amino acid at the five altered positions was represented by the three values of the scaled principal properties, t1–t3, which when combined into peptides resulting in 15 values representing a 15-dimensional principal property space. A library of 22 peptides with the most homogeneous distribution possible of selected amino acids at each position was chosen out of the suggested D-optimal designs while maximizing the volume spanned in the principal property space. Two peptides with molecular properties closest to the center of the principal property space were added as center points. The procedure used to select these 24 peptides ensured that the chosen library had high chemical diversity and information content (Table 2).

Synthesis, Biological Testing, and Preanalysis of Data. The peptides of the library (Table 2) were synthesized on solid phase as N- and C-terminal amides. Cleavage from solid support and purification by reverse-phase HPLC rendered the pure products, hereafter referred to as 1–22 and CP1–CP2. The binding
The strength of the peptides to the MHCII Aq molecule was studied in a competitive assay in which the peptides were evaluated for their ability to prevent the binding of a biotinylated CLIP reference peptide to Aq-transfected cells. The test peptides were incubated at seven different concentrations (750, 250, 83, 28, 9, 3, and 1 μM; designated concentrations 1–7, respectively) in duplicate, and the experiments were repeated twice. Structure–activity relationships were evaluated using the % inhibition at different concentrations of the peptides as the biological response.

The partial least-square projections to latent structures (PLS) method has been the most commonly used regression method for developing QSAR models of peptides binding to class I and class II MHC molecules, although a recent publication has shown promising QSAR results with support vector machine regression. Here, the 15 principal property values used for the library selection, i.e., the combination of the PCA score values t1–t3 for all the peptides (the X matrix), were correlated to the biological response (the Y matrix) by PLS regression.

The analysis of the X and Y data revealed that the responses at concentrations 2–4 (i.e., 250, 83, and 28 μM) contained the most information, since most peptides did not bind at the lower concentrations 5–7 and that solubility and/or toxicity problems were detected for some peptides at the highest concentration (750 μM, concentration 1). The response block (concentrations 2–4) had a linear relationship to the logarithm of the corresponding concentration, and nontransformed responses gave similar results as the logit-transformation (cf. pIC50). The response values were scaled to unit variance, which gave results similar to those for the Pareto scaling. Single, duplicate, and quadruplet samples showed good reproducibility both in and between experiments except for run 2 of peptides 1–9 where the results deviated from the others and hence were excluded from further analysis. The most robust models were obtained when averages of duplicate samples were used as the biological responses. Therefore, in the final QSAR model the biological responses of the peptides were represented by the average percentage inhibition of the pairs of duplicate samples at concentrations 2–4 (referred to as Y2, Y3, and Y4) except that peptides 1–9 were each represented by a single duplicate sample average (Table 2). Peptides 7 and 9 and the two center peptides (CP1 and CP2) yielded deviating data because of solubility problems and/or toxicity to the cells in the binding assay and were not included in the modeling.

QSAR models of peptides binding to class I and class II MHC molecules using PLS regression have been reported including only linear terms but also linear and cross-terms that account for interactions between side chains at relative positions 1–2 and 1–3. A comparison of the two models based on the data in this study revealed identical interpretation of the linear terms when considering the PLS weight vectors (regres...
A q. These validation peptides were tested on a separate, later test set (Figure 5 and Table 3). These six peptides, referred to as medium- and low-ranked binders) were selected as an external improved binders. Two peptides from each class (high-, medium-, and low-ranked binders) were selected as an external test set. The affinity to A q was predicted out of the variation in the biological response (\(R^2\) = 0.70; \(R^2\) = 0.47; \(R^2\) = 0.45). The dModX plot and the normal probability plot of the residuals did not reveal any outliers.

The QSAR model was validated and tested for its predictability using an external test set. The affinity to A\(^8\) was predicted by the QSAR model for the 3 200 000 possible virtual peptides. Three clusters, each consisting of 5000 peptides predicted to have high, average, and low A\(^8\) affinity, can be seen in the PLS discriminant analysis (PLS-DA) plot (Figure 5). Most library peptides belonged to the class of peptides predicted to have high or average affinity (peptides V1–V6), while the peptides predicted to have low affinity (V5 and V6) displayed similar

Peptides Binding to Molecule A\(^8\) Journal of Medicinal Chemistry, 2007, Vol. 50, No. 9

<table>
<thead>
<tr>
<th>Peptide</th>
<th>Pos1</th>
<th>Pos2</th>
<th>Pos3</th>
<th>Pos4</th>
<th>Pos5</th>
<th>Inhibition (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Ac-Ile</td>
<td>Met</td>
<td>Met</td>
<td>Phe</td>
<td>Lys</td>
<td>Ser</td>
</tr>
<tr>
<td>2</td>
<td>Ac-Ile</td>
<td>Met</td>
<td>Met</td>
<td>Phe</td>
<td>Lys</td>
<td>Val</td>
</tr>
<tr>
<td>3</td>
<td>Ac-Ile</td>
<td>Met</td>
<td>Met</td>
<td>Phe</td>
<td>Lys</td>
<td>Thr</td>
</tr>
<tr>
<td>4</td>
<td>Ac-Ile</td>
<td>Met</td>
<td>Thr</td>
<td>Phe</td>
<td>Lys</td>
<td>Met</td>
</tr>
<tr>
<td>5</td>
<td>Ac-Ile</td>
<td>Met</td>
<td>Thr</td>
<td>Phe</td>
<td>Lys</td>
<td>Ala</td>
</tr>
<tr>
<td>6</td>
<td>Ac-Ile</td>
<td>Met</td>
<td>Ser</td>
<td>Phe</td>
<td>Lys</td>
<td>Thr</td>
</tr>
<tr>
<td>7</td>
<td>Ac-Ile</td>
<td>Val</td>
<td>Val</td>
<td>Phe</td>
<td>Lys</td>
<td>Ser</td>
</tr>
<tr>
<td>8</td>
<td>Ac-Ile</td>
<td>Val</td>
<td>Ala</td>
<td>Phe</td>
<td>Lys</td>
<td>Thr</td>
</tr>
<tr>
<td>9</td>
<td>Ac-Ile</td>
<td>Val</td>
<td>Thr</td>
<td>Phe</td>
<td>Lys</td>
<td>Val</td>
</tr>
<tr>
<td>10</td>
<td>Ac-Ile</td>
<td>Val</td>
<td>Ser</td>
<td>Phe</td>
<td>Lys</td>
<td>Ala</td>
</tr>
<tr>
<td>11</td>
<td>Ac-Ile</td>
<td>Ala</td>
<td>Met</td>
<td>Phe</td>
<td>Lys</td>
<td>Met</td>
</tr>
<tr>
<td>12</td>
<td>Ac-Ile</td>
<td>Ala</td>
<td>Met</td>
<td>Phe</td>
<td>Lys</td>
<td>Ser</td>
</tr>
<tr>
<td>13</td>
<td>Ac-Ile</td>
<td>Ala</td>
<td>Val</td>
<td>Phe</td>
<td>Lys</td>
<td>Met</td>
</tr>
<tr>
<td>14</td>
<td>Ac-Ile</td>
<td>Ala</td>
<td>Ala</td>
<td>Phe</td>
<td>Lys</td>
<td>Ala</td>
</tr>
<tr>
<td>15</td>
<td>Ac-Ile</td>
<td>Ala</td>
<td>Thr</td>
<td>Phe</td>
<td>Lys</td>
<td>Thr</td>
</tr>
<tr>
<td>16</td>
<td>Ac-Ile</td>
<td>Thr</td>
<td>Thr</td>
<td>Phe</td>
<td>Lys</td>
<td>Met</td>
</tr>
<tr>
<td>17</td>
<td>Ac-Ile</td>
<td>Thr</td>
<td>Met</td>
<td>Phe</td>
<td>Lys</td>
<td>Val</td>
</tr>
<tr>
<td>18</td>
<td>Ac-Ile</td>
<td>Thr</td>
<td>Val</td>
<td>Phe</td>
<td>Lys</td>
<td>Val</td>
</tr>
<tr>
<td>19</td>
<td>Ac-Ile</td>
<td>Thr</td>
<td>Thr</td>
<td>Phe</td>
<td>Lys</td>
<td>Ser</td>
</tr>
<tr>
<td>20</td>
<td>Ac-Ile</td>
<td>Thr</td>
<td>Ser</td>
<td>Phe</td>
<td>Lys</td>
<td>Met</td>
</tr>
<tr>
<td>21</td>
<td>Ac-Ile</td>
<td>Ser</td>
<td>Ala</td>
<td>Phe</td>
<td>Lys</td>
<td>Thr</td>
</tr>
<tr>
<td>22</td>
<td>Ac-Ile</td>
<td>Ser</td>
<td>Thr</td>
<td>Phe</td>
<td>Lys</td>
<td>Ala</td>
</tr>
<tr>
<td>CP1</td>
<td>Ac-Ile</td>
<td>Val</td>
<td>Val</td>
<td>Phe</td>
<td>Lys</td>
<td>Val</td>
</tr>
<tr>
<td>CP2</td>
<td>Ac-Ile</td>
<td>Val</td>
<td>Ala</td>
<td>Phe</td>
<td>Lys</td>
<td>Ala</td>
</tr>
</tbody>
</table>

\(^*\) Peptides 1–22 were selected by D-optimal design while CP1 and CP2 represented center points. \(^\%\) % inhibition in which the average was obtained from duplicate samples in two separate runs for concentrations Y2, Y3, and Y4 corresponding to 250, 83, and 28 \(\mu\)M, respectively. Single dash (–) indicates that data from run 2 for peptides 1–9 were excluded because they were deviant. Double dashes (---) indicates deviating data due to solubility and/or toxicological problems.

Figure 4. Calculated versus experimentally determined inhibition values for one of the three responses (Y4) used in the multi-Y PLS regression. The QSAR model was based on 15 principal property values (t1–t3 at positions 1–5) for 20 peptides, and the three biological responses were represented as % inhibition at three different peptide concentrations (Y2, Y3, Y4).

The QSAR model. The final linear PLS model showed a good correlation between the experimental and calculated inhibition values (Figure 4). This two-component model explained 66% of the variation in the biological response (\(R^2\) = 0.59; \(R^2\) = 0.70; \(R^2\) = 0.70) with a cross-validated \(Q^2\) of 43% (\(Q^2\) = 0.37; \(Q^2\) = 0.47; \(Q^2\) = 0.45). The dModX plot and the normal probability plot of the residuals did not reveal any outliers.

The QSAR model was validated and tested for its predictability using an external test set. The affinity to A\(^9\) was predicted by the QSAR model for the 3 200 000 possible virtual peptides. Three clusters, each consisting of 5000 peptides predicted to have high, average, and low A\(^9\) affinity, can be seen in the PLS discriminant analysis (PLS-DA) plot (Figure 5). Most library peptides belonged to the class of peptides predicted to have average affinities, while several were predicted to have low affinity and only one was found in the high-affinity region (peptide 20), indicating that there is scope for designing improved binders. Two peptides from each class (high-, medium-, and low-ranked binders) were selected as an external test set (Figure 5 and Table 3). These six peptides, referred to as V1–V6, were synthesized and tested for binding strength to A\(^9\). These validation peptides were tested on a separate, later occasion with a different setup of biological material and modified protocol compared to the peptides used to build the

Figure 5. Calculated versus experimentally determined inhibition (%)
binding preferences and had lower binding affinity to A^a than V1. The results of the external validation confirmed the predictability of the QSAR model obtained.

A comparison of our model with the QSAR models developed for other mouse MHC class II molecules (I-A^b, I-A^k, I-A^d, I-A^s, I-E^d, and I-E^k) revealed that the R^2 and Q^2 are lower for our model (0.66 and 0.43 compared to 0.99 and 0.83). However, R^2 and Q^2 are internal validation criteria and external test sets are a superior alternative for evaluating the model quality. The predictive power of the models for the six class II molecules presented by Hattotuwagama et al. showed great variation but had similar high R^2 and Q^2 values of the models, where the A^a model had a very high predictivity while that of E^k was very poor. Our model successfully managed to predict the binding strength of six new peptides despite the relatively low internal statistical terms.

The interpretation of the influence of the different amino acid properties at the different positions is presented in detail below. The regression coefficients of the different peptide positions showed the same pattern for all three investigated concentrations, and the regression weight values (w × c plot) can be seen in Figure 6. A design guide of preferred directions in the score plots of the 20 amino acids is provided in the Supporting Information.

1. Interpretation of the QSAR Model. Positions 4 and 5 had the strongest influence on the peptide binding to A^a according to the regression model, as shown by their dominating regression weight values (Figure 6). The variables describing size and flexibility were the main contributors for position 4, as both t1 and t3 were strongly negatively correlated with the response. For a good binder the amino acid at position 4 should preferably be large and flexible, e.g., arginine. The original molecular descriptors responsible for these features were revealed, based on their PCA loading values, to be volume, surface area, and the Kier flexibility indexes 1K, 2K, and PHI (Figure 3). In addition, the positively correlated regression weight for t2 indicated that hydrogen bond donors/acceptors could be preferred over more lipophilic amino acids. Although the side chain of this residue has been shown to be very important for T-cell stimulations, our model shows that it also influences binding to the A^a. By comparison of the relative position in the principal property space of glutamic acid found in rat CII with that of aspartic acid found in mouse CII, it can be seen that glutamic acid is preferred to aspartic acid in position 4. This correlates well to experimental results suggesting that rat CII binds more strongly to the A^a molecule than mouse CII and correlates with the finding that exchanging glutamic acid by aspartic acid at position 266 in CII results in a more than 10-fold reduction in binding strength to A^a.

In position 5, just as in position 4, large and to some extent flexible amino acids were strongly preferred because the t1 and t3 variables were negatively correlated with the response (Figure 3 and 6). Interpretation of the PCA loading vectors p1 and p3 showed that the dominating molecular descriptors, as for position 4, were volume, surface area, and the Kier flexibility indexes. The lipophilicity, as explained by t2, had a moderate impact on binding, and the regression weight value indicates a binding preference for hydrophobic and aromatic amino acids, e.g., phenylalanine and methionine.

In position 3, the regression weight values of t2 and t3 variables were of moderate sizes and negatively correlated with the response. The dominating original molecular descriptors were LOGPC, AM1SM2, and the Kier indexes, indicating preferences for hydrophobic and flexible amino acids (Figure 3). However, the size of the side chains does not appear to be an important feature for binding within the limited studied chemical space (cf. size of Pos3: t1 in Figure 6). The most suitable coded amino acids for this position belong to any of the two adjacent clusters containing isoleucine, leucine, valine, methionine, or maybe even the aromatic phenylalanine.

The PLS regression weight values for the side chain properties at positions 1 and 2 described by t1 – t3 were low to medium, indicating that the variations made at these positions had no major effect on the binding of the peptide to the A^a molecule. All amino acids within the investigated area in the principal property space were tolerated. However, the moderate positive weight values of t1 and t3 indicate that small rigid amino acids were preferred (Figure 3), while the polarity of the amino acids did not seem to matter (cf. Pos1: t2 and Pos2: t2 in Figure 6). The preferred amino acids at these positions correspond well with those naturally occurring in the CII peptide, i.e., Ala261 and Gly262.

2. Summarizing the Model. Preferred amino acids and characteristics of the studied positions are summarized in Figure 7. A molecular property binding motif was discernible and easily transferred to sequence binding motifs. Besides the previously identified CII260 and CII263 anchor positions, positions 4 and 5 (corresponding to Glu266 and Gln267 in CII) were the most important for peptide binding to the A^a molecule. Residues at these two positions should preferably be large and flexible. In addition, the model indicates that residues at position 4 should contain hydrogen bond donors and acceptors to promote binding, while residues at position 5 should be hydrophobic. At positions 1 and 2 (corresponding to Ala261 and Gly262 in CII) small and rigid amino acids are favored, even though moderately sized amino acids were tolerated. Finally, position 3 (corresponding to Gly265 in CII) should harbor aliphatic, hydrophobic residues for optimal binding. These proposed preferences were supported by the strong A^a binding of peptide V1 from the external test set. For V1 the amino acids at all positions fulfilled the suggested requirements: Thr at position 1, Ala at position 2, Ile at position 3, Arg at position 4, and Thr at position 5. It should be noted that this designed peptide has no amino acids in common at any of the five varied amino acid positions with the octamer originating from CII. On the other hand, peptide V5, which was predicted to be a poor binder, had unfavorable amino acids at all of the investigated positions, effectively preventing its binding to the A^a molecule. The A^a binding motif
Peptides Binding to Molecule A


Table 3. Amino Acid Sequences and A<sup>a</sup> Binding Data for the Six Peptides Used for External Validation of the QSAR Model

<table>
<thead>
<tr>
<th>Peptide</th>
<th>Pos1</th>
<th>Pos2</th>
<th>Pos3</th>
<th>Pos4</th>
<th>Pos5</th>
<th>A&lt;sup&gt;a&lt;/sup&gt; Binding&lt;sup&gt;a&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>V1</td>
<td>Ac-Ile</td>
<td>Thr</td>
<td>Ala</td>
<td>Phe</td>
<td>Lys</td>
<td>Arg Trp Lys-NH&lt;sub&gt;2&lt;/sub&gt; ± + 1</td>
</tr>
<tr>
<td>V2</td>
<td>Ac-Ile</td>
<td>Ile</td>
<td>Gly</td>
<td>Phe</td>
<td>Lys</td>
<td>Ala Arg Met Lys-NH&lt;sub&gt;2&lt;/sub&gt; ± + m 2</td>
</tr>
<tr>
<td>V3</td>
<td>Ac-Ile</td>
<td>Ile</td>
<td>Gin</td>
<td>Phe</td>
<td>Lys</td>
<td>Thr Leu Arg Lys-NH&lt;sub&gt;2&lt;/sub&gt; m + m 2</td>
</tr>
<tr>
<td>V4</td>
<td>Ac-Ile</td>
<td>Thr</td>
<td>Asn</td>
<td>Phe</td>
<td>Lys</td>
<td>Gly Ser Arg Lys-NH&lt;sub&gt;2&lt;/sub&gt; m + m 4</td>
</tr>
<tr>
<td>V5</td>
<td>Ac-Ile</td>
<td>Tyr</td>
<td>Met</td>
<td>Phe</td>
<td>Glu</td>
<td>Pro Gly Lys-NH&lt;sub&gt;2&lt;/sub&gt; + + 5</td>
</tr>
<tr>
<td>V6</td>
<td>Ac-Ile</td>
<td>Gln</td>
<td>Met</td>
<td>Phe</td>
<td>Lys</td>
<td>Arg Ala Gly Lys-NH&lt;sub&gt;2&lt;/sub&gt; + + 6</td>
</tr>
</tbody>
</table>

* Definitions: +, strong binders; m, medium binders; −, poor binders. * Predicted class membership by the PLS-DA model. * Class membership as determined from binding data. * Relative ranking of peptides based on binding data.

Figure 6. PLS weight values (w × c) for the QSAR model based on 15 principal property values (t<sub>i</sub> to t<sub>3</sub> at positions 1–5) and three biological responses represented as % inhibition at three different peptide concentrations (Y2, Y3, Y4).

The longer immunodominant part of CII, i.e., CII<sub>256</sub>–<sub>270</sub>, binds to both of the MHC class II molecules A<sup>a</sup> (mouse) and DR4 (human).<sup>20.71.72</sup> It has been proposed that the DR4-binding motif of CII<sub>256</sub>–<sub>270</sub> is shifted by three amino acids compared to that of A<sup>a</sup>, resulting in Phe263 and Glu266 being at the anchor positions instead of Ile260 and Phe263.<sup>71.72</sup> A comparison of the suggested DR4 peptide-binding motifs with our results regarding the A<sup>a</sup>-binding motif supports this hypothesis, but too few amino acids have been altered within the DR4 epitope to warrant further comparative conclusions.

No experimentally determined 3D structure of the A<sup>a</sup> molecule is available, but indications about the properties of the A<sup>a</sup> binding site can be obtained from the postulated preferred molecular properties of the peptides (Figure 7). In addition to earlier indications of two large hydrophobic binding pockets (P1 and P4),<sup>20</sup> the preferences for small, rigid amino acids in the P2 and P3 binding pockets suggest that they are of limited size. One could also speculate that the presence of large amino acids in P2 and/or P3 could prevent the anchor residues from reaching their binding pockets. In P6 the QSAR model implies that there is a hydrophobic area in the A<sup>a</sup> molecule. The P7 pocket appears to be rather large with some polar surface area, while the P8 pocket is indicated to be a large, hydrophobic binding area. When these preferences were compared with the 3D comparative model of the A<sup>a</sup> molecule,<sup>20</sup> the models matched remarkably well.

In the preferred design directions, as indicated by the QSAR model, there are few amino acids with required characteristics among the coded amino acids. More suitable non-natural amino acids could possibly be identified to further enhance the binding to the A<sup>a</sup> molecule, which could also increase the metabolic stability of the peptides. In addition, the information presented in Figure 7 could be used to develop novel peptide mimetics based on the proposed preferred molecular properties. New modified peptides with variations in physicochemical properties and binding strength would be highly valuable in the development of effective immunization procedures for use in future vaccination studies.

Conclusions

A peptide library was designed, synthesized, and evaluated for binding to the mouse MHC class II molecule A<sup>a</sup>. The SMD approach made it possible to select a chemically diverse and informative library of 22 peptides and two center points out of 3 200 000 possible peptide combinations. This highly reduced set of peptides, together with inhibition data from a cell-based competitive assay at three different concentrations of the peptides, resulted in a high-quality QSAR model based on PLS modeling that was successfully validated with an external test set of six peptides.

A molecular property binding motif for peptides binding to the mouse MHC class II molecule A<sup>a</sup> was established on the basis of interpretation of the QSAR model of the five varied positions. The C-terminal positions of the peptide scaffold (corresponding to CII<sub>266</sub> and CII<sub>267</sub>, respectively) appeared to have the strongest influence on the A<sup>a</sup>/peptide interaction, while the positions corresponding to CII<sub>261</sub> and CII<sub>262</sub> are the ones that have the least influence according to the QSAR model. In addition, the model provided indications of the characteristics of the binding site of the A<sup>a</sup> molecule and the findings in the present investigation correlated well with a comparative 3D model of the protein.

The QSAR model provides novel information and insight regarding the A<sup>a</sup> molecule/peptide component of the ternary A<sup>a</sup>, glycopeptide, and T-cell receptor complex. This information
could facilitate attempts to develop new treatments of autoimmune diseases such as RA.

Experimental Section

Theoretical Characterization of Amino Acids. The structures of the 20 naturally occurring amino acids were generated using Spartan software and subsequently characterized by 11 molecular descriptors including properties related to size (surface area, molecular weight, volume), electronic features (dipole, HOMO, LUMO, maximum charge, minimum charge, partial charge of c-at cl), and lipophilicity (log P, SM2) extracted from semiempirical AM1 calculations within Spartan software. In addition to these 3D-based descriptors, 17 descriptors including functional group indexes (path 1 Keir shape index-1K, path 2 Keir shape index 2K), Kier flexibility index (PHI), and saturation ratios (SatGrade, NAT/SKC) were computed using the Dragon software. A table of the descriptors is given in the Supporting Information. These molecular descriptors were compressed by PCA using SIMCA software. The number of significant principal components was decided using their eigenvalues, a Scree plot, and chemical interpretation of the loadings for the corresponding components.

Statistical Molecular Design and Data Analytical Methods. 1. Selection of Peptides. Each amino acid at the altered positions was represented by the corresponding values of the scaled principal components. The score values were then scaled to unit variance for each dimension to avoid bias in the weighting of the varied positions or molecular properties due to differences in variance. D-optimal design was performed using MODDE software to generate 15 libraries, each with 22 peptides. The selection by D-optimal design maximized the volume spanned in the principal property space through maximization of the determinant (Det) of the X’X matrix. The final library to synthesize was chosen from the 15 generated libraries based on G efficiency, Log(Det of X’X), Norm.log(Det of XX), condition number, and the most homogeneous distribution of the selected amino acids at each position. The 50 peptides with the minimal Euclidean distance to the calculated center point in the principal property space were calculated, and two peptides were chosen as center points based on their synthetic feasibility and added to the library of 22 peptides.

2. Projections to Latent Structures by Means of Partial Least-Squares. The structure descriptor matrix (X), based on the combination of PCA score values representing the molecular properties of the amino acids, was related to the biological activity response matrix (Y) using the PLS regression method. The PLS method maximizes the covariance between the latent variables of the X and Y matrices (multi-Y) and correlates these latent variables through linear combinations to a regression model. Even though it is a linear method, a nonlinear relationship can be handled to some extent through transformation of Y, inclusion of nonlinear terms, and extraction of additional PLS components. The % inhibition at several different concentrations (multi-Y), scaled to unit variance, was used as the response. The use of correlated response variables as a multi-Y matrix enhanced the stability and reliability of the models because the biological data contained noise and, in some cases, missing and deviating data. The quality of the model was investigated by estimating the amounts of explained variation (R²) and cross-validated predicted variation (Q²) using seven cross-validation rounds. The validation of the final model was made by using an external test set. The interpretation of the influence of the different amino acid positions was based on the first PLS component weight vector, which provides the best estimate of the variable importance when only one response or highly correlated multiresponses are used. All PLS modeling was performed using SIMCA software.

Preanalysis of Biological Data. The X and Y data were preanalyzed to determine relevant concentrations, scaling preferences, transformations, included model terms, and influence of the experimental layout to assess the reproducibility of the assays and to detect peptide outliers. Relevant concentrations as multi-Y and

<table>
<thead>
<tr>
<th>Peptide pos.</th>
<th>Pos1 P2</th>
<th>Pos2 P3</th>
<th>Pos3 P6</th>
<th>Pos4 P7</th>
<th>Pos5 P8</th>
</tr>
</thead>
<tbody>
<tr>
<td>A pos.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Preferred binders</td>
<td>Molecular property</td>
<td>small</td>
<td>rigid</td>
<td>small</td>
<td>rigid</td>
</tr>
<tr>
<td>Poor binders</td>
<td>Molecular property</td>
<td>large</td>
<td>flexible</td>
<td>large</td>
<td>flexible</td>
</tr>
</tbody>
</table>

**Figure 7.** Summary of the interpretation of the QSAR model illustrating the preferred and rejected amino acids and of molecular properties to promote Aq binding. Pos1 – Pos5 represent the five varied positions in the peptide scaffold, and suggested characteristics at corresponding positions in the binding site of the Aq molecule (P2, P3, P6 – P8) are also shown. Amino acids are represented by their conventional one-letter code.
model terms to include in the final PLS-QSAR model and whether the data should be logit-transformed and scaled to unit variance, pareto-scaled or nonscaled, were assessed by considering the amounts of variation explained by the resulting PLS models ($R^2$), their cross-validation values ($Q^2$), and the normal probability plots of the residuals. Three sets of model terms were investigated: only linear terms, linear and all cross terms for relative positions 1–2 and 1–3, and linear and selected cross terms (1–2 and 1–3) with PLS coefficients larger than 0.05 for all investigated concentrations. In order to test the biological effects of the entire peptide library including center points and reference peptides as duplicated samples, the experiment was divided and tested in three different experiments (experiment 1, peptides 1–9; experiment 2, peptides 10–18; experiment 3, peptides 19 CP[1–22]). The experiments were repeated (runs 1 and 2), resulting in four samples/response curves per peptide. PCA of the $Y$ variables, for all four samples (duplicated samples and duplicated runs), was used to assess the reproducibility both in and between experiments. Standard deviations were calculated, and a PLS with indicator variables was performed to identify deviating experiments. Thus, the $X$ matrix was extended by six columns with either 1 or 0 as indicator variable, representing belonging or not belonging to a certain experiment and run. Deviating data due to solubility problems and/or toxicity to cells used in the binding assay were visually detected when running the FACS analysis and/or as outliers in the PCA.

**External Validation.** A representative test set for external validation was chosen from all possible 3 200 000 in silico generated peptides, based on the coded amino acids. The virtually combined peptides were represented by their values of the scaled principal properties score vectors, and the established QSAR model equation was used to predict their binding strength ($\%$ inhibition). Three ranking algorithms were then used to sort and select three sets of peptides: the 5000 predicted to be the strongest binders, the 5000 predicted to have the closest to average binding strength, and the 5000 predicted to be the weakest binders. Separation of the three classes of predicted peptides was enhanced using PLS-DA, and selections were made from each class. The test set was synthesized and biologically evaluated for $A^\alpha$ binding.

**Solid-Phase Peptide Synthesis.** The peptides were synthesized in a manually operated reactor or a Pioneer peptide synthesis system (Applied Biosystems, The Netherlands), using standard solid-phase peptide synthesis methodology on a Tentagel-S NH$_2$ resin (Rapp Polymere, Germany) in which the linker Fmoc-2,4-dimethoxy-4′-carboxymethoxy)benzhydrylamine (Rink) was first coupled to the resin. This rendered peptides as C-terminal amides after cleavage from the resin. N$\alpha$-Fmoc amino acids carrying standard side chain protecting groups (Bachem, Switzerland and NeoSystème, France, 4 equiv), as well as the Rink linker (4 equiv), were coupled to the resin in dimethylformamide (DMF), which was predestilled and used immediately or stored for a short time over 3 Å molecular sieves. In the manually operated reactor disopropyl carbodiimide (DIC, 3.9 equiv) was used as a coupling reagent in the presence of 1-hydroxybenzotriazole (HOBT, 6 equiv). The progress of the reaction was monitored by the naked eye using bromophenol blue as an indicator. Alternatively, coupling reactions were performed in the Pioneer peptide synthesis system utilizing 0.5 M HBTU and 0.5 M Dipea as coupling reagents with UV monitoring, all according to the manufacturer’s instructions. Fmoc protective groups were removed after each coupling cycle using 20% piperidine in DMF. After completion of the synthetic sequence, the N-terminal of the peptides was acetylated by incubation with Ac$_2$O/DMF (1:2) for 1 h. The peptides were cleaved and deprotected by incubation with TFA/H$_2$O/thioanisole/ethanediol (35:2:2:1) for 3 h at 40 °C. Following repeated concentration from HOAc, the peptides were precipitated from Et$_2$O and the crude products were freeze-dried.

Purification by reversed-phase HPLC and freeze-drying gave the homogeneous compounds (N-Lys, C-Lys, 1–22, CP[1, CP[2, V1–V6]). Analytical reversed-phase HPLC was performed using a Kromasil C-8 column (250 mm x 4.6 mm, 5 μm, 100 Å), elution with a linear gradient of MeCN (0 → 100% or 0 → 80% over 60 min), balance H$_2$O, both containing 0.1% TFA, and flow rate of 1.5 mL/min. Preparative reversed-phase HPLC was performed using a larger Kromasil C-8 column (250 mm x 20 mm, 5 μm, 100 Å) with the same eluents but a flow rate of 11 mL/min. In both cases the eluate was monitored by a UV detector at 214 nm. The identity of the peptides was confirmed by MS and their purity ($\geq 95\%$) by analytical HPLC.

**$A^\alpha$ Binding Assay.** The binding of peptides to $A^\alpha$ MHC class II molecules was preformed in a competitive assay using flow cytometry analysis. Briefly, the test peptides (1–22, CP[1, CP[2, V1–V6) and a reference peptide were incubated in 96-well plates at seven different concentrations (750, 250, 83, 28, 9, 3, and 1 μM, which are concentrations 1, 2, 3, 4, 5, 6, 7, respectively) for 2.5 h at 37 °C with a fixed concentration of biotinylated CLIP peptide (5 μM) and M12Q 14–7 cells transfected with H-2A$.^\alpha$. After being washed to remove excess peptide, the cells were stained with 0.2 μL of streptavidin–phycoerythrin (SAPE), which binds to the biotinylated CLIP peptide. The phycoerythrin (PE) dye was detected by flow cytometry analysis using FACSort (Becton Dickinson, San Jose, CA) and Becton Dickinson software. The “% inhibition” for each peptide was calculated from the gated mean fluorescence by comparison with the fluorescence from the positive control (no inhibiting peptide) after subtracting the signal from the negative control (no biotinylated CLIP peptide). The experiment was performed in duplicate and was repeated once. The biological evaluation of the validation peptides was performed in a similar way in duplicate but at a separate occasion and later than for the designed peptide library. Since the reference peptide induced a weak response, the validation peptides were subjected to an additional binding study with a modified protocol.

Recombinant $A^\beta$ molecules were captured by incubation at 4 °C overnight in a 96-well microtiter plate precoated with the mAb Y3-P and blocked with PBS containing 2% low fat milk. After washing, increasing concentrations of glycopolypeptides were added and incubated for 48 h at room temperature together with a fixed concentration of biotinylated CLIP peptide (“CLIPbio”, 2.5 μM). CLIPbio-MHC class II complexes were quantified using the dissociation-enhanced lanthanide fluorimunoassay (DELFIA) kit system based on the time-resolved fluorimunoassay technique with europium-labeled streptavidin (Wallac, Turku) according to the manufacturer’s instructions. The six validation peptides (V1–V6) and a reference peptide were tested at 0.8, 4, 20, 100, and 500 μM.

The full dose response curves for the second experiment were used to classify the validation peptides and to rank them in terms of relative binding strength.

**Determination of T-Cell Hybridoma Responses.** The response of each T-cell hybridoma line, i.e., the amount of IL-2 secreted following incubation with antigen-presenting spleen cells in the various concentrations of glycopolypeptides (N-Lys and C-Lys) was determined in a standard assay using the T-cell clone CTLL. Briefly, 5 × 10$^4$ hybridoma T cells were cocultured with 5 × 10$^5$ syngeneic spleen cells and antigen in a volume of 200 μL in 96-well flat-bottom microtiter plates. After 24 h, 100 μL aliquots of the supernatants were removed and frozen to kill any transferred hybridoma T cells. IL-2 sensitive CTLL T cells (1 × 10$^5$/mL, 100 μL/well) were added to the thawed supernatant. The CTLL cultures were incubated for 24 h, after which they were pulsed with 1 μCi of $[^3H]$thymidine and incubated for an additional 15–18 h. The cells were harvested on glass fiber sheets in a Filtermate cell harvester (Packard Instruments, Meriden, CT), and the amount of radioactivity in them was determined using a Matrix 96 direct β counter (Packard). All experiments were performed in duplicate.

Acknowledgment. This work was funded by grants from the Swedish Research Council, the Göran Gustafsson Foundation for Research in Natural Sciences and Medicine, and the program “Glycoconjugates in Biological Systems” (GLIBS) sponsored by the Swedish Foundation for Strategic Research. We thank Dr. Fredrik Pettersson for the ranking algorithms.
References and Notes


Peptides Binding to Molecule A


(50) UNIX Spartan 4.0, Wavefunction, Inc. (18401 Von Karman Ave, Suite 370, Irvine, CA 92612).

(51) Dragon: Tattle Srl (Via V. Pisan1, 13-02124 Milano, Italy).


(54) Simca-P 10.5: Umetrics (Box 7960, S-907 19 Umeå, Sweden).


(56) Modde 6.0: Umetrics (Box 7960, S-907 19 Umeå, Sweden).


(59) Petterson, F. Department of Bioinformatics, Wellcome Trust Centre, Oxford, U.K.; fredrikp@well.ox.ac.uk.


JM061209B

Journal of Medicinal Chemistry, 2007, Vol. 50, No. 9 2059