Research article

Statistical deconvolution of enthalpic energetic contributions to MHC-peptide binding affinity

Matthew N Davies*†1, Channa K Hattotuwagama†1, David S Moss2, Michael GB Drew3 and Darren R Flower1

Address: 1Edward Jenner Institute for Vaccine Research, Compton, Newbury, RG20 7NN, UK, 2School of Crystallography, Birkbeck College, London WC1E 7HX, UK and 3Structural and Computational Chemistry Group, University of Reading, Reading RG6 6AH, UK

Email: Matthew N Davies* - matthew.davies@jenner.ac.uk; Channa K Hattotuwagama - channa.hattotuwagama@jenner.ac.uk; David S Moss - d.moss@mail.cryst.bbk.ac.uk; Michael GB Drew - m.g.b.drew@reading.ac.uk; Darren R Flower - darren.flower@jenner.ac.uk

* Corresponding author    †Equal contributors

Abstract

Background: MHC Class I molecules present antigenic peptides to cytotoxic T cells, which forms an integral part of the adaptive immune response. Peptides are bound within a groove formed by the MHC heavy chain. Previous approaches to MHC Class I-peptide binding prediction have largely concentrated on the peptide anchor residues located at the P2 and C-terminus positions.

Results: A large dataset comprising MHC-peptide structural complexes was created by re-modelling pre-determined x-ray crystallographic structures. Static energetic analysis, following energy minimisation, was performed on the dataset in order to characterise interactions between bound peptides and the MHC Class I molecule, partitioning the interactions within the groove into van der Waals, electrostatic and total non-bonded energy contributions.

Conclusion: The QSAR techniques of Genetic Function Approximation (GFA) and Genetic Partial Least Squares (G/PLS) algorithms were used to identify key interactions between the two molecules by comparing the calculated energy values with experimentally-determined BL50 data. Although the peptide termini binding interactions help ensure the stability of the MHC Class I-peptide complex, the central region of the peptide is also important in defining the specificity of the interaction. As thermodynamic studies indicate that peptide association and dissociation may be driven entropically, it may be necessary to incorporate entropic contributions into future calculations.

Background

Major Histocompatibility Complex (MHC) molecules play a central role in the adaptive immune system, forming a complex with foreign antigenic peptides and displaying them on the cell surface for inspection by receptors of cytotoxic and helper T cells. In consort with the T cell receptor, the repertoire of MHC-peptide complexes shapes the specificity of the T cell response. Two different types of MHC molecules – Class I and Class II – are recognized by distinct sets of T-cells: CD8+ and CD4+ respectively. MHC Class I molecules, which predominantly present antigenic material derived from the cytosol, are composed of an α heavy chain, a light chain (β2-microglobulin or β2m) and a peptide between 8–15
Amino acids in length. The peptide is bound within a groove formed by the MHC heavy chain.

Analysis of MHC Class I-peptide binding interactions together with specific peptide-MHC crystal structures reveals several common trends. A conserved cluster of tyrosine residues are present at the N terminus of the peptide while a different set of MHC Class I residues form a series of hydrogen bonds and ionic interactions with the C terminus and peptide backbone. Previous attempts to predict potential epitopes have concentrated on binding motifs. Such motifs greatly emphasize so-called 'anchor' residues at peptide position P2 and the peptide C-terminus, which occupy the B and F pocket of the MHC Class I binding groove respectively [1,2]. The B pocket is composed of eight α heavy chain residues (9, 24, 45, 66, 67, 70 and 99) and the F pocket is composed of four (77, 80, 81 and 116). Previous analysis of HLA-A*0201 [3] found that residues such as leucine, isoleucine, valine and methionine are favoured for the P2 position while the optimal residues for the C-terminus are leucine and valine (although methionine, isoleucine, alanine and threonine can also occupy this position). In some cases, contributions made by the peptide residues at positions 1, 3, 6 and 7 can compensate for the presence of unfavoured residues at the P2 and C-terminal positions. In an attempt to discover novel T cell epitopes, many motif-based techniques have been used to predict the affinity of MHC-peptide interactions. Diverse methods, such as EpiVax [4], Artificial Neural Networks [5], Hidden Markov Models [6], Support Vector Machines [7] and Profiles [8], have all been used to create T-cell epitope prediction programs. Typically, prediction techniques have been calibrated by using experimentally-determined IC50 and BL50 data [9]. The model is fitted to a set of experimental binding parameters and is used to predict the affinity of novel ligands. This technique assumes that the free energy of binding correlates with a subset of energy components determined from the structure. In this paper, a technique similar to COMBINE is developed which incorporates both structural and data-driven methods in order to create a novel method for the analysis of peptide-MHC binding. This new method seeks to better understand the atomic interactions that determine binding affinities: static energy analysis calculations combined with QSAR analysis are used to calculate the relative contribution of interactions between the MHC Class I allele A*0201 and bound peptide residues, thus generating an accurate and predictive system for the A2 supertype.

Results

QSAR analysis was carried out on three individual models: total/electrostatic/van der Waals energy descriptors; electrostatic/van der Waals energy descriptors and total non-bonded energy descriptors. The matrix consisted of 118 peptide rows and 4860 descriptor (energy, electrostatic and van der Waals) columns. To make this analysis tractable, any columns consisting entirely of values of ~0 kcal mol⁻¹ (signifying that no contributions were made between the peptide and binding site) were omitted from the dataset, yielding a final matrix of 1504 descriptor columns (454 total energy descriptors, 331 electrostatic energy descriptors and 719 van der Waals energy descriptors). The resulting regression equation modeled the binding affinities of a training set of 83 peptides and was used to predict the binding affinities of an independent test set of 35 peptides.

The GFA and G/PLS algorithms were used to select the optimum number of descriptors for use in the regression analysis. Although the methods are similar, the application of G/PLS often allows the construction of larger QSAR equations while still avoiding over-fitting and eliminating most variables. Within the Cerius² software, the crucial variable is the initial number of descriptors chosen in the regression equation although this can be varied during the process by including a mutation probability for increasing or decreasing the equation length. The statistical indices for the best GFA and G/PLS algorithms for all 3 models are shown in Table 1. For all the QSAR models,
the number of terms in the initial and final equations was varied between 15–20 descriptors. It was found that raising or lowering the number of descriptors outside of this range caused a steady increase in the PRESS value and a lowering of the $r^2$. Models were created with 10, 25 and 30 descriptors and all exhibited a decrease in the quality of the model. The GFA Total + Electrostatic + VDW model is shown in Table 2 as example. It is likely that the use of too many descriptors or too few will cause overfitting or underfitting of the model respectively. 15–20 descriptors may represent a range where there is sufficient information to build a model without overemphasising specific interactions that do not represent general trends within the dataset. The best model was chosen based on both the $r^2$ value of the test set and also on the (CV)$r^2$ value from the training set. All selected models had the highest avail-

### Table 1: QSAR Results Based on full binding groove for GFA and G/PLS Total, Electrostatic + VDW and Total + Electrostatic + VDW models.

<table>
<thead>
<tr>
<th>Model</th>
<th>QSAR Analysis (Cerius)</th>
<th>Equation Length</th>
<th>STATISTICAL PARAMETERS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Terms + Constant</td>
<td>Training Set</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>LOF</td>
</tr>
<tr>
<td>Total + Electrostatic + VDW</td>
<td>GFA 15 0.286 0.766 0.602 22.501 0.466 0.217 46.266</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>16 0.340 0.734 0.601 22.538 0.311 0.097 60.902</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>17 0.308 0.772 0.583 23.557 0.776 0.602 34.625</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>18 0.311 0.781 0.578 23.840 0.436 0.190 54.676</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>19 0.279 0.813 0.663 19.074 0.252 0.063 809.185</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>20 0.409 0.741 0.523 26.946 0.031 0.001 540.451</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>G/PLS 15 - 0.565 0.351 36.78 0.186 0.035 60.451</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>16* - 0.779 0.590 13.87 0.740 0.547 39.552</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>17 - 0.609 0.167 47.06 0.461 0.213 51.299</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>18 - 0.572 0.216 44.32 0.048 0.002 69.939</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>19 - 0.620 0.195 45.50 0.155 0.024 64.633</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>20 - 0.612 0.113 50.13 0.235 0.055 61.639</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Electrostatic + VDW</td>
<td>GFA 15 0.382 0.687 0.487 29.336 0.087 0.008 68.676</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>16 0.316 0.753 0.487 29.336 0.087 0.008 68.676</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>17 0.320 0.763 0.311 71.12 0.514 0.264 59.054</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>18 0.351 0.753 -1.892 163.51 0.698 0.488 48.489</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>19 0.349 0.767 0.616 21.719 0.731 0.534 39.464</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>20 0.345 0.781 0.646 20.03 0.523 0.274 39.683</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>G/PLS 15* - 0.723 0.541 21.359 0.737 0.544 52.287</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>16 - 0.681 0.353 36.59 0.760 0.578 37.425</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>17 - 0.674 0.303 39.43 0.605 0.366 53.741</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>18 - 0.624 -1.892 163.51 0.698 0.488 48.489</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>19 - 0.704 0.323 38.27 0.593 0.352 42.913</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>20 - 0.686 0.440 31.67 0.650 0.423 40.518</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>GFA 15 0.362 0.703 0.252 42.292 0.198 0.039 286.799</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>16 0.349 0.728 0.568 24.420 0.200 0.048 121.370</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>17 0.403 0.701 0.473 29.811 -0.060 0.004 356.214</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>18* 0.200 0.861 0.751 11.071 0.781 0.610 51.105</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>19 0.408 0.727 0.534 26.348 0.487 0.237 65.488</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>20 0.388 0.755 0.584 23.517 -0.123 0.015 5580.648</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>G/PLS 15 - 0.665 0.299 39.63 0.796 0.634 40.751</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>16 - 0.648 0.344 37.11 0.393 0.155 53.478</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>17 - 0.646 0.353 36.60 0.375 0.141 55.306</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>18 - 0.640 0.131 49.14 0.656 0.430 39.995</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>19 - 0.686 0.420 32.81 0.571 0.327 45.867</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>20 - 0.662 0.395 34.19 0.253 0.064 61.058</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Indicates where outliers have been removed.
able test set $r^2$ value, except for the G/PLS Electrostatic + VDW model where the 15 descriptor model was chosen as it has a (CV)$r^2$ value significantly higher than the 16 descriptor model. Marked equations had a few anomalous outliers removed to consolidate the model. Further analysis was also carried out limiting the potential descriptors to those amino acids comprising previously-defined MHC anchor pockets: P2 and the C terminus [1]. In that model, a total of 138 residues were used (46 energy terms, 46 electrostatic terms and 46 van der Waals terms). Despite the reduced population size, the optimal number of descriptors still occurred within the 10–20 descriptor range (see Table 3). The reduced number of descriptors generally seems to decrease the predictivity of the models significantly. However, one exception (the G/PLS Total + Electrostatic + VDW model) did produce a test set $r^2$ value of 0.711, which is higher than any model produced using the full range of descriptors. The GFA and G/PLS models produced extremely poor results with a test set $r^2$ value of 0.044 and 0.263 respectively. The best individual models for both the GFA and G/PLS algorithms taken from both Tables 1 and 3 are presented in Table 4 and 5 respectively. Description sets taken from the best models generated by the GFA and G/PLS calculations were used to validate the QSAR analysis using the PLS module in SYBYL 6.9. The predictivity of GFA and G/PLS models compares well with the PLS method, indicating that the predictive ability of the models is consistent in terms of descriptor choice, despite differences in the form of the predictive equation in each case.

Table 6 shows the residues that make up the descriptors for each of the six Cerius2 QSAR models presented in Table 3. Each model's residue distributions around the peptide are shown in Figures 1, 2, 3, 4, 5, 6. The selected coefficients are generated from those residues which make a consistent contribution over the entire training set. An overview of the QSAR models would suggest that although specific MHC-peptide interactions are difficult to identify, certain regions of the $\alpha$ heavy chain seem to relate to specific peptide residues. P2 is associated with the $\alpha$60–70 region while the $\alpha$160–170 region is associated with P2-P4. The $\alpha$150–160 and $\alpha$140–145 regions relate to P5 and P7 specifically. These associations may be fundamental to the binding and are maintained irrespective of the sequence of the peptide. In several cases, the MHC residues interact with the main chain atoms of the peptide and therefore make energetic contributions independent of the peptide sequence. Conversely, regions $\alpha$1–4, $\alpha$26–33, $\alpha$46–58, $\alpha$101–111 and $\alpha$168–180 all make little or no energetic contribution due to their distance from the peptide residues. It is difficult to identify specific interactions from MHC residues, however, the following associations between specific MHC and peptide residues are all represented in three out of the six models: Met 5 interacting with P3, Asp 77 interacting with P9, Thr 143 interacting with P9, Val 152 interacting with P6 and Leu 156 interacting with P5. This might suggest that those five interactions all have an effect in determining the specificity of the interaction. Of the B pocket residues, Tyr 99 is the only residue to appear in two models while B pocket residues Ala 24, Met 45, Lys 66 and Tyr 116 only appear in one model each. Asp 77 appears in three models as is the only one of the four residues lining the F pocket which is used as a descriptor. Both Lys 66 and Asp 77 have been identified as forming conserved hydrogen bonds within the groove, respectively with the oxygen of the P2 residue and the nitrogen of the P9 residue [1]. Lys 66 is of particular interest as it has been identified as being a functional "hotspot" due to the formation of an ion pair with Glu 63 and an apparent dual role in both peptide binding and T cell receptor interaction [20-22]. In spite of this the residue is only used as a descriptor by G/PLS Total + Electrostatic + VDW model. The conserved bonds, Glu 63 - P2 amide nitrogen and Tyr 99-P3 amide nitrogen, are not

---

**Table 2: QSAR Results for GFA Total + Electrostatic + VDW model based on 10, 20, 25, 30 descriptors.**

<table>
<thead>
<tr>
<th>Model</th>
<th>QSAR Analysis (Cerius)</th>
<th>Equation Length Terms + Constant</th>
<th>STATISTICAL PARAMETERS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Training Set</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>LOF $r^2$</td>
</tr>
<tr>
<td>Total + Electrostatic + VDW</td>
<td>GFA</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td></td>
<td>0.321</td>
<td>0.670</td>
</tr>
<tr>
<td>15</td>
<td></td>
<td>0.371</td>
<td>0.696</td>
</tr>
<tr>
<td>20</td>
<td></td>
<td>0.288</td>
<td>0.818</td>
</tr>
<tr>
<td>25</td>
<td></td>
<td>0.357</td>
<td>0.832</td>
</tr>
<tr>
<td>30</td>
<td></td>
<td>0.389</td>
<td>0.871</td>
</tr>
</tbody>
</table>
represented at all. Both residues, however, are used as descriptors in relation to P1 and in relation to P2 and P4 respectively in separate models. Previous work [10] has identified residues 9, 97, 114 and 116 as key to defining the A2-supertype, hence representative of vital points of interaction within the groove. They are poorly represented by the QSAR models, the only equation that incorporates one of them is the G/PLS Electrostatic + VDW model, where 116 interacts with P5. However, they do all occur within regions of the groove identified as important to the binding, again suggesting that it is regions rather than specific residues that determine the stability of the complex.

**Discussion**

Overall, within our models, the distribution of the descriptors does not lend support to the view that burial

---

**Table 3: QSAR Results Based on Binding Pockets for GFA and G/PLS Total, Electrostatic + VDW and Total + Electrostatic + VDW models.**

<table>
<thead>
<tr>
<th>Model Analysis (Cerius)</th>
<th>Equation Length Terms + Constant</th>
<th>STATISTICAL PARAMETERS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Training Set</td>
</tr>
<tr>
<td></td>
<td></td>
<td>LOF</td>
</tr>
<tr>
<td>Total + Electrostatic + VDW</td>
<td></td>
<td>GFA</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GFA</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GFA</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GFA</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GFA</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GFA</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GFA</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GFA</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GFA</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GFA</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GFA</td>
</tr>
</tbody>
</table>

* Indicates where outliers have been removed.
Table 4: Best QSAR models for GFA and G/PLS based on full binding groove in comparison with the PLS method. *PLS Models based on Cerius GFA & G/PLS Models respectively.

<table>
<thead>
<tr>
<th>Model</th>
<th>QSAR Analysis</th>
<th>Equation Length</th>
<th>STATISTICAL PARAMETERS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>TRAINING SET</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>LOF</td>
</tr>
<tr>
<td>Total + Electrostatic + VDW</td>
<td>GFA¹</td>
<td>17</td>
<td>0.308</td>
</tr>
<tr>
<td></td>
<td>G/PLS¹</td>
<td>16</td>
<td>0.779</td>
</tr>
<tr>
<td></td>
<td>PLS² (GFA)</td>
<td>17</td>
<td>0.604</td>
</tr>
<tr>
<td></td>
<td>PLS² (G/PLS)</td>
<td>16</td>
<td>0.571</td>
</tr>
<tr>
<td>Electrostatic + VDW</td>
<td>GFA¹</td>
<td>19</td>
<td>0.349</td>
</tr>
<tr>
<td></td>
<td>G/PLS¹</td>
<td>15</td>
<td>0.723</td>
</tr>
<tr>
<td></td>
<td>PLS² (GFA)</td>
<td>19</td>
<td>0.675</td>
</tr>
<tr>
<td></td>
<td>PLS² (G/PLS)</td>
<td>15</td>
<td>0.558</td>
</tr>
<tr>
<td>Total</td>
<td>GFA¹</td>
<td>18</td>
<td>0.200</td>
</tr>
<tr>
<td></td>
<td>G/PLS¹</td>
<td>15</td>
<td>0.665</td>
</tr>
<tr>
<td></td>
<td>PLS² (GFA)</td>
<td>18</td>
<td>0.654</td>
</tr>
<tr>
<td></td>
<td>PLS² (G/PLS)</td>
<td>15</td>
<td>0.583</td>
</tr>
</tbody>
</table>
Table 5: Best QSAR models for GFA and G/PLS based on binding Pockets in comparison with the PLS method. * PLS Models based on Cerius GFA & G/PLS Models respectively.

<table>
<thead>
<tr>
<th>Model</th>
<th>QSAR Analysis</th>
<th>Equation Length</th>
<th>STATISTICAL PARAMETERS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Terms + Constant</td>
<td>TRAINING SET</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>LOF</td>
</tr>
<tr>
<td>Total + Electrostatic + VDW</td>
<td>GFA¹</td>
<td>13</td>
<td>0.204</td>
</tr>
<tr>
<td></td>
<td>G/PLS¹</td>
<td>16</td>
<td>0.789</td>
</tr>
<tr>
<td></td>
<td>PLS² (GFA)</td>
<td>13</td>
<td>0.593</td>
</tr>
<tr>
<td></td>
<td>PLS² (G/PLS)</td>
<td>16</td>
<td>0.704</td>
</tr>
<tr>
<td>Electrostatic + VDW</td>
<td>GFA¹</td>
<td>13</td>
<td>0.242</td>
</tr>
<tr>
<td></td>
<td>G/PLS¹</td>
<td>19</td>
<td>0.643</td>
</tr>
<tr>
<td></td>
<td>PLS² (GFA)</td>
<td>13</td>
<td>0.548</td>
</tr>
<tr>
<td></td>
<td>PLS² (G/PLS)</td>
<td>19</td>
<td>0.729</td>
</tr>
<tr>
<td>Total</td>
<td>GFA¹</td>
<td>20</td>
<td>0.280</td>
</tr>
<tr>
<td></td>
<td>G/PLS¹</td>
<td>20</td>
<td>0.528</td>
</tr>
<tr>
<td></td>
<td>PLS² (GFA)</td>
<td>20</td>
<td>0.627</td>
</tr>
<tr>
<td></td>
<td>PLS² (G/PLS)</td>
<td>20</td>
<td>0.654</td>
</tr>
</tbody>
</table>
### Table 6: Residue descriptors for the Best QSAR Results based on full binding groove.

<table>
<thead>
<tr>
<th>Total + Electrostatic + VDW</th>
<th>Peptide Pos.</th>
<th>P1</th>
<th>P3</th>
<th>P7</th>
<th>P9</th>
<th>P9</th>
<th>P2</th>
<th>P9</th>
<th>P1</th>
<th>P2</th>
<th>P5</th>
<th>P5</th>
<th>P5</th>
<th>P6</th>
<th>P6</th>
<th>P7</th>
<th>P9</th>
</tr>
</thead>
<tbody>
<tr>
<td>MHC Residue No.</td>
<td>63</td>
<td>5</td>
<td>141</td>
<td>14</td>
<td>48</td>
<td>39</td>
<td>115</td>
<td>51</td>
<td>69</td>
<td>130</td>
<td>133</td>
<td>144</td>
<td>75</td>
<td>148</td>
<td>140</td>
<td>115</td>
<td></td>
</tr>
<tr>
<td>MHC Residue</td>
<td>Glu Met Gln Arg Arg Asp Gln Trp Ala Leu Trp Lys Val Ala Ala Gln</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>G/PLS</td>
<td>Peptide Pos.</td>
<td>P1</td>
<td>P4</td>
<td>P5</td>
<td>P6</td>
<td>P8</td>
<td>P9</td>
<td>P9</td>
<td>P1</td>
<td>P2</td>
<td>P5</td>
<td>P5</td>
<td>P5</td>
<td>P7</td>
<td>P9</td>
<td>P9</td>
<td>P9</td>
</tr>
<tr>
<td>MHC Residue No.</td>
<td>171</td>
<td>164</td>
<td>157</td>
<td>66</td>
<td>19</td>
<td>72</td>
<td>77</td>
<td>123</td>
<td>171</td>
<td>156</td>
<td>99</td>
<td>66</td>
<td>66</td>
<td>73</td>
<td>155</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MHC Residue</td>
<td>Tyr Cys Arg Lys Gln Gln Asp Tyr Tyr Leu Tyr Lys Lys Thr Gln</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Electrostatic + VDW</td>
<td>Peptide Pos.</td>
<td>P1</td>
<td>P5</td>
<td>P7</td>
<td>P8</td>
<td>P8</td>
<td>P9</td>
<td>P9</td>
<td>P2</td>
<td>P2</td>
<td>P2</td>
<td>P4</td>
<td>P4</td>
<td>P5</td>
<td>P7</td>
<td>P9</td>
<td>P9</td>
</tr>
<tr>
<td>MHC Residue No.</td>
<td>10</td>
<td>35</td>
<td>21</td>
<td>99</td>
<td>117</td>
<td>95</td>
<td>126</td>
<td>60</td>
<td>73</td>
<td>160</td>
<td>45</td>
<td>99</td>
<td>130</td>
<td>23</td>
<td>8</td>
<td>117</td>
<td>134</td>
</tr>
<tr>
<td>MHC Residue</td>
<td>Thr Arg Arg Tyr Ala Val Leu Ala Trp Thr Leu Met Tyr Leu Ile Phe Ala Thr</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>G/PLS</td>
<td>Peptide Pos.</td>
<td>P1</td>
<td>P1</td>
<td>P7</td>
<td>P9</td>
<td>P2</td>
<td>P2</td>
<td>P2</td>
<td>P4</td>
<td>P5</td>
<td>P5</td>
<td>P5</td>
<td>P6</td>
<td>P7</td>
<td>P9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MHC Residue No.</td>
<td>43</td>
<td>115</td>
<td>22</td>
<td>159</td>
<td>24</td>
<td>68</td>
<td>167</td>
<td>163</td>
<td>116</td>
<td>156</td>
<td>158</td>
<td>152</td>
<td>93</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MHC Residue</td>
<td>Gin Gin Phe Tyr Ala Lys Trp Thr Tyr Leu Ala Val Thr His</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>Peptide Pos.</td>
<td>P1</td>
<td>P1</td>
<td>P1</td>
<td>P2</td>
<td>P2</td>
<td>P2</td>
<td>P2</td>
<td>P3</td>
<td>P3</td>
<td>P3</td>
<td>P3</td>
<td>P5</td>
<td>P5</td>
<td>P6</td>
<td>P6</td>
<td>P9</td>
</tr>
<tr>
<td>MHC Residue No.</td>
<td>25</td>
<td>61</td>
<td>156</td>
<td>33</td>
<td>64</td>
<td>68</td>
<td>163</td>
<td>5</td>
<td>34</td>
<td>162</td>
<td>168</td>
<td>152</td>
<td>156</td>
<td>115</td>
<td>152</td>
<td>77</td>
<td>143</td>
</tr>
<tr>
<td>MHC Residue</td>
<td>Val Asp Leu Phe Thr Lys Thr Met Val Gly Leu Val Leu Gln Val Asp Thr</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>G/PLS</td>
<td>Peptide Pos.</td>
<td>P1</td>
<td>P2</td>
<td>P2</td>
<td>P2</td>
<td>P3</td>
<td>P3</td>
<td>P3</td>
<td>P5</td>
<td>P6</td>
<td>P7</td>
<td>P9</td>
<td>P9</td>
<td>P9</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MHC Residue No.</td>
<td>25</td>
<td>33</td>
<td>62</td>
<td>119</td>
<td>5</td>
<td>74</td>
<td>98</td>
<td>156</td>
<td>152</td>
<td>143</td>
<td>14</td>
<td>74</td>
<td>77</td>
<td>83</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MHC Residue</td>
<td>Val Phe Gly Asp Met His Met Leu Val Thr Arg His Asp Gly</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
of anchor residues, within defined cavities, are mainly responsible for stabilising the peptide within the groove. Rather it indicates a fairly even distribution of MHC residue interactions along the length of the peptide. Attempts to improve the QSAR model by focusing on the MHC residues associated with the anchor residues were largely unsuccessful (see Table 3 and Table 5), in spite of the predictive accuracy of the G/PLS Total + Electrostatics + VDW model. Despite the fact that the anchor models generate a similar number of descriptors to those of the full groove, focusing the interaction towards the peptide termini at the expense of the variable region generally reduces the algorithm’s predictivity. It is possible that this reflects the comparative dominance of electrostatic interactions in determining binding affinity within both of the anchor pockets. Model predictivity also deteriorates after removal of each descriptor type compared to the full residue model, particularly in the case of models based on Total energy. Moreover, attempts to split the interaction energy into long range electrostatic and short range van der Waals forces also decreases correlation in the full residue model, suggesting that a single residue’s total energy contribution is primarily enthalpic in character. It should be noted that several of the residues highlighted by the models are distant from the peptide (see Figures 1, 2, 3, 4, 5, 6), demonstrating the importance of long-range electrostatic interactions, which can be up to 8Å length, in stabilizing the complex. This could also be an effect of superadditivity correlation (SAC) [23] where correlation between variables between two sites does not imply that the two sites directly interact. Mutational data has indicated that long range non-causal interactions can be stronger in magnitude than causal correlations induced by direct interactions [24,25].

The Total energy models showed the highest correlation in the test set for both the GFA and G/PLS model although the (CV)r² value of the G/PLS model dropped sharply in the training set. Both models placed great emphasis on the conserved P1-P3 region of the MHC Class I groove at the expense of the central and C-termini regions. The key interaction of Asp 77 with the P9 residue was incorporated into both the GFA and G/PLS model, indicating that it may be a vital descriptor for a successful predictive algorithm. Both sets of models where the energetic contributions were broken down into the electrostatic and van der Waals contributions showed a lower degree of correlation (although the PRESS values for the Test Set compared favourably with the Total Energy Models). Building comparative models using the program SYBYL (see Table 4 and 5) produced a consistently lower level of correlation compared to the Cerius² model but the apparent trends are comparable. The QSAR models also show a similar degree of correlation to models generated by the additive method [10] which produced a higher r² value of 0.954 but an equivalent (CV)r² value of 0.602 from a model containing only 6 components.

The limitations of motif-based methods and the potential of structural analysis may be in part due to the variable nature of the binding with the central region of the peptide. Where as the termini regions (P1-P3 and P8-P9) are heavily constrained by interactions with conserved MHC residues, the central region (P4-P7) typically exhibits a far...
more variable conformation. Analysis of the MHC Class I structure [1] has shown that the central region is less constrained within the groove than the termini and can dramatically alter its conformation in response to the peptide sequence. It is also the region which is most exposed to the T cell receptor and because of this may arch away from the floor of the groove. The conformational variability of the central region means that definite contacts with specific MHC residues are not obvious. This makes predicting the contribution to affinity of that region of the ligand for the MHC molecule difficult. Nevertheless, the models demonstrate that the central region does make an important contribution to the specificity of the interaction.

It is also possible that static energy analysis overlooks significant energetic interactions occurring during binding. The thermodynamic property, which we are trying to estimate with our scoring function, is the energy released when ligand and receptor bind; the so-called Gibbs free energy of binding: \( \Delta G_{\text{bind}} \). It can be represented as \( \Delta G_{\text{bind}} = \Delta H - T \Delta S \) where \( \Delta H \) is the enthalpic (internal) energy and \( T \Delta S \) is an entropy term which is indicative of the relative gain or loss of disorder upon binding [26]. The binding groove is hydrophobic in character, particularly in the region of the anchor residues, and the calculated energy interactions reflect this, particularly with the high incidence of charge-charge interactions within the groove. It is necessary for the solvent entropy from the burial of hydrophobic groups to offset the reduction in peptide conformational entropy that occurs upon binding. The measured interactions therefore reflect the hydrogen bonds and salt bridges that are formed when the peptide moves from the loosely packed, partially bound state to a being fully bound into a complex. Water is stripped from partial and full charges on the MHC and peptide residues and there is a reduction in the favourable hydrogen bond enthalpy associated from interactions with bound water molecules. In taking our work forward, it is therefore necessary to incorporate the entropic contribution to the energetic calculation into the descriptors in order to create a more accurate representation of the free energy change between the MHC molecules bound and unbound states. This may be done by use of the computationally efficient MM_PBSA/GBSA method which combines the molecular mechanical energies with the continuum solvent approaches [27,28]. It is hoped that the inclusion of these entropic terms may significantly increase the predictive capability of future QSAR analysis. However, the local interactions have always proved to be more relevant to determining the affinity of a given peptide than the general properties of the ligand-receptor interaction [29].

Previous work [10-16] has already indicated that motif-based analysis of the anchor residues within the MHC Class I binding groove can only provide limited accuracy in determining potential T-cell epitopes. Thus the QSAR models presented here strongly challenge the assumption that only the P2 and P9 residues of a nonameric peptide make a significant contribution to binding. While the majority of the descriptors do concentrate on the termini regions of the peptide, there is a clear, and substantial, energetic contribution made by the structurally variable central region. It is possible that more detailed free energy calculations of the MHC Class I-peptide binding interac-
tions, whereby the entropic contribution to the association of the peptide to the MHC Class I molecule is incorporated into the model, will, in the future, lead to the generation of more accurate predictive algorithms.

Methods

Energy minimization of HLA-A*0201 structure

An HLA-A*0201 dataset, consisting of 118 peptides of known sequence, was obtained from Doytchinova et al. [10]. The structure of the MHC Class I allele HLA-A*0201 bound to a nonameric melanocyte-melanoma tumor-antigen peptide elucidated by X-ray diffraction at a resolution of 2.15Å [2] (pdb code: 1JHT) was used as the template. Residues of the bound peptide were remodeled using the crystallographic modeling program ‘O’ [29] to generate 118 MHC Class I-peptide structures equivalent to the data set. Energy minimization was carried out on the structures using AMBER Version 8 [30]. The interatomic potential energy function uses the AMBER forcefield 03 for parameterisation. Hydrogen atoms were added to the structure, and the system was fully solvated using TIP3 waters [31]. This function was performed by the LEaP program [30]. The full structure of the HLA-A*0201-peptide complex was explicitly represented within the simulation. The energy of the solvated molecular complex was minimized using a steepest descent method that continued for 20,000 1 fs time steps or until the root mean square deviation between successive time steps had fallen below 0.01Å. The peptides were then remodeled and annealed by raising the temperature of the system from 0 to 500 K for a period of 8 picoseconds and maintaining the system at that temperature for a further 6 picoseconds. The system was cooled to 0.2 K over a period of 33 picoseconds before being rested at 1 K for a further 30 picoseconds. Both the minimization and the annealing were performed using the sander program and static energy analysis was then carried out on the minimised structures using the anal program [30]. Anal calculates the group-group interaction energies between different parts of the system based upon the position of their composite atoms. The interaction energies, which are measured in kcal mol⁻¹, between the nine peptide residues and the first 180 residues of the MHC Class I heavy chain (which incorporates the entire MHC Class I binding groove) were calculated. Three values were generated for each interaction; the electrostatic interaction energy, the van der Waals interaction energy and the total non-bonded interaction energy. Total energies are the sum of electrostatic and van der Waals terms; since these values can have different signs and be significantly different in magnitude, their sum gives rise to a non-trivial and informative descriptor.

QSAR regression

The BL50 binding affinities for the dataset, were extracted from the AntiJen database [32-34]. The half-maximal binding level (BL50) is the peptide concentration yielding the half-maximal Fluorescence Index (FI) of the reference peptide in each assay. These values were converted to pBL50 (-logBL50) and used as the dependent variables in the QSAR regression. Peptides were partitioned between training and test datasets in a ratio of 70:30. A program was developed which transformed the nine amino acid
peptide sequence into rows of a table. When choosing a test set, the program omits rows that contain the least contributing amino acid starting at position 1 and working systematically to positions 2, 3, through to position 9 until 30% of the dataset remains. Three different QSAR techniques were employed during the study: the Genetic Function Approximation (GFA) algorithm [35], Genetic Partial Least Squares (G/PLS) [36], which were both carried out using Cerius² [37], and Partial Least Squares (PLS), carried out using SYBYL 6.9 [37]. The genetic algorithm first selects equations that contain a randomly chosen subset of dependent variables (or descriptors) and, subsequently, improved equations are 'evolved' using a genetic crossover operation, which selects the best descriptors using multiple linear regression, and finally uses a least squares technique to produce the final model. G/PLS, an alternative to the GFA algorithm, uses GFA to select appropriate basis functions and PLS regression as the fitting technique to weight the basis functions’ relative contributions in the final model. In the case of both algorithms, the population size of the initial equation set was fixed at 100 while the number of generations of equations (genetic crossovers) was set at 5000. Once a QSAR calculation has been completed, a list of outliers is generated by the GFA and G/PLS algorithms. Outliers are removed from the training set, the QSAR is repeated again and new equations are generated, until, at most, 10-15% of the training set outliers are removed.

Within Cerius² the validation terms used were correlation coefficient ($r^2$), Friedman's Lack of Fit (LOF) [38] and the Cross-Validation correlation coefficient ((CV)$r^2$) using the Leave-One-Out (LOO) procedure [39]. Adding more terms to the regression model always reduces the normal least squares error (LSE) but does not necessarily reduce the LOF measure. Thus, as a new term reduces the LSE, it also tends to increase the value of LOF. Within SYBYL, the predictive power of the PLS Cross-Validation Leave-One-Out (CV-LOO) model is assessed using the following parameters: cross-validated coefficient ($q^2$) and the Standard Error of Prediction (SEP) [10-16]. The optimal number of components (NC) resulting from the CV-LOO is then used in the non-cross validated model which was assessed using standard Multiple Linear Regression (MLR) validation terms: explained variance ($r^2$) and Standard Error of Estimate (SEE) [10-16]. Increasing the number of components improves the fit between target and explanatory properties and corresponds to the best $q^2$.

Abbreviations

- GFA – Genetic Function Approximation
- G/PLS – Genetic Partial Least Squares
- MHC – Major Histocompatibility Complex
- COMBINE – COMparative BINding Energy
- SAC – Superadditivity Correlation
- PLS – Partial Least Squares
- LOF – Lack of Fit
- (CV)$r^2$ – Cross-Validation correlation coefficient
- LOO – Leave-One-Out
- SEP – Standard Error of Prediction
- NC – Number of Components
- SEE – Standard Error of Estimate
- MLR – Multiple Linear Regression

Authors’ contributions
Matthew Davies remodelled the MHC Class I structures, carried out the energy minimisation simulations and calculated the static energy interactions between the MHC molecule and its bound peptide

Channa Hattotuwagama carried out the QSAR analysis of the static energy interactions data using of Genetic Function Approximation (GFA) and Genetic Partial Least Squares (G/PLS) algorithms

David Moss supervised the molecular modeling work at Birkbeck College

Michael Drew supervised the QSAR analysis at the University of Reading

Darren Flower co-ordinated the analysis of the static energy interactions by QSAR methods and helped to interpret the data in an immunological context as well as overseeing the whole project

Acknowledgements
The authors would like to thank Dr. Mark T. Storr and Roger Robinson at the University of Reading for general help, advice, time and use of computer facilities in Computational Group as well as Dr Andy Purkiss and Dr David Houldershaw of Birkbeck Crystallography Department for their assistance with the energy minimisation simulations.

References


30. Hattotuwagama CK, Guan P, Doytchinova IA, Flowe DR: In Silico QSAR-Based Predictions of Class I and Class II MHC Epitopes. Immunoinformatics: Opportunities and Challenges of Bridging Immunology with Computer and Information Sciences. in press.


