BIOINFORMATICS

Vol. 22 no. 14 2006, pages e16–e24
doi:10.1093/bioinformatics/btl216

DynaPred: A structure and sequence based method for the prediction of MHC class I binding peptide sequences and conformations

Iris Antes[1,*], Shirley W. I. Siu[1] and Thomas Lengauer

MPI für Informatik, Stuhlsatzenhausweg 85, D-66123 Saarbrücken, Germany

ABSTRACT

Motivation: The binding of endogenous antigenic peptides to MHC class I molecules is an important step during the immunologic response of a host against a pathogen. Thus, various sequence- and structure-based prediction methods have been proposed for this purpose. The sequence-based methods are computationally efficient, but are hampered by the need of sufficient experimental data and do not provide a structural interpretation of their results. The structural methods are data-independent, but are quite time-consuming and thus not suited for screening of whole genomes. Here, we present a new method, which performs sequence-based prediction by incorporating information obtained from molecular modeling. This allows us to perform large databases screening and to provide structural information of the results.

Results: We developed a SVM-trained, quantitative matrix-based method for the prediction of MHC class I binding peptides, in which the features of the scoring matrix are energy terms retrieved from molecular dynamics simulations. At the same time we used the equilibrated structures obtained from the same simulations in a simple and efficient docking procedure. Our method consists of two steps: First, we predict potential binders from sequence data alone and second, we construct protein-peptide complexes for the predicted binders. So far, we tested our approach on the HLA-A0201 allele. We constructed two prediction models, using local, position-dependent (*DynaPred*^{POS}) and global, position-independent (*DynaPred*) features. The former model outperformed the two sequence-based methods used in our evaluation; the latter shows a much higher generalizability towards other alleles than the position-dependent models. The constructed peptide structures can be refined within seconds to structures with an average backbone RMSD of 1.53 Å from the corresponding experimental structures.

Contact: antes@mpi-sb.mpg.de

1 INTRODUCTION

The binding of antigenic peptides originating from pathogens to the major histocompatibility complex (MHC) class I is one of the crucial steps during the intracellular immunological response against the intruder (Paul *et al*., 1998). After a pathogen enters the host cell, proteins from the invading organism are cleaved into smaller peptide fragments by the proteasome. These fragments are transported into the endoplasmic reticulum by the TAP proteins, where they bind to MHC molecules. Afterwards the MHC-peptide complex is translocated to the cell surface. At the surface of the cell, pathogenic peptides are identified by T-cell receptors (TCRs) via TCR-MHC-peptide complex formation. This step initiates the immunological response against the pathogen. Peptides which can trigger such a response are called epitopes. Not all peptides binding to MHC molecules are epitopes, but all T-cell epitopes need to bind to MHC molecules. Thus, knowing which and understanding why certain peptides bind to a specific MHC is not only fundamental to the understanding of the immune system, but also a crucial step in vaccine and immunotherapeutic development. Experimental screening of peptides with respect to their MHC binding capabilities is very demanding due to the large number of possible peptide sequences and the high polymorphism of the MHC molecules. Thus there is a strong interest in computational methods for predicting the binding capabilities of peptides to MHC as a first step to select peptides for screening.

For the prediction of MHC (class I and II) binding peptides, sequence- and structure-based methods as well as their combinations were used for both classification and regression models. Classification models distinguish binders from non-binders, whereas regression methods try to predict the binding affinity of peptides to MHC molecules.

Sequence based prediction methods include binding motifs (Rammensee *et al*., 1999; Hammer, 1995; Reche *et al*., 2002; Peters *et al*., 2003), quantitative matrices (Parker *et al*., 1994; Southwood *et al*., 1998), data-derived matrices (Yu *et al*., 2002), and the combination of a motif based approach with Gibbs sampling (Nielsen *et al*., 2004). For the training, various machine learning techniques have been applied such as artificial neural networks (Brusic *et al*., 1998; Gulukota *et al*., 1997; Milik *et al*., 1998), hidden markov models (Mamitsuka 1998), classification trees (Segal *et al*., 2001), support vector machines (Dönnes and Elofsson, 2002; Zao *et al*., 2003; Bhasin *et al*., 2004), and biosupport vector machines (Yang and Johnson, 2005). These methods encode sequences as binary vectors or as numerical vectors based on their physiochemical property values. Due to the limited public availability of consistent quantitative binding data, most methods are trained for classification. Still, regression was performed so far in QSAR studies

*To whom correspondence should be addressed.
[1]Both authors contributed equally to this work.

(Doytchinova *et al*., 2002, 2004; Li *et al*., 2004) and using average relative binding matrices (Bui *et al*., 2005). Structural information has been used for prediction in the context of 3D-QSAR (Doytchinova *et al*., 2002, 2004) and docking (Bordner and Abagyan, 2006).

Most prediction methods are based on the so called 'additive model'. This model assumes that the overall binding affinity of a peptide can be approximated as the sum of the properties of the individual residues. Extensions of this model by including neighbor interactions have led only to slight or even no improvement of the prediction accuracy (Doytchinova *et al*., 2002; Peters *et al*., 2003). In the context of 3D-QSAR the additive model was compared to a model based only on 'global' structural features (Doytchinova *et al*., 2005), which were calculated for the whole peptide and not for the individual residues. This study showed that global features did not perform as well as local, residue based features for binding affinity prediction. The success of the additive model can be explained by the structure of the MHC binding groove, which consists of nine residue binding pockets located next to each other along the groove. The peptide is bound in an extended conformation with one residue of the peptide occupying exactly one binding pocket, thus the effect of the interaction between the neighboring side chains is minimal.

Several structural search algorithms for the identification of low energy peptide-binding conformations have been proposed. One class of methods is based on the observation that for each MHC allele there are certain conserved peptide 'anchor' residues which bind tightly to specific MHC binding pockets. These approaches (Rosenfeld *et al*., 2003, Tong *et al*., 2004, Logean *et al*., 2002) consist of two main steps: first, placing the anchor residues in the binding pocket and second, constructing the rest of the peptide based on the anchor positions. A different class of methods is based on the division of the peptides into backbone and side chains (Ota *et al*., 2001, Altuvia *et al*., 1995, Schueler-Furman *et al*., 1998). These methods use backbone conformations from experimental structures and predict the side chain conformations either by threading or the use of rotamer libraries. Another study uses dead-end elimination within a combinatorial build-up algorithm (Desmet *et al*., 1997). The method, which is closest to our proposed method, is a residue-based free-energy mapping approach (Sezermann *et al*., 1996). Two other studies use Monte-Carlo annealing approaches to dock peptides into the binding pocket (Liu *et al*., 2004) and use the docking scores for prediction (Bordner *et al*., 2006).

Comparing sequence and structure-based methods, the latter have the advantage that they are independent of the amount of experimental binding data, but are too time-consuming for the screening of large numbers of peptides. On the other hand, sequence-based prediction methods are fast, but are strongly dependent on the amount of binding data available for specific alleles. Thus currently they achieve high performance only for the intensively investigated alleles. This becomes even more serious for the quantitative prediction of binding affinities because for this purpose large screening experiments are necessary to produce comparable IC_{50} values for the training of the models. Although such efforts are ongoing, they will always be focused towards the most important alleles. Another drawback of sequence based methods is their limited structural interpretability, which is of crucial importance for the design of peptide mimicking vaccines and drug like molecules.

Here we present a combined two-step structure and sequence-based prediction method *DynaPred*, which allows at the same time a fast prediction of MHC class I binders and an efficient construction of docked peptide conformations. The prediction method uses two feature matrices derived from structural calculations as basis for support vector machine training: A local, position-dependent (*DynaPred*^{POS}) and a global, position-independent (*DynaPred)* matrix. The docking method is based on equilibrated, pre-calculated structures for each amino acid in each of the binding pockets. So far quantitative matrices used for the prediction of MHC-binding peptides are based on sequence data, partially including biophysical amino acid properties. Structure-based biophysical data were used in the context of 3D-QSAR, which, however, only considers the structural properties of the peptides, but not their interactions with the binding pocket. We based the choice of our scoring-matrix features on the linear energy approximation for the calculation of binding affinities. Linear energy models (Aqvist *et al*., 2002) were used in various studies and were successfully tested for predicting the binding affinities of tri-peptides to OppA (Wang *et al*., 2002). In the context of MHC-peptide binding, such approaches were applied for the scoring of docked peptides (Logean *et al*., 2002, Sezermann *et al*., 1996) and prediction based on docking results (Bordner *et al*., 2006).

However, to our knowledge this is the first time that structure-based interaction energy terms are used for a residue-based prediction approach for peptide binding. A residue-based docking method was presented for MHC-peptide complexes by Sezermann *et al*., 1996. However, its discrete rotamer-based search algorithm leads to many different peptide structures, all very similar in energy, and thus extensive post-processing of these structures is necessary to find the best conformer. We avoid this last step by the use of one equilibrated residue side chain conformation, which was calculated by molecular dynamics, instead of a discrete search algorithm.

We implemented and evaluated our approach for the most frequently occurring allele HLA-A*0201 with 9-mer peptides.

2 METHODS

2.1 General strategy

The basic strategy behind our method is to approximate the binding free energy of all 20 amino acids in each of the nine binding pockets of the MHC binding groove using energetic information obtained by molecular dynamics simulations. This information is used subsequently for the training of a sequence-based predictive model. In addition, the structural information obtained by the simulations is used for constructing the peptide-protein complexes of the predicted binders. Our algorithm is based on a single main assumption: The total binding affinity of a peptide can be approximated as the sum of the binding affinities of its individual amino acids, neglecting the effect of the neighboring residues (See 'Introduction' for the validity of this assumption). This allows us to simulate each amino acid individually in each binding pocket. Initial conformations of the individual residues bound to the MHC protein are constructed from crystal structures. To stabilize the peptide conformations, we extend the single residues to peptide-trimers and dimers, by adding a glycine residue at both sides (for terminating residues only on the non-terminating side). For side chains for which no bound conformation was available, existing residues are mutated to the corresponding amino acid. MD simulations are performed on the bound complexes as well as on the individual molecules in solution. Important energy terms reflecting the binding properties of the amino acids are calculated from the simulation results and subsequently used for the construction of a binding-free-energy-based scoring matrix (BFESM). This matrix contains energy terms for each residue in each binding pocket and forms the basis for the construction of two prediction models.

Our approach allows us to predict MHC-binding peptides with the speed of sequence-based methods, but on the basis of structurally derived energies. In addition, the equilibrated MD structures serve as templates to enable fast construction of the conformations predicted binding sequences.

Our proposed method can be summarized as follows:

(1) For a given MHC allele, the compatibility of each amino acid in each of the nine binding pockets is examined thoroughly by MD simulations.

(2) A Binding-Free-Energy-Based Scoring Matrix (BFESM) is produced by extracting values of energy terms important for binding from the simulations.

(3) The position-based bound conformations are extracted from the simulations for each amino acid type and saved in a data base.

(4) Experimental binding data together with the BFESM is used in the training process to generate the prediction models.

Finally, prediction is a two step process: First, the query sequence is classified as a binder or non-binder; then the bound conformation of a predicted binder is generated.

2.2 Scoring matrix

For the construction of the Binding-Free-Energy-based Scoring Matrix (BFESM) we use energy terms obtained by molecular dynamics simulations. According to the linear energy model (Aqvist *et al*., 2002), the binding free energy can be approximated by the difference between the interaction energies ΔG^{el} and ΔG^{np} of the ligand in the protein-ligand complex (bound state) and in solution (free state). We extend this model by adding the energy contributions of the protein and ΔG^{int} and $T\Delta S^{conf}$:

$$\Delta G^{bind} = \Delta G^{el} + \Delta G^{np} + \Delta G^{int} - T\Delta S^{conf} \quad (1)$$

(ΔG^{el} = electrostatic, ΔG^{np} = nonpolar, ΔG^{int} = internal, $T\Delta S^{conf}$= entropic contribution)

Thus the following energy terms are included in the BFESM:

(1) The electrostatic contribution, which consists of the electrostatic interaction energy between the peptide and the MHC molecule and a desolvation term:

$$\Delta G^{el} = \langle V^{el}_{bound.p-l}\rangle + (\langle V^{el}_{bound.p-sol}\rangle - \langle V^{el}_{free.p-sol}\rangle) + (\langle V^{el}_{bound.l-sol}\rangle - \langle V^{el}_{free.l-sol}\rangle) \quad (2)$$

(p = protein, l = ligand, sol = solvent, V^{el} = electrostatic energy)

(2) The non-polar (hydrophobic) contribution, which can be approximated by change in the Solvent Accessible Surface area (SAS) upon binding:

$$\Delta G^{np} \propto \Delta SAS \quad (3)$$

A change in the surface area by 1 Å^2 corresponds to approximately 10.45 kJ mol^{-1} (Chothia, 1974). The change in SAS can be calculated as the difference in surface area between the complex and its individual components in solution.

(3) Due to the restricted space in the binding pocket, the residue might be forced to adopt a higher-energy conformation in the binding pocket than in the solvent. This effect is accounted for by the differences in the bond angle and torsion energies between the free and the bound states:

$$\Delta G^{int} = (\langle V^{int}_{bound.p}\rangle - \langle V^{int}_{free.p}\rangle) + (\langle V^{int}_{bound.l}\rangle - \langle V^{int}_{free.l}\rangle) \quad (4)$$

(4) The loss in conformational entropy, $-T\Delta S^{conf}$, can be approximated using the empirical scale of Pickett and Sternberg (Pickett and Sternberg, 1993). This model assumes that a solvent-exposed side chain, whose relative accessibility (RA) is greater than 60%, can rotate freely; whereas a buried side chain (RA < 60%) is restrained to one rotamer. The RA is defined as:

$$RA = \frac{SAS^{sc}_{bound.l}}{SAS^{sc}_{free.l}} \quad (5)$$

The correspondence between RA and its energetic contribution was taken from (Pickett and Sternberg, 1993).

In summary, to estimate the change in free energy, the energy values at the right hand side of Eq. (2)—Eq. (5) are required. They are calculated for each amino acid in each binding pocket from the MD simulations and used to construct the BFESM.

Table 1. Crystal structures used for the initial backbone conformations of the pseudo-peptides.

PDB	Peptide Source	Sequence	Res. (Å)
1AKJ	HIV reverse transcriptase	ILKEPVHGV	2.65
1DUZ	HTLV-1 TAX protein	LLFGYPVYV	1.80
1HHG	HIV-1 GP120 envelope protein	TLTSCNTSV	2.60
1QRN	Altered HTLV-1 TAX peptide P6A	LLFGYAVYV	2.80

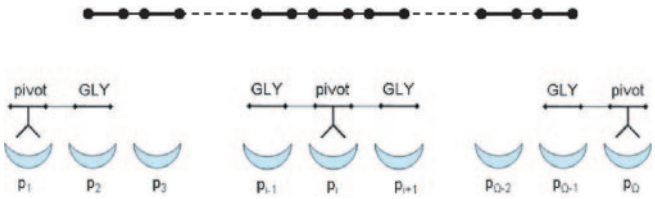

Fig. 1. Schematic representation of the pseudo-peptides used in the simulations. 3-mer or 2-mer pseudo-peptides are constructed depending on the pockets position (from p_1 to p_Ω).

2.3 Simulation setup

To calculate all energy contributions, simulations of all pseudo-peptide MHC complexes and of the MHC molecule and all amino acids in solution were performed. For the construction of the pseudo-peptides the PDB structures given in Table 1 were used. The structure of the MHC protein was taken from PDB structure 1AKJ. Each energy value is calculated as the ensemble average over the last 200ps of the trajectory after the system equilibrium is reached.

2.3.1. Pseudo-peptide generation The amino acid to be investigated (called the *pivot* residue) is embedded inside a short peptide (called *pseudo-peptide*), which is either a 2-mer or 3-mer (see Fig. 1). 2-mers are used for residues at the N and C-termini of the peptide, binding to the pocket 1 and 9. In 2-mers the pivot residue has one neighboring glycine residue. For all other binding pockets, 3-mers are used, consisting of the pivot residue and two neighboring glycines.

For the pseudo-peptide construction all structures in Table 1 were superimposed with respect to the MHC backbone surrounding the binding pocket (residue 1-180). The initial backbone conformations of the pseudo-peptides were extracted from these structures. For this purpose the bound peptide conformations were divided into di/trimers and the side chains of the first and last residue of the di/trimer were replaced by hydrogen, resulting in the two flanking GLY residues. For amino acids for which no experimental structures were available, we mutated existing residues using the program SCWRL3.0 (Canutescu *et al*., 2003).

2.3.2 Simulation conditions All molecular dynamics simulations were performed using GROMACS3.2 (Lindhal *et al*., 2001) and the OPLSAA/L force field and explicit SPC water. Long range electrostatic interactions were calculated using the Particle-Mesh Ewald method and bond constraints were applied using LINCS, and the time step was set to 2 fs. For each simulation, first a steepest-descent energy minimization was performed for 1000 steps. Then the system was solvated using a cubic box with a minimum distance of 0.7 nm between the box boundaries and the protein. The system was heated up from 0 to 300K in 100ps, before it was equilibrated at 300K using NPT ensemble (Berendsen thermo- and barostat). The total equilibration times were dependent on the flexibility of the side chains (800–3000ps). After equilibrium was reached the simulations were continued for another 200–400ps.

To approximate the constraining force of the remaining fragments of the 9-mer peptide on the pivot residue, we applied position restraints to certain atoms of the peptide during the simulations. The restraints were chosen such that the pseudo-peptide backbone was still able to move within a few Å to span the space occupied by the different backbones of the structures in Table 4 and to allow free rotation of the pivot residue. Thus, strong forces (1000 kJ/(mol*nm)) were applied only to the heavy atoms of the flanking glycine residues and weak forces (100 kJ/(mol*nm)) to the C- and N-backbone atoms of the pivot residue.

2.4 Training and testing of the prediction models

2.4.1 Binding-Free-Energy-based Scoring Matrix The Binding-Free-Energy-based Scoring Matrix (BFESM) is a quantitative matrix of dimension 20 *x (no-of-pockets) x (no-of-features)*. Each entry represents one feature of a particular amino acid in a particular binding pocket. The BFESM is used to generate the feature vectors for each given sequence in the training set, all vectors together produce the feature matrix for model generation and prediction.

2.4.2 Prediction models Two feature matrices were constructed from the BFESM: A local feature matrix, which uses all the residue and binding pocket positional information from the scoring matrix and is thus called 'the position-dependent feature set' (*DynaPred*^{POS}), and a global feature matrix, for which the information from the BFESM is reduced, assuming that the positional information can be neglected and that the same feature can be summed up over all residues to give one value for each feature for each peptide. This model, the 'position-independent feature set' (*DynaPred*), can best be compared to the global features used in (Doytchinova *et al*., 2005; Bordner and Abagyan, 2006). Both features sets were tested. For the training of the support vector machines a radial kernel function were applied. The models were implemented in R (R Devel. Core Team, 2005).

2.4.3 Data sets Two publicly available data sets were used in our study: MHCPEP and SYFPEITHI. MHCPEP is a static database of MHC peptide sequences (Brusic *et al*., 1998a). Non-binding data was obtained from the author upon request. SYFPEITHI (Rammensee *et al*., 1999) is an online database with over 4500 sequences and 250 motifs from naturally processed peptides and T-cell epitopes. Since we have focused on binary classification, all 9-mer binding sequences are considered as binders, regardless of their binding specificity. Duplicated or contradicting entries were removed. Since SYFPEITHI contains only binders, the non-binding sequences from the MHCPEP data sets were included for prediction. The training of the two models was performed on three data set combinations: MHCPEP (binder + non-binder), SYFPEITHI (binder) + MHCPEP (non-binder), and MHCPEP (binder + non-binder) + SYFPEITHI (binder).

2.4.4 Testing and Evaluation We evaluated the overall performance of the prediction models, the robustness against data set size, and the generalizability with respect to other alleles. To evaluate the overall performance leave-one-out cross validation was used. To test the robustness a certain number of sequences was drawn randomly from the MHCPEP data set and each model was tested on this data set by performing 10-fold cross validation. In order to obtain an average accuracy that reflects the performance of the method in that setting, we repeated the 10-fold cross validation 10 times.

Table 2. Data sets (HLA-A*0201 allele) used in this study.

Data set	Binders	Non-binders
MHCPEP	344	383
SYFPEITHI	243	0

To test the generalizability of the prediction models, binding sequences were extracted for other HLA-A-type alleles than A*0201 from the MHCPEP database. To ensure that the results are statistically reasonable, we selected only alleles for which more than 10 unseen sequences (sequences not in the HLA-A*0201 training set) were found in the data base. We collected data for 12 alleles; 6 of them were not subtype-specific.

For evaluation on an independent data set, we chose the HIV-genome and used the prediction models trained on the combined MHCPEP/SYFPEITHI data set. For the prediction we used the complete HIV genome of the HXB2 strain (GenBank accession number K03455). The 3150 residues were divided into MHC binding and non-binding regions according to the HIV-Epitope map (Korber *et al*., 2005). Binding sequences were extracted as indicated on the map, while the non-binding sequences were generated by chopping 9-mer sequences (with eight overlapping positions) from the non-binding regions, and deleting the duplicated entries. Only peptides for which all 9 residues were located in either region (epitope or non-epitope) were included in the performance evaluation.

2.5 Construction of peptide conformations

For the construction of the peptide conformations we calculated the average conformations of the pivot residues from the last 200ps of the simulations. To generate the docked conformation of the peptide, the saved conformations of each residue in the peptide sequence were linked together inside the MHC-binding pocket of PDB structure 1AKJ (same structure as used for the simulations). Then steepest-descent energy minimization was applied to relax first the backbone and then side chains of the peptide. Afterwards the potential energies of the energy-minimized peptide structures were compared to the potential energies of the corresponding experimental structures and the RMSD of the two was calculated.

3 RESULTS

3.1 Simulation results

Simulations were performed for all pseudo-peptide-MHC complexes and the free molecules in solution. One major concern about the use of molecular dynamics for the purpose of sampling side chain conformations is that, due to the limited ability of the MD approach to cross conformational barriers, the conformational space of the residue might not be sampled adequately. We observed that for cases in which the binding pocket size allowed changes in the conformations of the side chains, these conformational changes occurred during the equilibration period of the simulations (leading to all-atom side chain RMSD values up to 3.01 Å). This showed that our approach is capable of sampling the conformational space of the residues in the binding pockets. Nevertheless, after equilibrium was reached all residues were settled at their most favorable conformation. To examine quantitatively the stability of the final conformations of the pseudo-peptides after equilibration, single-linkage-clustering was performed for all pivot residue structures

Table 3. Overall performance of the four prediction models using different data sets (ACC = accuracy (TP + TN)/(TP + TN + FP + FN), SEN = sensitivity (TP/TP + FN), SPC = specificity (TN/FP + TN), AUC = area under the curve (ROC analysis), TP = true positive, TN = true negative, FP = false positive, FN = false negative predictions).

Data set	MHCPEP				SYF + MHCPEP:NB				MHCPEP + SYF			
	ACC	SEN	SPC	AUC	ACC	SEN	SPC	AUC	ACC	SEN	SPC	AUC
SVMHC	0.78	0.81	0.76	0.86	0.81	0.84	0.79	0.90	0.83	0.91	0.73	0.88
YKW0201	0.82	0.89	0.76	0.88	0.81	0.68	0.89	0.91	0.84	0.89	0.76	0.89
DynaPred	0.77	0.77	0.78	0.87	0.78	0.67	0.84	0.85	0.79	0.88	0.66	0.85
$DynaPred^{POS}$	0.85	0.84	0.86	0.91	0.88	0.84	0.91	0.93	0.87	0.90	0.83	0.92

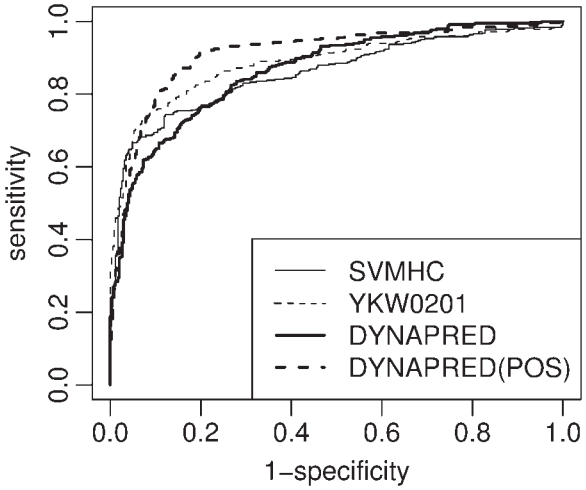

Fig. 2. ROC plots for overall performance evaluation using the MHCPEP data set.

sampled within the last 200ps of the simulations. Using a cutoff of 1.0 Å RMSD, only a single cluster was found for all pseudo-peptide simulations except for three cases. However, in these cases 194 to 199 structures out of 200 belonged to the first cluster and only 1–6 structures (0.5–3.0 %) were different. Since all additional clusters are under-represented, it is clear that the adoption of the corresponding conformations is only a rare event after the equilibrium is reached. Still, it shows that the MD approach is capable of sampling these conformations. Hence, we perceive that the average structure of the last 200ps represents the most favorable conformation of a bound pseudo-peptide in the binding pocket.

3.2 Prediction model

3.2.1 Overall performance To evaluate the performance of the models, we used the 10-fold cross validation or LOO (leave-one-out) techniques and calculated the Receiver Operating Characteristics Curve (ROC) (Sing *et al*., 2005). We compared our models to two models from the literature: the SVMHC model from (Dönnes *et al*., 2002) and the YKW0201 model from (Yu *et al*., 2002). We chose these two models for comparison, because in our method we use the quantitative-matrix approach combined with the SVM method. Thus it seemed sensible to compare our model to other methods using the same techniques, but no structural information. The SVMHC method uses SVM training of a simple binary vector approach, whereas the YKW0201 method uses a quantitative matrix, but no SVM. In addition, the YKW0201 model was previously compared to ANN and HMM methods and showed a comparable performance (Yu *et al*., 2002). Thus there was no need to include these methods into our comparison as well.

Table 3 and Fig. 2 depict the overall performance of the different methods obtained by LOO cross-validation. It can be observed that all methods perform well (>77% accuracy and >0.85 AUC). The SVMHC and YKW0201 methods show comparable performance on all data sets used. Because these methods are position-dependent—like all sequence based methods—they have to be compared to our position-dependent model. It can be observed that for all three data sets our position-dependent model, $DynaPred^{POS}$, outperforms all other models. The same can be seen in the ROC analysis as shown in Fig. 2. This shows that energetic data derived from structural studies are well suited as features for MHC-peptide binding prediction. The performance of the position-independent model, *DynaPred*, is only slightly lower than for the other three methods, despite the fact that no position information is included in this model. This shows that global structural features can be useful for binding prediction, although they do not perform as well as position-dependent models. Nevertheless, the position independent model is extremely robust with respect to the data set used (differs less than 1.5% ACC). The other methods show deviations of up to 5% accuracy (SVMHC) between different data sets.

3.2.2 Robustness Due to the high polymorphism of MHC molecules it is impossible to obtain large experimental binding data sets for all existing alleles. Thus it is important to test the performance of the methods with respect to their robustness against small data sets. In Fig. 3 the performance of the four models is given for different data set sizes. The results show that all methods except SVMHC have a comparably stable performance if the data set has more than 50 binders and 50 non-binders. On the contrary, the SVMHC model is highly dependent on the data volume, which is probably due to its simple binary encoding approach. Again the position-independent model shows the smallest variations above a data set size of 100 binders and 100 non-binders. Overall, the test shows that a data set with at least 100 binders and 100 non-binders is necessary for training a decent prediction model.

3.2.3 HIV-epitope prediction In the evaluation test on the HIV-genome the following accuracies were reached: SVMHC 82.72%,

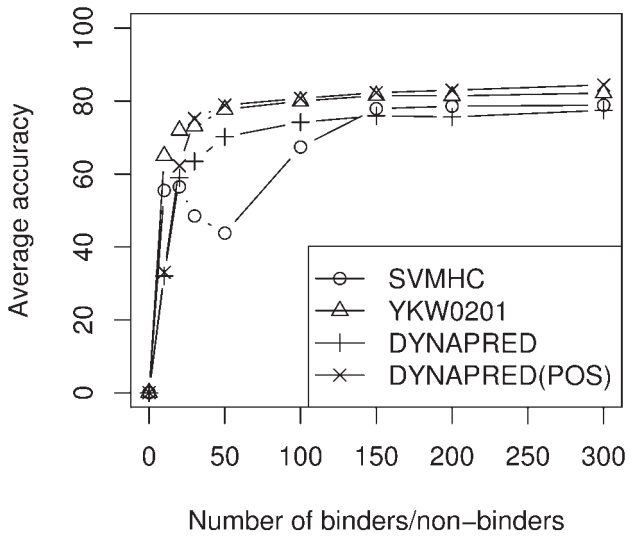

Fig. 3. Accuracy (%) of the 10-fold cross validation results for the training of the four models given in Table 3 using different data set sizes. The numbers correspond to the size of each of the two sets (binders (non-binders)).

YKW0201 82.11%, *DynaPred* 69.68%, *DynaPred*^{POS} 85.45%. Thus, the performance of the three position-dependent models is comparable to their performance on the training data set. The position-independent model shows a lower performance, which is surprising, because in all other tests it proved to be the more robust of our two models. Overall, the data shows that our models do perform nearly as well on independent data set as on the training data.

3.2.4 Generalizability The last test we performed on the four methods was a generalizability test on different HLA-A-type alleles. In Fig. 4 the percentage of correctly predicted binders by the models trained on the combined data set is given for the different alleles. It can be observed that in general the position-independent model considerably outperforms all other models, except for A2, which is the supertype of HLA-A*0201, and thus contains mainly HLA-A*0201 sequences. The prediction capabilities of SVMHC, YKW0201, and our position-dependent model are mostly between 10–30% implying that cross-allele prediction is not feasible for them.

3.3 Construction of peptide conformations

The last step of our prediction algorithm is the construction of bound peptide conformations for all predicted binding sequences. For testing this step, we generated bound conformations for all peptide sequences of the structures given in Table 4 by connecting the saved residue conformations from the simulation runs and performing a short energy minimization. At this point we abstained specifically from further structural refinement, because we wanted to evaluate two points crucial for our method: First, are we able to construct a decent peptide backbone structure by simply 'stitching' together the pivot residue conformations and subsequent energy minimization. This was not obvious at the beginning, because the pseudo-peptide backbone was still able to move within a few Å even with the restraints applied. Second, if we were able to do so, is the overall energy of the constructed peptides comparable to the energies of the experimental peptide structures. This would be a prove for the validity of our additive single residue approach and in addition, is a prerequisite for a possible use of the constructed peptides for further refinement and the calculation of binding affinities from the complex structures.

For this evaluation, we calculated the backbone RMSD and the differences in the potential energies between the constructed peptides in the binding groove and the peptides in the experimental structures. The RMSD values are given in Table 4. RMSD^{1AKJ} provides a measure for the difference between the backbones of the experimental structures and the backbone of 1AKJ. RMSD^{Gen} compares the backbone of the constructed peptides to the crystal structures. Comparing the data shows that the deviation of our constructed backbone structures from the experimental structures is comparable to the variation between the experimental structures. This proves that even with this rather simple approach we can generate decent backbone peptide structures based on our residue conformations.

To investigate the correlation between the energies of the constructed and experimental structures, we compared the potential energies of the bound peptide structures for both sets. The correlation plot is shown in Fig. 5. A correlation of 0.81 was found between the energies, validating that the energies derived from our single residue conformations are suited for prediction and ranking purposes. However, a general energy setoff can be observed in the plot. This is due to two reasons: First, a rather high average RMSD value (3.8 Å, data not shown) for the solvent exposed side chains was observed. The treatment of these residues posed also a problem for all previously reported structural studies, due to the lack of solvent. Thus in most studies the RMSD of the solvent exposed residues is either in the same range as reported here or these residues are placed according to known X-ray structure conformations. Second, during the minimization of the peptide backbone, the side chain conformations are distorted. Due to the simple refinement strategy used, the side chains might not re-equilibrate into their global, minimum conformation, but rather a local minimum. However, the average RMSD for the buried anchor side chains is only 1.1 Å. Thus, both RMSD values—backbone and buried side chains—are in the same range as in other docking approaches.

4 DISCUSSION

We present a new combined structure- and sequence-based method for the prediction of MHC-binding peptides, in which residue-based energy terms from MD simulations are used as features to train a position-dependent (*DynaPred*^{POS}) and a position-independent (*DynaPred*) prediction model for peptide-MHC class I binding using SVMs. The performance of the prediction models was tested successfully on the HIV genome as an independent test set. Our position-dependent model outperforms the two other sequence-based models in our evaluation, validating that structure based energies are well suited as features for binding prediction. The position-independent model showed a lower performance (∼5% accuracy) than the position-dependent models, but had a much higher generalizability towards other HLA-A-type alleles. This is in agreement with the performance of other prediction models based only on global features (Doytchinova *et al*., 2005; Bordner and Abagyan, 2006). The high generalizability of methods based on global features can be explained by the fact that for HLA-subtypes

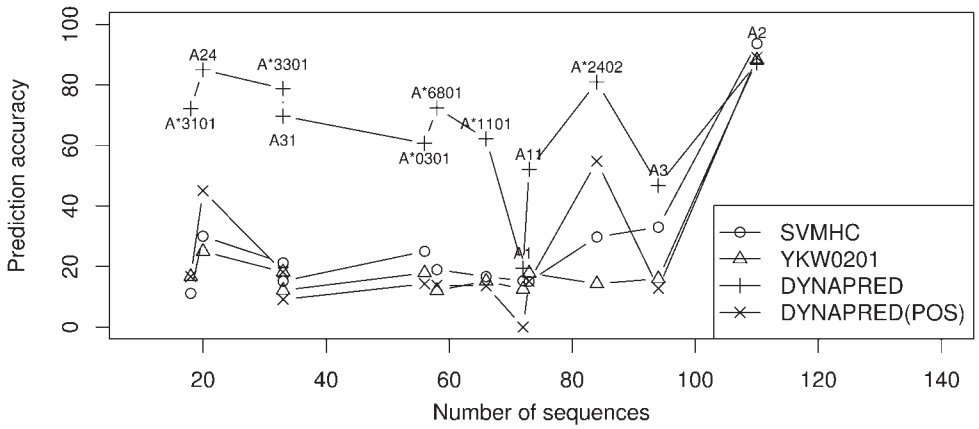

Fig. 4. Correctly predicted binders (%) by the four prediction models from Table 3 trained on the combined data set (MHCPEP + SYF) on various HLA-A-type alleles. The alleles are ordered according to the number of peptide sequences available for the specific allele.

Table 4. Backbone-RMSD (Å) between the generated peptides and the crystal structures ($RMSD^{Gen}$) and between the experimental structures and 1AKJ ($RMSD^{1AKJ}$).

PDB	Resolution	Sequences	$RMSD^{1AKJ}$	$RMSD^{Gen}$
1AKJ	2.65	ILKEPVHGV	0.00	1.18
1AO7	2.60	LLFGYPVYV	1.22	1.58
1B0G	2.50	ALWGFFPVL	1.23	1.41
1BD2	2.50	LLFGYPVYV	1.24	1.59
1DUZ	1.80	LLFGYPVYV	1.33	1.68
1HHG	2.60	TLTSCNTSV	1.67	1.38
1HHI	2.50	GILGFVFTL	1.38	1.67
1HHJ	2.50	ILKEPVHGV	0.52	1.29
1HHK	2.50	LLFGYPVYV	1.29	1.69
1I1F	2.80	FLKEPVHGV	0.50	1.32
1I1Y	2.20	YLKEPVHGV	0.66	1.41
1I7R	2.20	FAPGFFPYL	1.32	1.57
1I7T	2.80	ALWGVFPVL	1.17	1.62
1I7U	1.80	ALWGVFPVL	1.26	1.76
1IM3	2.20	LLFGYPVYV	1.29	1.61
1JHT	2.15	ALGIGILTV	1.46	1.32
1OGA	1.40	GILGFVFTL	1.62	1.67
1QRN	2.80	LLFGYAVYV	1.29	1.84
1QSE	2.80	LLFGYPRYV	1.17	1.33
1QSF	2.80	LLFGYPVAV	1.10	1.63
Average RMSD			1.14	1.53

often only one or two of the nine binding pockets have different binding site residues. These local differences do strongly affect position-dependent methods, but are averaged out by the use of global features. The generalizability of prediction methods is highly desirable because of the polymorphism of the MHC molecules and the need of 'supertype' MHC binders for purposes like vaccine design. In addition, there is still a severe lack of experimental binding data for less common HLA-types, thus preventing the training of prediction models for these types. This makes highly generalizable models, which also work for these HLA-types, an alternative. However, the generalizability comes at the price of lower accuracy (about 5% less). Thus our prediction approach,

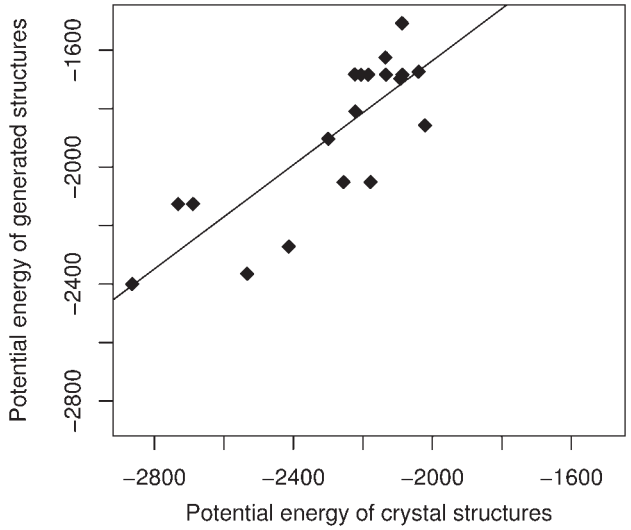

Fig. 5. Correlation between the potential energy of the constructed structures and the crystal structures (kJ/mol).

which uses the same features for position-dependent and position-independent prediction models, and thus allows using either model depending on the allele and purpose of the study, may be an attractive choice.

We showed that with our molecular-dynamics-based approach it is possible to sample the residues conformational space within each binding pocket adequately. Based on these simulations, we are able to construct decent conformations of bound peptides, which have RMSD values that are comparable with the results of other docking studies. In addition, the correlation between the potential energies of the constructed peptides and the potential energies for the corresponding experimental structures is as high as 0.81. This validates the use of our equilibrated residue structures for prediction as well as for peptide construction and is in agreement with a former study, in which a linear energy approach based on MD simulations performed very well for the calculation of free energies of binding of small peptides to OppA (Wang *et al.*, 2002). The low RMSD values and high energy correlation obtained

for the constructed peptides are very promising, especially considering that only a simple approach of concatenation and energy minimization was used. Due to the pre-calculation of the residue structures, the concatenation and minimization is extremely fast, compared with other docking methods. Thus our method provides a fast alternative to generate the initial docked structure which can be refined subsequently for binding affinity prediction. However, the efficiency of our method comes at a price, which is the necessity to pre-calculate the bound conformations of the single residues for each binding pocket. Thus to use our method for other protein targets these conformations must first be calculated for this target. This distinguishes our method considerably from other docking methods such as Liu *et al*. 2004 and Bordner *et al*. 2006. However, the purpose of this work was not to develop a general protein-peptide docking method, but to improve MHC/peptide binding prediction by the use of structural features. For this purpose the higher compute-time efficiency of our approach is more important than transferability. In addition, it is still an open question to what extend new simulations need to be performed to compute bound peptide conformations for other MHC alleles, especially if the allele-specific binding pocket mutations are conservative or only one or two side chains differ in the binding pocket. There are several other interesting topics to be investigated in the future: For example, like all other structural approaches, we are experiencing problems with the treatment of the solvent-exposed residues. To solve this problem, further refinement strategies should be investigated. In addition, the performance of our method for regression should be evaluated and it would be interesting to try to further improve the accuracy of the structure based position-independent model.

ACKNOWLEDGEMENTS

We thank V. Brusic for providing the non-binding data of the MHCPEP data base, K. Roomp and J. Rahnenführer for many inspiring and helpful discussions, and J. Büch for technical support. Funding was obtained from the IMPRS program of the Max-Planck-Society.

REFERENCES

Aqvist,J., Luzhkov,V.B. and Brandsdal,B.O. (2002) Ligand Binding Affinities from MD Simulations. *Acc. Chem. Res.,* **35**, 358–365.

Altuvia,Y., Schueler,O. and Margalit,H. (1995) Ranking potential binding peptides to MHC molecules by a computational threading approach. *J. Mol. Biol.* **249**, 244–250.

Bhasin,M. and Raghava,G.P.S. (2004) Prediction of CTL epitopes using QM, SVM and ANN techniques. *Vaccine,* **22**, 3195–3204.

Bordner,A.J. and Abagyan,R. (2006) Ab initio prediction of peptide-MHC binding geometry for diverse class I MHC allotypes. *Proteins*, in press.

Brusic,V., Rudy,G., Kyne,A.P. and Harrison,L. (1998a) MHCPEP, a database of MHC-binding peptides (updated 1997). *Nucl. Acid. Res.,* **26**, 368–371.

Brusic,V., Rudy,G., Honeyman,M., Hammer,J. and Harrison,L. (1998b) Prediction of MHC Class II binding peptides using an evolutionary algorithm and artificial neural network. *Bioinformatics,* **14**, 121–130.

Bui,H.H., Sidney,J., Peters,B.M., Sinichi,A., Purton,K.A., Mothé,B.R., Chisari,F.V., Watkins,D.I. and Sette,A. (2005) Automated generation and evaluation of specific MHC binding predictive tools: ARB matrix applications. *Immunogenetics,* **57**, 304–314.

Canutescu,A.A., Shelenkov,A.A. and Dunbrack,R.L.,Jr. (2003) A graph theory algorithm for protein side-chain prediction. *Protein Science,* **12**, 2001–2014.

Chothia,C. (1974) Hydrophobic bonding and accessible surface area in proteins. *Nature,* **248**, 338–339.

Desmet,J., Wilson,I.A., Joniau,M., Maeyer,M. and Lasters,I. (1997) Computation of the binding of fully flexible peptides to proteins with flexible side chains. *FASEB J.,* **11**, 164–172.

Doytchinova,I.A., Blythe,M.J. and Flower,D.R. (2002) Additive method for the prediction of protein-peptide binding affinity: Application to the MHC class I molecule HLA-A*0201. *J Proteome Res,* **1**(3), 263–272.

Doytchinova,I.A., Guan,P. and Flower,D.R. (2004) Quantitative structure-activity relationships and the prediction of MHC supermotifs. *Methods*, **34**, 444–453.

Doytchinova,I.A., Walshe,V., Borrow,P. and Flower,D.R. (2005) Towards the chemometric dissection of peptide—HLA-A*0201 binding affinity: Comparison of local and global QSAR models. *J. Comp-Aided Mol. Design,* **19**, 203–212.

Dönnes,P. and Elofsson,A. (2002) Prediction of MHC class I binding peptides using SVMHC. *BMC Bioinformatics,* **3**, 25.

Dunbrack,R.L., Canutescu,A.A. and Shelenkov,A.A. (2003) A graph theory algorithm for protein side-chain prediction. *Prot. Sci.* **12**, 2001–2014.

Gulukota,K., Sidney,J., Sette,A. and DeLisi,C. (1997) Two complementary methods for predicting peptides binding major histocompatibility complex molecules. *J. Mol. Biol.,* **267**, 1258–1267.

Hammer J. (1995) New methods to predict MHC-binding sequences within protein antigens. *Curr. Opin. Immunol.,* **7**, 263–269.

Korber,B.T.M., Brander,C., Haynes,B.F., Koup,R., Moore,J.P., Walker,B.D. and Watkins,D.I. (ed.) (2005) *HIV Molecular Immunology 2005*, Los Alamos National Laboratory, Theoretical Biology and Biophysics, Los Alamos, New Mexico. LA-UR 06-0036.

Lindahl,E., Hess,B. and Spoel,D. (2001) GROMACS 3.0: A package for molecular simulation and trajectory analysis. *J. Mol. Modelling* **7**, 306–317.

Liu,Z., Dominy,B.N. and Shakhnovich,E.I. (2004) Structural Mining: Self-Consistent Design on Flexible Protein-Peptide Docking and Transferable Binding Affinity Potential. *J. Am. Chem. Soc.,* **126**(27), 8515–8528.

Logean,A. and Rognan,D. (2002) Recovery of known T-cell epitopes by computational scanning of a viral genome. *J. Comp-Aided Mol. Design,* **16**, 229–243.

Mamitsuka,H. (1998) Predicting peptides that bind to MHC molecules using supervised learning of Hidden Markov Models. *Proteins: Structure, Function and Genetics,* **33**, 460–474.

Milik,M., Sauer,D., Brunmark,A.P., Yuan,L., Vitiello,A., Jackson,M.R., Peterson,P.A., Skolnick,J. and Glass,C.A. (1998) Application of an artificial neural network to predict specific class I MHC binding peptide sequences. *Nat. Biotech.,* **16**(8): 753–756.

Nielsen,M., Lundegaard,C., Worning,P., Hvid,C.S., Lamberth,K., Buus,S., Brunak,S. and Lund,O. (2004) Improved prediction of MHC class I and class II epitopes using a novel Gibbs sampling approach. *Bioinformatics,* **20**(9), 1388–1397.

Ota,N. and Agard,DA. (2001) Binding mode prediction for a flexible ligand in a flexible pocket using multi-conformation simulated annealing pseudo crystallographic refinement. *J. Mol. Biol.,* **314**, 607–617.

Parker,K.C., Bednarek,M.A. and Coligan,J.E. (1994) Scheme for ranking potential HLA-A2 binding peptides based on independent binding of individual peptide side chains. *J. Immunol.,* **152**, 163–175.

Paul,W.E. (ed.) (1998) Fundamental Immunology, 4th Edn. Raven Press, New York, NY.

Peters,B., Tong,W., Sidney,J., Sette,A. and Weng,Z. (2003) Examining the independent binding assumption for binding of peptide epitopes to MHC-I molecules. *Bioinformatics,* **19**(14), 1765–1772.

Pickett,S.D. and Sternberg,M.J. (1993) Empirical scale of side-chain conformational entropy in protein folding. *J. Mol. Biol.,* **231**, 825–839.

Rammensee,H.G., Bachman,J., Philipp,N., Emmerich,N., Bachor,O.A. and Stevanovic,S. (1999) SYFPEITHI: a database for MHC ligands and peptide motifs. *Immunogenetics,* **50**, 3–9.

Devel R. Core Team. (2005) R: A Language and Environment for Statistical Computing. Vienna, Austria. http://www.R-project.org.

Reche,P.A., Glutting,J.P. and Reinherz,E.L. (2002) Prediction of MHC class I binding peptides using profile motifs. *Hum. Immunol.,* **63**, 701–709.

Rosenfeld,R., Zheng,Q., Vajda,S. and DeLisi,C. (1993) Computing the structure of bound peptides—Application to antigen recognition by class I major histocompatibility complex receptors. *J. Mol. Biol.,* **234**, 515–521.

Schueler-Furman,O., Elber,R. and Margalit,H. (1997) Knowledge-based structure prediction of MHC class I bound peptides: A study of 23 complexes. *Fold. Des.* **3**, 549–564.

Segal,M.R., Cummings,M.P. and Hubbard,A.E. (2001) Relating amino acid sequence to phenotype: Analysis of peptide-binding data. *Biometrics,* **57**, 632–642,

Sette,A., Buus,S., Appella,E., Smith,J.A., Chesnut,R., Miles C.,Colon SM. and Grey HM. (1989) Prediction of major histocompatibility complex binding regions

of protein antigens by sequence pattern analysis. *Proc. Natl Acad. Sci. USA,* **86**, 3296.

Sezerman,U., Vajda,S. and DeLisi,C. (1996) Free energy mapping of class I MHC molecules and structural determination of bound peptides. *Prot. Sci.,* **5**, 1272–1281.

Sing,T., Sander,O., Beerenwinkel,N, and Lengauer,T. (2005) ROCR: visualizing classifier performance in R, *Bioinformatics,* **21**(20), 3940–3941.

Southwood,S., Sidney,J., Kondo,A., Guercio,M., Appella,E., Hoffman,S., Kubo,R.T., Chesnut,R.W., Grey,H.M. and Sette,A. (1998) Several common HLA-DR types share largely overlapping peptide binding repertoires. *J. Immunol.,* **160**, 3363–3373.

Ting,W. and Wade,R.C. (2002) Comparative Binding Energy (COMBINE) Analysis of OppA-Peptide Complexes Relate Structure to Binding Thermodynamics. *J. Med. Chem.,* **45**, 4828–4837.

Tong,J.C., Tan,T.W. and Ranganathan,S. (2004) Modeling the structure of bound peptide ligands to Major Histocompability Complex. *Prot. Sci.,* **13**, 2523–2532.

Yang Z.R. and Johnson, F.C. (2005) Prediction of T-Cell Epitopes Using Biosupport Vector Machines. *J. Chem. Inf. Model.,* **45**, 1424–1428.

Yu,K., Petrovsky,N., Schonbach,C., Koh,J.Y. and Brusic, V. (2002) Methods for prediction of peptide binding to MHC molecules: A comparative study. *Molecular Medicine*, **8(3)**, 137–148.

Zhao,Y., Pinilla,C., Valmori,D., Martin,R. and Simon,R. (2003) Application of support vector machines for T-cell epitopes prediction. *Bioinformatics,* **19**(15), 1978–1984.

Zhihua,L., Yuzhang,W., Bo, Z., Bing,N. and Li,W. (2004) Toward the Quantitative Prediction of T-Cell Epitopes: QSAR Studies on Peptides Having Affinity with the Class I MHC Molecular HLA-A*0201. *J. Comp. Biol.,* **11**(4), 683–694.