Research paper

Prediction of supertype-specific HLA class I binding peptides using support vector machines

Guang Lan Zhang a,b, Ivana Bozic c, Chee Keong Kwoh b, J. Thomas August d, Vladimir Brusic e,⁎

a Institute for Infocomm Research, 21 Heng Mui Keng Terrace, Singapore 119613, Singapore
b School of Computer Engineering, Nanyang Technological University, Block N4, Nanyang Avenue, Singapore 639798, Singapore
c Faculty of Mathematics, University of Belgrade, Belgrade, Serbia and Montenegro
d Department of Pharmacology and Molecular Sciences, Johns Hopkins School of Medicine, Baltimore, MD 21205, USA
e Cancer Vaccine Center, Dana-Farber Cancer Institute, Boston, MA 02115, USA

Received 27 November 2006; accepted 20 December 2006
Available online 25 January 2007

Abstract

Experimental approaches for identifying T-cell epitopes are time-consuming, costly and not applicable to the large scale screening. Computer modeling methods can help to minimize the number of experiments required, enable a systematic scanning for candidate major histocompatibility complex (MHC) binding peptides and thus speed up vaccine development. We developed a prediction system based on a novel data representation of peptide/MHC interaction and support vector machines (SVM) for prediction of peptides that promiscuously bind to multiple Human Leukocyte Antigen (HLA, human MHC) alleles belonging to a HLA supertype. Ten-fold cross-validation results showed that the overall performance of SVM models is improved in comparison to our previously published methods based on hidden Markov models (HMM) and artificial neural networks (ANN), also confirmed by blind testing. At specificity 0.90, sensitivity values of SVM models were 0.90 and 0.92 for HLA-A2 and -A3 dataset respectively. Average area under the receiver operating curve (AROC) of SVM models in blind testing are 0.89 and 0.92 for HLA-A2 and -A3 datasets. AROC of HLA-A2 and -A3 SVM models were 0.94 and 0.95, validated using a full overlapping study of 9-mer peptides from human papillomavirus type 16 E6 and E7 proteins. In addition, a large-scale experimental dataset has been used to validate HLA-A2 and -A3 SVM models. The SVM prediction models were integrated into a web-based computational system MULTIPRED1, accessible at antigen.i2r.a-star.edu.sg/multipred1/.

© 2007 Elsevier B.V. All rights reserved.

Keywords: T-cell epitope; Human Leukocyte Antigen supertype; Promiscuous binding peptide; Support vector machines

1. Introduction

Cellular immunity in vertebrates is mediated by T cells of the immune system which generate highly specific and lasting immune responses to pathogens (Fabbri et al., 2003). T-cell-based immune responses are mediated by antigenic peptides presented by major histocompatibility complex (MHC) molecules (Pamer and Cresswell, 1998; Yewdell and Bennink, 2001). Antigenic peptides bind MHC molecules and form peptide/MHC complexes. Peptide/MHC complexes shown to be recognized by T cells are called T-cell
epitopes. Identifying promiscuous peptides that bind multiple Human Leukocyte Antigen (HLA, human MHC) alleles is a basis for T cell epitope mapping and epitope-based vaccine development (Berzofsky et al., 2001; Srinivasan et al., 2004a; De Groot, 2006). HLA genes are the most polymorphic human genes known (Williams, 2001), with more than 2400 allelic variants identified in the human population as of July 2006 (Williams et al., 2001; Srinivasan et al., 2004a; De Groot, 2006). Because of the high HLA polymorphism, identifying promiscuous peptides that bind more than one HLA allele is essential for the development of vaccines with a broad and unbiased coverage of the human population. HLA alleles that share sequence similarity and that bind largely overlapping sets of peptides define HLA supertypes (Sette and Sidney, 1999; Doytchinova et al., 2004; Lund et al., 2004). Promiscuous peptides have been reported in the context of HLA supertypes (Threlked et al., 1997; Wilson et al., 2003; Srinivasan et al., 2004b). Epitope-based vaccines show great potential in fighting infectious diseases (Sette et al., 2000; Ada, 2003; Wilson et al., 2003), and they are also investigated for control of cancers, allergy, autoimmunity, and even dementia (Alexander et al., 2002; Durrant and Ramage, 2005; Quintana and Cohen, 2005; Verhagen et al., 2005; Wisniewski and Frangione, 2005; De Groot, 2006).

Experimental validation of peptide binding to HLA molecules is time-consuming and costly, and thus not applicable to large scale screening across multiple HLA alleles. Computational methods are instrumental for systematic large-scale identification of MHC-binding peptides (Schirle et al., 2001; Brusic et al., 2004). One type of methods is structure-based approach that relies on structural conservation observed in 3D structure of peptide–MHC complexes (Schueler-Furman et al., 2000; Bui et al., 2006; Tong et al., 2006). These methods are computationally intensive, and have mainly been applied to MHC molecules with known crystal structures. Data-driven approaches include statistical methods based on experimental peptide binding measurements. These methods include binding motifs (Rammensee et al., 1993), quantitative matrices (Parker et al., 1994; Singh and Raghava, 2003; Reche and Reinherz, 2005; Peters and Sette, 2005), artificial neural networks (ANN) (Honeyman et al., 1998; Christensen et al., 2003), hidden Markov models (HMM) (Mamitsuka, 1998; Brusic et al., 2002), decision trees (Savoie et al., 1999; Segal et al., 2001), discriminant analysis (Mallios, 2001), multivariate regression (Lin et al., 2004), ensemble classifier (Xiao and Segal, 2005), support vector machines (SVM) (Donnes and Elofsson, 2002; Zhao et al., 2003; Bhasin and Raghava, 2004; Riedesel et al., 2004; Bozic et al., 2005; Liu et al., 2006; Cui et al., 2007), and biosupport vector machine which is modified from a conventional support vector machine by introducing a biobasis function so that the non-numerical attributes of amino acids can be recognized without a feature extraction process (Yang and Johnson, 2005). Recently a structure- and sequence-based method was reported, in which residue-based energy terms from the molecular dynamics simulations are used as features to train SVM prediction models for peptide/MHC class I binding (Antes et al., 2006).

SVM-based models showed higher accuracy than other prediction methods in studies of peptide binding to a single HLA molecule. We have employed SVM models with a novel data representation, which captures information of the interaction between a peptide and an HLA molecule and allows the use of a single model for prediction of peptide binding to a multiplicity of alleles that belong to a particular HLA supertype. Earlier we reported the application of HMM (Brusic et al., 2002) and ANN (Zhang et al., 2005b) for prediction of peptide binding to the HLA-A2 supertype. A web-based prediction system, MULTIPRED (Zhang et al., 2005a), was developed using HMM and ANN models. In this study we extended MULTIPRED by applying SVM models. The SVM-MULTIPRED was applied to prediction of HLA class I supertype-specific promiscuous binding peptides in the context of HLA-A2 and -A3. Extensive testing, including blind testing and 10-fold cross-validation, were performed to assess the performance of the prediction models. Validation of the models was conducted using experimental data from human papillomavirus (HPV) type 16 E6 and E7 proteins and a large-scale experimental dataset made available recently by Peters et al. (2006). The performance of the SVM models were compared with that of HMM and ANN models. MULTIPRED1 is the updated version of MULTIPRED (Zhang et al., 2005a). MULTIPRED1 is accessible at antigen.i2r.a-star.edu.sg/multipred1/.

2. Materials and methods

2.1. Data and data representation

Nine-mer peptide data were extracted from the MHCEP database (Brusic et al., 1994), published articles, and a set of HLA non-binding peptides (Brusic, V. unpublished data). The HLA-A2 supertype dataset, named as Dataset1, has 3050 peptides (664 binders and 2386 non-binders) related to 15 alleles (Table 1) of
HLA-A2 supertype and the HLA-A3 supertype dataset, named Dataset2, has 2216 peptides (680 binders and 1536 non-binders) related to eight alleles (Table 2) of HLA-A3 supertype. Nine-mer peptides were used in building models because the predominant length of peptides that bind HLA-A2 and -A3 (class I) alleles is nine-amino-acid long (Rammensee et al., 1993). The datasets are available for download at antigen.i2r.a-star.edu.sg/multipred1/data.

To model the interaction between a peptide and a HLA molecule, a peptide/HLA interaction is represented by a virtual peptide composed of peptide residues and HLA residues that come in contact with the peptide (Zhang et al., 2005b). The HLA protein sequences were extracted from the IMGT/HLA Sequence Database, release 2.4.0 (http://www.anthonynolan.org.uk/HIG/). To simplify the data representation and eliminate redundant information, we considered only those contact residues that vary across the HLA-A2/A3 alleles appearing in Dataset1/Dataset2, and discarded the residues that are conserved across those HLA-A2/A3 alleles. The contact residues conserve across the alleles will not provide any useful information to feature vectors, as they are same in all feature vectors. There are in total 48 peptide contact residues of which 29 are non-conserved across HLA-A variants (Chelvanayagam, 1996; Brusic et al., 2002). Of 29 non-conserved residues, 12 are non-conserved across the 15 HLA-A2 alleles, see Table 3. By combining the 9-mer peptide and the 12 non-conserved contact residues we defined a virtual peptide. The virtual peptide has 21 amino acids capturing the interaction information between a peptide and a HLA-A2 allele: P1-P2-P3-P4-P5-P6-P7-P8-P9-R9-R62-R63-R66-R70-R73-R74-R95-R97-R99-R152-R156. P denotes peptide residues and R denotes HLA-A contact residues. Of the 29 non-conserved residues, 12 are non-conserved across the eight HLA-A3 alleles, see Table 4. There are 20 amino acids and each of them can be encoded as a binary string of length 20 with a unique position set to “1” and other positions set to “0.” For example, the first two amino acids, alanine (A) and cysteine (C) are encoded by 10000000000000000000 and 01000000000000000000 respectively, and the last amino acid tyrosine (Y) is encoded by 00000000000000000001. A 21-amino-acid-long peptide can be represented as a binary string of length 420.

### Table 1
Number of 9-mer peptides related to 15 HLA alleles belonging to A2 supertype in Dataset1

<table>
<thead>
<tr>
<th>HLA-A2 allele</th>
<th>Binders</th>
<th>Non-binders</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>A*0201</td>
<td>440</td>
<td>1999</td>
<td>2439</td>
</tr>
<tr>
<td>A*0202</td>
<td>45</td>
<td>25</td>
<td>70</td>
</tr>
<tr>
<td>A*0203</td>
<td>46</td>
<td>7</td>
<td>53</td>
</tr>
<tr>
<td>A*0204</td>
<td>23</td>
<td>224</td>
<td>247</td>
</tr>
<tr>
<td>A*0205</td>
<td>16</td>
<td>40</td>
<td>56</td>
</tr>
<tr>
<td>A*0206</td>
<td>43</td>
<td>37</td>
<td>80</td>
</tr>
<tr>
<td>A*0207</td>
<td>4</td>
<td>11</td>
<td>15</td>
</tr>
<tr>
<td>A*0208</td>
<td>0</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>A*0209</td>
<td>5</td>
<td>1</td>
<td>6</td>
</tr>
<tr>
<td>A*0210</td>
<td>3</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>A*0211</td>
<td>4</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>A*0214</td>
<td>8</td>
<td>1</td>
<td>9</td>
</tr>
<tr>
<td>A*0217</td>
<td>2</td>
<td>4</td>
<td>6</td>
</tr>
<tr>
<td>A*6802</td>
<td>23</td>
<td>31</td>
<td>54</td>
</tr>
<tr>
<td>A*6901</td>
<td>2</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>Total</td>
<td>664</td>
<td>2386</td>
<td>3050</td>
</tr>
</tbody>
</table>

### Table 2
Number of 9-mer peptides related to eight HLA alleles belonging to A3 supertype in Dataset2

<table>
<thead>
<tr>
<th>HLA-A3 allele</th>
<th>Binders</th>
<th>Non-binders</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>A*0301</td>
<td>107</td>
<td>89</td>
<td>196</td>
</tr>
<tr>
<td>A*0302</td>
<td>146</td>
<td>259</td>
<td>405</td>
</tr>
<tr>
<td>A*1101</td>
<td>142</td>
<td>223</td>
<td>365</td>
</tr>
<tr>
<td>A*1102</td>
<td>142</td>
<td>211</td>
<td>353</td>
</tr>
<tr>
<td>A*3101</td>
<td>44</td>
<td>54</td>
<td>98</td>
</tr>
<tr>
<td>A*3301</td>
<td>35</td>
<td>62</td>
<td>97</td>
</tr>
<tr>
<td>A*3303</td>
<td>5</td>
<td>0</td>
<td>5</td>
</tr>
<tr>
<td>A*6801</td>
<td>59</td>
<td>638</td>
<td>697</td>
</tr>
<tr>
<td>Total</td>
<td>680</td>
<td>1536</td>
<td>2216</td>
</tr>
</tbody>
</table>

2.2. Support vector machines

SVMs are popular due to attractive features and their promising performance. SVMs implement a simple idea: they map pattern vectors to a high-dimensional feature space where a best separating hyperplane can be constructed (Webb, 2002). It provides a linear separation in an augmented space, by means of defined kernels. A SVM is an implementation of the Structural Risk Minimization (SRM) principle, which minimizes the upper bound on the expected generalization error (Vapnik, 1998). The SRM principle was shown to be superior to traditional Empirical Risk Minimization (ERM) principle, employed by conventional neural networks. SRM minimizes an upper bound on the expected risk, as opposed to ERM that minimizes the error on the training data. This difference enables SVMs to generalize well, which is a statistical learning goal (Gunn, 1998).

Here we consider the binary classification task in which we have a set of training patterns \( \{x_i, i = 1, \ldots, n\} \) assigned to one of two classes, \( \omega_1 \) and \( \omega_2 \), with corresponding labels \( g(x) = 0 \). In this study, \( x_i \) is a binary
string representing a 9-mer peptide and class ω₁ and ω₂ correspond to HLA binders and non-binders. The linear discriminate function is:

\[ g(x) = w^T x + w_0 \]  

with decision rule

\[ w^T x + w_0 \begin{cases} < 0 & \text{if } \omega_1 \text{ with corresponding numeric value } y_i = +1 \\ > 0 & \text{if } \omega_2 \text{ with corresponding numeric value } y_i = -1 \end{cases} \]  

(2)

All training points are correctly classified if

\[ y_i (w^T x_i + w_0) \geq 0 \text{ for all } i \]

There may be more than one separating hyperplanes. The maximal margin classifier determines the hyperplane for which the margin — the distance to two parallel hyperplanes on each side of the separating hyperplane — is the largest. The assumption is that the larger the margin, the better the generalization error of the linear classifier defined by the separating hyperplane. To aid generalization, a margin, \( b > 0 \), is introduced. Solution must satisfy the condition:

\[ y_i (w^T x_i + w_0) \geq b \]  

Without loss of generality, a value \( b = 1 \) may be taken, defining canonical hyperplane:

\[ H_1 : \ w^T x_i + w_0 = +1 \] \text{ and } \[ H_2 : \ w^T x_i + w_0 = -1, \] \text{ and have} \[ w^T x + w_0 \begin{cases} \geq +1 & \text{if } \omega_1 \text{ with corresponding numeric value } y_i = +1 \\ \leq -1 & \text{if } \omega_2 \text{ with corresponding numeric value } y_i = -1 \end{cases} \]  

(4)

The distance between each of these two hyperplanes and the separating hyperplane, \( g(x) = 0 \), is \( 1/|w| \) and is termed margin. Maximizing the margin means that we seek a solution that minimize \(|w|\) subject to the constraints:

\[ y_i (w^T x_i + w_0) \geq 1 \text{ for all } i \]  

A standard approach to optimization problems that have equality and inequality constraints is the Lagrange
formalism (Fletcher, 1987). This formalism leads to the primal form of the objective function, \( L_p \), given by

\[
L_p = \frac{1}{2} w^T w - \sum_{i=1}^{n} \alpha_i (y_i (w^T x_i + w_0) - 1)
\]

where \( \{\alpha_i, i=1,...,n; \alpha_i \geq 0\} \) are the Lagrange multipliers. The solution to the problem of minimizing \( w^T w \) subject to constraints (5) is equivalent to determining the saddle point of the function \( L_p \), at which \( L_p \) is minimized with respect to \( w \) and \( w_0 \) and maximized with respect to the \( \alpha_i \). Differentiating \( L_p \) with respect to \( w \) and \( w_0 \) and equating to zero yields:

\[
\sum_{i=1}^{n} \alpha_i y_i = 0 \quad w = \sum_{i=1}^{n} \alpha_i y_i x_i
\]

Substituting into Eq. (6) gives the dual form of the Lagrangian:

\[
L_p = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j x_i^T x_j
\]

which is maximized with respect to the \( \alpha_i \) subject to:

\[
\alpha_i \geq 0 \quad \sum_{i=1}^{n} \alpha_i y_i = 0
\]

The support vector algorithm may be applied in a transformed feature space, \( \phi(x) \), using a nonlinear function. The discriminate function shown in Eq. (1) now changes to:

\[
g(x) = (w^T \phi(x) + w_0)
\]

Acceptable kernels must be expressible as an inner product in a feature space, or \( k(x,y) = \phi^T(x) \phi(y) \). Frequently used kernels are polynomial:

\[
k(x,y) = (1 + x^T y)^d
\]

and radial basis (Gaussian) kernel:

\[
k(x,y) = \exp(-||x-y||/(g^2))
\]

Both polynomial and Gaussian kernels have been tried in our experiments.

2.3. Training, testing and validation

Training of support vector machines was carried out using SVMlight (Joachims, 1999). An input to this package is a binary vector indicating whether a particular amino acid is presented at a particular position and a label with value 1 or 0 indicating if the peptide is a HLA binder or non-binder. Three kernel functions were examined (linear, polynomial and radial basis), and three existing parameters were optimized (trade-off \( c \), power \( d \) for polynomial kernel, and parameter \( g \) for radial kernel). SVM models with kernels and parameter settings were trained and evaluated for each of HLA-A2 and -A3 dataset. Parameter values used during this process were the following: \( c \) (trade-off between training error and margin) was varied from 0.01 to 20, \( d \) (degree in polynomial kernel) from 1 to 10 and \( g \) (as shown in Gaussian kernel) from 0.001 to 1. Models that showed the best performance (one for each supertype) were used for final testing and validation. Although by default threshold 0 represents the separating hyperplane, data available are often imbalanced and not randomly distributed. Moving the decision boundary, which is equivalent to choosing a different threshold, is often used for remedying the imbalanced training-data problem (Wu and Chang, 2003). In our experiments, the decision boundary (threshold) was chosen using the performance measures of the SVM models.

The predictive performance was assessed using sensitivity (SE) and specificity (SP), and receiver operating characteristic (ROC) analysis. TP is the number of true positives (experimental binders predicted as binders), FP the number of false positives (experimental non-binders predicted as binders), TN the number of true negatives (experimental non-binders predicted as non-binders) and FN the number of false negatives (experimental binders predicted as non-binders). SE= TP/(TP+FN) indicates the percentage of correctly predicted binders; SP= TN/(TN+FP) stands for the percentage of correctly predicted non-binders. The ROC curve is a plot of SE against (1-SP) at various classification thresholds (Swets, 1988). The area under the ROC curve (\( A_{ROC} \)) is a measure for the overall prediction performance. Values of \( A_{ROC}<0.7 \) represent poor predictions; \( A_{ROC}>0.8 \) represents good and \( A_{ROC}>0.9 \) represents excellent predictions, while \( A_{ROC}=0.5 \) represents random guessing (Swets, 1988).

Table 5

<table>
<thead>
<tr>
<th>HLA-A2 allele</th>
<th>Training data</th>
<th>Testing data</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Binders</td>
<td>Non-binders</td>
</tr>
<tr>
<td>A*0201</td>
<td>224</td>
<td>378</td>
</tr>
<tr>
<td>A*0202</td>
<td>619</td>
<td>2361</td>
</tr>
<tr>
<td>A*0204</td>
<td>641</td>
<td>2162</td>
</tr>
<tr>
<td>A*0205</td>
<td>648</td>
<td>2346</td>
</tr>
<tr>
<td>A*0206</td>
<td>621</td>
<td>2349</td>
</tr>
</tbody>
</table>
The statistical significance of the comparisons was assessed by t-test. The t-test assesses whether the means of two groups are statistically different from each other (Pagano and Gauvreau, 2000) and is suitable for evaluation of statistical differences for small number of measurements (30 or less in a sample).

Cross-validation is a method for error rate estimation. It implements a simple idea: the dataset of size n samples is partitioned into two parts, the model parameters are estimated using one set and the goodness-of-fit criterion evaluated on the second set. The cross-validation estimates the goodness-of-fit, which identifies how well a statistical model fits a set of observations. In our experiments, 10-fold cross-validation was performed to evaluate the performance of the classifiers. The dataset was randomly divided into 10 sets with approximately equal size. For each “fold”, the classifier is trained using all but one of the 10 groups and then tested on the remaining “unseen” group. This procedure is repeated for each of the 10 groups. The final $A_{ROC}$ is calculated by micro-averaging the results obtained from the 10 runs (TP, TN, FP, and FN were summed up before the calculation of $A_{ROC}$).

In addition to cross-validation, we also performed blind testing for the assessment of the performance of SVM models for prediction of promiscuous HLA-A2 and -A3 binding peptides. For testing purposes, a model was built for each allele. The test set for each model included all peptides related to the allele, while the training set consisted of all peptides related to other HLA alleles from the same supertype. Thus prediction of peptides that bind one HLA allele was performed without any prior knowledge of the peptides binding or not binding to this allele. Because the actual prediction model incorporates all data, the testing results are likely to represent an underestimate of the actual performance. We trained five HLA-A2 and seven HLA-A3 models — the selected molecules are shown in Tables 5 and 6. Other alleles were not used for testing, as there was insufficient experimental data related to them for testing to be valid. The performance of SVM was compared to performances of existing methods, HMM and ANN models.

Stratified 10-fold cross-validation was performed on peptides related to A*0201 and A*0302 molecules. In stratified 10-fold cross-validation, the dataset is randomly divided into 10 sets with approximately equal size and class distributions. And the peptides in the training data that were similar (only one amino acid different) to any peptide from the test set were removed.

A set of 240 9-mer peptides of HPV type 16 E6 and E7 proteins with experimentally identified binding affinity were used for model validation (Kast et al., 1994). Experiments revealed that there are four 9-mer peptides, E6-7, E6-18, E6-26 and E6-52, bind to A*0201 in HPV E6 and seven A*0201 9-mer binders, E7-7, E7-11, E7-12, E7-66, E7-82, E7-85 and E7-86 in HPV E7. There are nine 9-mer peptides, E6-7, E6-33, E6-42, E6-59, E6-75, E6-89, E6-93, E6-125 and E6-143, that bind A*0301 in HPV E6 and one weak A*0301 binder, E7-89 in HPV E7. The training datasets were checked for the duplicate 9-mers peptides pertaining to E6 and E7 proteins and were removed. After removing duplicates, there were 3027 9-mer peptides (651 binders and 2376 non-binders) and 2164 9-mer peptides (653 binders and 1511 non-binders) in the training datasets for HLA-A2 and HLA-A3 models respectively. The comparative performance with other servers for prediction of promiscuous HLA class I peptides was not possible because of unmatched allelic variants that are covered by models in different servers.

A large experimental dataset of quantitative MHC–peptide binding data was recently made available.

---

**Table 6**

Blind testing on seven HLA-A3 alleles: numbers of peptides in training and testing sets

<table>
<thead>
<tr>
<th>HLA-A3 allele</th>
<th>Training data</th>
<th>Testing data</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Binders</td>
<td>Non-binders</td>
</tr>
<tr>
<td>A*0301</td>
<td>573</td>
<td>1447</td>
</tr>
<tr>
<td>A*0302</td>
<td>534</td>
<td>1277</td>
</tr>
<tr>
<td>A*1101</td>
<td>538</td>
<td>1313</td>
</tr>
<tr>
<td>A*1102</td>
<td>538</td>
<td>1325</td>
</tr>
<tr>
<td>A*3101</td>
<td>636</td>
<td>1482</td>
</tr>
<tr>
<td>A*3301</td>
<td>645</td>
<td>1474</td>
</tr>
<tr>
<td>A*6801</td>
<td>621</td>
<td>898</td>
</tr>
</tbody>
</table>

**Table 7**

Number of 9-mer peptides in the Peters dataset related to HLA-A2 alleles; number of overlapping peptides between Dataset1 and the Peters dataset; number of non-overlapping peptides (binders/non-binders) used for testing of the HLA-A2 SVM model

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>The Peters dataset</td>
<td>3089</td>
<td>1447</td>
<td>1443</td>
<td>1437</td>
<td>1434</td>
<td>833</td>
</tr>
<tr>
<td>Overlapping</td>
<td>240</td>
<td>54</td>
<td>48</td>
<td>47</td>
<td>47</td>
<td>0</td>
</tr>
<tr>
<td>Non-overlapping</td>
<td>2849</td>
<td>1393</td>
<td>1395</td>
<td>1390</td>
<td>1387</td>
<td>833</td>
</tr>
<tr>
<td></td>
<td>(1024/1825)</td>
<td>(611/782)</td>
<td>(600/795)</td>
<td>(480/910)</td>
<td>(387/1000)</td>
<td>(86/747)</td>
</tr>
</tbody>
</table>
Using this dataset, they compared the performance of different bioinformatics approaches, including MULTIPRED ANN and HMM models, in predicting MHC binding peptides. The performance of these models has been reported in Tables 2 and 3 of the paper. We used the same dataset for evaluation of our HLA-A2/A3 SVM models. In the Peters dataset, the IC50 value used to separate binders and non-binders is 500 nM, whereas in the Dataset1 and Dataset2, the IC50 value used to separate binders and non-binders is 5000 nM. The difference in the threshold setting results in some discrepancy between our datasets and the Peters dataset. Some identical peptides between Dataset1, Dataset2 and the Peters dataset have been classified into different classes. The binding affinities of such peptides in Dataset1 and Dataset2 were modified to the binding affinities in the Peters dataset and these overlapping peptides were removed from the test sets. The number of 9-mer peptides in the Peters dataset related to HLA-A2/A3 alleles, the number of overlapping 9-mer peptides between Dataset1/Dataset2 and the Peters dataset and the number of non-overlapping peptides used for testing of the HLA-A2/A3 SVM models are shown in Tables 7 and 8 respectively. The testing datasets are available for download at antigen.i2r.a-star.edu.sg/multipred1/data.

### 3. Results

#### 3.1. Cross-validation results

Ten-fold cross-validation was performed on all three methods. The $A_{ROC}$ of ANN models is 0.93 for HLA-A2 dataset and 0.89 for HLA-A3 dataset. The $A_{ROC}$ of HMM models are 0.77 for HLA-A2 dataset and 0.71 for HLA-A3 dataset. The $A_{ROC}$ of SVM models are 0.95 for HLA-A2 dataset and 0.97 for HLA-A3 dataset. The sensitivities of the models were calculated at three levels: SP values of 0.8, 0.9 and 0.95 (Tables 9 and 10). For both HLA-A2 and HLA-A3 datasets, SVM models are of the highest sensitivity. Especially when specificity threshold is high (at 0.95), the sensitivity of SVM models is markedly higher that that of ANN and HMM models. The $A_{ROC}$ values of the models in predicting peptides binding to HLA-A2/A3 alleles are listed in Tables 11 and 12. “Average” is the average of the $A_{ROC}$ values for the individual tests and “Std. dev” is the standard deviation of the measurements. SVM models perform the best on all alleles with average $A_{ROC} = 0.914$ for HLA-A2 alleles and $A_{ROC} = 0.951$ for HLA-A3 alleles. Figs. 1 and 2 show the specificity and sensitivity relationship in 10-fold cross-validation.

#### 3.2. Blind testing results

The $A_{ROC}$ values of the prediction models in blind testing are shown in Tables 13 and 14. The $A_{ROC}$ values of predictions for peptide binding to A*0201, A*0204, A*0205, A*0301, A*1101, A*1102, A*3301, A*6801 were equal to or higher than 0.90. Overall, predictions models performed well ($A_{ROC} = 0.892$ for HLA-A2

### 2.4. Comparison of prediction methods

The performance of the SVM method was compared to the previously reported MULTIPRED models based on HMM (Brusic et al., 2002) and ANN (Zhang et al., 2005b). All models were retrained with the current dataset and assessed using the same train/test partitions.

#### Table 9

<table>
<thead>
<tr>
<th>Specificity</th>
<th>Sensitivity SVM</th>
<th>ANN</th>
<th>HMM</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.80</td>
<td>0.96 (-0.82)</td>
<td>0.95</td>
<td>0.69</td>
</tr>
<tr>
<td>0.90</td>
<td>0.90 (-0.52)</td>
<td>0.84</td>
<td>0.55</td>
</tr>
<tr>
<td>0.95</td>
<td>0.76 (-0.02)</td>
<td>0.55</td>
<td>0.42</td>
</tr>
</tbody>
</table>

The values are given for three levels of specificity, 0.8, 0.9 and 0.95.
alleles and $A_{ROC} = 0.924$ for HLA-A3 alleles). The significance of these results is that training sets in this particular assessment did not include any of the peptides from test sets which contain all the peptides related with an allele at one time. The results corroborate that this method can be used for prediction of good or excellent accuracy of peptide binding to complete supertypes, even for those alleles where no binding data are available.

The $A_{ROC}$ values for A*0202, A*1102, A*3101 and A*3301 prediction models were improved markedly using SVM models (Tables 13 and 14). This might be explained by the fact that sets of peptides related to these molecules are relatively small (less than 100 peptides). SVM are known to outperform other prediction methods on smaller datasets. HMM performed better on A*0201 and A*0301 (however, SVM still has reasonably high $A_{ROC}$ values for these two molecules), and both HMM and ANN models of A*0206 molecule performed better then the corresponding SVM model.

T-test was applied to determine whether the difference in SVM, HMM and ANN predictive performance is statistically significant. The results showed that the performance of SVM is statistically better than the performances of other two methods on HLA-A3 alleles ($P<0.05$). However, the same conclusion could not be drawn for HLA-A2 alleles. This might be due either to similar performance of studied predictive models or to the imbalance in the HLA-A2 dataset. SVM often do not perform the best on imbalanced datasets (Wu and Chang, 2003), and HLA-A2 Dataset1 is imbalanced in two ways: binders/non-binders ratio is close to 1:4, and peptides related to A*0201 present 80% (2439 of 3050), while peptides related to all other alleles present only 20% of the total dataset. The HLA-A3 Dataset2 is more balanced in both aspects, and in this case we observed statistically better performance of SVM than other prediction methods.

3.3. Validation using HPV E6 and E7 proteins

The $A_{ROC}$ values of SVM, HMM, and ANN predictions for HPV E6 and E7 proteins are shown in

![Fig. 1. Sensitivity vs. specificity in 10-fold cross-validation on Dataset1 using SVM, ANN and HMM models.](image1)

![Fig. 2. Sensitivity vs. specificity in 10-fold cross-validation on Dataset2 using SVM, ANN and HMM models.](image2)
Overall, SVM models show better performance on prediction of peptides derived as a full overlapping set from viral proteins E6 and E7 than ANN and HMM. SVM show the highest A\textsubscript{ROC} for HLA-A3 supertype predictions, and SVM and ANN show the same excellent performance on predictions for HLA-A2 supertype.

In summary, we report the improved overall performance of SVM models for prediction of peptide binding to multiple molecules of HLA-A3 supertype, and at least as good performance for HLA-A2 as supported it with exhaustive testing. We have shown that SVM models can predict peptide binding to HLA-A2 and HLA-A3 molecules with good performance, including those alleles for which no experimental binding data are currently available.

### 3.4. Validation using the Peters dataset

The A\textsubscript{ROC} values of HLA-A2 SVM model for predictions of binding peptides for HLA-A2 supertype alleles (A*0201, A*0202, A*0203, A*0206, A*6802 and A*6901) using the Peters dataset (Table 16) show marked improvement for all studied alleles. We note that HLA-A2 SVM model performs well in predicting A*6901 binding peptides (A\textsubscript{ROC}=0.81) although only four peptides related to A*6901 were used for model training. In addition, HLA-A2 SVM model showed better performance in predicting peptides binding to all HLA-A2 alleles than all the external tools evaluated in (Peters et al., 2006) and is equal or better for three of five studied HLA-A3 alleles than the best external tool. We note that only five A*3301-related peptides were in Dataset2 for training of HLA-A3 SVM model.

In Dataset2, there were 697 peptides related to A*6801 with 8.5% of them being binders and 91.5% being non-binders. To understand why HLA-A3 SVM model performed poorly in predicting peptide binding to A*6801, we performed additional analysis at the peptide sequence level. The poor performance of the HLA-A3 SVM model in predicting peptides binding to A*6801 was because the limited number of binders in training dataset do not represent the diverse pattern of A*6801 binding motif. For example, in the training data all binders had K or R at the anchor position 9, while Y, which is present in some binders in the Peters dataset was not present in some binders in the Peters dataset was not represented in the training data. This emphasizes the need for use of representative datasets for development of MHC-binding prediction servers. While the accuracy of predictions will improve by adding new data to the training sets, it is important to note that about a quarter of MHC class I binding peptides lack canonical binding motifs (A. Sette, personal communication). These peptides are underrepresented in the current datasets, which are used both for training and testing, and therefore the values of accuracy of predictive methods are likely to be somewhat lower than reported, across all comparative studies.

### 3.5. Validation using the Gero dataset

The A\textsubscript{ROC} values of HLA-A3 SVM model for predictions of binding peptides for HLA-A3 supertype alleles (A*0301, A*0302, A*1101, A*1102, A*3101, A*3301, A*6801) using the Gero dataset (Table 17) show marked improvement for all studied alleles. We note that HLA-A3 SVM model performs well in predicting A*6801 binding peptides (A\textsubscript{ROC}=0.81) although only four peptides related to A*6801 were used for model training.

### Tables

#### Table 13

<table>
<thead>
<tr>
<th>HLA-A2 allele</th>
<th>SVM</th>
<th>ANN</th>
<th>HMM</th>
</tr>
</thead>
<tbody>
<tr>
<td>A*0201</td>
<td>0.90</td>
<td>0.87</td>
<td>0.93</td>
</tr>
<tr>
<td>A*0202</td>
<td>0.81</td>
<td>0.76</td>
<td>0.73</td>
</tr>
<tr>
<td>A*0204</td>
<td>0.93</td>
<td>0.88</td>
<td>0.92</td>
</tr>
<tr>
<td>A*0205</td>
<td>0.97</td>
<td>0.93</td>
<td>0.94</td>
</tr>
<tr>
<td>A*0206</td>
<td>0.85</td>
<td>0.91</td>
<td>0.88</td>
</tr>
<tr>
<td><strong>Average</strong></td>
<td>0.892</td>
<td>0.87</td>
<td>0.88</td>
</tr>
<tr>
<td><strong>Std.dev.</strong></td>
<td>0.063</td>
<td>0.066</td>
<td>0.087</td>
</tr>
</tbody>
</table>

#### Table 14

<table>
<thead>
<tr>
<th>HLA-A3 allele</th>
<th>SVM</th>
<th>ANN</th>
<th>HMM</th>
</tr>
</thead>
<tbody>
<tr>
<td>A*0301</td>
<td>0.93</td>
<td>0.89</td>
<td>0.94</td>
</tr>
<tr>
<td>A*0302</td>
<td>0.86</td>
<td>0.84</td>
<td>0.86</td>
</tr>
<tr>
<td>A*1101</td>
<td>0.96</td>
<td>0.91</td>
<td>0.91</td>
</tr>
<tr>
<td>A*1102</td>
<td>0.96</td>
<td>0.86</td>
<td>0.86</td>
</tr>
<tr>
<td>A*3101</td>
<td>0.87</td>
<td>0.69</td>
<td>0.66</td>
</tr>
<tr>
<td>A*3301</td>
<td>0.92</td>
<td>0.63</td>
<td>0.58</td>
</tr>
<tr>
<td>A*6801</td>
<td>0.97</td>
<td>0.96</td>
<td>0.95</td>
</tr>
<tr>
<td><strong>Average</strong></td>
<td>0.924</td>
<td>0.83</td>
<td>0.8229</td>
</tr>
<tr>
<td><strong>Std.dev.</strong></td>
<td>0.044</td>
<td>0.144</td>
<td>0.120</td>
</tr>
</tbody>
</table>

#### Table 15

<table>
<thead>
<tr>
<th>HLA-A2</th>
<th>SVM</th>
<th>ANN</th>
<th>HMM</th>
</tr>
</thead>
<tbody>
<tr>
<td>A*0201</td>
<td>0.94</td>
<td>0.94</td>
<td>0.88</td>
</tr>
<tr>
<td>A*0202</td>
<td>0.95</td>
<td>0.80</td>
<td>0.91</td>
</tr>
</tbody>
</table>

#### Table 16

<table>
<thead>
<tr>
<th>HLA-A*</th>
<th>0201</th>
<th>0202</th>
<th>0203</th>
<th>0206</th>
<th>6802</th>
<th>6901</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM \textsubscript{A\textsubscript{ROC}}</td>
<td>0.91</td>
<td>0.83</td>
<td>0.82</td>
<td>0.79</td>
<td>0.74</td>
<td>0.81</td>
</tr>
<tr>
<td>ANN \textsubscript{A\textsubscript{ROC}}</td>
<td>0.88</td>
<td>0.79</td>
<td>0.79</td>
<td>0.74</td>
<td>&lt;0.64</td>
<td>&lt;0.64</td>
</tr>
<tr>
<td>HLA-A*</td>
<td>0301</td>
<td>1101</td>
<td>3101</td>
<td>3301</td>
<td>6801</td>
<td></td>
</tr>
<tr>
<td>SVM \textsubscript{A\textsubscript{ROC}}</td>
<td>0.87</td>
<td>0.89</td>
<td>0.83</td>
<td>0.79</td>
<td>0.76</td>
<td></td>
</tr>
<tr>
<td>ANN \textsubscript{A\textsubscript{ROC}}</td>
<td>0.85</td>
<td>0.87</td>
<td>&lt;0.83</td>
<td>&lt;0.81</td>
<td>&lt;0.77</td>
<td></td>
</tr>
</tbody>
</table>

\textsuperscript{A\textsubscript{ROC}} values of ANN MULTIPRED are taken from Peters et al. (2006).
3.5. MULTIPRED1 — an online computational system for prediction of promiscuous HLA binding peptides

Previously we developed MULTIPRED (Zhang et al., 2005a), a web-based computational system for prediction of peptide binding to multiple molecules belonging to HLA class I A2, A3 and class II DR supertypes. It uses HMM and ANN as predictive engines. With SVM models integrated, an updated version of the system — MULTIPRED1 — was developed.

In MULTIPRED1, the prediction scores produced by SVM models were rescaled to map them into the range from one to nine. The mapping of scores was done according to equation, \( Score_N = \frac{Score - Score_{\text{min}}}{(Score_{\text{max}} - Score_{\text{min}})} \times 8 + 1 \), where \( Score_N \) denotes the normalized score, \( Score \) denotes the raw prediction score produced by SVM models, and \( Score_{\text{max}} \) and \( Score_{\text{min}} \) denote the possible minimum and maximum values of the raw scores respectively. The values for \( Score_{\text{max}} \) and \( Score_{\text{min}} \) were obtained through extensive testing. More than 650 randomly selected protein sequences from the NCBI protein database (contained more than 220,000 9-mer peptides) were used for prediction using the SVM models. Since the testing data contains huge number of 9-mer peptides, the highest and lowest predicted score from the testing data were taken as reasonable maximum and minimum scores for normalization.

4. Discussion and conclusion

Throughout training and testing, we examined different kernels (linear, radial basis and polynomial) and optimized SVM parameters (trade-off \( c \), parameter \( g \) for radial basis kernel and degree \( d \) for polynomial kernel). The models that had the best performances (in terms of the highest average blind testing of \( A_{\text{ROC}} \) value for molecules within a given supertype) were: Gaussian kernel with \( g = 0.1 \) and \( c = 0.5 \) for HLA-A2 and Gaussian kernel with \( g = 0.1 \) and \( c = 2 \) for HLA-A3 supertype.

We report that Gaussian kernel performs best on both HLA-A2 and -A3 molecules. This is in contrast to the report by (Zhao et al., 2003), where linear kernel showed the best performance for T-cell epitope prediction. However, their model was trained only for a single allele (A*0201) and their dataset contained a smaller number of peptides (203). Our results also differ from those reported by Donnes and Elofsson (2002), who trained a different SVM model, including different selection of kernels and parameters, for each allele. These results indicate that future improvements of SVM performance will likely be driven by the increase of datasets, rather than optimization of kernel function and SVM parameters. The main difference between this work and other studies that employed SVM for prediction of HLA binders is that each SVM model built is for prediction of peptides that bind molecules from an entire supertype (HLA-A2 or -A3). In addition, our models can predict peptide binding to HLA-A2 and -A3 alleles for which no binding data are available. For example, HLA-A2 SVM model performs reasonably well in predicting A*6901 binding peptides (\( A_{\text{ROC}}=0.81 \)) when validated using the Peters dataset in spite of only four A*6901-related peptides being used for training the model. However, the limited number of peptides in training dataset may affect the performance of the prediction models, as in the case of A*6801.

Development of epitope-based vaccines is progressing rapidly. Selection of peptides for candidate subunit vaccines is one of the greatest obstacles in the development of vaccines with broad and non-ethnically-biased coverage of the human population. Computational systems that can predict promiscuous peptides complement experimental approaches for overcoming this obstacle.

Acknowledgements

This project has been funded in part (GLZ, JTA, and VB) with the USA Federal funds from the NIAID, NIH, Department of Health and Human Services, under Grant No. 5U19AI56541 and Contract No. HHSN266200400085C.

References

evidence of multiple registers in class II binding peptides.
Bioinformatics 22, 1232.
allergy-directed immunotherapy. Expert Opin. Ther. Targets 9, 217.
and the histocompatibility laboratory. J. Mol. Diagnostics 3, 98.
Wilson, C., McKinney, D., Anders, M., MaWhinney, S., Forster, J.,
Crimi, C., Southwood, S., Sette, A., Chesnut, R., Newman, M.,
Livingston, B., 2003. Development of a DNA vaccine designed to
induce cytotoxic T lymphocyte responses to multiple conserved
Wisniewski, T., Frangione, B., 2005. Immunological and anti-
chaperone therapeutic approaches for Alzheimer disease. Brain
Pathol. 15, 72.
transformation for imbalanced-data learning. Proceedings of the
Twentieth International Conference on Machine Learning (ICML-
2003), Washington, DC.
epitope profiles of HIV-1: classifier choice and peptide represen-
Zhang, G.L., Khan, A.M., Srinivasan, K.N., August, J.T., Brusic, V.,
2005a. MULTIPRED: a computational system for prediction of
promiscuous HLA binding peptides. Nucleic Acids Res. 33, W172.
Zhang, G.L., Khan, A.M., Srinivasan, K.N., August, J.T., Brusic, V.,
2005b. Neural models for predicting viral vaccine targets.
Application of support vector machines for T-cell epitopes