Available online at www.sciencedirect.com

ScienceDirect

Journal of Molecular Graphics and Modelling 26 (2007) 246–254

Journal of Molecular Graphics and Modelling

www.elsevier.com/locate/JMGM

QSAR method for prediction of protein-peptide binding affinity: Application to MHC class I molecule HLA-A*0201

Chunyan Zhao [a], Haixia Zhang [a,*], Feng Luan [a], Ruisheng Zhang [b], Mancang Liu [a], Zhide Hu [a], Botao Fan [c]

[a] Department of Chemistry, Lanzhou University, Lanzhou 730000, China
[b] Department of Computer Science, Lanzhou University, Lanzhou 730000, China
[c] Université Paris 7-Denis Diderot, ITODYS 1, Rue Guy de la Brosse, 75005 Paris, France

Received 5 December 2006; accepted 5 December 2006
Available online 9 December 2006

Abstract

The support vector machine (SVM), which is a novel algorithm from the machine learning community, was used to develop quantitative structure–activity relationship (QSAR) models for predicting the binding affinity of 152 nonapeptides, which can bind to class I MHC HLA-A*201 molecule. Each peptide was represented by a large pool of descriptors including constitutional, topological descriptors and physical–chemical properties. The heuristic method (HM) was then used to search the descriptor space for selecting the proper ones responsible for binding affinity. The four descriptors were obtained to build linear models based on HM and nonlinear models based on SVM method. The best results are found using SVM: root mean-square (RMS) errors for training, test and whole data set were 0.383, 0.385 and 0.384, respectively. This paper allow the prediction of the binding affinity of new, untested peptides and, through the analysis of contribution of each parameter of different residue at specific position of peptidic ligands, to understand nature of the forces governing binding behavior and suggest new ideas for further synthesis of high-affinity peptides.

Keywords: HLA-A*0201; Binding affinity; Amino acids property; QSAR; Heuristic method (HM); Support vector machine (SVM)

1. Introduction

Major histocompatibility complex (MHC) molecules are highly polymorphic cell surface molecules that present peptidic ligands to cells of the T-cell compartment of the immune system and play a critical role in initiating and regulating immune responses. The immune system distinguishes body cells from invading antigens (class I MHC proteins) and immune system cells from other cells (class II MHC proteins) [1–3]. They form stable complexes with proteolytically digested protein fragments composed of approximately 8–10 amino acids. This complex together with the peptide is transferred to the cell surface and can be recognized by T-cells via the T-cell-receptor (TCR) [4,5]. Without presentation of peptides, no immune response against viruses can be initiated which leads to death of the organism and is the strategy of many pathogens [6]. The binding behavior depends on so-called anchor-amino-acids, which bind often with low specificity to the MHC, leaving the residual peptide exposed to the TCR [7,8]. It is a prerequisite for recognition by the T-cells, but only certain peptides can bind to any given MHC molecule. Determining which peptides can bind to specific MHC molecules is fundamental to understanding the basis of immunity and for identifying of candidates for the design of vaccines and immunotherapeutic drugs.

Peptides for class I MHC proteins are typically, but not exclusively, derived from intracellular proteins, which are targeted to the proteasome. Class I MHC molecules bind peptides that are 10–30 amino acids long with a core region of 13 amino acids containing a primary anchor residue [9,10]. Analysis of binding motifs suggests that only a core of nine amino acids within a peptide is essential for peptide/MHC binding [11]. The class I MHC proteins are mainly encoded by three separate but homologous genetic loci, HLA-A, HLA-B, and HLA-C [12]. Among them, the allele HLA-A*0201 is one of the most frequent

* Corresponding author. Tel.: +86 931 891 2578; fax: +86 931 891 2582.
E-mail addresses: zhaocy0225@hotmail.com (C. Zhao), zhanghx@lzu.edu.cn (H. Zhang).

1093-3263/$ – see front matter © 2006 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmgm.2006.12.002

class I alleles in many different populations, which has been demonstrated to play a critical role in antigen presentation of both viral antigens [14] and tumor antigens from a variety of cancers [15–18]. In this paper, a large set of nonapeptides having affinity with the class I MHC HLA-A*0201 molecule was used as objects for studying binding behavior.

Generally, identification of the peptides was done by binding assays in vitro after all possible peptides have been synthesized and tested [19,20]. This is an extremely expensive approach, because even a very small virus encodes a considerable number of medium size proteins. For each of these proteins hundreds of peptides have to be synthesized and their ability to bind at the MHC must be probed in experiment. This often shows that only very few peptides can indeed bind to the MHC and that from thousands of screened peptides only one or two bind with high affinity, which is required for a functional immune response. To simplify the searching process and to explore the behavior adequately, some computer-based methods were generated aiming to reduce the number of peptides, which have to be tested in vitro. There are structure based and sequence based approaches to predict the ability of peptides to bind at the MHC. The former uses X-ray structures of MHC or even better of the MHC-peptide complex as a starting point to model the binding geometry of different peptides [21]. It has the advantage to require only knowledge of one or at most a few crystal structures to study the peptide binding and provides a deeper understanding of the importance of specific inter-actions between peptides and the MHC. However, obviously, a structural approach is limited to MHC types with a known structure. More recently, based on sequence information, a number of methods and algorithms have also been introduced to identify and characterize the T-cell epitopes, including binding motif scheme to matrix scoring schemes [22–24], decision trees [25], evolutionary algorithms [26], hidden Markov [27], CoMFA (comparative molecular field analysis)/CoMSIA (comparative molecular similarity indices analysis) [28], multiple regression [29] and neutral networks [30]. As far as we know, there are few works detailing the problem of epitope prediction in a quantitative way [31–33]. Doytchinova and Flower [28] applied CoMFA and CoMSIA analysis to perform 3D QSAR (quantitative structure-activity relationships) study on MHC/epitope binding affinity and Lin et al. [29] built a linear function for predicting binding affinity of nonapeptides. These may be the most important works in computational vaccinology to initiate the prediction epitope in a quantitative way. However, they have their own disadvantages, e.g., for CoMFA/CoMSIA analysis, the latent parameters do not have an explicit significance, and are not suitable to deal with the large and diverse data set and the activity data in vivo. And, as we know, the relationship between the receptors and ligands are often complex and nonlinear, thus, it is difficult to express them with a certain linear function. And the predictive results of the linear method were poor considering the large number of outliers in their predictive models [29]. Consequently, it is necessary to construct a more convenient and robust models for HLA-A*0201 binding, allowing us to predict the affinity of new, untested peptides in a quantitative manner and, suggest new ideas for further synthesis of high-affinity peptides.

In this study, QSAR method, as a powerful quantitative methodology, was used to correlate structural variation with variation in binding affinity for 152 peptides having affinity with the class I MHC HLA-A*201 molecule. QSAR method was proved to be one of the most promising and successful methods for rapidly predicting the biological activity and/or toxicity of chemicals [29,32,34]. Many different technologies can be applied for the development of a quantitative relationship between the structural descriptors and the properties, including linear methods, e.g., multiple linear regression (MLR) [29], partial least squares (PLS) [35], and heuristic method (HM) [36], or nonlinear methods, e.g., neural networks [37]. For these methods, as we know, linear method is much limited for a complex biological system. The flexibility of neural networks enables them to discover more complex nonlinear relationships in experimental data. However, these neural systems have some problems inherent to its architecture, such as overtraining, overfitting, and network optimization. And, other problems with the use of neural networks concern the reproducibility of results, due largely to random initialization of the networks and variation of stopping criteria. Owing to these reasons above there is a continuing need for the application of more accurate and informative techniques in QSAR analysis. The support vector machine (SVM) is a new algorithm developed from the machine learning community [38,39]. Due to its remarkable generalization performance, the SVM have attracted attention and gained extensive application, e.g., isolated handwritten digit recognition [40], object recognition [41], drug design [42], protein structure prediction [43], genes identification [44], and diseases diagnosis [35], etc. In most of these cases, the performance of SVM modeling is significantly better than that of traditional machine learning approaches. It has a number of interesting properties, including an effective avoidance of overfitting, which improves its ability to build models using large numbers of molecular property descriptors with relatively few experimental results in the training set.

In the present work, the heuristic method and the support vector machine were utilized to establish linear and nonlinear relationship for HLA-A*0201 binding, based on a large pool of descriptors, including constitutional, topological and physical–chemical properties. This would allow the prediction of the binding affinity of new, untested peptides and, through the analysis of contribution of each parameter of different residue at specific position of peptidic ligands, to understand nature of the forces governing binding behavior and suggest new ideas for further synthesis of high-affinity peptides.

2. Methodology

2.1. Peptides and binding affinities

In this paper, a set of 152 peptides having affinity with the class I MHC HLA-A*0201 molecule was used. All the peptides chosen consisted of nine amino acids. The sequences of peptides and their binding affinities were obtained from Ref. [24] (shown in supporting information). The peptides in the study included 65 high-affinity peptides ($\mathrm{pIC}_{50} \geq 7.301$), 59 intermediate affinity ($7.301 > \text{pIC}_{50} \geq 6.301$), and 28 low-affinity peptides ($\text{pIC}_{50} < 6.301$). All the peptide is positioned with the N-terminus to the left and the C-terminus to the right as it is oriented in the binding cleft on the HLA-A*0201 molecule. The magnitude of measured binding affinity ranges over almost four orders: from 5.146 to 8.770 in log units. Such broad representation over the data space is important to ensure predictive capability of the QSAR models.

2.2. Descriptor calculation

Charactering the structure of a molecule is a very important task in QSAR technique studies. The descriptors used in this study included two categories: (1) 38 constitutional and 38 topological descriptors based on nonapeptides structures. The structures of the nonapeptides were drawn using the sequence editor of Hyperchem and saved as the hin files. Then their hin files were transferred into software CODESSA, developed by the Katritzky group [45,46], to calculate constitutional and topological descriptors, which has been successfully used in various QSPR/QSAR researches. In the present work, only the constitutional and topological descriptors were calculated because their calculation does not need the optimization of the molecular structure, which is time-consuming. (2) Fifty-one physical–chemical properties of the nonapeptides based on amino acids sequence and physical–chemical properties, including 49 two-dimensional properties and 2 three-dimensional properties. The values of 51 descriptors for each amino acid can be obtained from Refs. [13,47].

Thus, a large pool of 127 descriptors was obtained presenting 2D and 3D information of the nonapeptides, and imported into CODESSA software for further selection using heuristic method.

2.3. The heuristic method [36]

The heuristic multilinear regression procedures available in the framework of the CODESSA program were used to perform a complete search for the best multilinear correlations with a multitude of descriptors. These procedures provide collinearity control (i.e., any two descriptors intercorrelated above 0.8 are never involved in the same model) and implement heuristic algorithms for the rapid selection of the best correlation, without testing all possible combinations of the available descriptors. The heuristic method of the descriptor selection proceeds with a pre-selection of descriptors by eliminating (i) those descriptors that are not available for each structure, (ii) descriptors having a small variation in magnitude for all structures, (iii) descriptors that give a F-test's value below 1.0 in the one-parameter correlation, and (iv) descriptors whose t-values are less than the user-specified value, etc. The descriptors were ordered by decreasing correlation coefficient when used in one-parameter correlations. The next step involves correlation of the given property with (i) the top descriptor in the above list with each of the remaining descriptors and (ii) the next one with each of the remaining descriptors, etc. The best pairs, as evidenced by the highest F-values in the two-parameter correlations, are chosen and used for further inclusion of descriptors in a similar manner.

The heuristic method usually produces correlations two to five times faster than other methods, with comparable quality. The rapidity of calculations from the heuristic method renders it the first method of choice in practical research. Thus, we chose this method for building the linear model.

2.4. SVM model development [38,39]

SVM algorithms are mainly developed by Vapnik's and Burges's work. The main advantage of SVM is that it adopts the structure risk minimization (SRM) principle, which has been shown to be superior to the traditional empirical risk minimization (ERM) principle, employed by conventional neural networks. SRM minimizes an upper bound of the generalization error on Vapnik–Chernoverkis (VC) dimension, as opposed to ERM that minimizes the training error.

For the case of regression approximation, suppose there are a given set of data points $G = \{(x_i, d_i)\}_i^n$ (x_i is the input vector, d_i is the desired value, and n is the total number of data patterns) drawn independently and identically from an unknown function, SVMs approximate the function with three distinct characteristics: (i) SVMs estimate the regression in a set of linear functions, (ii) SVMs define the regression estimation as the problem of risk minimization with respect to the ε-insensitive loss function, and (iii) SVMs minimize the risk based on the SRM principle whereby elements of the structure are defined by the inequality $||w||^2 \leq \text{constant}$. The linear function is formulated in the high dimensional feature space, with the form of function (1):

$$y = f(x) = w\phi(x) + b \tag{1}$$

where $\phi(x)$ is the high dimensional feature space, which is nonlinearly mapped from the input space x. Characteristics (ii) and (iii) are reflected in the minimization of the regularized risk function (2) of SVMs, by which the coefficients w and b are estimated. The goal of this risk function is to find a function that has at most ε deviation from the actual values in all the training data points, and at the same time is as flat as possible:

$$R_{\text{SVMs}}(C) = C\frac{1}{n}\sum_{i=1}^{n} L_\varepsilon(d_i, y_i) + \frac{1}{2}||w||^2 \tag{2}$$

$$L_\varepsilon(d, y) = \left\{ \begin{array}{ll} |d-y| - \varepsilon, & |d-y| \geq \varepsilon, \\ 0 & \text{otherwise.} \end{array} \right\} \tag{3}$$

The first term $C(1/n)\sum_{i=1}^{n} L_\varepsilon(d_i, y_i)$is the so-called empirical error (risk), which is measured by the ε-insensitive loss function (3). This loss function provides the advantage of using sparse data points to represent the designed function (1). The second term $(1/2)||w||^2$, on the other hand, is called the regularized term. ε is called the tube size of SVMs, and C is the regularization constant determining the trade-off between the empirical error and the regularized term. Introduction of the positive slack variables ξ, ξ^* leads Eq. (4) to the following constrained function:

$$\text{minimize}: \quad R_{\text{SVMs}}(w, \xi^{(*)}) = \frac{1}{2}||w||^2 + C\sum_{i=1}^{n}(\xi_i + \xi_i^*) \qquad (4)$$

where i represents the data sequence, with $i = n$ being the most recent observation and $i = 1$ being the earliest observation. Finally, by introducing Lagrange multipliers and exploiting the optimality constraints, decision function (5) takes the following form:

$$f(x, a_i^*) = \sum_{i=1}^{n}(a_i - a_i^*)K(x, x_i) + b \qquad (5)$$

where a_i and a_i^* are the introduced Lagrange multipliers.

So far, by exploiting the Karush–Kuhn–Tucker (KKT) conditions, only a number of coefficients among a_i and a_i^* will be nonzero, and the data points associated with them could be referred to support vectors. In this equation, K refers to kernel function, including linear, polynomial, splines, and radial basis function. In support vector regression, the Gaussian radial basis function (6) is commonly used, which has the following form:

$$\text{radial basis function(RBF)}: \quad k(\overline{x_i}, \overline{x_j}) = \exp(-\gamma||\overline{x_i} - \overline{x_j}||^2) \qquad (6)$$

2.5. Model validation

Validation of the models was required to test the predictive ability and generalization of the methods as well as to enable comparison between them. The whole data set was divided into a training set (102 peptides) for model development/calibration and an independent test set (50 peptides) for external prediction. The construction of the test set was accomplished by insisting two features. First, the members of the test set should be representative of all members of the training set in terms of the ranges of pIC_{50}. For this study, data set included high-affinity, intermediate-affinity and low-affinity peptides. The cases in the test set should cover all the three types of the peptides. Second, that each amino acid at each position in the test set should also be present at that position in the training set. It means that for the cases in the test set, the amino acid residues in each position, P1 for instance, should include all the types that appear in the same position of the cases in the training set. Based on the two criteria, training and test set were picked from the original data set. For the training set, the predictive models underwent a leave-two-out (LTO) procedure. The stability of the correlations was tested against the cross-validated coefficient R^2_{cv}, which describes the stability of a model obtained by focusing on the sensitivity of the model to the elimination of any single data point. The goodness of the correlation is also tested by root-mean-square (RMS) error. They are computed by following formula:

$$R^2_{\text{cv}} = 1 - \frac{\sum_{k=1}^{n_s}(y_k - \widehat{y_k})^2}{\sum_{k=1}^{n_s}\left[y_k - \left(\sum_{k=1}^{n_s} y_k\right)/n_s\right]^2}$$

$$\text{RMS} = \sqrt{\frac{\sum_{i=1}^{n_s}(y_k - \widehat{y_k})^2}{n_s}}$$

where y_k is the desired output and $\widehat{y_k}$ is the actual output of the model, and n_s is the number of compounds in the analyzed set.

2.6. Heuristic method and SVM method implementation and computation environment

All calculation programs implementing heuristic method were performed in CODESSA software. All calculation programs implementing SVM were written in an R-file based on the R script for SVM. All statistical works were completed by SPSS software. The scripts were compiled using an R 1.7.1 compiler running operating system on a Pentium IV PC with 256 M RAM.

3. Results

3.1. Results of the heuristic method

As shown above, the obtained 127 descriptors were imported into CODESSA software. A variety of subset sizes was investigated to determine the optimum number of the descriptors in models. When adding another descriptor did not improve significantly the R^2_{cv} of a model, it was determined that the optimum subset size had been achieved. To avoid the ‘‘over-parametrization’’ of the model, an increase of the R^2_{cv} value of less than 0.02 was chosen as the breakpoint criterion. The result showed that the four descriptors appear to be sufficient for successful models. The multilinear analysis of the binding affinity values for the peptides of the training set resulted in the four-parameter model were summarized in Table 1, and the correlation matrix of these descriptors was shown in Table 2. The linear correlation coefficient value of each two descriptors is <0.70 (Table 2), which means the descriptors were independent in this multilinear analysis. For this model, the cross-validated coefficient R^2_{cv} was 0.615. The produced RMS error were 0.535 binding affinity units for the training set, 0.511for the test set, and 0.523 for the whole data set, the corresponding correlation coefficients (R^2) were 0.615, 0.667, and 0.627, respectively (Table 3). Fig. 1 showed a plot of the predicted versus experimental binding affinity for all of the 152 peptides studied.

3.2. Results of the SVM method

From Table 3 and Fig. 1, it can be seen that the model of the heuristic method was not sufficiently accurate and the prediction ability was not satisfactory, showing the factors influencing the binding affinities of these peptides were complex and not all of them were linear correlations with the binding affinity. So, after the establishment of the linear model by HM, the nonlinear prediction model by SVM was built to further discuss the correlation between the amino acid sequences and the binding affinity based on the same subset of descriptors.

Table 1
Descriptors, coefficients, standard error, and *t*-values for the linear model

Descriptor	Chemical meanings	Coefficient[a]	S.E.[b]	Beta	*t*-Test[c]
Intercept	Intercept	0.831	0.974		0.853
$-T\Delta S_k$	Unfolding entropy change of hydration	0.532	0.148	0.508	6.213
$-T\Delta S_c$	Unfolding entropy change of chain	0.781	0.126	0.349	3.585
pH$_j$ (P*i*)	Isoelectric point	0.713	0.202	0.267	3.524
ECI	Electronic charge index of the side chain	−0.485	0.229	−0.165	−2.121

$N = 102$; $F = 38.745$; $s^2 = 0.294$; $R^2_{\text{cv}} = 0.615$.

[a] Coefficients of the descriptors in linear function.

[b] Standard error of the each descriptor.

[c] *t*-Test the difference between a sample mean and a known or hypothesized value.

Table 2
Correlation matrix of the four descriptors

	$-T\Delta S_k$	$-T\Delta S_c$	pH$_j$	ECI
$-T\Delta S_k$	1.000	0.093	0.582	0.557
$-T\Delta S_c$		1.000	0.373	0.042
pH$_j$			1.000	0.692
ECI				1.000

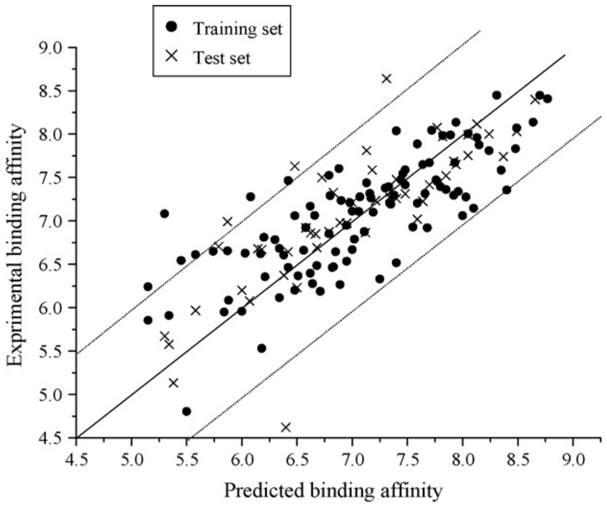

Fig. 1. Predicted vs. experimental binding affinity (pIC$_{50}$) (HM).

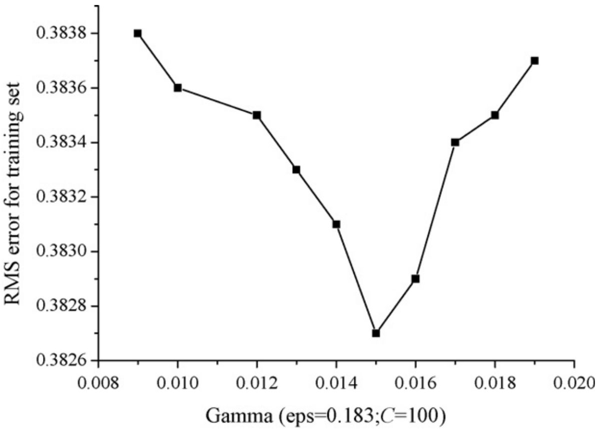

Fig. 2. The gamma vs. RMS error (ε = 0.183; *C* = 100).

In this work, SVM training included the selection of capacity parameter *C*, ε of ε-insensitive loss function and the corresponding parameters of the kernel function. Firstly, the kernel function should be decided, which determines the sample distribution in the mapping space. The radial basis function (RBF) is commonly used in many studies because of its good general performance and few parameters to be adjusted [48]. In this work, the radial basis function was used, the form of which in R is as follows:

$$\exp(-\gamma * |u - v|^2) \tag{9}$$

Where γ is a parameter of the kernel; u, v are two independent variables.

Secondly, corresponding parameters, i.e. γ of the kernel function greatly affect the number of support vectors, which has a close relation with the performance of the SVM and training time. Too many support vectors could produce overfitting and increase the training time. In addition, γ controls the amplitude of the RBF function and, therefore, controls the generalization ability of SVM. The plot of γ versus RMS on the LOO cross-validation is shown in Fig. 2. As can be seen from the figure, the optimal γ was 0.015.

Table 3
Statistical results of different QSAR models

	Training set		Test set		Whole data set				
	RMS	R^2_{cv}	RMS	R^2	RMS	R^2	*F*	*t*-Test	Number of outliers
HM[a]	0.535	0.615	0.511	0.667	0.523	0.627	252.536	15.891	12
SVM[b]	0.383	0.805	0.385	0.800	0.384	0.804	410.945	20.272	1

[a] Model of HM.

[b] Model of SVM.

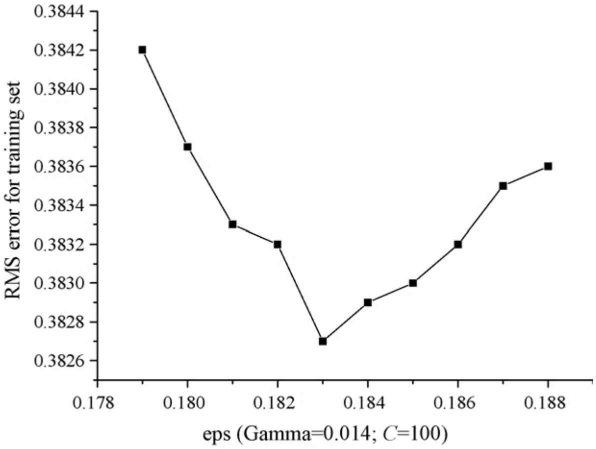

Fig. 3. The epsilon vs. RMS error (Gamma = 0.015; $C = 100$).

Parameter ε-insensitive prevents the entire training set meeting boundary conditions and so allows for the possibility of sparsity in the dual formulation's solution. The optimal value for ε depends on the type of noise present in the data, which is usually unknown. The RMS error of LOO cross-validation on different epsilon is recorded in Fig. 3 and the optimal value was found to be 0.183.

Lastly, the effect of capacity parameter C was tested. It controls the trade-off between maximizing the margin and minimizing the training error. If C is too small then insufficient stress will be placed on fitting the training data. If C is too large then the algorithm will overfit the training data. However, Ref. [49] indicated that prediction error was scarcely influenced by C. To make the learning process stable, a large value should be set up for C initially (e.g., $C = 100$). The plot of RMS error versus C value is shown in Fig. 4 with values $\gamma = 0.015$, $\varepsilon = 0.183$. The optimal value of C was 100.

Therefore, the best choices for γ, ε and C were 0.015, 0.183 and 100, with the support vector number of 88. For the optimal model, the cross-validated coefficient R^2_{cv} was 0.805. It gave RMS error of 0.383 binding affinity units for the training set, 0.385 for the test set, and 0.384 for the whole set, and the corresponding correlation coefficients (R^2) were 0.805, 0.800, and 0.804, respectively (Table 3). Fig. 5 shows the predicted versus experimental values of binding affinity for the whole data set.

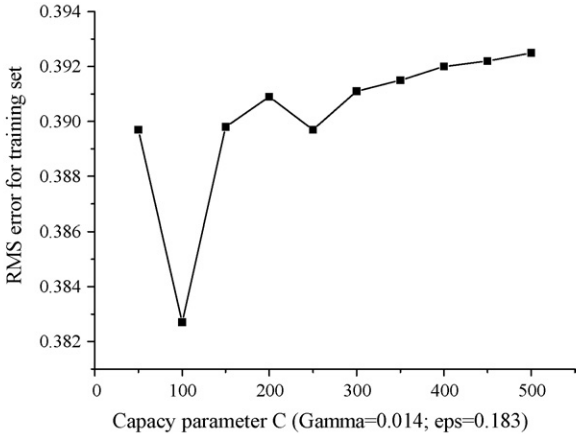

Fig. 4. The capacity parameter C vs. RMS error (Gamma = 0.015; $\varepsilon = 0.183$).

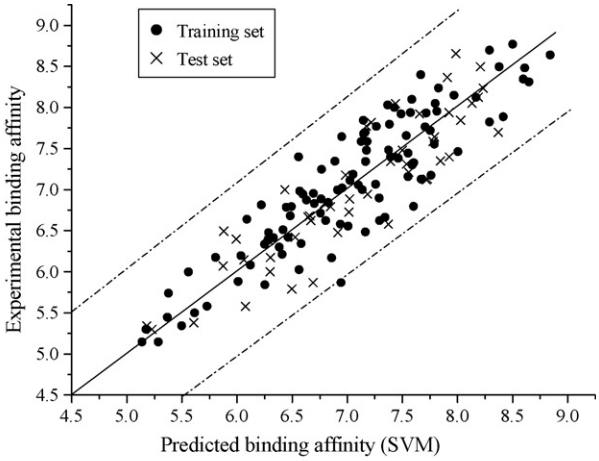

Fig. 5. Predicted vs. experimental binding affinity (pIC_{50}) (SVM).

4. Discussion

The calculated binding affinities obtained from two predictive models were listed in supporting information. Table 3 showed the statistical parameters of the results obtained from two studies for the same set of peptides. The RMS errors of the SVM model for the training, the test, and the whole data set were much lower than that of the models proposed in the heuristic method. The correlation coefficient (R^2) given by the SVM model was higher than that of the models in the heuristic method. And the outliers of HM model were 12, while that of SVM model was only 1. Through a regression analysis on the experimental and the calculated binding affinity obtained by different methods for the whole data set, the results of F-test and t-test were obtained and also shown in Table 3. From the table, it can be seen that the SVM model gives the highest F and t values, so this model gives the most satisfactory results, compared with the results obtained from the heuristic methods. Consequently, this SVM approach currently constitutes the most accurate method for predicting the binding affinity of peptides, thus, for search of new drugs.

4.1. Outlier analysis

In this study, peptides with residuals between experimental and predicted pIC_{50} values above 1 log unit were considered as outliers. As shown in supporting information and Table 3, although SVM model yielded some good statistical results, there is still one outlier (peptide 10, ALPYWNFAT), existing in both two predictive models. Three possible reasons can explain why the peptide is outlier: an incorrectly measured experimental value, a different binding conformation, or a significant difference in the physicochemical properties. For the peptide 10, it seems to be important that this peptide includes a single presentation of a particular amino acid in the training set amino acids (Thr at the 9th position in peptide 10). This may lead to

Table 4
Frequency of the amino acid residues at different positions in the peptides in the high-affinity peptides

Position	Previous studies		Present study	
	Favorite structural features[a]	Favorite amino acids[a]	Favorite amino acids[b]	Frequency[b]
P1	Aromatic side chains	Tyr, Ile	Ile	17
P2	Long side chains and hydrophobic side chains	Val, Phe, Met, Leu, Ile	Leu	48
P3	Aromatic side chains and hydrophobic side chains	Phe, Tyr, Trp	Tyr	9
P4	Non-branched side chains and hydrophilic side chains	Glu,Gln	Gln	18
P5	Branched, aromatic and hydrophilic side chains	Tyr, Asn, Arg, Ser, Thr, Val	Val	17
P6	Carrying H-bond donor groups	Thr, Tyr,	Pro	30
P7	Short side chains and hydrophobic side chains	Ala, Val, Pro	Val	19
P8	Short side chains and hydrophilic side chains	Ser, Thr	Ser	20
P9	Short side chains and hydrophobic side chains	Ala, Val	Val	41

[a] Favorite structural features and amino acids obtained from previous studies.
[b] Statistical results of the favorite amino acids and their frequencies in the present study.

lack-training of the model, thus resulting the large diverse. Obviously, this is a weakness of the investigated set but not of the method. The growth of the database will decrease the number of missing amino acids at particular positions, thus increase the predictive accuracy.

4.2. Discussion of descriptors

Generally, the interaction between receptor and ligand is related to the contribution of hydrophobic property, stereo property and electronic property of the ligands. As a measure of the binding ability, four quantitative descriptors, $-T\Delta S_k$, $-T\Delta S_c$, pH_j and ECI were used, characterizing different aspects of the protein interaction with nonapeptide ligands (Table 1). By interpreting these descriptors, it is possible to give us some insight into factors that are likely to govern the binding behavior of the nonapeptides, thus, help us understand which interaction may play an important role in the binding process.

For a certain protein, the change in stability of the protein–ligand complex is often categorically labeled as entropic change of the ligand. Descriptor $-T\Delta S_k$ was defined as unfolding entropy change of hydration reflecting hydrophobic property. Another one, $-T\Delta S_c$ (unfolding entropy change of chain) is related to stereo effect for the entropy change, especially for side chain of the ligands. As seen in Table 1, the coefficient for $-T\Delta S_k$ and $-T\Delta S_c$ were positive. It indicates that, in this study, the entropic effect may play a positive contribution to protein–ligand process. The third one, isoelectric point (pH_j) is the pH value, at which a molecule carries no net electrical charge, and a substance in a solution is electrically neutral and has its unique properties. The isoelectric point of an amino acid is a very important physical–chemical property of amino acids, which is necessary to distinguish amino acids because of a special neutral property of amino acids at the isoelectric point. The last descriptor, electronic charge index of the side chain (ECI) is a measure of the local polarity in the side chain proposed by Collantes [47]. The ECI value for natural amino acids used in this study is calculated as the sum of the absolute values of the charges q for each atom i present in the side chain of the amino acids:

$$\mathrm{ECI} = \sum |q_i|$$

A significant ECI contribution to activity may indicate the presence of a dipolar interaction of the side chain with the receptor. It reflects the electric and stereo property of the peptide ligands.

From the above discussion, it can be seen that each descriptor involved in the model has an explicit physical meaning, and can account for the structural features responsible for the drug protein binding. According to the analysis of the corresponding regression coefficient (Table 1), unfolding entropy change of hydration, unfolding entropy change of chain and isoelectric point present positive contributions for binding affinity, whereas electronic charge index of the side chain present negative contribution.

4.3. Frequency analysis of amino acid

In addition, previous studies [28,37,50] proved that each amino acid in the peptide ligand binds to HLA-A2*0201 independently of one another to enhance or detract from the overall binding affinity, and the contribution to the binding affinity of each amino acid residue in different position of the peptide is different. In order to identify peptide residues that may be important for the structure and biological function of the protein binding, some investigation should be performed for the high-affinity peptides of this data set to character the features of certain amino acid in certain position. Table 4 showed the favorite structural features and the favorite amino acids which have high frequency in different positions for the high-affinity peptides. The analyses of contribution of each residue at specific position of peptidic ligands are helpful to understand nature of the forces governing binding behavior.

As seen in Table 4, for position 1, CoMFA study [28] indicates that bulk side chains bearing a hydrogen-bond-forming group are preferred at this position. The most suitable amino acid for this position seems to be Tyr and Ile, which have an aromatic side chains together with a hydrogen-bond acceptor near the N-terminal. For the high affinity nonapeptides in this work, as shown in Table 4, Ile is the favorite one. It is in a good agreement with the 3D-QSAR studies. For position 2, large hydrophobic side chains are preferred and Val, Phe, Leu, Ile, and Met are well accommodated. A same analysis for the high-affinity nonapeptides were also applied for this position and Leu is proved to be the favorite one in this data set. For position 3, previous study showed that amino acids with hydrophobic aromatic rings would enhance the binding affinity of the peptides. Phe, Trp and Tyr will be suitable, because of its ability to form hydrogen bonds. Frequency of amino acid in this position also indicated that Tyr was well accepted at this position. Also for positions 4–9, the favorite structural features and the favorite amino acids at these positions suggested by 3D study were listed in Table 4, together with frequency analyses of the amino acids. Thus, based on the analyses above, we can ‘‘construct’’ some ideal nonapeptides which should have high binding affinities, e.g., ILYQVPFSV, ILWQVPFSV and IVWQVPFSV. Observing these fictitious peptides, we can find that peptide 102 (ILWQVPFSV) in this data set is just such an ‘‘ideal’’ one, which has the highest binding affinity ($pIC_{50} = 8.770$). It proved that this analysis is valuable for us to suggest new ideas for searching peptides which can indeed bind to the MHC molecule before hundreds of peptides were synthesized.

5. Conclusion

The heuristic method and the support vector machine were used to construct the linear and nonlinear quantitative relationships for the prediction of the affinity of a diverse set of 152 peptides binding to the HLA-A*0201 molecules. A large pool of descriptors including constitutional, topological and physical–chemical properties was calculated, and four descriptors were selected by the heuristic method to construct QSAR models. Inspecting this work, we can conclude that: (1) the four selected descriptors can represent the features of the peptides binding behavior. The heuristic linear model could identify and provide some insight into which structural information is most related to the binding affinity. (2) The QSAR method delineated those areas within the binding site that favor or disfavor the presence of a group. This suggests concerning the occupation of each of the nine positions in the nonamer peptide by favorable amino acid residues. The prediction of epitopes is vital to our goal of developing computational vaccinology. (3) SVM method proved to be a very promising tool in the prediction of the affinity of nonapeptides binding to the HLA-A*0201 molecules. It is due to that SVM method embodies the structural risk minimization principle which minimizes an upper bound of the generalization error rather than minimizes the training error. This eventually leads to better generalization than neural networks, which implement the empirical risk minimization principle and while the neural network may not converge to global solutions.

Although the prediction of MHC binding is only one component of the process leading to T-cell activation, it probably forms the most selective filter. We would expect computational vaccinology to have a similar effect on the search for new vaccines as molecular modeling and QSAR have had on search for new drugs.

Acknowledgements

The authors thank the National Natural Science Foundation of China (NSFC) Fund (NO.20305008) for financial support.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.jmgm.2006.12.002.

Reference

[1] D.N. Garboczi, P. Ghosh, U. Utz, Q.R. Fan, W.E. Biddison, D.C. Wiley, Structure of the complex between human T-cell receptor, viral peptide and HLA-A2, Nature 384 (1996) 134–141.

[2] Y. Guilloux, S. Lucas, V.G. Brichard, A. Van Pel, C. Viret, E. De Plaen, F. Brasseur, B. Lethe, F. Jotereau, T. Boon, A peptide recognized by human cytolytic T lymphocytes on HLA-A2 melanomas is encoded by an intron sequence of the *N*-acetylglucosaminyltransferase V gene, J. Exp. Med. 183 (1996) 1173–1183.

[3] A. Lanzavecchia, P.A. Reid, C. Watts, Irreversible association of peptides with class II MHC molecules in living cells, Nature 357 (1992) 249–252.

[4] K.C. Garcia, M. Degano, L.R. Pease, M. Huang, P.A. Peterson, L. Leyton, I.A. Wilson, Structural basis of plasticity in T cell receptor recognition of a self peptide-MHC antigen, Science 279 (1998) 1166–1172.

[5] L. Stolze, A.K. Nussbaum, A. Sijts, N.P. Emmerich, P.M. Kloetzel, H. Schild, The function of the proteasome system in MHC class I antigen processing, Immunol. Today 21 (2000) 317–319.

[6] D. Tortorella, B.E. Gewurz, M.H. Furman, D.J. Schust, H.L. Ploegh, Viral subversion of the immune system, Annu. Rev. Immunol. 18 (2000) 861–926.

[7] M. Bouvier, D.C. Wiley, Importance of peptide amino and carboxyl termini to the stability of MHC class I molecules, Science 265 (1994) 398–402.

[8] K. Falk, O. Rotzschke, S. Stefanovic, G. Jung, H.G. Rammensee, Allele-specific motifs revealed by sequencing of self peptides eluted from MHC molecules, Nature 351 (1991) 290–296.

[9] R.M. Chicz, R.G. Urban, J.C. Gorga, D.A. Vignali, W.S. Lane, J.L. Strominger, Specificity and promiscuity among naturally processed peptides bound to HLA-DR alleles, J. Exp. Med. 178 (1993) 27–47.

[10] T.S. Jardetzky, J.H. Brown, J.G. Gorga, L.C. Stern, R.G. Urban, J.L. Strominger, D.C. Wiley, Crystallographic analysis of endogenous peptides associated with HLA-DR1 suggests a common, polyproline II-like conformation for bound peptides, Proc. Natl Acad. Sci. U.S.A. 93 (1996) 734–738.

[11] H.G. Rammennsee, T. Friede, S. Stevanovic, MHC ligands and peptide motifs: first listing, Immunogenetics 41 (1995) 178–228.

[12] K.C. Garcia, Molecular interactions between extracellular components of the T-cell receptor signaling complex, Immunol. Rev. 172 (1999) 73–85.

[13] K. Tomii, M. Kanehisa, Analysis of amino acid indices and mutation matrices for sequence comparison and structure prediction of proteins, Protein Eng. 9 (1996) 27–36.

[14] A.J. McMichael, P. Parham, F.M. Brodsky, J.R. Pilch, Influenza virus-specific cytotoxic T lymphocytes recognize HLA-molecules. Blocking by monoclonal anti-HLA antibodies, J. Exp. Med. 152 (1980) 195–203.

[15] D.J. Schendel, B. Gansbacher, R. Oberneder, M. Kriegmair, A. Hofstetter, G. Riethmuller, O.G. Segurado, Tumor-specific lysis of human renal cell carcinomas by tumor-infiltrating lymphocytes. I. HLA-A2-restricted recognition of autologous and allogeneic tumor lines, J. Immunol. 151 (1993) 4209–4220.

[16] Y. Rongcun, F. Salazar-Onfray, J. Charo, K.J. Malmberg, K. Evrin, H. Maes, C. Hising, M. Petersson, O. Larsson, L. Lan, E. Appella, A. Sette, E. Celis, R. Kiessling, Identification of new HER2/neu-derived peptide epitopes that can elicit specific CTL against autologous and allogeneic carcinomas and melanomas, J. Immunol. 163 (1999) 1037–1044.

[17] L. Rivoltini, Y. Kawakami, K. Sakaguchi, S. Southwood, A. Sette, P.F. Robbins, F.M. Marincola, M.L. Salgaller, J.R. Yannelli, E. Appella, S.A. Rosenberg, Induction of tumor-reactive CTL from peripheral blood and tumor-infiltrating lymphocytes of melanoma patients by in vitro stimulation with an immunodominant peptide of the human melanoma antigen MART-1, J. Immunol. 154 (1995) 2257–2265.

[18] M.R. Parkhurst, E.B. Fitzgerald, S. Southwood, A. Sette, S.A. Rosenberg, Y. Kawakami, Identification of a shared HLAA*0201-restricted T-cell epitope from the melanoma antigen tyrosinase-related protein 2 (TRP2), Cancer Res. 58 (1998) 4895–4901.

[19] R.A. Henderson, H. Michel, K. Sakaguchi, J. Shabanowitz, E. Apella, D.F. Hunt, V.H. Engelhard, HLA-A21 associated peptides from a mutant cell line: a second pathway of antigen presentation, Science 255 (1992) 1264–1266.

[20] M. Regner, M.H. Claesson, S. Bregenholt, M. Ropke, An improved method for the detection of peptide-induced upregulation of HLA-A2 molecules on TAP-deficient T2 cells, Exp. Clin. Immunogenet. 13 (1996) 30–35.

[21] R. Rosenfeld, Q. Zheng, S. Vajda, C. Delisi, Computing the structure of bound peptides application to antigen recognition by class-I major histocompatibility complex receptors, J. Mol. Biol. 234 (1994) 515–521.

[22] R. Bertoni, J. Sidney, P. Flower, R.W. Chesnut, F.V. Chisari, A. Sette, Human histocompatibility leukocyte antigen-binding supermotifs predict broadly cross-reactive cytotoxic T lymphocyte responses in patients with acute hepatitis, J. Clin. Invest. 100 (1997) 503–513.

[23] H.G. Rammensee, T. Friede, S. Stevanoviic, MHC ligands and peptide motifs: first listing, Immunogenetics 41 (1995) 178–228.

[24] H.G. Rammensee, J. Bachmann, N.P. Emmerich, O.A. Bachor, S. Stevanovic, SYFPEITHI: database for MHC ligands and peptide motifs, Immunogenetics 50 (1999) 213–219.

[25] C.J. Savoie, N. Kamikawaji, T. Sasazuki, S. Kuhara, Use of BONSAI decision trees for the identification of potential MHC class I peptide epitope motifs, Pac. Symp. Biocomput. 4 (1999) 182–189.

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

[27] H. Mamitsuka, Predicting peptides that bind to MHC molecules using supervised learning of hidden Markov models, Protein 33 (1998) 460–474.

[28] I.A. Doytchinova, D.R. Flower, Toward the quantitative prediction of T-cell epitopes: CoMFA and CoMSIA studies of peptides with affinity for the class I MHC molecule HLA-A*0201, J. Med. Chem. 44 (2001) 3572–3581.

[29] Z.H. Lin, Z. Wu, Y.L. Wei, B. Ni, B. Zhu, L. Wang, A rapid method for quantitative prediction of high affinity CTL epitopes: QSAR studies on peptides having affinity with the class I MHC molecular HLA-A.0201, Lett. Peptide Sci. 10 (2003) 15–23.

[30] M.C. Honeyman, V. Brusic, N.L. Stone, L.C. Harrison, Neural network-based prediction of candidate T-cell epitopes, Nat. Biotechnol. 16 (1998) 966–969.

[31] I.A. Doytchinova, D.R. Flower, Quantitative approaches to computational vaccinology, Immunol. Cell Biol. 80 (2002) 270–279.

[32] V. Brusic, K. Bucci, C. Schonbach, N. Petrovsky, J. Zeleznikow, J.W. Kazura, Efficient discovery of immune response targets by cyclical refinement of QSAR models of peptide binding, J. Mol. Graph. Model. 19 (2001) 405–411.

[33] I.A. Doytchinova, D.R. Flower, Physiochemical explanation of peptide binding to HLA-A*0201 major histocompatibility complex: a three dimensional quantitative structure–activity relationship study, Proteins 48 (2002) 505–518.

[34] A.R. Katritzky, V.S. Lobanov, M. Karelson, CODESSA: Training Manual, University of Florida, Gainesville, 1995.

[35] A.R. Katritzky, V.S. Lobanov, M. Karelson, CODESSA: Reference Manual, University of Florida, Gainesville, 1994.

[36] Z.Y. Xiao, S. Varma, Y.D. Xiao, A. Tropsha, Modeling of p38 mitogen-activated protein kinase inhibitors using the Catalyst^{TM} HypoGen and *k*-nearest neighbor QSAR methods, J. Mol. Graph. Model. 23 (2004) 129–138.

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

[38] A.R. Katritzky, V.S. Lobanov, M. Karelson, Comprehensive Descriptors for Structural and Statistical Analysis. Reference Manual, Version 2.13, 1995–1997.

[39] C. Wan, P.B. Harrington, Self-configuring radial basis function neural networks for chemical pattern recognition, J. Chem. Inform. Comput. Sci. 39 (1999) 1049–1056.

[40] V. Vapnik, Estimation of Dependencies Based on Empirical Data, Springer, Berlin, 1982.

[41] C.J.C. Burges, A Tutorial on Support Vector Machines for Pattern Recognition, Data Min. Knowl. Disc. 2 (1998) 121–167.

[42] Y.L. Cun, L. Jackel, L. Bottou, C. Cortes, J. Denker, H. Drucker, I. Guyon, U. Müller, E. Säckinger, P.Y. Simard, V. Vapnik, Learning algorithms for classification: a comparison on handwritten digit recognition, neural networks. Neural networks: the statistical mechanics perspective, World Scientific (1995) 261–276.

[43] V. Blanz, B. Schölkopf, H. Bülthoff, C. Burges, V. Vapnik, T. Vetter, Comparison of view-based object recognition algorithms using realistic 3D models, in: C.V.D. Malsburg, W.V. Seelen, J.C. Vorbrüggen, B. Sendhoff (Eds.), Artificial Neural Networks—ICANN'96, Springer. Lect. Notes Comput. Sci. 1112 (1996) 251–256.

[44] R. Burbidge, M. Trotter, B. Buxton, S. Holden, Drug design by machine learning: support vector machines for pharmaceutical data analysis, Comput. Chem. 26 (2001) 5–14.

[45] Y.D. Cai, X.J. Liu, X.B. Xu, K.C. Chou, Prediction of protein structural classes by support vector machines, Comput. Chem. 26 (2002) 293–296.

[46] L. Bao, Z.R. Sun, Identifying genes related to drug anticancer mechanisms using support vector machine, FEBS Lett. 521 (2002) 109–114.

[47] C.Y. Zhao, R.S. Zhang, H.X. Liu, C.X. Xue, S.G. Zhao, X.F. Zhou, M.C. Liu, B.T. Fan, Diagnosing anorexia based on partial least squares, back propagation neural network, and support vector machines, J. Chem. Inform. Comput. Sci. 44 (2004) 2040–2046.

[48] C. Bishop, Neural Networks for Pattern Recognition, Clarendon Press, Oxford, 1997.

[49] W.J. Wang, Z.B. Xu, W.Z. Lu, X.Y. Zhang, Determination of the spread parameter in the Gaussian kernel for classification and regression, Neurocomputing 55 (2003) 643–663.

[50] E.R. Collantes, W.J. Dunn, Amino acid side chain descriptors for quantitative structure-activity relationship studies of peptide analogues, J. Med. Chem. 38 (1995) 2705–2713.