Available online at www.sciencedirect.com
ScienceDirect
European Journal of Medicinal Chemistry 44 (2009) 1144–1154
EUROPEAN JOURNAL OF MEDICINAL CHEMISTRY
http://www.elsevier.com/locate/ejmech

Original article

A set of new amino acid descriptors applied in prediction of MHC class I binding peptides

Guizhao Liang [a,*], Li Yang [a], Zecong Chen [b], Hu Mei [a], Mao Shu [a], Zhiliang Li [a,b]

[a] College of Bioengineering, Chongqing University, Shapingba Road 174#, Chongqing 400030, PR China
[b] College of Chemistry and Chemical Engineering, Chongqing University, Chongqing 400030, PR China

Received 20 October 2007; received in revised form 10 June 2008; accepted 13 June 2008
Available online 26 June 2008

Abstract

A set of new amino acid descriptors, namely factor analysis scales of generalized amino acid information (FASGAI) involving hydrophobicity, alpha and turn propensities, bulky properties, compositional characteristics, local flexibility and electronic properties, was proposed to resolve the representation of peptide structures. FASGAI vectors were then used to represent the structures of 152 HLA-A*0201 restrictive T-cell epitopes with 9 amino acid residues. The features that are closely related to binding affinities were selected by genetic arithmetic, and the model based on partial least squares was developed to predict binding affinities. The model revealed promising predictive power, giving relatively high predictions for training and test samples. Further, the PreMHCbinding program at significantly lower computational complexity was exploited to predict MHC class I binding peptides. Quantitative structure–affinity relationship analyses demonstrated the bulky properties and hydrophobicity of the 3rd residue, bulky properties of the 2nd residue, hydrophobicity of the 9th residue that provided high positive contribution to the binding affinities, and that the hydrophobicity of the 4th residue and local flexibility of the 3rd residue were negative to binding affinities. The results showed that FASGAI vectors can be further utilized to represent the structures of other functional peptides; moreover, it has thus showed us further direction into the potential applications on relationship between structures and functions of proteins.

Keywords: FASGAI; HLA-A*0201; T-cell epitopes; QSAR; PLS; GA–PLS

1. Introduction

T-cells of human immune system continuously probe the existence of unwanted foreign antigens. Antigenic peptides called T-cell epitopes are present on the cell surface for recognition by T-cells through their receptors [1,2]. Major histocompatibility complex (MHC) molecules, which bind epitopes and express them to T-cells, play a crucial role in this process. The ability of the immune system to respond to a particular antigen varies between individuals according to their different patterns of MHC genes. The polymorphism of MHC molecules contributes to the diverse specificities of immune responses. It is pivotal to understand the specificity of a given MHC molecule in order to understand the nature of MHC bound peptides and to design effective T-cell vaccines. The MHC molecule has a peptide-binding groove with which it binds peptides in a highly promiscuous manner, and which is comprised by two α-helices supported by a β-sheet. The principles of peptide interacting with MHC class I and MHC class II are rather similar [3]. The whole length of the binding peptide is accommodated inside the binding groove of the MHC class I molecule [4]. The peptides presented by MHC class I molecules mainly contain 8–11 amino acid residues, with a small number of exceptions [5]. The peptide-binding groove's interaction is affected by primary and secondary anchors, which are the positions that provide the highest contribution to peptide binding within the peptide [6]. The presence of anchors is necessary, but not sufficient, for high-affinity binding. As an important determinative factor for MHC binding, identification of specific

* Corresponding author.
E-mail address: sdqdlgz@163.com (G. Liang).

0223-5234/$ - see front matter © 2008 Elsevier Masson SAS. All rights reserved.
doi:10.1016/j.ejmech.2008.06.011

motif has triggered considerable interest [7]. Although plenty of peptides have already been synthesized and tested, yet basic characters of affinity interaction between peptides and MHC molecules have not completely been understood.

The characterization of this affinity interaction is the key to understanding immunity. One or two peptides in every 200 random peptides bind with high affinity to a given MHC class I molecule [8,9]. For the most common MHC class I ligands with 9-mer peptides, the sequence variation reaches $20^9 = \sim 5 \times 10^{11}$. It is time-consuming and costly to determine their binding affinities entirely by experiments. Therefore, a reliable theoretical method to rapidly extract information about MHC binding properties of peptides will be of great practical utility. Besides, identification of various affinity motifs [10], thousands of specific allelic genes, promiscuous MHC binders and T-cell epitopes [11] provides a great deal of abundant information for utilizing computers to predict the MHC-affinity peptides by computational means. Several approaches such as simple/linear motifs [12], quantitative matrices [13], artificial neural networks [14], fuzzy neural networks [15], support vector machines [16], fuzzy classifier with the SWEEP operator method [17], and Hidden Markov Models [18,19] have already been used to predict peptide–MHC binding. Molecular dynamics simulations [20] and homology modeling [21] have also been applied to investigate MHC–peptide interactions. Nevertheless, most of them mainly concentrated on the description of calculational approaches. In this regard, a few researchers deeply investigated the reasonable structural representation of diversified peptides, which may result in little useful information about specific functional motifs.

An outstanding computational approach should extract not only position specific information, but also some string patterns on peptides [19,22]. As we all know positional, compositional and physicochemical characteristics of amino acids in a peptide are responsible for its functional properties. Therefore, it is important to rationally represent the structure of a given peptide in order to acquire helpful information. Amino acid descriptors are often used to describe the structures of peptides. The first amino acid descriptors were proposed by Sneath [23], who used physicochemical semiquantitative data to derive them for the 20 coded amino acids. Since then, a number of descriptors have been generated [24–27]. It should be mentioned that the z descriptors reflecting hydrophobicity, size, and electronegativity were obtained with principal component analysis from 29 physicochemical parameters of the 20 coded amino acids by Hellberg et al. [25], and have been successfully applied to investigate the qualitative structure–activity relationships of oligopeptides. Another set of amino acid descriptors was principal component score vectors of hydrophobic, steric, and electronic properties which were advocated by Mei et al. [28], and which have been employed to study the qualitative structure–activity relationships of bitter tasting dipeptides, angiotensin-converting enzyme inhibitors, and bradykinin-potentiating pentapeptides.

The present work was to develop a set of new amino acid descriptors. Then, it were used to characterize the structural characteristics of 152 peptides binding to the human MHC allele HLA-A*0201. Variable selection technique based on genetic algorithm was applied to remove redundant descriptors. A quantitative structure–affinity relationship (QSAR) model by partial least squares (PLS) was established to relate affinities of peptides to their structures, and to explore some diversified properties influencing binding affinities. Further, the PreMHCbinding program at significantly lower computational complexity was exploited to predict MHC class I binding peptides.

2. Materials and methods

2.1. Factor analysis scales of generalized amino acid information

Functions and structures of peptides or proteins are determined by the information contained in the amino acid sequence [29]. Hence, 335 significative properties (Table S1 in Supplementary material) of the 20 coded amino acids, representing alpha and turn propensities, beta propensity, hydrophobicity properties, physicochemical properties, composition properties and so on, were culled from the AAindex database [26,30,31] based on relative loading coefficients and communalities of the variables from initial factor analysis, along with relative ease of interpretation and perceived structural importance.

Exploratory factor analysis [32], as a powerful statistical procedure, was used to produce a subset of numerical variables that would summarize the entire constellation of all 335 properties. Factor analysis simplifies high-dimensional data by generating a smaller number of ''factors'' that describe the structure of highly correlated variables [32]. The resultant factors are linear functions of the original data, fewer in number than the original, and reflect clusters of covarying variables that describe the underlying or ''latent structure'' of the variables. Factor analysis models assume that observation i denoted as $x_i \in \mathbf{R}^p$, can be decomposed into $x_i = \Lambda f_i + u_i$, where Λ is a loading matrix which represents the relative weight given to each factor, $\mathbf{R}^k \rightarrow \mathbf{R}^p$ is linear, $f_i \sim N_k(0, I_k)$, f_i is a factor matrix, $u_i \sim N_p(0, \psi)$, where u_i is a matrix which represents the residual part, ψ is a diagonal matrix [=diag(ψ_1,…, ψ_n)] and ψ_i is the relative variance in the ith property that cannot be accounted for the factors extracted, all f_i and u_j are independent, and $k < p$. The new set of inferred variables f_i is called common or latent factors, whereas u_j are called unique factors. Factor analysis differs from principal component analysis in that the latter does not distinguish between common and unique variances; with principal components, all $u_i = 0$.

A number of different extraction methods, including maximum likelihood, principal component, principal axis extraction and so on, can be used [33]. Here, principal component method was used to extract and obtain several clusters of highly intercorrelated physiochemical variables. The factor coefficient contained in the Λ matrix is a regression coefficient quantifying the relationship between the variable and the

common factor. The interpretability of factors is improved through rotation. Such rotations maximize the loading of each variable on the extracted factors whilst minimizing the loading on all other. There are two major categories of rotations, orthogonal rotations, which produce uncorrelated factors, and oblique rotations, which produce correlated factors [33]. Here, because there was a certain correlation between diversified properties of proteins or peptides, promax algorithm with Kaiser normalization, as an oblique solution, was used to obtain correlated factors to simple structure to improve their interpretation.

Interpretation on loadings demonstrated that the 1st 6 factors (Tables S1 and S2 in Supplementary material) posses straightforward and physiochemical information, reflecting hydrophobicity, alpha and turn propensities, bulky properties, composition characteristics, local flexibility and electronic properties, respectively. The 5th and the 6th factors were still considered because they obviously represent physiochemical information, although they explained relatively little variance. Six factors accounted for an 83.5% variance of these 335 variables according to the relationship between component number and eigenvalues (Table 1). Communality values (Table S1 in Supplementary material), the sum of the squared factor coefficients for each property, reflect the portion of the common variation in a variable by 6 factors. Communality value for each variable is all larger than 0.500; especially, many of the properties express high communality values (>0.9), suggesting that they have high factor coefficients on at least one factor and that the 6-factor model is sufficient. Factor analysis produces a new set of synthetic traits called factor scores (Table 1) that are linear combinations of the original variables. Here, these 6-factor score vectors are tentatively called factor analysis scales of generalized amino acid information (FASGAI). FASGAI summarized most information of the 335 properties, so it can be utilized to represent the structural features of peptides or proteins. Each residue in a sequence is described by 6 FASGAI vectors according to the varied amino acid position. Accordingly, the structural characteristics of a sequence with n amino acid residues were represented by the concatenation of $6 \times n$ FASGAI vectors.

2.2. Epitope data

The structures of 152 HLA-A*0201 restrictive T-cell epitopes and their logarithmic values of $1/IC_{50}$ (pIC_{50}) (Table S3 in Supplementary material) were collected by Doytchinova and Flower [34]. The binding affinities (IC_{50}) are assessed by a quantitative assay based on the inhibition of binding of a radiolabeled standard peptide, i.e. FLPSDYFPSV, to detergent-solubilized MHC molecules. An epitope with 9 amino acid residues was represented by the concatenation of 54 (6×9) FASGAI vectors.

2.3. Variable selection

Here, variable selection is completed by using genetic arithmetic–partial least squares (GA–PLS) as an effective variable selection tool nowadays, which is a sophisticated hybrid approach that combines GA as a powerful optimization method with PLS as a robust statistical method for variables' selection [35,36]. In GA–PLS, the chromosome and its fitness in the species correspond to a set of variables and internal predictive ability of the derived PLS model, respectively. The fitness of each chromosome is evaluated by the internal

Table 1
Factor scores for 335 property parameters of 20 coded amino acids

Amino acid (denotation in single letter)	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5	Factor 6
A	0.207	0.821	−1.009	1.387	0.063	−0.600
R	−1.229	0.378	0.516	−0.328	−0.052	2.728
N	−1.009	−0.939	−0.428	−0.397	−0.539	−0.605
D	−1.298	−0.444	−0.584	−0.175	−0.259	−1.762
C	0.997	0.021	−1.419	−2.080	−0.799	0.502
Q	−0.880	0.381	−0.044	−0.455	−0.040	0.405
E	−1.349	1.388	−0.361	0.213	0.424	−1.303
G	−0.205	−2.219	−1.656	1.229	−1.115	−1.146
H	−0.270	0.461	−0.024	−1.407	0.001	0.169
I	1.524	0.536	0.809	0.734	−0.196	0.427
L	1.200	1.128	0.703	1.904	0.536	−0.141
K	−1.387	0.572	0.285	0.333	−0.169	1.157
M	0.886	1.346	0.277	−0.913	0.007	−0.265
F	1.247	0.293	1.336	−0.026	0.012	−0.015
P	−0.407	−2.038	−0.564	−0.128	3.847	−1.008
S	−0.495	−0.847	−1.079	0.582	0.035	−0.068
T	−0.032	−0.450	−0.610	0.341	0.117	0.577
W	0.844	−0.075	2.069	−1.360	−0.810	−0.380
Y	0.329	−0.858	1.753	−0.479	−0.835	0.289
V	1.332	0.545	0.029	1.026	−0.229	1.038
Eigenvalues	131.67	54.32	39.88	24.01	16.49	13.30
Percent_explained	39.3	16.2	11.9	7.2	4.9	4.0
Cumulative percent_explained	39.3	55.5	67.4	74.6	79.5	83.5

predictive ability of the PLS model derived from a binary bit pattern. The internal predictive performance of the model is expressed in terms of a cross-validation square of multiple correlation coefficient value (hereafter, denoted by Q^2_{cv}) by the leave-one-out procedure as follows:

$$Q^2_{cv} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad (1)$$

where y_i and $\hat{y}_i$ represent the observed value and the predicted value of the dependent variable, respectively; $\bar{y}$ indicates the mean observed value of the dependent variable; n is the number of samples.

2.4. Partial least squares (PLS) modeling

PLS is mainly used for modeling linear regression between multi-dependent variables and multi-independent variables [37,38]. PLS also has the desirable property that the precision of the model parameters is improved with the increasing number of relevant variables and observations. The PLS regression algorithm consists of outer relations (x and y blocks individually) and an inner relation linking both blocks:

$$x_{ik} = \sum_{a=1}^{A} t_{ia} p_{ak} + e_{ik} \quad (2)$$

$$x_{im} = \sum_{a=1}^{A} u_{ia} c_{am} + g_{im} \quad (3)$$

The t and u latent variables are correlated through the inner relation given below which leads to the estimation of y from x.

$$\hat{u} = bt \quad (4)$$

2.5. Model validation

In the present work, the leave-one-out cross-validation method was used to prove the internal predictive ability of a model obtained. The predictive performance of the model obtained was assessed by the prediction values of Q^2_{cv} (Eq. (1)). External validation can only be achieved by splitting the total data set into a training set for establishing the model and a test set for evaluating the predictive performance of the model. The external prediction power of the model developed was evaluated by an external cross-validated correlation coefficient (hereafter, denoted by Q^2_{ext}) as follows [36]:

$$Q^2_{ext} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_{tr})^2} \quad (5)$$

where y_i and $\hat{y}_i$ are the observed and predicted values over test set of the dependent variable, respectively; $\bar{y}_{tr}$ is the mean value of the dependent variable of the training set; n is the number of test set.

For comparison, the 102 peptides in the training set of Ref. [34] were also treated as training set of this study to construct a QSAR model and the remaining 50 peptides were regarded as test set to validate the external prediction power of the model.

2.6. One-way analysis of variance

In this study, one-way analysis of variance [32] was implemented to analyze the variables corresponding to each site in the three-group peptides classified by different intervals of binding affinities (low affinity: $pIC_{50} < 6.301$ ($IC_{50} > 500$ nM), intermediate affinity: $7.301 > pIC_{50} > 6.301$ (50 nM $< IC_{50} <$ 500 nM), high affinity: $pIC_{50} > 7.301$ ($IC_{50} <$ 50 nM)). These group sizes are 28, 59 and 65, respectively (Table S3 in Supplementary material). Test of homogeneity of variances was firstly carried through. If it was homogeneity of variances, then Fisher analysis of variances was performed; or else, Brown–Forsythe analysis of variances was done. If means of these corresponding variables differed significantly, then multiple comparisons would be implemented. Significance level is 0.05.

2.7. Software used

Factor analysis and one-way analysis of variance were implemented using the SPSS 13.0 statistical software. SIMCA-P (version 10.5, Umetrics AB, 2004) software was employed to perform the PLS analysis. GA–PLS program was written in M-file based on software Matlab (version 6.1.0.450 release 12.1, MathWorks, Natick, MA, 2001).

3. Results and discussion

3.1. Analysis of PLS modeling

Empirical parameters influencing the performance of GA–PLS were determined by experience from the series of GA–PLS studies. Parameters were set as follows: the number of population was 200, the maximum number of generations was 200, the generation gap was 0.8, the crossover frequency was 0.5, the mutation rate was 0.005, and the fitting function was Q^2_{cv}. An optimal model was obtained from 10 models trained. Finally, 25 FASGAI variables (Table 2) were selected to establish the model using PLS. If the increase of the cumulative Q^2_{cv} was less than 0.097 when a principal component was added to the model, then the component was excluded from the model for it could not explain any significant structure–activity trend. As a result, two significant components, accounting for 68.3% of variance, were given. Standard error of estimation for training set was $SE_{cum} = 0.510$. A 52.5% variance was cumulatively explained by the leave-one-out cross-validation procedure. Binding affinities of the test set were predicted by the PLS model; consequently, Q^2_{ext} of 0.620 and standard error of prediction (SP_{ext}) of 0.567 for test set were obtained.

Table 2
Description, coefficient, and variable importance in projection (VIP)[a] of the PLS model

No.	Description of variables	Coefficients	VIP
–	Constant	8.255	–
1	Bulk properties of the 1st residue	0.124	1.044
2	Local flexibility of the 1st residue	−0.104	0.653
3	Hydrophobicity of the 2nd residue	0.128	0.784
4	Alpha and turn propensities of the 2nd residue	0.148	0.815
5	Bulk properties of the 2nd residue	0.182	0.997
6	Local flexibility of the 2nd residue	0.06	0.308
7	Hydrophobicity of the 3rd residue	0.186	0.979
8	Bulk properties of the 3rd residue	0.254	1.318
9	Local flexibility of the 3rd residue	−0.205	1.088
10	Hydrophobicity of the 4th residue	−0.269	1.485
11	Alpha and turn propensities of the 4th residue	−0.066	0.909
12	Local flexibility of the 4th residue	0.084	1.051
13	Electronic properties of the 4th residue	−0.156	1.279
14	Bulk properties of the 5th residue	0.127	0.782
15	Electronic properties of the 5th residue	0.170	0.886
16	Hydrophobicity of the 6th residue	0.106	0.993
17	Alpha and turn propensities of the 6th residue	−0.086	1.384
18	Composition characteristics of the 6th residue	0.022	0.575
19	Electronic properties of the 6th residue	0.041	0.978
20	Hydrophobicity of the 7th residue	0.127	1.144
21	Bulk properties of the 7th residue	0.109	0.801
22	Electronic properties of the 7th residue	−0.126	1.236
23	Hydrophobicity of the 8th residue	−0.159	0.958
24	Hydrophobicity of the 9th residue	0.181	0.927
25	Bulky properties of the 9th residue	−0.131	0.774

[a] Variable importance in projection (VIP) is the sum of the variable influence over all model dimensions and is a measure of variable importance.

Scores of the PLS model (Fig. 1) show that the high-dimensional properties of independent variables of two samples may be similar to each other when they relatively approach. It can be seen that other sample dots locate in Hotelling T^2 ellipse except the 14th, the 44th, and the 63rd samples, which demonstrate that the high-dimensional properties of the independent variables of these three samples may obviously be different from those of other samples. Loadings of the PLS model (Fig. 2) show that loadings of the 8th and the 20th variables are higher (>0.250) than those of other variables in the 1st principal component, and are positively correlated with affinities (Y). These variables refer to bulky properties of the 3rd residue and hydrophobicity of the 7th residue, respectively.

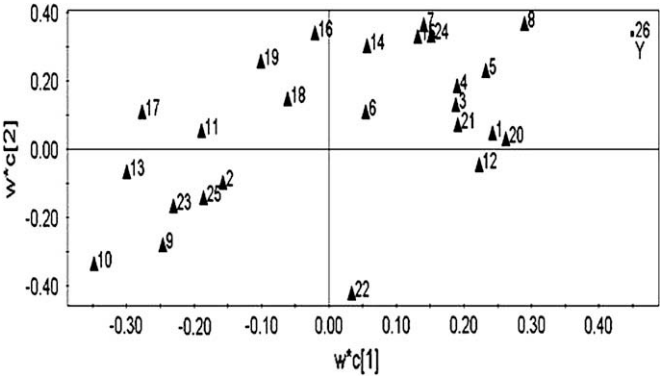

Fig. 2. Loadings of the PLS model.

Loadings of the 10th, the 13th, and the 17th variables are all less than −0.250, which shows hydrophobicity and electronic properties of the 4th residue, and alpha and turn propensities of the 6th residue are negative to binding affinities, respectively. In loadings of the 2nd principal component, the 7th, the 8th, the 14th, the 15th, the 16th, the 19th and the 24th variables, i.e., hydrophobicity and bulky properties of the 3rd residue, bulky properties and electronic properties of the 5th residue, hydrophobicity and electronic properties of the 6th residue, and hydrophobicity of the 9th residue, make relatively large contribution to them; on the contrary, the 9th, the 10th and the 22nd variables, involving local flexibility of the 3rd residue, hydrophobicity of the 4th residue, and electronic properties of the 7th residue, respectively, are negatively correlated with binding affinities. Other variables with little contribution to loadings of the 1st two principal components are not analyzed here. The PLS model was further validated by response permutations. The 20-random-permutation validation of the PLS model (Fig. 3) indicates that the intercepts of cumulative multiple correlation coefficient (R^2_{cum}) and Q^2_{cv} are 0.151 and −0.249, respectively. Hence, it was considered that the relatively large R^2_{cum} and Q^2_{cv} were not a result of accidental factors. (For a model to be valid, the desirable intercept limits should be $R^2_{cum} < 0.300$ and $Q^2_{cv} < 0.050$ [39].)

Prediction errors for training set show that the errors for the 3rd sample (HLESLFTAV), the 5th sample (LLSCLGCKI) and the 7th sample (TLLVVMGTL) are all larger than 1.000

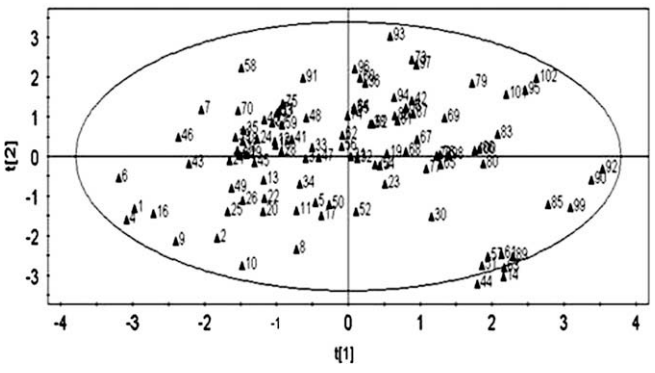

Fig. 1. Scores of the PLS model.

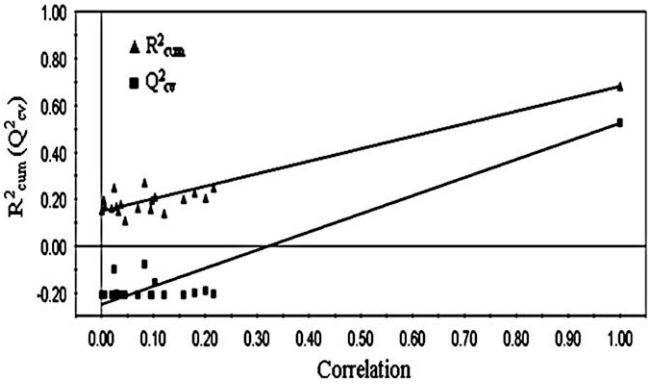

Fig. 3. The 20-random-permutation validation of the PLS model.

Table 3
Comparison between different QSAR models

Descriptors	Correlation method	Data size	Outliers	Fields	A[a]	R^2_{cum}[b]	SE_{cum}[c]	Q^2_{cv}[d]	R^2_{ext}[e]	Q^2_{ext}[f]	SP_{ext}[g]
CoMFA[h]	PLS	102/50[i]	4	Steric + electrostatic	6	0.911	nd[j]	0.480	<0.5	nd	nd
CoMSIA[h]	PLS	102/50	4	Hydrophobic + steric + H-bond donor	5	0.870	0.294	0.542	0.679	nd	0.390
FASGAI	PLS	102/50	0	54 Variables	2	0.659	0.510	0.352	0.574	0.576	0.567
FASGAI	(GA)PLS	102/50	0	25 Variables	2	0.683	0.492	0.525	0.619	0.620	0.538

[a] The number of principal components.
[b] Cumulative multiple correlation coefficient for the regression of predicted and observed activities of training set.
[c] Standard error of estimation for training set.
[d] A cross-validation square of multiple correlation coefficient value.
[e] Coefficient of determination for the regression of predicted and observed activities of test set.
[f] An external cross-validation correlation coefficient.
[g] Standard error of prediction for test set.
[h] The method was implemented using Sybyl 6.6 molecular modeling software and corresponding results can be found in Ref. [33].
[i] The two numbers separated by slashes denote the numbers of compounds in the training and test sets, respectively.
[j] Not determined.

(logarithm value) (Table S3 in Supplementary material). There may be three possible reasons for large errors of these peptides: an incorrectly measured experimental value, a different binding conformation, or a significant difference in the physicochemical properties [34]. We can see that the 1st residue of the 3rd peptide is histidine (H), and that there is only one peptide (HLAVIGALL) whose 1st position is histidine residue in the training set. Therefore, it can be presumed that a significant difference in the physicochemical properties resulting from histidine residue with positive charge may lead to the large error of the 3rd sample. By comparison, the prediction error for the 5th sample (LLSCLGCKI) in Ref. [34] was also relatively large, so we speculate that an incorrect T-cell measured experimental value may result in the large error of this sample. The 7th residue of the 7th sample (TLLVVMGTL) is the same glycine (G) residue as the 7th residue of the only sample (VLLDYQGML) in the training set. Consequently, it can be forecasted that the specific effect of the steric conformation of glycine causes the large error of this peptide.

3.2. Models' comparison and evaluation

A training set of 102 peptides with four outliers eliminated, that is, only 98 peptides with affinities for the class I MHC HLA-A*0201 molecule was investigated to develop affinity prediction models using comparative molecular field analysis (CoMFA) and comparative molecular similarity indices analysis (CoMSIA) by Doytchinova and Flower [34].

Although the results by CoMSIA were slightly better than those by this work in a certain degree (Table 3), it is comparatively complicated to optimize the structure of a specific peptide, and obtain perfect parameters of modeling by CoMSIA [40] or CoMFA [41]. Markedly, FASGAI vectors cherish straightforward physicochemical information and strong characterization repertoire besides much convenience and easy manipulation.

Besides, modeling results by the FASGAI representation and PLS modeling process (FASGAI–PLS) were inferior to those of the combined approach involving the FASGAI representation, variable selection by GA–PLS, and PLS modeling measure (Table 3), which shows variable selection by GA–PLS is valid for this object studied. Therefore, it can be concluded that the model based on the FASGAI–GA–PLS procedure is reliable and robust, and that it can be further utilized to predict MHC class I binding peptides.

3.3. Specificity analysis

Higher VIP values indicated good correlation between the variable and the data. It can be seen from Table 2 that VIPs of 9 variables are all larger than 1.000. Coefficient values of these variables in the PLS model show that the bulky properties and hydrophobicity of the 3rd residue, bulky properties of the 2nd residue, and hydrophobicity of the 9th residue make relatively high positive contribution to binding affinities, reversely, the hydrophobicity of the 4th residue and local flexibility of the 3rd residue produce relatively high negative contribution to binding affinities. That is to say, the larger the absolute values of the variable properties are, the higher the influences from the variable properties on binding affinities are. Preferred amino acids at positions that are important for antimicrobial activity of peptides containing 9 residues are tabulated in Table 4. Peptides with high binding activity can be obtained by the alteration of these important amino acid residues.

Some anchors have been described as leucine at position 2, leucine or valine at the C-terminal end [42], and prominent roles for several other positions (1, 3, and 7), so-called secondary anchor residues have also been demonstrated [4,43,44].

Table 4
Preferred amino acids at positions which are important for activities of T-cell epitopes

Amino acid positions[a]	Properties	Preferred amino acids	Contribution to peptide activity
The 3rd residue	Bulky properties	W, Y, F	Positive
The 3rd residue	Hydrophobicity	K, E, D, R, N	Positive
The 2nd residue	Bulky properties	W, Y, F	Positive
The 9th residue	Hydrophobicity	K, E, D, R, N	Positive
The 4th residue	Hydrophobicity	K, E, D, R, N	Negative
The 3rd residue	Local flexibility	P, L, E	Negative

[a] The 1st amino acid position starts from N-terminal of the peptide.

Table 5
Analysis of variance of 25 variables included in the PLS model

No	Description of variables	Mean	Mean	Mean	Test of homogeneity of variances				ANOVA			Robust tests of equality of means				
					Levene statistic	df1[a]	df2[b]	Sig.	*F*	Sig.	Result	Statistic	df1[a]	df2[b]	Sig.	Result
1	Bulk properties of the 1st residue	−0.131	0.169	0.640	0.103	149	2	0.902	6.282	0.002	Significant	6.394	114.898	2	0.002	
2	Local flexibility of the 1st residue	−0.070	−0.129	−0.348	0.275	149	2	0.760	3.524	0.032	Significant	3.772	120.478	2	0.026	
3	Hydrophobicity of the 2nd residue	0.796	1.027	1.101	16.341	149	2	3.83E−07	4.222	0.016		3.205	59.277	2	0.048	Significant
4	Alpha and turn propensities of the 2nd residue	0.548	0.853	0.992	7.803	149	2	0.0006	6.518	0.002		4.963	59.331	2	0.010	Significant
5	Bulk properties of the 2nd residue	0.119	0.420	0.530	11.603	149	2	0.00002	5.972	0.003		4.891	69.913	2	0.010	Significant
6	Local flexibility of the 2nd residue	0.324	0.343	0.385	5.106	149	2	0.007	0.269	0.764		0.158	37.488	2	0.854	Non-significant
7	Hydrophobicity of the 3rd residue	0.122	0.120	0.380	0.134	149	2	0.875	1.504	0.226	Non-significant	1.466	103.928	2	0.236	
8	Bulk properties of the 3rd residue	−0.259	−0.243	0.437	2.836	149	2	0.062	8.050	0.0005	Significant	9.171	142.411	2	0.0002	
9	Local flexibility of the 3rd residue	0.126	−0.098	−0.186	0.281	149	2	0.755	2.248	0.109	Non-significant	1.877	73.210	2	0.160	
10	Hydrophobicity of the 4th residue	0.295	−0.237	−0.519	15.436	149	2	8.08E−07	9.113	0.0002		7.738	76.103	2	0.0009	Significant
11	Alpha and turn propensities of the 4th residue	0.096	−0.082	−0.665	8.373	149	2	0.0004	6.812	0.001		8.023	146.230	2	0.0005	Significant
12	Local flexibility of the 4th residue	−0.038	0.180	0.792	18.737	149	2	5.51E−08	4.391	0.014		5.532	137.101	2	0.005	Significant
13	Electronic properties of the 4th residue	0.373	0.212	−0.415	3.653	149	2	0.028	11.331	0.00003		12.776	137.503	2	8.15E−06	Significant
14	Bulk properties of the 5th residue	−0.208	−0.274	−0.081	0.121	149	2	0.886	0.571	0.566	Non-significant	0.587	116.200	2	0.557	
15	Electronic properties of the 5th residue	−0.315	0.030	0.065	0.681	149	2	0.508	1.806	0.168	Non-significant	1.896	121.360	2	0.155	
16	Hydrophobicity of the 6th residue	−0.061	0.350	−0.028	2.605	149	2	0.077	4.627	0.011	Significant	4.813	121.555	2	0.010	
17	Alpha and turn propensities of the 6th residue	−0.473	−0.248	−1.013	0.382	149	2	0.683	6.439	0.002	Significant	6.193	100.978	2	0.003	
18	Composition characteristics of the 6th residue	0.398	0.371	0.131	12.992	149	2	6.29E−06	1.450	0.238		1.374	95.379	2	0.258	Non-significant
19	Electronic properties of the 6th residue	−0.314	−0.169	−0.428	1.013	149	2	0.365	1.875	0.157	Non-significant	1.653	80.470	2	0.198	
20	Hydrophobicity of the 7th residue	0.315	0.625	0.758	1.044	149	2	0.355	2.969	0.054	Non-significant	2.718	89.278	2	0.071	
21	Bulk properties of the 7th residue	0.040	0.049	0.291	0.351	149	2	0.705	1.250	0.289	Non-significant	1.168	93.429	2	0.316	
22	Electronic properties of the 7th residue	0.014	0.034	0.088	0.840	149	2	0.434	0.151	0.860	Non-significant	0.152	110.868	2	0.859	
23	Hydrophobicity of the 8th residue	0.420	0.327	−0.069	6.346	149	2	0.002	5.532	0.005		5.204	96.535	2	0.007	Significant
24	Hydrophobicity of the 9th residue	1.105	1.165	1.177	2.221	149	2	0.112	0.286	0.752	Non-significant	0.224	61.249	2	0.800	
25	Bulky properties of the 9th residue	0.357	0.299	0.075	3.875	149	2	0.023	3.908	0.022		3.854	107.908	2	0.024	Significant

[a] df1: Degree of freedom within groups.
[b] df2: Degree of freedom between groups.

Table 6
Analysis of variance of 29 variables excluded from the PLS model

No.	Description of variables	Mean	Mean	Mean	Test of homogeneity of variances				ANOVA			Robust tests of equality of means				
					Levene statistic	df1[a]	df2[b]	Sig.	*F*	Sig.	Result	Statistic	df1[a]	df2[b]	Sig.	Result
1	Hydrophobicity of the 1st residue	0.294	0.536	0.546	0.026	149	2	0.974	0.997	0.371	Non-significant	1.011	113.135	2	0.367	
2	Alpha and turn propensities of the 1st residue	0.082	0.097	−0.045	0.024	149	2	0.976	0.396	0.674	Non-significant	0.393	109.327	2	0.676	
3	Compositional characteristics of the 1st residue	0.589	0.443	0.322	1.142	149	2	0.322	0.814	0.445	Non-significant	0.748	91.038	2	0.476	
4	Electronic properties of the 1st residue	−0.135	0.009	0.226	0.047	149	2	0.954	3.476	0.033	Significant	3.547	112.976	2	0.032	
5	Compositional characteristics of the 2nd residue	1.011	1.453	1.445	1.797	149	2	0.169	2.966	0.055	Non-significant	2.658	81.802	2	0.076	
6	Electronic properties of the 2nd residue	0.128	0.025	−0.035	8.965	149	2	0.0002	1.739	0.179		1.421	68.855	2	0.248	Non-significant
7	Alpha and turn propensities of the 3rd residue	0.219	0.029	0.173	0.601	149	2	0.550	0.464	0.630	Non-significant	0.439	97.626	2	0.646	
8	Compositional characteristics of the 3rd residue	0.593	0.377	0.149	0.703	149	2	0.497	1.876	0.157	Non-significant	2	127.701	2	0.14	
9	Electronic properties of the 3rd residue	−0.233	−0.436	−0.299	1.649	149	2	0.196	0.978	0.378	Non-significant	0.94	100.802	2	0.394	
10	Bulky properties of the 4th residue	−0.019	−0.212	−0.476	1.585	149	2	0.208	4.250	0.016	Significant	3.852	87.981	2	0.025	
11	Compositional characteristics of the 4th residue	0.212	0.149	0.112	1.734	149	2	0.18	0.138	0.871	Non-significant	0.129	94.214	2	0.879	
12	Hydrophobicity of the 5th residue	0.388	0.355	0.396	0.542	149	2	0.583	0.034	0.967	Non-significant	0.033	103.363	2	0.967	
13	Alpha and turn propensities of the 5th residue	0.208	−0.076	−0.217	0.481	149	2	0.619	1.298	0.276	Non-significant	1.256	100.976	2	0.289	
14	Compositional characteristics of the 5th residue	0.841	0.589	0.668	0.156	149	2	0.856	0.687	0.505	Non-significant	0.655	98.982	2	0.522	
15	Local flexibility of the 5th residue	−0.153	−0.259	−0.234	0.596	149	2	0.553	0.280	0.756	Non-significant	0.301	117.594	2	0.741	
16	Bulky properties of the 6th residue	−0.497	−0.19	−0.258	3.022	149	2	0.052	1.350	0.262	Non-significant	1.393	119.229	2	0.252	
17	Local flexibility of the 6th residue	0.55	0.746	1.679	13.865	149	2	3.00E−06	5.500	0.005		5.922	126.829	2	0.003	Significant
18	Alpha and turn propensities of the 7th residue	−0.057	0.329	0.042	1.408	149	2	0.248	2.382	0.096	Non-significant	2.583	127.862	2	0.079	
19	Compositional characteristics of the 7th residue	0.257	0.584	0.406	2.050	149	2	0.132	1.270	0.284	Non-significant	1.274	111.328	2	0.284	
20	Local flexibility of the 7th residue	−0.228	0.12	0.371	7.261	149	2	0.001	3.404	0.036		4.633	120.216	2	0.012	Significant
21	Alpha and turn propensities of the 8th residue	0.128	0.121	−0.339	1.388	149	2	0.253	4.465	0.013	Significant	4.012	83.021	2	0.021	
22	Bulky properties of the 8th residue	−0.185	−0.232	−0.476	4.034	149	2	0.020	1.497	0.227		1.412	96.922	2	0.249	Non-significant

(*continued on next page*)

Table 6 (*continued*)

No.	Description of variables	Mean	Mean	Mean	Test of homogeneity of variances				ANOVA			Robust tests of equality of means				
					Levene statistic	df1[a]	df2[b]	Sig.	*F*	Sig.	Result	Statistic	df1[a]	df2[b]	Sig.	Result
23	Compositional characteristics of the 8th residue	0.770	0.397	0.413	3.530	149	2	0.0318	1.543	0.217		1.473	99.697	2	0.234	Non-significant
24	Local flexibility of the 8th residue	0.074	−0.0005	0.244	0.853	149	2	0.428	1.299	0.276	Non-significant	1.289	99.025	2	0.280	
25	Electronic properties of the 8th residue	−0.129	−0.013	−0.054	1.623	149	2	0.201	0.315	0.73	Non-significant	0.299	98.303	2	0.742	
26	Alpha and turn propensities of the 9th residue	0.730	0.812	0.661	5.475	149	2	0.005	3.467	0.034		3.16	89.520	2	0.047	Significant
27	Compositional characteristics of the 9th residue	1.230	1.312	1.137	6.714	149	2	0.002	1.343	0.264		1.309	102.510	2	0.275	Non-significant
28	Local flexibility of the 9th residue	0.091	0.137	−0.068	12.974	149	2	6.39E−06	6.272	0.002		6.007	100.541	2	0.003	Significant
29	Electronic properties of the 9th residue	0.413	0.268	0.606	0.064	149	2	0.938	4.618	0.011	Significant	4.262	89.828	2	0.017	

[a] df1: Degree of freedom within groups.
[b] df2: Degree of freedom between groups.

Studies showed that local hydrophobicity, steric bulk and hydrogen-bond donor ability of these anchors are crucial to the product of high binding affinities. By comparison, we found that most variables with high positive and negative coefficients in the PLS model were those related to anchor positions, and they produced remarkable impact on binding affinities. These variables reflected hydrophobicity, local secondary structural conformation, some physicochemical properties and so on. However, it should be noted that only the variable, i.e., the hydrophobicity of the 4th residue, is not related to these anchors which have been known nowadays. Is the 4th residue a potential anchor site? This is an open question which should be continued to be investigated.

The *F* test, the homoscedasticity, the significance level value, degrees of freedom, other corresponding parameters resulting from the analysis of variance for 25 variables included in the PLS model and 29 variables excluded from the PLS model are tabulated in Table 5 and Table 6, respectively. Investigations on the specificity difference between the low-, intermediate- and high-affinity peptides, and exploration of the magnitude of the difference will provide with reliable references for searching high-affinity peptides and understanding the binding mechanism.

3.4. The PreMHCbinding program

On the basis of this study, the PreMHCbinding program was exploited to predict MHC class I binding peptides with 9 amino acid residues.

The user interface of the program is displayed in Fig. 4. The program is practically easy to be operated. It contains four main parts including description, usage, attention, and text area for the sequence input and the result output and color pictures. Thereinto, the ''description'' part introduces the process on which the program is based, the ''usage'' and the ''attention'' parts recommend how the program is operated, and the ''text area'' part is used to input the sequence and obtain the prediction result, and color pictures provide favorable vision for the operator.

The PreMHCbinding program can be available via e-mail.

4. Conclusions

Prediction of MHC class I binding peptides is helpful for determining high-affinity peptides binding with particular MHC molecules, for understanding the immune mechanism and for designing and synthesizing efficient peptide bacterins. One desirable approach not only can give better predictions but also can provide considerable useful information. Multidimensional properties of the 20 coded amino acids were collected to acquire FASGAI vectors using multivariate statistical analyses. The structures of the peptides were represented by FASGAI vectors, and the features closely related to binding affinities were selected by GA–PLS. Afterward, the prediction model of MHC class I binding peptides was constructed using PLS. The properties tightly related to binding affinities were deeply analyzed, and the difference among

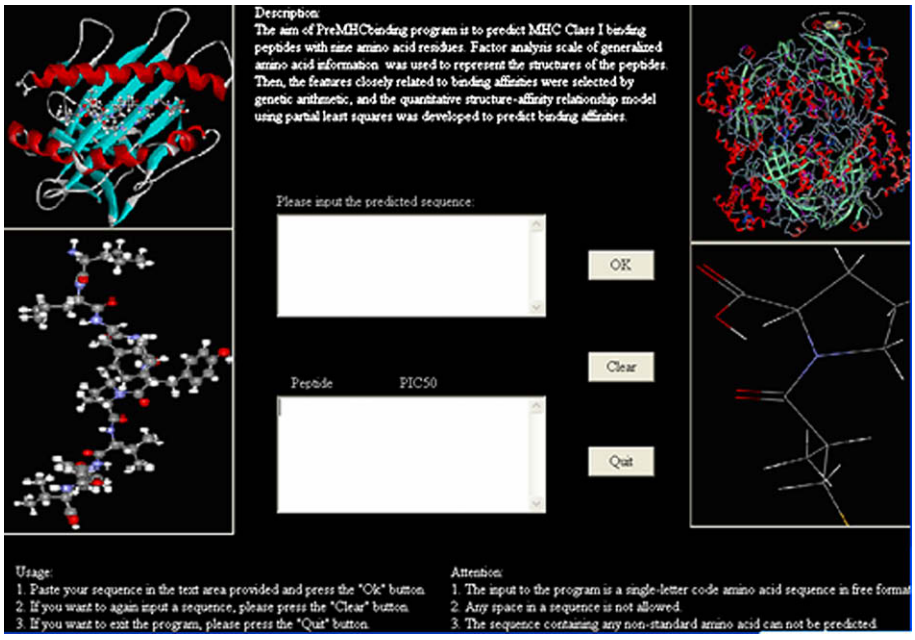

Fig. 4. The user interface of the PreMHCbinding program.

the properties corresponding to each residue was investigated by one-way analysis of variance, which brings a promising idea for predicting their affinities and designing high-affinity peptides, and which provides a possible solution to the difference between the sequence features, physiochemical aspects and affinities. The results indicated that the combination based on the FASGAI–GA–PLS procedure is robust and reliable for predicting MHC class I binding peptides; particularly, FASGAI vectors have many preponderant characteristics such as plentiful structural information, easy manipulation and high characterization competence. It can be further applied to investigate the relationship between structures and functions of other peptides, even for various proteins.

Acknowledgments

This work was supported by the National High-tech Research Program (the "863" Program) (2006AA02Z312), National 111 Programme of Introducing Talents of Discipline to Universities (0507111106) and Innovative Group Program for Graduates of Chongqing University, Science and Innovation Fund (200711C1A0010260).

Appendix. Supplementary material

Supplementary data associated with this article can be found in the online version, at doi:10.1016/j.ejmech.2008.06.011.

References

[1] H.G. Rammensee, K. Falk, O. Rotzschke, Annu. Rev. Immunol. 11 (1993) 213–244.
[2] P. Cresswell, Annu. Rev. Immunol. 11 (1994) 259–293.
[3] H.G. Rammensee, Curr. Opin. Immunol. 7 (1995) 85–96.
[4] D.R. Madden, D.N. Garboczi, D.C. Wiley, Cell 75 (1993) 693–708.
[5] V. Brusic, G. Rudy, L.C. Harrison, Nucleic Acids Res. 26 (1998) 368–371.
[6] V. Brusic, V.B. Bajic, N. Petrovsky, Methods 34 (2004) 436–443.
[7] H.-P. Adams, J.A. Koziol, J. Immunol. Methods 185 (1995) 181–190.
[8] K. Udaka, K.-H. Wiesmuller, S. Kienle, G. Jung, P. Walden, J. Biol. Chem. 270 (1995) 24130–24134.
[9] K. Udaka, K.-H. Wiesmueller, S. Kienle, G. Jung, P. Walden, J. Exp. Med. 181 (1995) 2097–2108.
[10] H.-G. Rammensee, J. Bachmann, N.P.N. Emmerich, O.A. Bachor, S. Stevanovic, Immunogenetics 50 (1999) 213–219.
[11] L.J. Stern, J.H. Brown, T.S. Jardetzky, J.C. Gorga, R.G. Urban, J.L. Strominger, D.C. Wiley, Nature 368 (1994) 215–221.
[12] H.G. Rammensee, T. Friede, S. Stevanoviic, Immunogenetics 41 (1995) 178–228.
[13] J. Hammer, E. Bono, F. Gallazzi, C. Belunis, Z. Nagy, F. Sinigaglia, J. Exp. Med. 180 (1994) 2353–2358.
[14] M.C. Honeyman, V. Brusic, N.L. Stone, L.C. Harrison, Nat. Biotechnol. 16 (1998) 966–969.
[15] H. Noguchi, T. Hanai, H. Honda, L.C. Harrison, T. Kobayashi, J. Biosci. Bioeng. 92 (2001) 227–231.
[16] M. Bhasin, G.P. Raghava, Bioinformatics 20 (2004) 421–423.
[17] H. Takahashi, H. Honda, J. Biosci. Bioeng. 101 (2006) 137–141.
[18] K. Udaka, H. Mamitsuka, Y. Nakaseko, N. Abe, J. Biol. Phys. 28 (2002) 183–194.
[19] H. Mamitsuka, Proteins. 33 (1998) 460–474.
[20] S.Z. Wan, P. Coveney, D.R. Flower, J. Comput. Chem. 25 (2004) 1803–1813.
[21] A. Kosmopoulou, M. Vlassi, A. Stavrakoudis, C. Sakarellos, M. Sakarellos-Daitsiotis, J. Comput. Chem. 27 (2006) 1033–1044.
[22] V. Brusic, G. Rudy, M. Honeyman, J. Hammer, L. Harrison, Bioinformatics 14 (1998) 121–130.
[23] P.H. Sneath, J. Theor. Biol. 12 (1966) 157–195.
[24] A. Kidera, Y. Konishi, M. Poka, T. Ooi, H.A. Scheraga, J. Protein Chem. 4 (1985) 23–55.
[25] S. Hellberg, M. Sjöström, B. Skagerberg, S. Wold, J. Med. Chem. 30 (1987) 1126–1135.

[26] K. Nakai, A. Kidera, M. Kanehisa, Protein Eng. 2 (1988) 93–100.
[27] J.L. Fauchere, M. Charton, L.B. Kier, A. Verloop, V. Pliska, Int. J. Pept. Protein Res. 32 (1988) 269–278.
[28] H. Mei, Z.H. Liao, Y. Zhou, S.Z. Li, Biopolymers 80 (2005) 775–786.
[29] C.B. Anfinsen, Science 181 (1973) 223–230.
[30] S. Kawashima, M. Kanehisa, Nucleic Acids Res. 28 (2000) 374.
[31] K. Tomii, M. Kanehisa, Protein Eng. 9 (1996) 27–36.
[32] R.A. Johnson, D.W. Wichern, Applied Multivariate Statistical Analysis, Prentice Hall, Upper Saddle River, New Jersey, 2002.
[33] A. Field, Discovering Statistics Using SPSS, second ed. Sage, London, 2005.
[34] I.A. Doytchinova, D.R. Flower, J. Med. Chem. 44 (2001) 3572–3581.
[35] C.B. Lucasius, G. Kateman, Trends Anal. Chem. 10 (1991) 254–261.
[36] A. Tropsha, P. Gramatica, V.K. Gombar, QSAR Comb. Sci. 22 (2003) 69–77.
[37] S. Wold, M. Sjöström, L. Eriksson, Chemom. Intell. Lab. Syst. 58 (2001) 109–130.
[38] R. Leardi, R. Boggia, M. Terrile, J. Chemometr 6 (1992) 267–281.
[39] P.M. Andersson, M. Sjostrom, T. Lundstedt, Chemom. Intell. Lab. Syst. 42 (1998) 41–50.
[40] G. Klebe, U. Abraham, T. Mietzner, J. Med. Chem. 37 (1994) 4130–4146.
[41] R.D. Cramer, D.E. Patterson, J.D. Bunce, J. Am. Chem. Soc. 110 (1988) 5959–5967.
[42] K. Falk, O. Rotzschke, S. Stevanovic, G. Jung, H.G. Rammensee, Nature 351 (1991) 290–296.
[43] J. Ruppert, J. Sidney, E. Celis, R.T. Kubo, H.M. Grey, A. Sette, Cell 74 (1993) 929–937.
[44] D.R. Madden, Annu. Rev. Immunol. 13 (1995) 587–622.