A New Set of Amino Acid Descriptors for the Development of Quantitative Sequence-Activity Modelings of HLA-A*0201 Restrictive CTL Epitopes

Gui-Zhao Lianga*( ) , Mao Shua ( ) and Sheng-Shi ZL. Li b ( )

a College of Bioengineering, Chongqing University, Chongqing 400030, P. R. China
b College of Chemistry and Chemical Engineering, Chongqing University, Chongqing 400030, P. R. China

A new set of amino acid descriptors, SZOTT (Scores of Zero dimension, One dimension, Two dimension, Three dimension) was derived from principal component analysis on a matrix of 1369 descriptors including 0D, 1D, 2D and 3D information of 20 coded amino acids. SZOTT scales were then employed to represent structures of HLA-A*0201 restrictive CTL epitopes, and further, quantitative sequence-activity modelings (QSAMs) were constructed based on partial least squares and back-propagation artificial neural networks. Satisfying results show that structural information related to biological activities of HLA-A*0201 restrictive CTL epitopes can be preferably represented by SZOTT scales, which may be a useful structural representation technique for study of QSAMs of peptides.

Keywords: Amino acids; CTL epitopes; SZOTT; QSAM; Principal component analysis; Partial least squares; Back-propagation artificial neural networks.

INTRODUCTION

Known as critical elements in life science, peptides have attracted considerable biological, pharmacological and medicinal chemical interest in recent years. With the development of the peptide library, thousands of peptides have been designed and synthesized. A QSAM provides a practical tool for the analysis of biological data just as UV, IR, MS and NMR technologies do. The idea behind a QSAM is that structural features can be correlated with biological activities, i.e., biological activities can be modeled as a function of molecular structures. As we all know, a precise amino acid sequence for a peptide is required for certain particular functions or biological activity. A QSAM will then indicate how the changes in peptide sequences are correlated with the variation in biological activities and how to modify the sequences to achieve the improved activities. In the sense of those, it is highly important to develop and investigate quantitative amino acid descriptors. Since some amino acid descriptors were used to construct some quantitative sequence-activity modelings of oxytocin-vasopressine analogues by Sneath, a number of quantitative amino acid descriptors have been put forward in the past few years. Particularly, a recent development in the QSAM field of peptides was the use of amino acid “z scores” which were scales of hydrophilicity (z1), bulk (z2) and electronic (z3) properties by principal component analysis relying on 29 physicochemical variables of 20 coded amino acids. Subsequently, the z scores were revised by adding new data to the multi-property matrix and eventually were extended to non-coded amino acids. The z scores have been proven to be powerful for describing structural characteristics of small peptides related to their activities with good results. More recently, VHSE (principal components score vectors of hydrophobic, steric, and electronic properties) was derived from a total of 50 physicochemical variables including 18 hydrophobic properties, 17 steric properties, and 15 electronic properties of 20 coded amino acids using principal component analysis by Mei et al. Then a comparison of the results to those obtained with z scores and other 2D or 3D descriptors showed the scales are comparable for parameterizing the structural variability of some oligopeptides. An excellent descriptor which is crucial to the success of QSAMs should contain as much structural information as possible related to biological ac-
activities. On the basis of those works above, a new set of amino acid descriptors, namely SZOTT (Scores of Zero dimension, One dimension, Two dimension, Three dimension), was proposed in the present study. SZOTT scales were then applied to express structures of HLA-A*0201 restrictive CTL epitopes. QSAMs were created by partial least squares and back-propagation artificial neural network. The results showed that SZOTT scales have strong competence for structure description in QSAMs of peptides.

**PRINCIPLES AND METHODS**

**Descriptor preparation**

A total of 1369 descriptors were collected to describe structural diversities for amino acid. Categories of the descriptors are as follows: (a) 0D-31 constitutional descriptors,10 molecular weight, sum of atomic properties, etc.; (b) 1D-69 descriptors including functional groups counts,10 atom centered fragments,11 molecular properties,11 etc.; (c) 2D-640 descriptors including topological electronegativity descriptors,12,13 topological charge indices,14 work and path counts,15 BCUTs descriptors,10 Galvez indices,16 autocorrelations,17 connectivity index,18 information index,16 eigenvalue-based indices,19 etc.; and (d) 3D-629 descriptors including Randić molecular profile,20 geometric descriptors,21 RDF descriptors,22 MoRSE descriptors,23 WHIMs descriptors,24 GETAWAY descriptors,25 etc. Matrices of various descriptors for 20 coded amino acids were available as supporting information.

**SZOTT scales generation**

Because the molecular descriptors collected may be highly correlated with each other, principal component analysis (PCA),26 as a useful tool for dimensionality reduction technique, was employed to compress the number of the descriptors, and then to obtain non-correlated linear combinations of the descriptors. As a result, the first 13 PCs with corresponding eigenvalues beyond 1 accounted for 96.19% of variable dispersion. Therefore, the first 13 PC scores could explain most information in the matrix of 1369 descriptors. So, the original data matrix could be replaced by the matrix including these 13 PCs scores. Here, these 13 score vectors were tentatively called SZOTT (Scores of Zero dimension, One dimension, Two dimension, Three dimension) (Table 1). Accordingly, each peptide is described by 13 SZOTT scales according to various amino acid positions. Therefore, a set of peptides varied in n positions can be expressed by the concatenation of 13 × n variables.

**Table 1. SZOTT (ti) scales for 20 coded amino acids**

<table>
<thead>
<tr>
<th>AA</th>
<th>Abbr.</th>
<th>t₁</th>
<th>t₂</th>
<th>t₃</th>
<th>t₄</th>
<th>t₅</th>
<th>t₆</th>
<th>t₇</th>
<th>t₈</th>
<th>t₉</th>
<th>t₁₀</th>
<th>t₁₁</th>
<th>t₁₂</th>
<th>t₁₃</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ala</td>
<td>A</td>
<td>-34.27</td>
<td>8.26</td>
<td>-5.60</td>
<td>5.73</td>
<td>-3.88</td>
<td>-3.64</td>
<td>2.37</td>
<td>-0.15</td>
<td>0.86</td>
<td>-0.11</td>
<td>4.20</td>
<td>-7.54</td>
<td>1.57</td>
</tr>
<tr>
<td>Gln</td>
<td>Q</td>
<td>3.87</td>
<td>-5.66</td>
<td>12.25</td>
<td>-3.74</td>
<td>0.28</td>
<td>0.60</td>
<td>1.80</td>
<td>6.48</td>
<td>-1.78</td>
<td>-3.44</td>
<td>0.36</td>
<td>-4.60</td>
<td>-2.27</td>
</tr>
<tr>
<td>Asp</td>
<td>D</td>
<td>-7.95</td>
<td>4.08</td>
<td>12.01</td>
<td>-11.02</td>
<td>2.62</td>
<td>1.87</td>
<td>-2.50</td>
<td>9.24</td>
<td>3.60</td>
<td>5.12</td>
<td>1.19</td>
<td>-5.92</td>
<td>5.35</td>
</tr>
<tr>
<td>Thr</td>
<td>T</td>
<td>-15.66</td>
<td>-1.88</td>
<td>-3.96</td>
<td>-7.52</td>
<td>-6.86</td>
<td>0.22</td>
<td>-1.35</td>
<td>-2.51</td>
<td>-6.64</td>
<td>-1.14</td>
<td>6.36</td>
<td>1.96</td>
<td>1.58</td>
</tr>
<tr>
<td>His</td>
<td>H</td>
<td>15.31</td>
<td>7.02</td>
<td>8.35</td>
<td>-4.85</td>
<td>-8.02</td>
<td>-6.50</td>
<td>6.07</td>
<td>1.41</td>
<td>6.48</td>
<td>6.71</td>
<td>1.14</td>
<td>0.90</td>
<td></td>
</tr>
<tr>
<td>Tyr</td>
<td>Y</td>
<td>28.81</td>
<td>8.70</td>
<td>-8.05</td>
<td>2.32</td>
<td>10.52</td>
<td>7.74</td>
<td>9.17</td>
<td>-6.89</td>
<td>-1.27</td>
<td>-2.45</td>
<td>-1.62</td>
<td>0.54</td>
<td>1.71</td>
</tr>
<tr>
<td>Ile</td>
<td>I</td>
<td>15.31</td>
<td>7.02</td>
<td>8.35</td>
<td>-4.85</td>
<td>-8.02</td>
<td>10.05</td>
<td>-1.66</td>
<td>-4.14</td>
<td>18.14</td>
<td>-4.70</td>
<td>5.93</td>
<td>0.86</td>
<td>0.91</td>
</tr>
<tr>
<td>Pro</td>
<td>P</td>
<td>-12.00</td>
<td>1.96</td>
<td>-10.48</td>
<td>0.81</td>
<td>20.48</td>
<td>6.56</td>
<td>9.29</td>
<td>-11.98</td>
<td>-11.58</td>
<td>0.24</td>
<td>-7.63</td>
<td>-9.35</td>
<td>-6.11</td>
</tr>
<tr>
<td>Val</td>
<td>V</td>
<td>-11.85</td>
<td>-13.58</td>
<td>-16.85</td>
<td>-2.68</td>
<td>-4.65</td>
<td>3.23</td>
<td>0.64</td>
<td>1.40</td>
<td>-0.74</td>
<td>1.77</td>
<td>-4.08</td>
<td>-1.86</td>
<td></td>
</tr>
</tbody>
</table>
Data collected and structure parameterization

HLA-A*0201 as an important histocompatibility complex (class I MHC) has been proved to be crucial to antigen presentation of both viral antigens and tumor antigens from a variety of cancers. Its 3D structure, with a limited number of different peptide ligands, namely CTL epitopes, has been explored by X-ray crystallography.\textsuperscript{27} CTL epitopes bonding to class I MHC play prominent roles in organism immunity. Studies on QSAMs of HLA-A*0201 restrictive CTL epitopes may be favorable for further synthesis of high-affinity peptides and shorten development periods of bacterins. SZOTT scales were used to characterize structures of HLA-A*0201 restrictive CTL epitopes. The data used in this work were reported by Doytchinova et al.\textsuperscript{28} In that study, a training set of 102 peptides with four outliers eliminated, that was, only 98 peptides with affinity for the class I MHC HLA-A*0201 molecule were investigated to develop 3D QSAMs using comparative molecular field analysis (CoMFA)\textsuperscript{29} and comparative molecular similarity indices analysis (CoMSIA).\textsuperscript{30} A test set containing 50 peptides was applied to explore external predictivity of models. The results by CoMSIA were satisfying. The sequences of 152 peptides and their binding affinities (IC\textsubscript{50}) with HLA-A*0201 were assessed by a quantitative assay based on the inhibition of binding of a radiolabeled standard peptide, i.e., FLPSDYFPSV, to detergent-solubilized MHC molecules, and the log values of 1/IC\textsubscript{50} (pIC\textsubscript{50}) treated as Y for QSAMs are shown in Supplementary Table 1. In this work, structural characteristics of every peptide with nine residues are represented by 117(13 × 9) SZOTT scales.

Variable selection

Redundant descriptors should be deleted from the developed model in order to promote its robustness and predictive capability, especially when the number of the variables is very large. Several variable selection methods such as stepwise multiple regression, genetic algorithm,\textsuperscript{31} generalized simulated annealing\textsuperscript{32} have been widely used to eliminate irrelevant variables. Here, variable selection was completed by genetic algorithm-partial least squares (GA-PLS), as a popular 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 variable selection. In GA-PLS, the chromosome and its fitness in the species correspond to a set of variables and internal predictivity of the derived PLS model, respectively. GA includes five steps: (1) The initial population of chromosomes is created by setting all bits in each chromosome to a random value; (2) The fitness of each chromosome is evaluated by the internal predictivity of PLS model derived from a binary bit pattern. The internal predictivity of the model is expressed in terms of a cross-validated $R^2$ (multiple coefficient of determination) value (hereafter, denoted by $Q^2_{cv}$, seen from the formula (1)) by the leave-one-out procedure; (3) The chromosome with the least number of the variables and the highest fitness is marked as an informative chromosome; (4) GA manipulation including crossover, mutation, and replication is carried out; and (5) The cycle of the above four steps (from step 2 to 4) is repeated until the optimal chromosomes are achieved. A detailed description of GA-PLS is given in a reference.\textsuperscript{33}

Internal test and external validation

An excellent QSAM not only should have favorable estimation ability for any internal sample, but also should have outstanding predictive ability for any external sample. The most usual method to prove a QSAM to have excellent internal predictivity is a cross-validation method. In the present work, leave-one-out cross validation for internal validation criteria was used. Predictive performance of the model is assessed by the prediction values of $Q^2_{cv}$, as follows:

$$Q^2_{cv} = 1 - \frac{PRESS}{SSY}$$

where $Q^2_{cv}$ is a cross-validated $R^2$ value by the leave-one-out procedure; SSY is the sum of the squared deviation of dependent variable values from their mean; PRESS is the predicted sum of squares obtained from the leave-one-out method.

In general, the model with high $Q^2_{cv}$ would have a high predictive ability. However, predictivity of a QSAM only by the leave-one-out procedure was insufficient as some authors have shown in some recent reports.\textsuperscript{34,35} Those models with apparent high predictivity, highlighted only by internal validation methods, could be unpredictable when they are used to verify new chemicals not used in developing QSAMs. Thus, for a stronger evaluation of model applicability for prediction of new chemicals, external validation of the model is required. External validation can only be achieved by splitting the total data set into a training set for establishing QSAMs and a test set for evaluating the performance of the models obtained. The external predictive power of a QSAM is evaluated by $Q^2_{ext}$ as follows:\textsuperscript{35}
where $y_i$ and $\hat{y}_i$ are respectively the observed and calculated (over test set) values of the dependent variables, and $\overline{y}_p$ is the averaged value of the dependent variables for the training set; the summations cover all compounds in the test set.

For comparison, the same 102 peptides as in the training set in the reference also were treated as a training set which was utilized to construct QSAMs and the same remaining 50 peptides were regarded as a test set in order to validate the external prediction power of the model.

**Partial least squares modeling**

Partial least squares (PLS) is the most used latent regression method for relating two data matrices, $X$ and $Y$, by a linear multivariate model. Compared to traditional methods, PLS can analyze the data with many, noisy, collinear, and even incomplete variables in both $X$ and $Y$. The desirable property of PLS is that the precision of the model parameters improve with the increasing number of relevant variables and observations. The PLS regression algorithm consists of outer relations ($X$ and $Y$ block individually) and an inner relation linking both blocks:

$$x_{ik} = \sum_{a=1}^{d} l_{ai} p_{ak} + e_{ik}$$

$$y_{im} = \sum_{a=1}^{d} h_{am} c_{am} + g_{im}$$

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

$$\hat{u} = bt$$

**Back-propagation artificial neural network modeling**

Back-propagation artificial neural network (BP ANN) is a methodology related to nonlinear regression techniques. In the present study, BP ANN was utilized to develop a nonlinear QSAM in order to investigate applicability of the SZOTT scales. Generally, BP ANN consists of three layers: the input layer, the hidden layer and the output layer. There are many advantages of techniques of BP ANN in contrast to traditional linear modeling methods. A challenge in training BP ANN, however, is the choice of appropriate network architecture, i.e., the number of neurons in the hidden layer. But there are no available rigorous theoretical rules which may be relied on to determine the number of neurons. Here, it is predetermined by means of the mean square error (MSE) which is used as an error function, and which is computed according to the following equation:

$$MSE = \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n-2}$$

where $y_i$ and $\hat{y}_i$ are the observed and predicted values of the dependent variables for training set, respectively, $n$ is the number of samples in training set.

Although it has been demonstrated that this technique was often superior to traditional linear approaches such as multiple linear regression, there was a danger of over-fitting. In order to avoid over-fitting, it should be noted that the size of the training set must be large enough to allow the learning process to converge, and proper external validation is needed.

**Software used**

One thousand three hundred and sixty-nine descriptor parameters of 20 coded amino acids were calculated by the Dragon program (free download at http://www.disat.unimib.it/chm/). PCA for data exploration was performed by Matlab 7.0 after the variables of the $X$ matrix were autoscaled. The GA-PLS and BP ANN were run in the environment of Matlab 7.0 as well. The variables of the $X$ matrix were mean centered and scaled to unit variants prior to PLS regression, which was implemented by SIMCA-P 10.0 software.

**RESULTS AND DISCUSSION**

**Results of variable selection**

Every peptide with nine amino acid residues is expressed by 117(13 x 9) SZOTT descriptors. GA-PLS was used on the data set to eliminate the autocorrelation vectors and optimize their descriptive power. The values of empirical parameters influencing the performance of GA-PLS were determined by experience from a series of GA-PLS studies. Parameters selected were as follows: the number of populations was 500, the maximum number of generations was 200, the generation gap was 0.8, the crossover frequency was 0.5, and the mutation rate was 0.005. The fitness of each chromosome was expressed by $Q^2_{cv}$, which showed the internal predictivity of the PLS model. The best
model with the highest $Q^2_{cv}$ was found (the fourth model) in 10 trained models (Supplementary Table 2), when 45 variables were selected to build following models based on PLS or BP ANN.

**QSAM studies by PLS modeling**

The 45 variables selected were employed to develop a model for the training set by PLS. The model with $R^2 = 0.778$ and $Q^2_{cv} = 0.515$ was obtained. The first two remarkable PCs $Y$ took part in the foundation of the model. The observed and calculated activities of binding affinities are gathered in Supplementary Table 1.

Regression between calculated and observed activities of CTL epitopes shows (Fig. 1) the simulation ability of model was satisfactory. PLS scores of CTL epitopes (Fig. 2) indicate their high dimensional properties of the independent variables may be similar to each other when the two samples are quite near. It can be seen that other samplings located in the Hotelling T2 ellipse, except for the 33rd sample, show that the high dimensional properties of the independent variables for this sample may be obviously different from those of other samples. The distance to the PLS model in the $X$ space is depicted in Fig. 3 in order to investigate efficiency on recombination for every sample. It can be seen that the 7th, 8th, 22nd, 33rd, 37th, 39th, 64th, 84th, 87th and 94th samples can not well be recombined on $X$ by two PCs because the normalized distance to $X$ over-

---

**Table 2. Comparison between different QSAMs of CTL epitopes**

<table>
<thead>
<tr>
<th>Descriptor</th>
<th>Data set size</th>
<th>Correlation method</th>
<th>Outliers</th>
<th>$R^2$</th>
<th>$Q^2_{cv}$</th>
<th>SEP</th>
<th>$R^2_{ext}$</th>
<th>$Q^2_{ext}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>CoMFA</td>
<td>102/50</td>
<td>PLS</td>
<td>4</td>
<td>0.827</td>
<td>0.477</td>
<td>nd</td>
<td>&lt; 0.500</td>
<td>nd</td>
</tr>
<tr>
<td>CoMSIA</td>
<td>102/50</td>
<td>PLS</td>
<td>4</td>
<td>0.890</td>
<td>0.542</td>
<td>0.563</td>
<td>0.679</td>
<td>nd</td>
</tr>
<tr>
<td>SZOTT</td>
<td>102/50</td>
<td>PLS</td>
<td>0</td>
<td>0.778</td>
<td>0.515</td>
<td>0.403</td>
<td>0.728</td>
<td>0.730</td>
</tr>
<tr>
<td>SZOTT</td>
<td>102/50</td>
<td>BP ANN</td>
<td>0</td>
<td>0.822</td>
<td>0.557</td>
<td>0.362</td>
<td>0.678</td>
<td>0.680</td>
</tr>
</tbody>
</table>

*The two numbers separated by slashes denote the numbers of compounds in the training set and in the test set, respectively.*

$R^2$: Multiple coefficients of determination for training set.

$Q^2_{cv}$: Multiple coefficients of determination for training set by the leave-one-out procedure.

SEP: Standard error of prediction.

$R^2_{ext}$: Multiple coefficients of determination for test set.

$nd$: Not determined.

---

Fig. 1. Regression between calculated and observed activities by PLS model.

Fig. 2. Scores by PLS model.

Fig. 3. Distance in $X$ space modeled by PLS model.
runs the critical value of 1.253 (significance level = 5%). It can be concluded from the plot of loadings by PLS (Fig. 4) that loadings of \(v_7, v_{14}, v_{16}, v_{17}\) and \(v_{38}\) in the first PC (> 0.200) are larger than those of other variables and show a positive correlation to \(Y(pIC_{50})\). Meanwhile, loading contributions of \(v_4, v_5, v_9, v_{14}, v_{17}, v_{20}, v_{37}\) and \(v_{43}\) (> 0.200) in the second PC to \(Y\) are larger than those of other variables. It should be noted that \(v_7, v_{14}\) and \(v_{17}\) gave more contributions to \(Y\) than the others did in two PCs. On the contrary, six variables, i.e., \(v_{24}, v_{27}, v_{30}, v_{40}, v_{42}\) and \(v_{45}\) result in obvious negative contributions to the first PC relative to other variables, and three variables, i.e., \(v_8, v_{21}, v_{25}\) and \(v_{33}\) lead to more negative contributions to the second PC than those of others. Table 2 summarizes some important statistical parameters of QSAMs of CTL epitopes. It can be seen clearly that, although the internal predictive ability of the model derived from PLS was slightly inferior to that derived from CoMSIA representation and PLS modeling (CoMSIA-PLS), its external predictivity was clearly superior to that by CoMSIA-PLS, and the standard error of prediction (SEP) was less than that of CoMSIA-PLS as well.

**QSAM studies by BP ANN modeling**

PCA was used to impress the number of variables in order to cut short the number of input dimensions for BP ANN. As a result, 17 PCs with corresponding eigenvalues larger than 1 were obtained. These PCs explained 74.03% variance of data matrix involving variables selected by GA-PLS. A three-layer BP ANN model was constructed in this work. The selected parameters of BP ANN were as follows: The learning rate \(\rho\) was randomly selected between 0 and 1, and the momentum \(\delta\) was varied from 0 to 1 at random. The transfer functions for the hidden layer and the outer layer were sigmoid and linear, respectively. A Nguyen-Widrow algorithm was implemented in the neural-network simulator for a suitable set of initial starting weights. The number of neurons in the hidden layer was an important factor determining the performance of BP ANN. It was optimized as mentioned below: when the number of iterations was fixed on 1000, the ANN configuration was optimized through an investigation on changes of the value of MSEs for the test set corresponding to the number of neurons in the hidden layer. The plot of changes of MSEs with the number of neurons are pictured in Fig. 5. It can be seen that there is the lowest MSE when the number of neurons is 3. Consequently, the best architecture of BP ANN with 17-3-1 was employed to set up the models. The over-training effect of ANN must be avoided in order to ensure the ability to predict. Therefore, a strategy was adopted to tackle this embarrassment aside from splitting the samples into a training set and a test set. Since the MSE by the PLS model was 0.164, as a reference, the MSEs for the test set were computed at 0.01 intervals between 0.1 and 0.2. Each step was simulated 20 times in order to remove happen-chance. The plot of changes of MSEs for the test set with those of the training set is depicted in Fig. 6. It can be seen that the lowest MSE for the test set is 0.244 when the MSE of the training set is 0.130. So, ANN was trained when the MSE of the test set was 0.244. Regression between observed and calculated activities of CTL epitopes (Fig. 7) shows the predictive results of the model obtained from BP ANN were gratifying. Statistical parameters (Table 2) indicate the internal predictivity of the model by BPANN was superior to that by PLS, but its external predictivity was worse than that by PLS. Meanwhile, its internal and external predictivity took on a slight superiority toward those of the model through CoMSIA-PLS.

![Fig. 4. Plot of loadings by PLS model.](image)

![Fig. 5. Plot of changes for the MSE of the test set with the number of hidden neurons.](image)
CONCLUSIONS

In this work, SZOTT scales were proposed to represent structures of HLA-A*0201 restrictive CTL epitopes. Afterward, QSAMs for them were developed by PLS and BP ANN. It was clear that SZOTT scales could contribute to the development and proposal of some simpler and validated QSAMs for HLA-A*0201 restrictive CTL epitopes. With SZOTT representation, PLS and BP ANN modeling for QSAMs performed better and with more comprehensive results in comparison with CoMSIA-PLS. There have already been many structure description techniques for QSAMs in computer-aided drug design. It should be kept in mind that, however, any measure for structure description could not reliably ensure its availability when aiming at the entire universe of compounds as some references have indicated. Therefore, its disadvantages, deficiencies and applied domains must be made clear before it is made full use of. When it came to a conclusion, however, SZOTT scales are simpler and easier descriptors which contained much structural information related to the activities of peptides, and they are expected to be a useful structural expression methodology for QSAMs of polypeptides. Further research on SZOTT scales on QSAMs is in progress.

ACKNOWLEDGEMENTS

The authors thank the National High-tech Research Program (the “863” Program) (Grant no. 2006AA02Z312) and Innovative Group Program for Graduates of Chongqing University, Science and Innovation Fund (200711C1A0010260).

Received January 12, 2006.

REFERENCES