TAP-Binding Peptides Prediction by QSAR Modeling Based on Amino Acid Structural Information

Yuanqing Wang1,2,§, Xiaoming Cheng3,§, Yong Lin2, Haixia Wen2, Li Wang4, Qingyou Xia*,1 and Zhihua Lin*,2

1Institute of Sericulture and Systems Biology, Southwest University, Chongqing, 400715, P.R. China
2College of Pharmacy and Bioengineering, Chongqing University of Technology, Chongqing, 400050, P.R. China
3Institute of Respiratory Diseases, Xinqiao Hospital, Third Military Medical University, Chongqing, 400037, P.R. China
4Institute of Immunology of PLA, Third Military Medica University, Chongqing, 400038, P.R. China

Abstract: The transporter associated with antigen processing (TAP) is essential for peptide delivery from the cytosol into the lumen of the endoplasmic reticulum (ER), where these peptides are loaded on a major histocompatibility complex (MHC) I molecules and form peptide-MHC complex. The peptide-MHC leaves the ER and displays their antigenic cargo on the cell surface to cytotoxic T cells. In this study, 89 physicochemical properties of amino acid were collected from AAIndex database, and used to characterize the peptides which were binding to TAP. Then, the stepwise regression (STR) was used to optimize the parameters which characterized the TAP binding peptides, and the multiple linear regression (MLR) was used to construct the quantitative structural activity relationship (QSAR) model based on optimized parameters. The quantitative models had good reliability and predictive ability: the Q² of “leave one out” validation is 0.676 and R² of test dataset is 0.722 respectively. Additionally, the standardized coefficients of the models could demonstrate the attributions for each position of epitope and determine which special amino acid is suitable at any position of the peptide. Therefore, the QSAR model constructed by STR-MLR has many advantages, such as, easier calculation and explanation, good performance, and definite physicochemical indication, which could be used to guide the design and modification of the TAP binding peptide.

Keywords: Transporter associated with antigen processing (TAP), major histocompatibility complex (MHC), TAP binding peptide, structural descriptors, stepwise regression (STR), multiple linear regression (MLR), quantitative structural activity relationship (QSAR), structural characterization.

1. INTRODUCTION

The mechanisms of presentation of intracellular antigens by major histocompatibility class I (MHC-I) molecules are well documented, and several of these steps have been described previously [1-3]. First, antigenic proteins are degraded into short peptides by different cytosolic complexes (the most highly implicated and frequently described being the proteasome). Second, peptides are translocated by the transporter associated with antigen-processing (TAP) into the endoplasmic reticulum (ER). Within the ER, these short peptides may undergo N-terminal trimming, whereas their C terminus is kept intact, and they bind newly synthesized empty MHC-I molecules. Third, stable peptide/MHC-I complexes are exported via the Golgi apparatus to the cell surface, where they can be recognized by CD8+ T cells expressing a specific T-cell receptor [4, 5]. Although the steps in antigen processing which determine a dominant T-cell response are still poorly understood, the TAP obviously plays a pivotal role in the cellular immune response. Therefore, elucidating the mechanisms that determine the antigenicity of epitopes for MHC-I presentation is critical for peptide-designed vaccines or immunotherapies that limit viral spread [6-8].

TAP transporter can transfer peptides of 8 to 40 amino acids, especially for 8 to 11 amino acids. Beside length preference, the nature of peptides also influences the peptide selectivity [9]. TAP prefers to transfer peptides with hydrophobic C terminus. Specifically, the peptides with hydrophobic residues at position 3 (P3), and those that are charged and hydrophobic residues at position 2 (P2) are favorable for transfer, but peptides with aromatic and acidic residues at position 1 (P1) are deleterious for transfer by TAP [10, 11].

In recent years, QSAR modeling becomes an important tool for drug design and structural optimization [12-14]. There is a variety of computational methods have been developed to predict TAP-binding peptides. These methods include following catalog: (1) artificial neural networks was used to construct quantitative model of TAP binding peptides [15-17]; (2) quantitative models were derived based on scoring function or quantitative matrix [18-21]; (3) support vector machine was applied to construct model for prediction of TAP affinity [19-22]. Additionally, some predictive systems of TAP binding peptide were provided...
online [17, 21], such as TAPPred, PREDTAP, and so on, and these websites could provide dataset, sequence and affinity prediction of TAP binding peptides.

In this study, 89 amino acid parameters come from AA index database [23, 24] were used to characterize the TAP binding peptides with 9 amino acids, which collected from ‘AntiJen’ [25, 26], then the stepwise regression (STR) was used to screen the structural parameters, and the predictive models were constructed by multiple linear regressions (MLR) based on screened parameters. The normalized regression coefficients (NRCs) could reflect the contributions of screened parameters, and the summation of NRCs could indicate the importance of any position, because the parameters screened to characterize every position were different. The model built here had good performance for TAP binding affinity prediction; also the models could be used to demonstrate the mechanism of interaction between TAP and peptides very well. More importantly, the QSAR model constructed by STR-MLR had more advantages, such as good correlation, definite physicochemical meaning, easier use and explanation, which could be used for QSAR research of other peptides.

2. MATERIALS AND METHODS

Data Source

The TAP binding peptides bound to MHC molecule from rat species were collected from ‘AntiJen’ database [25, 26]. The TAP binding affinity of peptides was expressed as relative IC\textsubscript{50}, the standard peptide and its concentration were ‘RRYNASTEL’ and 250 nM, more details of binding experiments was described in literature [27-29]. The sequence and TAP binding affinity (pIC\textsubscript{50}) are presented in the supplementary file. The peptides were divided into two groups. The training set, which is used to construct a predictive model validated by “leave-one-out” (LOO) validation; the test set, which is used to validate the predictive ability of the model. The test set was selected using the following steps: All sequences were sorted in descending order by their binding affinity. One in every five sequences was selected as the test dataset. Table 1 presents the MHC species, number of training and test set.

Parameters of Amino Acid

TAP is a heterodimer; the two subunits TAP1 and TAP2 each have an N-terminal membrane domain and a C-terminal nucleotide-binding domain (NBD). The membrane domains of TAP form the peptide-binding site and presumably a pore through which peptides are translocated, whereas ATP hydrolysis by the NBDs energizes the peptide translocation by the membrane domains [3]. The binding affinity between peptides and TAP relies on the hydrogen bonds, electrostatic and hydrophobic force. In this study, a set of amino acid physicochemical indices were used to characterize TAP binding peptides, which were collected from the AA Index database [24], including 544 indices. A total of 89 parameters which reflected the interaction between peptides and TAP, were selected to characterize the peptides, including 58 hydrophobicities, 22 physicochemical properties and 9 electrostatic properties.

Structural Characterization

There is an accepted assumption, in the quantitative structural activity relationship research of polypeptide; the activity of peptide relies on the total contribution of each amino acid. The characterization of peptide is usually performed by characterizing the amino acid of the peptide. Therefore, if the amino acid is characterized with ‘m’ parameters, the vector with ‘9*m’ elements was obtained and can be used for the structural characterization of non-peptide. For example, in this paper, 89 parameters were selected to characterize the amino acid of epitope peptide, which namely, \( V_1, V_2, \ldots, V_{89} \); thus, the structural parameters of a non-peptide could be presented as \( V_\text{p} \), where \( j (j=1, 2, \ldots, 9) \) is the position of amino acid and \( k (k=1, 2, \ldots, 89) \) is the order of parameters. These selected 89 parameters are listed in the supplementary file.

Method of Modeling

The most important structural variables that contributed to activity were screened by STR. The screened structural variables were then used to construct QSAR model by MLR. As mentioned previously in the MLR modeling, the dataset was divided into the training and test sets, which were used to construct and validate the predictive models respectively. The variables screened by STR were optimized by regulating the threshold of introducing (\( F_1 \)) and eliminating (\( F_2 \)), and the goals of optimization were to improve Q\textsuperscript{2} of LOO and R\textsuperscript{2} of test set for MLR modeling. Here, the method of modeling is called STR-MLR and all calculation was completed using Matlab on PC.

3. RESULTS AND DISCUSSION

The QSAR model was constructed for TAP binding peptides bound to rat MHC molecule by STR and MLR methods, based on amino acid structural information. The model had good reliability and predictive ability, the R\textsuperscript{2} and Q\textsuperscript{2} for training set were 0.738 and 0.676 respectively, and the R\textsuperscript{2} for test set was 0.722. To increase the reliability of QSAR model, the relative IC\textsubscript{50} of peptides were selected for modeling was determined using the same reporter peptide and standard concentration. The parameters used for characterization were optimized by STR based on their contribution for activity; thus, the parameters had definite physicochemical meaning. The optimized parameters were then used to construct QSAR model by MLR, and the model could reflect the contribution of parameters for activity directly. Additionally, the STR and MLR are general accepted methods used in the QSAR research and chemoinformatics, as they are simple and easy to use.

TAP binding peptides, pIC\textsubscript{50}, relative error, and NRCs were presented in the supplementary file, and the correlation plot diagrams of the observed and calculated binding affinity are shown in Fig. (1). Fig. (1) indicates the QSAR model has not only a good estimative ability for training set, but also a good predictive ability for the test set. The maximal relative error for calculated or predicted affinity was nearly 20%; most of them were within 10%. Although the database provided more than one affinity for some samples, in order to ensure the QSAR model was reasonable, one of them was...
selected using QSAR research randomly, which might cause some samples to have a larger relative error of about 15%. In general, the QSAR model can be used to predict affinity novel TAP binding peptide accurately.

![Graph showing observed and calculated affinity](image)

**Fig. (1).** Plot of observed and calculated (predicted) affinity of TAP binding peptides.

The physicochemical parameters of amino acids screened by STR and their NRCs are listed in Table 1. The NRCs of parameters can reflect their contributions for binding affinity, and their absolute values are related to significant of parameters directly. For example, larger ‘smoothed upsilon steric parameter’ and smaller ‘steric parameter’ at position 2 are favorable for binding affinity, because of positive and negative NRC respectively. The ‘smoothed upsilon steric parameter’ has the largest NRC, thus it has the most significant contribution for binding affinity.

In order to investigate the importance of amino acids at any position for binding affinity, the relative contribution are calculated and listed in Table 2 by following steps: (a) normalization of every parameter, (b) calculation the relative contribution (normalized parameter multiply corresponding NCR) of every parameter, (c) summation of relative contribution of parameters for each position. Table 2 indicates the importance of amino acids for binding affinity at any position, and the dominant amino acids have higher relative contribution. For example, Trp(W), Val(V) and Thr(T) are dominant amino acids for position 2. Conversely, there have Pro(P), Asn(N) and Asp(D) at position 2 are unfavorable for binding affinity.

Fig. (2) is the bar chart of summation of NRCs, which is calculated by summation of the positive and negative NRCs from the model (listed in Table 2). The contributions of every parameter at any position are denoted by NRCs of the model, and the summation of the NRC can reflect the contribution of every position of peptide for binding affinity. Fig. (2) shows the position 1, 2, 3 and 9 are more important than other positions, these are similar to the anchor site of MHC binding peptides. Although position 4, 5, 6, 7, 8 have influence for binding affinity, the contributions were not significant. The result is in conformity with that of literatures [27, 28]. Fig. (2) can provide some clues for design and modification of TAP binding peptide, which puts the natural (or non-natural) amino acids with the appropriate physicochemical properties at position 1, 2, 3, or 9 improving the binding affinity. According to the mechanism of epitope bound to MHC, modification of amino acid at position 1, 2, 3 and 9 is important for increasing binding affinity. Therefore, natural (or non-natural) amino acids with corresponding properties are put the position 1, 2, 3 and 9, based on TAP and MHC binding peptide QSAR research, which could increase binding affinity for TAP and MHC simultaneously, and then promote the transport rate of TAP and the binding strength of MHC. Consequently, the immune response would be enhanced significantly.

**CONCLUSION**

In this study, 89 indices were screened from the AA index database and used to characterize TAP binding peptides from rat species, and then the quantitative model was constructed by STR-MLR, indicating that the variables were optimized by STR, and optimized parameters were used to construct quantitative model by MLR. The

<table>
<thead>
<tr>
<th>Access Number</th>
<th>Description of Parameter</th>
<th>Reference</th>
<th>NRCs</th>
<th>Position</th>
</tr>
</thead>
<tbody>
<tr>
<td>FAUJ880102</td>
<td>Smoothed upsilon steric parameter</td>
<td>Fauchere et al., 1988</td>
<td>0.546</td>
<td>2</td>
</tr>
<tr>
<td>NOZY710101</td>
<td>Transfer energy, organic solvent/water</td>
<td>Nozaki-Tanford, 1971</td>
<td>0.524</td>
<td>9</td>
</tr>
<tr>
<td>WARP780101</td>
<td>Average interactions per side chain atom</td>
<td>Warme-Morgan, 1978</td>
<td>-0.422</td>
<td>2</td>
</tr>
<tr>
<td>EISD860103</td>
<td>Direction of hydrophobic moment</td>
<td>Eisenberg-McLachlan, 1986</td>
<td>-0.338</td>
<td>3</td>
</tr>
<tr>
<td>SUEM840102</td>
<td>Zimm-Bragg parameter</td>
<td>Sueki et al., 1984</td>
<td>-0.308</td>
<td>1</td>
</tr>
<tr>
<td>ZIMJ680102</td>
<td>Bulkiness</td>
<td>Zimmerman et al., 1968</td>
<td>-0.275</td>
<td>1</td>
</tr>
<tr>
<td>CHAM810101</td>
<td>Steric parameter</td>
<td>Charton, 1981</td>
<td>0.182</td>
<td>5</td>
</tr>
<tr>
<td>KRIW790101</td>
<td>Side chain interaction parameter</td>
<td>Krigbaum-Komoriya, 1979</td>
<td>0.167</td>
<td>9</td>
</tr>
<tr>
<td>CHAM820102</td>
<td>Free energy of solution in water, kcal/mole</td>
<td>Charton-Charton, 1982</td>
<td>-0.166</td>
<td>3</td>
</tr>
<tr>
<td>FAUJ880103</td>
<td>Normalized van der Waals volume</td>
<td>Fauchere et al., 1988</td>
<td>-0.146</td>
<td>8</td>
</tr>
</tbody>
</table>

Note: the table was sorted by absolute value of NRCs descending.
quantitative models had good reliability and predictive ability, and the $Q^2$ is 0.676 and $R^2$ of the test set is 0.722. The NRCs of the quantitative model could be used to determine the important position of the peptide and dominant amino acids for binding affinity. The summation of NRCs indicated the contributions of position 1, 2, 3 and 9 were more important than those of position 4, 5, 6, 7 and 8, which are similar to the anchor site of peptide bound to MHC. The QSAR model is not only used to predict binding affinity of novel peptide, but also guide design and modification of TAP binding peptide. Importantly, the QSAR model constructed by STR-MLR has some advantages, such as definite physicochemical meaning, easier to explain and use, which could be of use to QSAR research other peptides.

**SUPPLEMENTARY MATERIAL**

Supplementary material is available on the publishers Web site along with the published article. The details are as follows:

Parameter: parameter name and reference used in this paper.

Parameter Values: value of the parameters.

Characterization of Sequence: Sequences are characterized by screened structural parameters.

Datasets: dataset of TAP binding peptides for RAT species used to construct (training set) and test (test set) quantitative model.
ACKNOWLEDGEMENTS

This study was supported by National Natural Science Foundation of China (No. 60873103), Key Project of National Natural Science Foundation of China (No. 30830090), Create New Drugs National Major Projects (No. 2009ZX09503-005), Scientific Research Foundation of Education Committee of Chongqing (No. KJ100801), and Natural Science Foundation Project of CQ CSTC (No. CSTC, 2010BB5306).

REFERENCE