Article

Computational Prediction of the Specificities of Proteasome Interaction with Antigen Protein

Tao Liu[1], Wei Liu[1, 2], Zhe Song[1], Chunbo Jiao[1], Minghua Zhu[1] and Xiaogang Wang[1]

In the processing and presentation of antigenic peptides bound by the major histocompatibility complex (MHC) class I molecule, the ubiquitin-proteasome system of the eukaryotic cells plays an important role in proteolysis and degradation. The ubiquitinated protein substrate is delivered into the 26S proteasome to be digested and degraded. The proteasome degrading substrate is actually protein-protein interactions. Some researches of predicting proteasome cleave site rarely gave the information of the proteasome interacting with its substrate, and so the accuracy and reliability of these proteasome cleavage predictive methods still need to be improved. This paper used support vector machine method (SVM) to predict the proteasomal cleavage sites, and the predictive accuracy of the model was 82.8%. We showed analytically that the cleavage specificities of the cleavage sites " | " and its adjacent positions, and gave the information about the proteasome interacting with its substrate from our research results. It demonstrates that the proteasome cleaving to target protein is selective, but not random. *Cellular & Molecular Immunology*. 2009;6(2):135-142.

Key Words: proteasome, major histocompatibility complex, support vector machine

Introduction

In the processing and presentation of antigenic peptides bound by the major histocompatibility complex (MHC) class I molecule, the ubiquitin-proteasome system of the eukaryotic cells plays an important roles in proteolysis and degradation, and the process is that the antigen protein binds many ubiquitin molecules by covalent bond to form ubiquitins-protein, namely ubiquitination. Then the ubiquitinated protein substrate is delivered into the 26S proteasome to be digested and degraded (1). The 26S proteasome is an ATP-dependent proteolytic complex which is formed by the 20S catalytic particle (CP) and two 19S regulatory particles (RP). 19S RP is responsible for recognizing of ubiquitinated proteins, releasing the free ubiquitin, unfolding the protein substrate, and then translating the substrate into the 20S CP (2). 20S CP constructed by four staggered rings is the proteolytic core of 26S proteasome. Each of the two outer rings contains seven different α subunits from α_1 to α_7, and each of the two inner rings is composed of seven different β subunits from β_1 to β_7, and the four rings enclose the central chamber (3). The narrow opening surrounded by the two α subunits is the channel through which protein substrate appears to the proteolytic core, usually the opening is closed by N-terminal residues of α subunits to prevent proteins from entering the 20S CP and being degraded. The combination of 19S RP with 20S CP induces the change of the conformation of α subunits and the channel opening, and then the protein substrates would enter the central chamber of 20S CP. Only the protein substrates which entered the 20S CP could be degraded. The N-terminal amino acid Thr1 on the β subunit is the centre of the active site of the proteasome, and different β subunits have different activities which have the ability of cleavage to one kind of peptide bond of the substrates. After digesting and degrading the substrates, the 20S CP would release the peptide fragments, remove and release the ubiquitin molecules (4). And in recent years there have been some studies on the prediction of proteasomal cleavage site with experiment data of the cleavage products of proteasome, for example, PAProC uses *in vitro* degradation data of human as well as yeast proteasome as training set, and predicts proteasome cleavage sites using an one hidden layer neural network model trained by an evolutionary algorithm (5); MAPPP (6) is a software package which contains proteasome cleavage prediction part FragPredict and MHC binding prediction part. And FragPredict predicts proteasome cleavage sites based on a statistical analysis of cleavage motifs in experiment and a kinetic model of the 20S proteasome; NetChop is trained on *in vitro* degradation data

[1]College of Advanced Science and Technology, Dalian University of Technology, Dalian 116023, China;

[2]Correspondence to: Wei Liu, College of Advanced Science and Technology, Dalian University of Technology, Dalian 116023, China. Tel: +86-411-8470-7872, E-mail: jchjys@dlut.edu.cn

Received Dec 3, 2008. Accepted Mar 4, 2009.

of 20S proteasome and MHC class I molecular ligand data, and predicts proteasome cleavage sites by artificial neural network model (7). The proteasome degrading substrate is actually protein-protein interactions, and the studies above and other interrelated research work rarely analyze and discuss the information about the proteasome interacting with its substrate, therefore, the accuracy and reliability of these proteasome cleavage predictive methods still need to be improved (8, 9). This paper used support vector machine method (SVM) to study the proteasomal cleavage specificity by reducing the noise from the experiment data of proteasome cleavage and extracting the useful information, and we should try to give the information about the proteasome interacting with its substrate from our research results. This would help us to understand the MHC class I molecule processing and presentation antigen pathway deeply and be useful to design and develop of tumor vaccine.

Materials and Methods

Cleavage data set and non-cleavage data set

In the processing and presentation of antigenic peptides, the proteasome can accurately cleavage the C-terminal of antigenic peptides, but the N-terminal of antigenic peptides usually are further trimmed by other enzymes in cells (4). In the amino acid sequence of the source protein of the antigenic peptide, when the cleavage occurs on the peptide bond of amino acid on the position P1, the sequence positions of flanking amino acids of the position P1 are described as (PL…P2 P1 | P1' P2'…PL'), and the cleavage site is signed as the symbol " | ". The amino acid sequence window size (Window Size) of cleavage or non-cleavage sample is determinate in its source protein. This paper would mainly pay attention to the C-terminal of the antigenic peptide and its flanking regions for the cleavage specificity. We assign the C-terminal of the human leukocyte antigen (HLA) class I molecular ligands as the cleavage site, and the amino acid sequence of the cleavage samples (positive set) is collected from expanding the HLA class I molecular ligands in itself source protein. We assign the middle position of the HLA class I molecular ligands as the non-cleavage site, and the amino acid sequence of the non-cleavage samples (negative set) is also collected from expanding the HLA class I molecular ligands in itself source protein. Goldberg et al. indicated it would be erroneous to assign each peptide-bonds within T-cell epitope as a non-cleavage site because many T-cell epitopes are cleaved by the proteasome (10). We would give the assumption that all peptide-bond within the epitope were sub-cleavage sites, and the middle position of the epitope would have less the cleavage probability than the others.

Four thousand nine hundred and fifteen HLA class I molecular ligands are extracted from the AntiJen database (http://www.jenner.ac.uk/AntiJen). They are associated with 22 HLA-A, 21 HLA-B and 3 HLA-C molecules originated from 307 human proteins (11). By removing ligands which are redundant, which of the sequence receive number in Swissprot database are not available, which of the amino acid sequence length more than 12 amino acids (AAs) or less than 8 AAs and which are included in Saxova's test set, finally 1275 HLA class I-bound ligands are available (12). Then 1275 cleavage samples and 1275 non-cleavage samples are obtained. For the purpose of conserving the most experimental information, it was held in cleavage samples set if there is the cleavage sample as well as the non-cleavage sample, and it was discarded in non-cleavage samples set. Finally 1275 cleavage samples and 1185 non-cleavage samples are available. Samples were represented using sparse binary encoding (7).

Support vector machine

SVM is a new type of supervised machine-learning techniques. The SVM algorithm is based on the theory of the Structural Risk Minimization principle (13). Here we briefly review the basis of the algorithm of SVM in classification problems. Suppose we are given a set of labeled training samples $(\vec{x}_1, y_1), \cdots, (\vec{x}_n, y_n)$. The $\vec{x}_i$ corresponds to the encoded representation of amino acid sequence of the antigenic peptide to SVM, and the y_i represents the cleavage value (positive sample $y_i = 1$ or negative sample $y_i = -1$). Each training sample $\vec{x}_i = \left[x_{i1}, x_{i2}, \cdots, x_{im}\right]^T \in R^m$ ($i = 1, 2, \cdots, n$) belongs to either of two classes and is given a label $y_i \in \{-1, 1\}$ ($i = 1, 2, \cdots, n$). In amino acid sequence of a sample, each position possibly appears one of 20 different AAs, so correspondingly 20 variables are introduced to denote 20 AAs in some position of amino acid sequence. The *m* presents the number of corresponding variables of one sample Window Size; the *n* denotes the sample number of data set. The SVM maps the input vectors $\vec{x}_i$ into the high-dimensional feature space Z by a nonlinearity map function $\phi(\vec{x})$ and then find a separating hyperplane $\vec{\omega} \cdot \phi(\vec{x}) + b = 0$ that maximizes the margin between two classes in the feature space Z. The $\vec{\omega} = \left[\omega_1, \omega_2, \cdots, \omega_m\right]^T$ is the weight vector which represents the normal of the separating hyperplane. If the data set is linearly separable in the feature space Z, the problem constructing the optimal separating hyperplane can be described as minimize $\frac{1}{2}\|\vec{\omega}\|$ subjecting to constraints $y_i\left(\vec{\omega} \cdot \vec{x}_i + b\right) \geq 1$, $i = 1, 2, \cdots, n$; if not, it can be reformulated as $\min_{\omega, b, \xi} \frac{1}{2}\|\vec{\omega}\|^2 + C\sum_{i=1}^{n} \xi_i$ subjecting to constraints $y_i\left(\vec{\omega} \cdot \phi(\vec{x}_i) + b\right) \geq 1 - \xi_i$, $\xi_i \geq 0$, $i = 1, 2, \cdots, n$, where ξ_i is slack variable measuring the degree of misclassification for the *i*th sample and C is the penalty parameter controlling the trade-off between margin maximization and degree of misclassification. The dimension of the feature space Z may be very high or even infinite, and the form of the map function $\phi(\vec{x})$ usually is unknown, and the problem is difficult in direct solution. So the problem is translated to its

Table 1. The performance of the predicting model based on SVM with different kernel functions

Window Size (AAs)		Function type	Function parameter		Penalty parameter C		Predictive performance			
Range	Best Value		Range	Best Value	Range	Best Value	Sen (%)	Spe (%)	Acc (%)	MCC
4~28	20	Poly	2~6	2	10^{-2}~10^{6}	1000	84.0	82.1	83.1	0.661
4~28	24	Rbf	10^{-4}~10^{4}	2	10^{-2}~10^{6}	1000	83.2	82.1	82.6	0.653
4~28	20	Sigmoid	10^{-3}~10^{3}	0.02	10^{-2}~10^{6}	50000	83.3	81.9	82.6	0.652
4~28	20	Linear			10^{-2}~10^{6}	1	80.0	80.3	80.2	0.603

dual problem through Lagrangian formulation (13). The objective function of dual problem is

$$L(\vec{\alpha}) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j \phi(\vec{x}_i) \cdot \phi(\vec{x}_j) \qquad (1)$$

Where $\vec{\alpha} = [\alpha_1, \alpha_2, \cdots, \alpha_n]^T$, α_i is a nonnegative Lagrange multiplier. Note that there is only the form of dot product between vector $\phi(\vec{x}_i)$ and $\phi(\vec{x}_j)$ in equation (1). According to the functional theory, a dot product between vector $\phi(\vec{x}_i)$ and $\phi(\vec{x}_j)$ in feature space Z corresponds to a kernel function $K(\vec{x}_i, \vec{x}_j)$ which satisfying the Mercer condition in input space (13). There are some useful kernel functions:

Linear kernel function:

$$K(\vec{x}_i, \vec{x}_j) = \vec{x}_i \cdot \vec{x}_j \qquad (2)$$

Polynomial kernel function:

$$K(\vec{x}_i, \vec{x}_j) = \left[(\vec{x}_i \cdot \vec{x}_j) + 1 \right]^q \qquad (3)$$

Radial based kernel function:

$$K(\vec{x}_i, \vec{x}_j) = \exp\left(-\frac{\left| \vec{x}_i - \vec{x}_j \right|^2}{\sigma^2} \right) \qquad (4)$$

Two layers neural network kernel function:

$$K(\vec{x}_i, \vec{x}_j) = \tanh\left(v(\vec{x}_i \cdot \vec{x}_j) - 1 \right) \qquad (5)$$

So, the objective function (1) can be written as:

$$L(\vec{\alpha}) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j K(\vec{x}_i, \vec{x}_j) \qquad (6)$$

the $\vec{\alpha}$ is solved by using the sequential minimal optimization (SMO) algorithm (14). The kernel function of the linear SVM is described by formulation (2). Finally the classification function of predicting the cleavage site of proteasome is:

$$f(\vec{x}) = \mathrm{sgn}\left(\vec{\omega}^* \cdot \phi(\vec{x}) + b^* \right) = \mathrm{sgn}\left(\sum_{SV} \alpha_i^* y_i \phi(\vec{x}_i) \cdot \phi(\vec{x}) + b^* \right) \qquad (7)$$

Where the sgn ( ) is sign function; the $\vec{\omega}^*$ is weight vector of optimal plane in feature space Z; the α_i^* is the support vector coefficient; and the b^* is the threshold value of classification.

Evaluation of predictive performance of the model

The performance of the model is evaluated by using the parameters: sensitivity (Sen), specificity (Spe), accuracy (Acc) and Matthews coefficient of correlation (MCC) (12).

$$\mathrm{Sen} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \times 100\% \qquad \mathrm{Spe} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}} \times 100\%$$

$$\mathrm{Acc} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{FP} + \mathrm{TN} + \mathrm{FN}} \times 100\%$$

$$\mathrm{MCC} = \frac{\mathrm{TP} \times \mathrm{TN} - \mathrm{FN} \times \mathrm{FP}}{\sqrt{(\mathrm{TN} + \mathrm{FN})(\mathrm{FN} + \mathrm{TP})(\mathrm{TP} + \mathrm{FP})(\mathrm{FP} + \mathrm{TN})}} \qquad (8)$$

where TP, FP, TN, FN is the number of predicted true positives, the number of predicted false positives, the number of predicted true negatives and the number of predicted false negatives, respectively. The Sen value gives the percentage of cleavage samples that are predicted correctly, and the Spe value gives the percentage of non-cleavage samples that are predicted correctly. The Acc and MCC values are the estimative measurement of predicting cleavage or non-cleavage samples.

Results

The performance of predictive model of proteasomal cleavage sites

A 10-fold cross-validation technique was used to evaluate the performance of classifiers with different parameters. The parameters mentioned here include Windows Size of sample, function type, function parameter and penalty parameter C. The Range and the Best Value of these parameters are shown in Table 1. When the sensitivity and specificity of the model are approximately equal, the result given by the model is the most satisfying and reliability (15). We select the biggest MCC value if the sensitivity and the specificity are approximately equal. According to this rule the best parameters were determined finally: the Windows Size is 20 AAs, the kernel function is Polynomial kernel function

Table 2. The comparison of performances between our predictive model and others

Method	N	Sen	Spe	MCC
PAProC	217	45.6	30.0	-0.25
FragPredict	231	83.5	16.5	0.00
NetChop1.0	231	39.8	46.3	-0.14
NetChop2.0	231	73.6	42.4	0.16
SVM_dut	231	75.3	70.1	0.46

$\left[\left(\vec{x}_i \cdot \vec{x}_j\right)+1\right]^2$, the penalty parameter C is 1000, the performance of the model is the accuracy 83.1% and MCC 0.661, which is shown in Table 1.

Comparing this predictive model with other models

In order to compare our predictive model (SVM_dut) with other models (5-7), we use the sample test set which is provided by Ref 12. The result of comparison shows that predictive performance of our model is most satisfying, which is shown in Table 2. And the performances of PAProC (5), MAPPP (6), NetChop (7) were provided by Ref 12.

The weight coefficients of the linear SVM

The proteasome cleavage antigenic peptides not only relate to the cleavage site, but also to amino acids in adjacent regions of the cleavage sites " | " (16). Because each weight coefficient ω^* respectively corresponds to the variable of input sample $\vec{x}_i$, the value of the weight coefficient ω^* can be recognized as the impact of 20 kinds of amino acids in each position of amino acid sequence on cleavage sites " | ". When the Windows Size of the sample's amino acid sequence is 20 AAs, we use linear SVM method to solve the weight vector $\vec{\omega}^*$ in the formula (7), the values of the weight coefficients ω^* are shown in Figure 1. On the P10~P10' positions of sample's amino acid sequence, the height of each amino acid symbol on each position shows the value of the weight coefficient ω^* which is the contribution of the corresponding amino acid to the cleavage sites " | ", e.g., the ω^* value of leucine is larger than that of lysine in P1 position. And amino acid symbol and upside-down one show the positive and negative ω^* value respectively. The height of the all amino acid symbol stacked on each position shows the absolute value of the sum of weight coefficients of all amino acid contributing to the cleavage sites " | ". Twenty amino acids are divided into four classes according to their physicochemical characteristics, i.e., acidic, red; basic, blue; neutral and hydrophobic, black; neutral and polar, green.

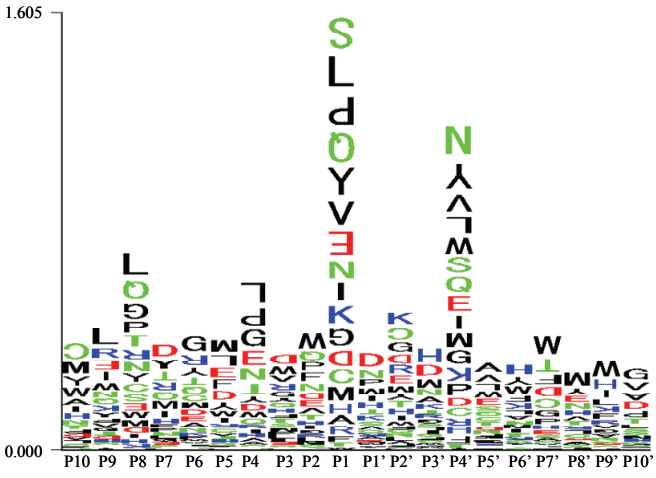

Figure 1. The weight coefficients of amino acids on each position when Windows Size is 20 AAs. Amino acids are color coded according to their physicochemical characteristics. Neutral and polar, green; basic, blue; acidic, red; neutral and hydrophobic, black. Amino acid symbol and upside-down one show the positive and negative weight coefficient respectively. P10 to P10' are presenting in the X-axis. The height of each amino acid symbol on each position is determined by the weight coefficient of its amino acids on the position, and the height of the all amino acid symbol stacked on each position is determined by the sum of the absolute value of weight coefficients of all amino acids on the position.

Table 3. The distances between active sites of mammalian 20S proteasome

PDB ID	β ring	(Thr1) β1	(Thr1) β2	(Thr1) β5
1IRU.pdb	(Thr1) β1	-	2.8 nm	6.2 nm
	(Thr1) β2	2.8 nm	-	6.3 nm
	(Thr1) β5	6.2 nm	6.3 nm	-

PDB ID	β ring / β* ring	(Thr1) β1	(Thr1) β2	(Thr1) β5
1IRU.pdb	(Thr1) β*1	2.9 nm	5.0 nm	5.8 nm
	(Thr1) β*2	5.0 nm	6.6 nm	4.1 nm
	(Thr1) β*5	5.8 nm	4.1 nm	4.9 nm

Discussion

The β_1, β_2 and β_5 subunits play a crucial role in cleavage catalysis within 20S CP. During the 20S assembly, after deletion of the N-terminal prosequence of the β subunit, the Thr1 residue which is the center of the active site of the proteasome is exposed. The Thr1 residue resides in the inner surface, to endow the β subunit with the catalysis character like serine-proteasome (16). The mammalian 20S proteasome has two β rings with three active sites respectively which not only cleave the peptide-bonds of the protein substrates, but also cleave them by cooperation of two sites (17). A protein substrate entering the proteasome is fully unfolded (3). On the opinion of "molecular ruler" proposed by the Ref 17, the

Table 4. The amino acids nonbonding contact with active sites of mammalian 20S proteasome

Amino acids nonbonding contact with Thr1 of β1		Amino acids nonbonding contact with Thr1 of β2		Amino acids nonbonding contact with Thr1 of β5	
	Thr2		Thr2		Thr2
	Ile3		Ile3		Thr3
	Asp17		Asp17		Asp17
	Arg19		Lys33		Arg19
	Lys33		Ala46		Lys33
	Arg45		Met127		Met45
Subunit β1	Ser46	Subunit β2	Gly128	Subunit β5	Gly129
	Gly129		Ser129		Ser130
	Ser130		Gly130		Gly131
	Gly131		Asp166		Asp167
	Asp167		Gly168		Tyr169
	Ser169		Ser169		Ser170
	Ser170				

Electropositive and basic amino acids are colored blue, electronegative and acidic amino acids are colored red, and neutral amino acids are colored black.

distance between the active sites of proteasome is related to the length of the cleavage product. In this paper, the distances between the active sites of the mammalian 20S proteasome are computed, which are shown in Table 3. In fact, the distances are got from the distances between the Thr1 Oγ of the β subunits under the three dimension space structure (PDB code: 1IRU). According to the relationship between the length of unfolded peptide chain and the number of amino acids in the peptide, that is found by Coux et al., the length of 2.8 nm of the cleavage product equals to the 7-8 AAs and the length of 6.6 nm equals to 16-19 AAs (3). Kisselev et al. discovers that the amino acid sequence of cleavage products of mammalian proteasome range in length from 3 to 22 AAs, and their abundance decreases with increasing length according to a log-normal distribution (18). About 70 percent of cleavage products are too short (< 7 AAs) to available in antigen presentation. Only less than 15 percent of the products' length is from 8 to 9 AAs. Therefore, there is not one-to-one correspondence relationship between the distance of proteasome active sites and the length of amino acid sequence of cleavage products. Dick's experiment demonstrates that the cleavage products which are generated by two kind β subunits of proteasome together will be cleavage again through the release-and-recapture pathway, as the result, the amino acid sequence of most of the products generated by proteasome is less than 10 AAs (19). It is known from Table 1 that the accuracy is 83.1% when the Windows Size is 20 AAs. The accuracy indicates the Windows Size 20 AAs of cleavage samples and non-cleavage samples could include most of information of cleavage products which are generated by 20S proteasome. It also demonstrates that the predictive model in this paper is reasonable when Windows Size is 20 AAs. Thus the performance of our model is superior to that of other model (5-7), which is shown in Table 2.

It is known from Figure 1 that the height of each amino acid symbol on each position from P10 to P10' shows the value of the weight coefficient ω^* which is the contribution of the corresponding amino acid in the samples to the cleavage sites " | ", and the height of the all amino acid symbols stacked on each position from P10 to P10' shows the absolute value of the sum of the weight coefficients of all amino acids on the position which is the contributing to the cleavage sites " | ". The amino acids on adjacent position of the cleavage site " | ", i.e., P1, P4', P4 and P8 positions, have distinct cleavage specificities, which shows that the process of proteasome cleaving to target protein is selective, but not random.

In order to explain the amino acid specificities of cleavage site " | " and its adjacent positions which are shown in Figure 1, the crystal structure of mammalian 20S proteasome (PDB code: 1IRU) are used to seek amino acids which could nonbonding contact with amino acid Thr1 of active sites. Two residues are defined to be nonbonding contact if the distance of two atoms of these residues were smaller than 0.4 nm (20). And the amino acids nonbonding contacting with the Thr1 of active sites are presented in Table 4. It can be seen from Table 4 that the amino acids surrounding the active site of β_1 subunit are three electropositive and basic amino acids (Arg19, Lys33 and Arg45) and two electronegative and acidic amino acids (Asp17 and Asp167); and the amino acids surrounding the active site of β_2 subunit are only one electropositive and basic amino acid (Lys33) and two electronegative and acidic amino acids (Asp17 and Asp166); and the amino acids surrounding the active site of β_5 subunit are two electropositive and basic amino acids (Arg19 and Lys33) and two electronegative and acidic amino acids (Asp17 and Asp167). Therefore, it could be considered simply that 1) amino acids surrounding the active site of β_1 subunit are electropositive as a whole, and the active site cleave the peptide-bond of electronegative and acidic amino acids easily; 2) amino acids surrounding the active site of β_2 subunit are electronegative as a whole, and the active site cleave the peptide-bond of electropositive and

basic amino acids easily; 3) amino acids surrounding the active site of β_5 subunit are neutral. The previous report was confirmed that β_1 subunit could cleave substrate after acidic residues; β_2 subunit could cleave substrate after basic residues; β_5 subunit could cleave substrate after large hydrophobic residues (3). Thus it could be seen that the reason that β_1, β_2 and β_5 subunit could cleave peptide-bond of different amino acid residues relates to the physicochemical characteristics of amino acids surrounding the active sites Thr1 of β subunit. The cleavage specificities of the position P1 in the Figure 1 reflect the preference of each active site of the β subunit to the amino acids of the cleavage position " | ".

It can be seen from the P1 position of sample's amino acid sequence in Figure 1 that hydrophobic amino acids Leu, Tyr, Val and Ile have the largest positive weight coefficients, and they make positive contribution to proteasome cleaving substrate. Hydrophilic amino acids Asn, Ser, Gln, Thr, acidic ones Asp, Glu and basic one His have negative weight coefficients, and they make negative contribution to cleavage. It can be seen from the crystal structure of mammalian 20S proteasome (PDB code: 1IRU) that active site Thr1 Oγ atom is in the center of the inner cavity wall of β subunit. Hydrophobic residues would fall into the inner cavity wall easily, and this could make Thr1 Oγ atom of β subunit direct interact with peptide-bond of substrate backbone and cleave the target protein with its surrounding residues together. After peptide-bond being cleaved, hydrophobic amino acids Leu, Tyr and Val in P1 form the C terminal of peptides. It consists with the fact that Leu and Val are the ligand primary motifs anchored in Pocket F of HLA-A2 molecule (21). It can be seen from Figure 1 that hydrophobic amino acids Gly and Pro have negative weight coefficients. Because their hydrophobicity is weakly, they would hardly fall into the inner cavity wall where the active site Thr1 Oγ atom located. Nussbaum et al. have confirmed that when study the cleavage performance by the wild type yeast 20S proteasome, Leu as a P1 residue has the highest signification and Gly is disfavored (22). Therefore, our result is consistent with the conclusions of Nussbaum et al. Thus, in the P1 position, hydrophobic amino acids are recognized by β subunit active sites easily and benefit proteasomal cleavage. In contrary, hydrophilic amino acids are recognized by β subunit active sites hardly and are deleterious for proteasomal cleavage.

It can be seen from the P4' position of sample's amino acid sequence in Figure 1 that hydrophilic amino acids Asn, Ser, Gln, Cys, acidic ones Asp, Glu, basic one His and weak hydrophobic ones Gly, Pro have positive weight coefficients, and they make positive contribution to 20S proteasome cleaving substrate. Hydrophobic amino acids Leu, Val, Tyr and Ile have negative weight coefficients, and they make negative contribution to 20S proteasome cleaving substrate. It is interesting that the specificities of P1 and P4' position related to the cleavage site " | " are very reverse. The amino acid Thr1 of active site of 20S proteasome β subunit can act on the peptide-bond of amino acid of P1 position on the sample's sequence, but how to explain the amino acid specificities of P4' position? Wenzel et al. have a hypothesis that a stretch of the unfolded polypeptide chain of the substrate protein binds to an extended groove which is located at the interface of α and β subunits (17). The crystal structure of mammalian 20S proteasome (PDB code: 1IRU) is used to calculate minimum distances between Thr1 Oγ atoms of β_1, β_2, β_5 subunits and atoms of α ring respectively, and the average value of minimum distances is 2.03 nm (minimum distances are 2.17 nm, 1.86 nm, and 2.07 nm respectively). And the average minimum distance that is almost equal to the length of 4-5 AAs corresponds to the length of the unfolded polypeptide between the cleavage site " | " and P4' position. So that the amino acid specificities of P4' position reflect the character of the amino acid residues on P4' position interacting with the amino acid residues on the interface of α and β rings in 20S proteasome.

It can be seen from the P4 position of sample's amino acid sequence in Figure 1 that hydrophilic amino acids Asn, Gln, weak hydrophobic one Gly, acidic ones Glu, Asp and basic ones His, Trp have positive weight coefficients, and they make positive contribution to 20S proteasome cleaving substrate. Hydrophobic amino acids Leu, Pro, Tyr and Trp have negative weight coefficients, and they make negative contribution to 20S proteasome cleaving substrate. Nussbaum et al. have confirmed that when the wild type yeast 20S proteasome cleaving the substrate, Pro has the highest signification in the P4 position (22). But this paper has the opposite result, and this is probably caused by the difference of structures of different species 20S proteasome. Unno et al. has confirmed that the α subunits of human proteasome have some obvious structure differences with the α subunits of yeast proteasome, and the β subunits of human proteasome also have some structure differences with the β subunits of yeast proteasome (23).

The cleavage and non-cleavage samples used in this paper are extended from peptide binding to HLA class I molecule, and then the amino acid specificities of P4 position of sample's amino acid sequence may contain the binding specificity of the amino acid anchored in Pocket C of HLA class I molecule, but we could not find the peptide binding specificity in previous reports (24-26). So, how to explain the amino acid specificities of P4 position? The following information is demonstrated in the crystal structure of mammalian 20S proteasome (PDB code: 1IRU). 1) The distances between Thr1 Oγ atom of β_1 subunit located on one β ring and atoms located of β^*_1, β^*_7 subunits located on the other β ring are 1.62 nm and 1.08 nm, respectively; 2) the distances between Thr1 Oγ atom of β_2 subunit located on one β ring and atoms located of β^*_7, β^*_6 subunits located on the other β ring are 1.49 nm and 1.00 nm, respectively; 3) the distances between Thr1 Oγ atom of β_5 subunit located on one β ring and atoms located of β^*_4, β^*_3 subunits located on the other β ring are 1.56 nm and 1.09 nm respectively. Thus, the distance between the active sites of β ring and the other β ring atoms located on the interface of two β rings are 1.00~1.62 nm, which corresponding to the length of 3-4 AAs

of unfolded peptide, and is the length between cleavage site " | " and the P4 position of sample's amino acid sequence in the unfolded protein. So the amino acid specificities of P4 position in Figure 1 reflect the character of proteasome cleavage and information of interacting between P4 position residues of the substrate and residues on the interface of two β rings of the proteasome.

It can be seen from the P8 position of sample's amino acid sequence in Figure 1 that hydrophobic amino acids Leu, Pro, Tyr and Trp have the largest positive weight coefficients, and they make positive contribution to 20S proteasome cleaving substrate. Hydrophilic amino acids Gln, Thr and Asn, weak hydrophobic one Gly, acidic ones Glu, Asp and basic ones Arg, His and Lys have negative weight coefficients, and they make negative contribution to 20S proteasome cleaving substrate. How to explain the amino acid specificities of P8 position of sample's amino acid sequence? It is shown in Table 3, the distance between different β_1 subunits in different β rings is 2.9 nm, and it is about 7-8 AAs in the unfolded protein. The distance between P8 and P1 position is also the length of 8 AAs in the sample's sequence, and the amino acid specificities of P8 and P1 positions are a little similar. This might be the cleavage specificity come from the interaction between two β_1 subunit active sites in different β rings of 20S proteasome on the substrate. Amino acid Leu is a primary motif anchored in Pocket B of HLA-A2 molecule (24), and the cleavage specificity of P8 position of sample's amino acid sequence might determine the binding specificity of the ligand motif anchored in Pocket B of HLA molecule.

Additionally, according to the study of Nussbaum et al., after inhibiting the activities of the subunits β_1, β_2, the cleavage specificity of the proteasome only with β_5 is given, the ranking of the contribution of hydrophobic amino acids in P1 position to cleavage is same as which in Figure 1 (22). It indicates that the peptide bond of hydrophobic amino acids in P1 position is mainly cleaved by the subunit β_5. The ranking of contribution of basic amino acids in P1 position to cleavage in Figure 1 is same with the selection preference of subunit β_2 to P1 position (22), which indicates that the peptide bond of basic amino acids in P1 position is mainly cleaved by the subunit β_2. Furthermore, it can be seen from Figure 1 that the weight value of hydrophobic amino acids is larger than which of other amino acids. It indicates that the substrate may be cleaved by subunit β_5 (chymotrypsin-like, ChT-L) firstly, and then cleaved by subunit β_2 (trypsin-like, T-L) and subunit β_1 (peptidylglutamyl-peptide-hydrolase, PGPH; caspase-like).

In a word, the predictive model of proteasomal cleavage sites is rational and feasible when Windows Size of sample data is 20 AAs. The amino acids in the adjacent positions of cleavage sites " | " show cleavage specificities obviously, i.e., hydrophobic amino acids on the P1 are easily recognized by active sites of β subunit, and this would benefit for active sites cleaving substrate; cleavage specificities of amino acids in P4' and P4 demonstrate the information of proteasome interacting with substrate.

References

1. Adams J. The proteasome: structure, function, and role in the cell. Cancer Treat Rev. 2003;29:3-9.
2. Smith D, Benaroudj N, Goldberg AL. Proteasomes and their associated ATPases: A destructive combination. J Structural Biol. 2006;156:72-83.
3. Coux O, Tanaka K, Goldberg AL. Structure and functions of the 20S and 26S proteasomes. Annu Rev Biochem. 1996;65:801-847.
4. Heinemeyer W, Ramos PC, Dohmen RJ. The ultimate nanoscale mincer: assembly, structure and active sites of the 20S proteasome core. Cell Mol Life Sci. 2004;61:1562-1578.
5. Nussbaum AK, Kuttler C, Hadeler KP, Rammensee HG, Schild H. PAProC: a prediction algorithm for proteasomal cleavages available on the WWW. Immunogenetics. 2001;53:87-94.
6. Holzhütter HG, Kloetzel PM. A kinetic model of vertebrate 20S proteasome accounting for the generation of major proteolytic fragments from oligomeric peptide substrates. Biophysical J. 2000;79:1196-1205.
7. Kesmir C, Nussbaum AK, Schild H, Detours V, Brunak S. Prediction of proteasome cleavage motifs by neural networks. Protein Eng. 2002;15:287-296.
8. Bhasin M, Raghava GPS. Pcleavage: an SVM based method for prediction of constitutive proteasome and immunoproteasome cleavage sites in antigenic sequences. Nucleic Acids Res. 2005; 33:202-207.
9. Ginodi I, Vider-Shalit T, Tsaban L, Louzoun Y. Precise score for the prediction of peptides cleaved by the proteasome. Bioinformatics. 2008;24:477-483.
10. Goldberg AL, Cascio P, Saric T, Rock KL. The importance of the proteasome and subsequent proteolytic steps in the generation of antigenic peptides. Mol Immunol. 2002;39:147-164.
11. Blythe MJ, Doytchinova IA, Flower DR. JenPep: a database of quantitative functional peptide data for immunology. Bioinformatics. 2002;18:434-439.
12. Saxova P, Buus S, Brunak S, Kesmir C. Predicting proteasomal cleavage sites: a comparison of available methods. Int Immunol. 2003;15:781-787.
13. Burges CJC. A tutorial on support vector machines for pattern recognition. Data Mining and Knowledge Discovery. 1998;2: 121-167.
14. Keerthi SS, Shevade SK, Bhattacharyya C, Murthy KRK. Improvements to platt's SMO algorithm for SVM classifier design. Neural Comput. 2001;13:637-649.
15. Bhasin M, Raghava GP. Prediction of CTL epitopes using QM, SVM and ANN techniques. Vaccine. 2004;22:3195-3204.
16. Altuvia Y, Margalit H. Sequence signals for generation of antigenic peptides by the proteasome: implications for proteasomal cleavage mechanism. J Mol Biol. 2000;295:879-890.
17. Wenzel T, Eckerskorn C, Lottspeich F, Baumeister W. Existence of a molecular ruler in proteasomes suggested by analysis of degradation products. FEBS Lett. 1994;349:205-209.
18. Kisselev AF, Akopian TN, Woo KM, Goldberg AL. The sizes of peptides generated from protein by mammalian 26 and 20 S proteasomes. Implications for understanding the degradative mechanism and antigen presentation. J Biol Chem. 1999;274: 3363-3371.
19. Dick LR, Moomaw CR, DeMartino GN, Slaughter CA. Degradation of oxidized insulin B chain by the multiproteinase complex macropain (proteasome). Biochemistry. 1991;30:2725-2734.

20. Altuvia Y, Margalit H. A structure-based approach for prediction of MHC-binding peptides. Methods. 2004;34:454-459.
21. Falk K, Rotzschke O, Stevanovic S, Jung G, Rammensee HG. Allele-specific motifs revealed by sequencing of self-peptides eluted from MHC molecules. Nature. 1991;351:290-296.
22. Nussbaum AK, Dick TP, Keilholz W, et al. Cleavage motifs of the yeast 20S proteasome β subunits deduced from digests of enolase 1. Proc Natl Acad Sci U S A. 1998;95:12504-12509.
23. Unno M, Mizushima T, Morimoto Y, et al. The structure of the mammalian 20S proteasome at 2.75 A resolution. Structure. 2002;10:609-618.
24. Song Z, Liu T, Liu W, Zhu MH, Wang XG. The QSAR model study of interaction between peptides and MHC molecules. Acta Phys Chim Sin. 2007;23:198-205.
25. Liu T, Song Z, Liu W, Wang XY, Qiu XM. Prediction of CTL epitopes based on modified artificial neural network. J Dalian Univ Technol. 2007;47:473-478.
26. Song Z, Liu T, Wang XY, Liu W. Application of partial least square method in studying of the quantitative structure-activity relationship of T cells epitopes. Mian Yi Xue Za Zhi. 2007;23: 166-171.