Science in China Series C: Life Sciences

Structural parameterization and functional prediction of antigenic polypeptome sequences with biological activity through quantitative sequence-activity models (QSAM) by molecular electronegativity edge-distance vector (VMED)

LI ZhiLiang[1,2†], WU ShiRong[1,2], CHEN ZeCong[1,2], YE Nancy[1,2], YANG ShengXi[1,2], LIAO ChunYang[1,2], ZHANG MengJun[1,2,3†], YANG Li[1,2], MEI Hu[1,2,4], YANG Yan[1,2], ZHAO Na[1,2], ZHOU Yuan[1,2], ZHOU Ping[1,2], XIONG Qing[1,2], XU Hong[1,2], LIU ShuShen[1,2], LING ZiHua[1,2], CHEN Gang[1,2,4†] & LI GenRong[1,2]

[1] College of Chemistry and Chemical Engineering/Key Laboratory for Chemobiomedical Science and Engineering under Chongqing Municipality, College of Life Science and Biological Engineering/Key Laboratory for Biomechanics and Tissue Engineering under Ministry of Education, Chongqing University, Chongqing 400044, China;
[2] State Key Laboratory for Chemobiosensors and Chemobiometrics under MOST at Hunan University, Changsha 410012, China;
[3] Department of Medical Analysis/PLA Center of Bioinformatics Immunology, Surgeon Third University, Chongqing 400031, China;
[4] Technology Centre for Life Sciences, Singapore Polytechnic, 500 Dover Road, Singapore 139651, Singapore

Only from the primary structures of peptides, a new set of descriptors called the molecular electronegativity edge-distance vector (VMED) was proposed and applied to describing and characterizing the molecular structures of oligopeptides and polypeptides, based on the electronegativity of each atom or electronic charge index (ECI) of atomic clusters and the bonding distance between atom-pairs. Here, the molecular structures of antigenic polypeptides were well expressed in order to propose the automated technique for the computerized identification of helper T lymphocyte (Th) epitopes. Furthermore, a modified MED vector was proposed from the primary structures of polypeptides, based on the ECI and the relative bonding distance of the fundamental skeleton groups. The side-chains of each amino acid were here treated as a pseudo-atom. The developed VMED was easy to calculate and able to work. Some quantitative model was established for 28 immunogenic or antigenic polypeptides (AGPP) with 14 (1—14) A^d and 14 other restricted activities assigned as "1"(+) and "0"(–), respectively. The latter comprised 6 A^b(15-20), 3 A^k(21-23), 2 E^k(24-26), 2 H-2^k(27 and 28) restricted sequences. Good results were obtained with 90% correct classification (only 2 wrong ones for 20 training samples) and 100% correct prediction (none wrong for 8 testing samples); while contrastively 100% correct classification (none wrong for 20 training samples) and 88% correct classification (1 wrong for 8 testing samples). Both stochastic samplings and cross validations were performed to demonstrate good performance. The described method may also be suitable for estimation and prediction of classes I and II for major histocompatibility antigen (MHC) epitope of human. It will be useful in immune identification and recognition of proteins and genes and in the design and devel-

Received Sepptember 30, 2006; accepted June 14, 2007
doi: 10.1007/s11427-007-0080-7

†Corresponding author (email: zlli-cqu@163.com; zlli2662@163.com; gchen@sp.edu.sg)

Supported by National High-Tech R&D Programme of China (863) (Grant No. 2006AA02Z312), National 111 Programme Introducing Talents of Discipline to Universities (Grant No. 0507111106), National Chunhui Project (Grant No. 990404+00307), State New Drug Project (Grant No. 1996ND1035A01), Fok Ying Tung Educational Foundation (Grant No. 980706), State Key Laboratory of Chemo/Biosensing and Chemometrics Foundation (KCBCF0501201), Chongqing University Innovation Fund (CUIF030506), Chongqing Municipality Applied Science Fund (Grant No. CASF01-3-6), and Momentous Juche Innovation Fund for Tackle Key Problem Items (MJIF 03-5-6+04-10-10)

opment of subunit vaccines. Several quantitative structure activity relationship (QSAR) models were developed for various oligopeptides and polypeptides including 58 dipeptides and 31 pentapeptides with angiotensin converting enzyme (ACE) inhibition by multiple linear regression (MLR) method. In order to explain the ability to characterize molecular structure of polypeptides, a molecular modeling investigation on QSAR was performed for functional prediction of polypeptide sequences with antigenic activity and heptapeptide sequences with tachykinin activity through quantitative sequence-activity models (QSAMs) by the molecular electronegativity edge-distance vector (VMED). The results showed that VMED exhibited both excellent structural selectivity and good activity prediction. Moreover, the results showed that VMED behaved quite well for both QSAR and QSAM of poly- and oligopeptides, which exhibited both good estimation ability and prediction power, equal to or better than those reported in the previous references. Finally, a preliminary conclusion was drwan: both classical and modified MED vectors were very useful structural descriptors. Some suggestions were proposed for further studies on QSAR/QSAM of proteins in various fields.

molecular electronegativity distance-edge vector (VMED), antigenic polypeptide (AGPP) sequences, bioactive oligopeptide (BAOP) chains, quantitative sequence-activity models (QSAM), theoretically computational descriptors (TCD)

In modern sciences and current technologies, of course including life sciences and biological technologies, there exists a tendency from qualitative description to quantitative regularity. Quantitative structure activity relationship (QSAR) is always a very active field of scientific researches, especially in recent study for biological macromolecules such as polypeptides, proteins, genes or nuclei acids. Great progress has been achieved with performing and finishing human genome project (HGP)[1−3]. How to establish quantitative structure activity relationships (QSARs) between biological sequences and functional activities, i.e. quantitative sequence-activity models (QSAM) has drawn much attention in biological, medical, pharmaceutical and related fields[4−20], and great achievements have been obtained in three-dimensional (3D) structural prediction. In 1961, Anfinsen et al.[21] thought that the primary structure completely determined their higher or 3-D structures, which became one of the most important topics. In 1993, Martin et al.[22] further demonstrated that the higher structural information was still entirely contained in the primary structures of proteins. There exist many successful stories[6−24] although there are still extreme difficulties. Recent comprehensive approaches[22,33] have been proposed and preliminary achievements have been obtained; among them, many description variables are based on the amino acid side-chains. As compositive segments of different proteins, various peptides are very important in all living systems. They act as hormones, enzyme inhibitors, antibodies, olfaction and taste receptors, antimicrobial compounds or agents, and other biological functions. Hence, they have attracted considerable pharmacological interest in recent years[1−35]. With development of peptide library, thousands of different peptides have been designed, synthesized and then subjected to a range of screening procedures and biological assays. To effectively use the peptide library, biological data can be analyzed with multivariate quantitative structure-activity relationships. For properties of peptides a precise amino acid sequence is required for a particular function or biological activity. A QSAR model will then indicate how the change in peptide sequence is correlated with the variation in biological activity and how to modify the sequence to achieve the improved activity. The basic assumption in QSAR is that the biological activity within a set is related to the structural variation of the compounds, i.e., biological activity can be modeled as a function of molecular structure. In this context, quantitative amino acid descriptors have shown to be valuable. Since the pioneering work of Sneath[23], who derived amino acid descriptors from semiqualitative physicochemical data for the 20 coded amino acids and used them in a quantitative sequence-activity model analysis of oxytocin-vasopressine analogues, many amino acid descriptors have been proposed for the 20 coded amino acids[24−35]. A notable development in QSAR is the use of amino acid "*z* scores" obtained by principal components analysis (PCA) based on 29 physicochemical variables of 20 coded amino acids[30−32]. Three resulting principal components (PCs), the so-

called principal properties, are linear combinations of the original parameters and primarily represent hydrophobicity, side-chain bulk, and electronic effect of amino acids. The *z* scores have proven to be useful for modeling some biological activities of small peptides as a function of the *z* scores. By using only 12 physicochemical variables, Hellberg et al.[31] took a first step toward expanding these scales to encompass 35 non coded amino acids. More recently, the same approach was expanded to more parameters for a larger set of amino acids (20 coded + 67 noncoded). Application of PCA resulted in a set of 5 orthogonal variables termed *zz* scores, among which the first three corresponded to the original *z* scores. The *zz* scores were applied with good results obtained[30,32] to two peptide data sets, both elastase substrates and neurotensin analogues. However, all the amino acid descriptors mentioned above are derived by PCA from data matrix comprised of hydrophobic, steric, and electronic properties of amino acids. Thus each principal component is still a linear combination of different properties limited to definite physiochemical meanings. In 1985, Kidera et al.[30] collected 188 properties of the 20 natural amino acids and applied factor analysis on these to obtain 10 orthogonal factors that are most important for determining the three-dimensional structure of protein. In 1987, Hellberg et al.[31] developed principal properties PP, or *z*-scores, for 20 natural and more than 110 unnatural amino acids. The *z*-scores were extracted through PCA from a collected experimental data on various peptides, such as HPLC retention times, pK_a's, NMR-derived properties, and other measurable variables related to hydrophobicity, size, and electronic features. By using *z*-scores and multivariate statistics, some good regression models were generated for peptide QSARs on oxytocin, bradykinin and substance P receptors or on sweetener peptides by PLS[24−33]. In 1995, Collantes et al.[34] established good 3D-QSAR models by using three-dimensional descriptors, both Isotropic Surface Area (ISA) and Electronic Charge Index (ECI). In 1999, Zaliani et al.[35] performed QSAR studies on dipeptides with good results from the extracted and condensed steric and electrostatic 3D- properties of the natural amino acids based on 36 statistics indexes. Particularly, Raychaudhury et al.[36] constructed descriptor to perform QSAR studys from the primary structure of polypeptides, and created a well- performed QSAR model. But most of hese successful reports were involved with complex calculations in structural characterization of the peptides. In our laboratories[37−40], based on both the electronegativity of each atom and the distance between these atoms, a new set of descriptors, called the molecular electronegativity distance/edge vector (VMED[39−48]/VMEE[49,50]) to describe the molecular structure of peptides, was proposed only from the primary structure of peptides. Several good quantitative structure activity relationship models were proposed on biological activity of 58 angiotensin converting enzyme (ACE) inhibitors, of 48 bitter tasting dipeptides (BTD), of 31 bradykinin-potentiating pentapeptides (BPP), of 24 tachykinin heptapeptide sequences(TAH) with *rabbit pulmonary artery* (RPA) activity and of 152 antigenic nonapeptides (AGN) with binding affinities related to HLA-A*0201 restrictive CTL epitopes[52−58], and here polypeptides were equal to and/or larger than decapeptides by multiple linear regression (MLR).

In order to explain the ability to characterize the molecular structure of polypeptides, QSAR modeling was performed for functional prediction of polypeptide sequences with antigenic activity through quantitative sequence-activity models by the molecular electronegativity distance-edge vector. The obtained results showed that VMED exhibited both eximious structural selectivity and excellent activity prediction. Besides, molecular structure of antigenic polypeptides required to be well expressed in order to propose the automated technique for the computerized identification and/or recognition of helper T lymphocyte (HTL, Th) epitopes. Furthermore, based on both electronic charge index and relative bonding distance of the fundamental skeleton groups as a pseudo-atom, here the side-chain of each amino acids, a modified MED vector was proposed from the primary structure of polypeptides. The developed VMED would be very useful in structural characterization and activity prediction of biological molecules including HTL polypeptide sequences with major histocompatibility antigen (MHC) activity because it was easy to calculate and able to work. Some quantitative model was established for 28 HLA-A*0201 restrictive CTL epitopes or immunogenic or antigenic polypeptides (AGPP) with 14 (1−14) A^d and 14 other restricted activities assigned as "1"(+) and "0"(–), respectively, latter covering 6 A^b(15−20), 3 A^k(21−23), 2 E^k(24−26), 2 H−2^k(27 and 28) restricted sequences. Stochastic sampling and cross vali-

dations were performed to demonstrate good performance. The proposed method may suit for estimation and prediction of both classes I and II for major histocompatibility antigen epitopes. It would be useful in immune recognition of proteins and helpful to design and development of subunit vaccines. Besides, the obtained good results showed that VMED behaved quite well for QSAR and QSAM of poly/oligopeptides, which exhibited good estimation ability and fine prediction capability, equal to or better than those reported in the previous references[11−20].

1 Principle and methodology: structural parameterization and molecular modeling

It is a general rule in chemistry and physics that the molecular structure determines its property and the molecular property reflects its structure. For the biological molecules in chemobiology, chemogenetics, chemoimmunology and chemopharmacy, their bioactivities are also determined by their molecular structures. The suitable characterization of biological molecular structure is one of the most important fundamental elements in quantitative structure-bioactivity relationships (QSBR). Bioactivities and properties of compounds depend on the types of both composing atoms and bonding conjunctions and furthermore, reflect the results of all atoms' micro-interactions, mainly electronic interaction. Compared to point charge in physics, the interactions between involved atoms can be defined in the following equation:

$$E = q_i q_j / 4\pi\varepsilon d_{ij}^2 \propto (q_i q_j / d_{ij}^2),$$
$$E \propto (q_i * q_j)/d_{ij} \text{ or } E \propto (q_i * q_j), \quad (1)$$

where q_i and q_j refer to relative Pauling electronegativity (X_P) of the ith and jth atoms versus carbon atom, and d_{ij} refers to the bond-conjunction distance, adding up the number of bond, between the ith and jth atoms, and ε stands for the dielectric constant. Pauling's electronegativity is one of the most important and useful concepts in chemistry and the related fields. It was not defined nor given here due to its wide application and frequent appearance in many textbooks and monographs (see some references, such as (a) Pauling, L, *J Am Chem Soc.* 1932, 54: 3570-3582; (b) Pauling, L. *The Nature of the Chemical Bond*, 3rd ed. Ithaca, New York: Cornell University, 1960; and references therein). Its scales were only provided for the related atoms: 2.55 for C, 3.04 for N and 3.44 for O, respectively, which are all non-hydrogen atoms in peptides. When atoms i and j are the carbon atoms, both q_i and q_j refer to unit: $q_i = q_j = X_{P(C)}/X_{P(C)} = 2.55/2.55 = 1$; when either atom i or j is not carbon atom, then either q_i or q_j is not unit: for the nitrogen atom (N), $q_i = q_j = X_{P(N)}/X_{P(C)} = 3.04/2.55 = 1.192$ and for the oxygen atom (O), $q_i = q_j = X_{P(O)}/X_{P(C)} = 3.44/2.55 = 1.349$. Add up all interactions between atoms within the equal bond-conjunction distance. That is, adding up the interactions between atoms whose bond-conjunction distance is one, adding up the interactions between atoms whose bond-conjunction distance is two, and so on. Thus a new set of parameters, VMEE (***v*** in short) have been proposed to characterize the molecular structure of peptides.

$$v_k = \sum_{i,j}^{n} (q_i q_j / d_{ij}^2)$$
$$(i = 1 \sim n, j = 1 \sim n, j \geqslant i; k = 1 \sim m, m = n(n+1)/2), \quad (2)$$

where v_k refers to the kth descriptor belonging to the ***v*** vector for $d_{ij} = k$. Generally, the farther the distance between atoms is, the weaker the interaction between them will be. So it seems enough to select ten elements with the farthest being ten bond-conjunction distance to characterize the molecular structure.

As for oligo-/poly-peptide structures, such as another dipeptide AG with a molecular graph omitted, the procedure of creating molecular electronegative edge vector[21−23] is briefly stated, in a similar way as illustrated above, as follows: The distance matrix of non-hydrogen atom in the GA molecule is not shown yet, the number of a certain distance can be known clearly. In the sample molecule AG, for the first element of the v vector there are altogether 9 groups of interactions between adjacent atoms: 3 groups of carbon-carbon interactions between 1st−2nd, 2nd−4th, 7th−8th atoms, 3 groups of carbon-nitrogen interactions between 2nd−3rd, 4th−6th, 6th−7th atoms, 3 groups of carbon-oxygen interactions between 4th−5th, 8th−9th, 8th−10th atoms. Next, for the second element of the ***v*** vector, there are altogether 11 groups of interactions between atoms with bond-conjugation distance being two: 2 groups of carbon-carbon interactions between 1st−4th, 4th−7th atoms, 4 groups of carbon-nitrogen interactions between the 1st−3rd, 2nd−6th, 3rd−4th, 6th−8th atoms, 3 groups of carbonoxygen interactions between 2nd−5th, 7th−9th, 7th−10th atoms, one group of nitrogen-oxygen interaction between 5th−6th atoms. Then, for the third to sixth non-zero elements, the calculation is done in the

same way, and for the seventh to tenth elements, all zero-valued elements are obtained due to the path account of seven through ten being zero. So all elements, from the first to tenth, of the v vector can be calculated as $\nu_1 = 12.6235$; $\nu_2 = 4.3109$; $\nu_3 = 1.9641$; $\nu_4 = 0.9990$; $\nu_5 = 0.53368$; $\nu_6 = 0.1556$; $\nu_7 = \nu_8 = \nu_9 = \nu_{10} = 0$. Therefore the v vector of sample molecule GA is: v=(12.6235, 4.3109, 1.9641, 0.9990, 0.5368, 0.1556, 0, 0, 0, 0). These elements values of the v vector of any other peptides can also be obtained with similar method.

As one of the most complicated and diverse immune systems, MHC is also called the HLA (humanleukocytoantigen, HLA) system. The HLA system possesses 4 types; among them type I consists of HLA-A, HLA-B and HLA-C and widely exists in various tissue cells. All type I MHC molecules are expressed in almost eukaryotic cells including CTL (Cytotoxic T lymphocyte) and infected by outer microbiologics. In order to explain the ability to characterize molecular structure of polypeptides, molecular modeling was further performed for functional prediction of polypeptide sequences with antigenic activity through QSAMs by VMED after extension. The results showed that VMED exhibited both excellent structural selectivity and ascendant activity prediction. Besides, the molecular structure of antigenic polypeptides requires to be well expressed in order to propose the automated technique for the computerized identification and/or recognition of HTL, Th epitopes. Furthermore, by considering the side chain of each amino acid as a pseudo-atom here, a modified MED vector was proposed from the primary structure of polypeptides, based on ECI and relative bonding distance (RBD) of the fundamental skeleton groups. The developed VMED would be very useful in structural characterization and activity prediction of biological molecules because it was easy to calculate and able to work. Some quantitative model was established for 28 AGPP with 14 (1－14) A^d and 14 other restricted activities, assigned as "1"(+) and "0"(–), respectively, latter covering 6 A^b(15－20), 3 A^k(21－23), 2 E^k(24－26), 2 H－2^k(27 and 28) restricted sequences. VMED is now extended as non-hydrogen atoms i and j, regarded as amino acid side chains; and their corresponding distance d_{ij} is considered as the length of both side chains. The electric charge of non-hydrogen atoms i is placed by its ECI[31－39], which is obtained by calculating the equation ECI=$\Sigma|q_i|$, where q_i is atomic local charge on each side chain calculated by quantum chemical method, CNDO/2. So it is not difficult to calculate for obtaining VMED of every polypeptides. Now take two antigenic polypeptides HLCGSHLVEAL and FESNFNTEATNR as examples to demonstrate calculation of the modified MED vector. First write the distance matrix of antigenic polypeptides and then know the interaction terms of amino acid side chains with a given distance in the main chain of antigenic undecapeptides. For instance 1, i.e. antigenic polypeptides 1 "HLCGSHLVEAL", there are 10 terms including H-L, L-C, C-G, G-S, S-H, H-L, L-V, V-E, E-A and A-L, and all of these terms are summed to give the first element of the MED vector:

$$\varpi_1 = \sum_{i=1}^{n}(q_i q_j / d_{ij}^2) = \sum_{i=1}^{10}(q_i q_j / \mathrm{I}) = v = \sum_{i<j}^{n}(q_i q_j)/d_{ij}$$

$$= \sum_{i<j}^{10}(q_i q_j) = 0.56\times0.10+0.10\times0.15+0.15\times0.02+0.02\times 0.56+0.56\times0.56+0.56\times0.10+0.10\times0.07+0.07\times1.31+1.31\times0.05 +0.05\times0.10=0.6240.$$

The other elements can be calculated in the similar way. So it is quite easy to obtain VMED for exampling antigenic polypeptide 1, HLCGSHLVEAL, as ν_1= (0.624, 0.1355, 0.1042, 0.0724, 0.0178, 0.0088, 0.0037, 0.0118, 0.0005, 0.0006). As for instance 2, i.e. antigenic dodecapeptide 22, "FESNFNTEATNR", there are 11 terms including F-E, E-S, S-N, N-F, F-N, N-T, T-E, E-A, A-T, T-N and N-R, and all of these terms are summed to give the first element of the MED vector:

$$\varpi_1 = \sum_{i=1}^{n}(q_i q_j / d_{ij}^2) = \sum_{i=1}^{11}(q_i q_j / 2^2) = v = \sum_{i<j}^{n}(q_i q_j)/ d_{ij} = \sum_{i<j}^{11}(q_i q_j / 2^2) = 0.14\times1.31+1.31\times0.56+0.56\times0.14 +0.14\times1.31+1.31\times0.14+0.14\times1.31+1.31\times0.65+0.65\times1.31 +1.31\times0.05+0.05\times0.65+0.65\times1.31+1.31\times1.69=6.884.$$

The other elements can be calculated in the similar way. So it is quite easy to obtain VMED for exampling antigenic polypeptide 22, FESNFNTEATNR, as ν_{22}= (6.8838, 1.861, 0.4915, 0.4837, 0.1896, 0.1412, 0.0524, 0.0595, 0.034, 0.024). For another antigenic undecapeptide 24 "LTALGAILKKK", there are also 10 terms including L-T, T-A, A-L, L-G, G-A, A-I, I-L, L-K, K-K, and K-K, and all of these terms are summed to give the first element of the MED vector:

$$\varpi_{24} = \sum_{i<j}^{n}(q_i q_j)/d_{ij} = \sum_{i<j}^{11}(q_i q_j) = 0.10\times0.65 +0.65\times 0.05+0.05\times0.10+0.10\times0.02+0.024\times0.05+0.05\times0.09+ 0.09\times0.10+0.10\times0.53+0.53\times0.53+0.53\times0.53=0.7338.$$

Similarly, it is quite easy to obtain VMED for this decapeptide 24, as ν_1= (0.7338, 0.1161, 0.0182, 0.0084,

0.0063, 0.0046, 0.0089, 0.0066, 0.0049, 0.0005).

For various antigenic polypeptides i (i=1,2,..,n), the biological activity measured at unrelated conditions, $y(i)$, can be described as a linear combination of the descriptor vector x (i, k) (k=1,2,...,m) correspondingly expressing different features:

$$y(i) = b(0) + \sum_k x(i,k)b(k) + e(i) = b0 + xb + e(i), \quad (3)$$

where $e(i)$ is the statistical residual or measurement noise; and here the descriptor vector $x(i,k)$(k=1,2,...,m) is justly the above-mentioned ν vector. In this case, the descriptor matrix X refers to the independent descriptive variables and the biological activity matrix Y to the dependent variables or functions. The calibration parameters or combination coefficients, $b(k)$, are usually obtained by indirect calibration methods, i.e., a calibration set consisting of n samples with known measured activities Y (matrix or vector) and the descriptors X (matrix), is used to build up the calibration model or to model the calibration process. Multiple linear regression (MLR) is the most frequently applied direct calibration method for this purpose. Stepwise multiple regression (SMR) is an alternative method. In MLR, a direct regression of X against Y is built up. MLR is performed to solve the above equation to give the calibration model (4) and a prediction model (5):

$$B = (X'X)^{-1}XY, \quad (4)$$

$$Y_{\text{un}} = BX_{\text{un}}, \quad (5)$$

where B is the calibration modeling coefficients; Y_{un} and X_{un} are the unknown biological activities and the calculated descriptor vector or matrix, respectively.

2 Experimental

2.1 Data sets

All the selected 28 AGPP, from undecopeptide through docosapeptide, are taken from refs. [31−39] (see Table 1 for details). There are both 14 (1−14) A^d and 14 other, non-A^d, restricted activities assigned as "1"(+) and "0"(−), respectively, latter covering 6 A^b (15−20), 3 A^k (21−23), 2 E^k (24−26) and H-2^k (27 and 28) restricted sequences.

2.2 Computational softwares

Calculation of VMED for QSAMs was done on a personal computer with the computational programs called MED-LCBMP, written domestically in Turbo C or Visual Basic languages.

3 Results and discussion

3.1 QSAM modeling of 28 antigenic polypeptides

In order to explain the ability to characterize molecular structure of polypeptides, QSAM modeling was performed for functional prediction of polypeptide sequences with antigenic activity. Before VMED $\boldsymbol{\nu}$ was employed to establish a QSAM equation or QSAR model by the multiple linear regression (MLR) technique, all biological activity data should be pretreated as a discrete variable "1" and "0" due to the original activities, "+" for active and "–" for inactive given in literature[35−37]. In the regression computations, the immunogenic polypeptides were estimated and/or predicted through VMED as active and inactive when the calculated values were near to "1" and "0", respectively (see Table 1 for the results).

The obtained results show that VMED exhibits both excellent structural selectivity and prominent activity prediction. Besides, the molecular structure of antigenic polypeptides was well expressed in order to develop automated technique for the computerized identification and/or recognition of HTL, Th epitopes. Furthermore, a modified MED vector was proposed from the primary structure of polypeptides, based on ECI and RBD of the fundamental skeleton groups and/or side chain of each amino acid as a pseudo-atom. The developed VMED would be very useful in structural characterization and activity prediction of biological molecules due to easy calculation and good performance.

Some quantitative models were established and tested by both stochastic sampling and cross validations in order to demonstrate preeminent modeling characterization. Stochastic sampling validations were done by arbitrarily selecting 8 samples (see Table 1 for those with the "*"symbol) as the testing prediction set and remaining 20 samples as the training calibration set from all 28 antigenic polypeptides[22]. The estimated and predicted results are shown in the 5th and 11th columns of Table 1. Only 2 samples (Nos. 5 and 24) were wrongly classified for the case that 20 samples were taken as the training set and were in full agreement with the situation that all 28 samples were taken as the training set, which indicated that the developed model possessed good estimation stability. Besides, all remaining 8 samples were

Table 1 Various sequences, descriptive vectors, observed and calculated activities of antigenic polypeptides with MHC restriction

No.	Note	Restricted	Name/type of peptide	Peptides/amino acid sequence	Length	v_1	v_2	v_3	v_4	v_5	v_6	v_7	v_8	v_9	v_{10}	y_{exp}	y_{est}	y_{prd}
1		(a) A^d	bovine insulin B-chain 5-15	HLCGSHLVEAL	11	0.6240	0.1355	0.1042	0.0724	0.0178	0.0088	0.0037	0.0118	0.0005	0.0006	+	+	+
2			sperm whale myoglobin 106-118	FISEAIIHVLHSR	13	2.2873	0.3753	0.1024	0.0809	0.0591	0.0129	0.0215	0.0181	0.0320	0.0108	+	+	+
3			cytochrome bovine 13-25	KCAQCHTVEKGGK	13	1.6601	0.5181	0.3392	0.0933	0.0928	0.0417	0.012	0.0138	0.0124	0.0004	+	+	+
4	*		chicken ovalbumin 323-339	ISQAVHAAHEINEAGR	16	3.1269	0.7711	0.5855	0.2255	0.0758	0.0921	0.0277	0.0569	0.0141	0.0211	+	+	+
5			staph nuclease 61-80	FTKKMVENAKKIEVEFDKGQ	20	4.5088	1.7228	0.7095	0.3155	0.2798	0.1197	0.1111	0.0548	0.0476	0.0437	-	–	–
6			lambda repressor 12-24	LEDARRLKAIYEK	13	6.7269	1.0101	0.6467	0.2227	0.1098	0.1182	0.0689	0.0310	0.0323	0.0245	+	+	+
7	*		flu haemaglutinin 130-142	HNTNGVTAACSHE	13	3.6834	0.7717	0.2235	0.0662	0.0584	0.0433	0.021	0.0211	0.0358	0.0190	+	+	+*
8	*		bovine insulin A-chain 1-21	GIVEQCCASVCSLYQLENYCN	21	6.6151	1.3637	0.7151	0.3181	0.1277	0.1149	0.0594	0.0369	0.0369	0.0374	+	+	+*
9			sheep insulin A-chain 1-21	GIVEQCCAGVCSLYQLENYCN	21	6.5503	1.3232	0.6724	0.2689	0.0838	0.0935	0.0573	0.0257	0.0282	0.0335	+	+	+
10			porcine insulin A-chain	GIVEQCCTSICSLYQLENYCN	21	7.0553	1.4028	0.8162	0.3893	0.1339	0.1292	0.0769	0.0383	0.0468	0.0453	+	+	+
11			equine insulin A-chain 1-21	GIVEQCCTGICSLYQLENYCN	21	6.6557	1.3623	0.7736	0.3401	0.0901	0.1078	0.0748	0.0271	0.0381	0.0414	+	+	+
12	*		HSV_Sd 245-260	APYSTLLPPELSETP	15	3.0927	0.6026	0.2764	0.1014	0.0667	0.0350	0.0330	0.024	0.0205	0.0151	+	+	+*
13			[Ala 14] insulin A-chain 1-14	GIVEQCCASVCSLA	14	2.3381	0.2190	0.0949	0.0624	0.0395	0.0123	0.0212	0.0141	0.0031	0.0013	+	+	+
14			[Ala 12/13] insulin A-chain 1-14	GIVEQCCASVCAAY	14	2.2391	0.2102	0.0739	0.062	0.0514	0.0109	0.0089	0.0042	0.0131	0.0095	+	+	+
15		(b) A^b	cytochrome horse 45-58	GFTYTDANKNKGIT	14	4.1241	1.3042	0.4112	0.2561	0.1297	0.0749	0.0304	0.0222	0.0074	0.0054	–	–	–
16	*		hen egg lysozyme 78-93	IPCSALLSSDITASVN	16	1.5974	0.7380	0.1085	0.1538	0.0442	0.0819	0.0254	0.0223	0.0052	0.0057	–	–	–*
17			staph nuclease 91-110	YIRADFKMVNEALVRQGLAK	20	5.1798	1.3283	0.2981	0.3752	0.3007	0.1192	0.0721	0.0641	0.0367	0.0417	–	–	–
18			pigeon cytochrome C 45-58	GFSYTDANKNKGIT	14	4.0467	1.2891	0.3987	0.2558	0.125	0.0736	0.028	0.0215	0.0074	0.0053	–	–	–
19			lamda repressor 73-88	VEEFSPSIAREIYEMY	15	6.4182	1.0015	0.6267	0.2997	0.149	0.0823	0.0701	0.0752	0.0401	0.0164	–	–	–
20	*		herp glycoprotein D 1-20	KYALADASLKMADPNRFRGK	20	4.0904	1.8924	0.6412	0.2378	0.2397	0.0552	0.0861	0.0418	0.0348	0.0371	–	–	–*
21		(c) A^k	hen lysozyme 46-60	NTDGSTDYGILQINS	15	4.9118	1.3967	0.3063	0.2552	0.1808	0.115	0.0802	0.0287	0.0368	0.0147	–	–	–
22			hen lysozyme 34-45	FESNFNTEATNR	12	6.8838	1.861	0.4915	0.4837	0.1896	0.1412	0.0524	0.0595	0.0341	0.0240	–	–	–
23			malaria circums. protein 326-343	PSDKHIEQYLKKIKNSIS	18	6.7081	1.1973	0.5645	0.3587	0.1863	0.1236	0.0922	0.0596	0.0333	0.0244	–	–	–
24	*	(d) E^k	sp. whale myoglobin 69-78	LTALGAILKKK	11	0.7338	0.1161	0.0182	0.0084	0.0063	0.0046	0.0089	0.0066	0.0049	0.0005	–	+	+*
25			staph nucl. 81-100	RTDKYGRGLAYIYADGKMVN	20	3.6395	1.8518	0.4172	0.3833	0.1377	0.1504	0.0399	0.0628	0.0263	0.0408	–	–	–
26			cytochrome moth 89-103	NERADLIAYLKQATK	15	5.5423	1.5097	0.4551	0.2082	0.0477	0.0728	0.0611	0.0346	0.0493	0.0326	–	–	–
27		(e) $H\text{-}2^k$	hepat. B suface antigen 140-154	TKPSDGNCTCIPIPS	15	1.8079	0.8233	0.2661	0.1379	0.0678	0.0484	0.0206	0.0227	0.0053	0.0095	–	–	–
28	*		hepat. B pre S 120-132	MQWNSTTFHQTLQ	14	6.8821	1.4996	0.6759	0.2722	0.1372	0.1142	0.0831	0.0568	0.0400	0.0183	–	–	–*

a) y_{obsd} refers to the observed data; b) y_{repd} to the reported data in references; c) y to the estimated values by modeling with all samples; d) y to the predicted values by modeling with randomly sampling; (e) y to the validated values by cross validation with the leave-one-out procedure. "+" stands for the active sample; "–" for the inactive sample. "*" represents the samples in the testing set.

rightly predicted for the testing set, which indicated that the developed model possessed good prediction capability. In ref. [36], although all the 20 training samples were correctly classified, one testing sample from the remaining 8 ones was mistakenly predicted, which indicated that the referenced model had lower predicting ability. In order to further evaluate validation performance, cross validation with leave-one-out procedure was made by the proposed method with quite good results (see the 5th and 17th columns in Table 1 for details). All these results illustrate that the created model (F) possesses both satisfactory estimation stability and excellent prediction power. Additionally, classification of Th epitopes demonstrates that both active and inactive samples are not all repulsive. In other words, some antigentic peptides with E^k restriction do not at all mean its no A^d-restriction. Actually, in refs. [52−58], various subtypes of HMC (DR1, DR7, DR5) can identify the same antigenic peptide, i.e. some epitopes can be both one-restriction (E^k) and another-restriction (A^d). Therefore, the sample (No. 24) "wrongly" classified here, an epitope with E^k-restriction, may also behave A^d-restriction. Of course, this needs the further validation with experiments. So, it does not mean at all that the classification of reference is unadvisable; but in opposition, this is just a characteristic.

The described method, with good results (Y_{model}, Y_{test}) being very close to those reported in refs. [31−37] through a much simpler method than the ones in refs. [31−37], may be very suitable for both estimation and prediction of classes I and II for MHC epitope of human. It may also be useful and helpful to immune identification and antigenic recognition of both proteins and genes and to design and development of various subunit vaccines. Moreover, the obtained results show that VMED behaves quite well for both QSAR and QSAM of poly- and oligopeptides, which exhibit both transcendent estimation ability and prominent prediction power, equal to or better than those reported in the previous references. Finally, a preliminary conclusion may be drawn: both the classical and modified MED vectors are very useful structural descriptor parameters. Some suggestions were proposed for further studies on QSAR/QSAM of proteins and nuclei acids in various fields.

3.2 QSAR modeling of various oligopeptides

Only from the primary structure of peptides, based on Pauling's electronegativity of each atom and the distance between atoms, a new set of descriptors, VMEE, was proposed in our laboratories. Several QSAR models were proposed on biological activity of 58 ACE inhibitors, 48 BTT dipeptides, 31 BPP agents, and 24 rapidus surge kinetin (RSK) heptapeptides, by various samples but effective molecular modeling such as MLR, stepwise multivariate regression (SMR), principal component regression (PCR) and so on. In order to explain the ability to characterize the molecular structure of peptides, a further investigation was carried out on modeling quantitative structure activity relationship of 152 CTL epitopes (antigenic oligopeptides, nonapeptides). Here, the main factors were extracted based on standard regression coefficients of each element and the results were close to or better than literature (see Table 2 and Figure 1, also see refs. [20−26]). Simultaneously, some information of advanced structure can be found from the main influent factors extracted based on the standard regression coefficients. Besides, the developed novel VMEDν has also excellent structural selectivity and ascendant activity estimation and this novel molecular electronegative edge vector, because it can be calculated easily only from the primary structure without requirements of other knowledge about electrostatic or electronic, geometry-steric or stereoscopic, hydrophobic or lipophilic parameters for residues of amino acids, will be useful in structural characterization and activity prediction of biological macromolecules[4−20], such as proteins and nuclei acids, due to its high structure selectivity, fine activity correlativity, good computation performance. The related work is in progress.

Furthermore, for polypeptides, it will be very heavy, complicated and cockamamie for atom-based fabrication, due to a too large number of atoms and a novel method of VMED ν is then developed by using residue-based construction. Certainly, there are some open problems requiring further consummation: 1) novel methods of both skeleton- and residue-based VMED ν are required to be deeply investigated for molecular structure expression, especially for the case with multifunctional groups, the approach described here needs to be improved further; and local characterization through a given skeleton is worthy to further examination; 2) The developed MED vectors seem suitable for structure expression and QSAR study of oligopeptides based on their primary structure with accurate prediction. However, prediction

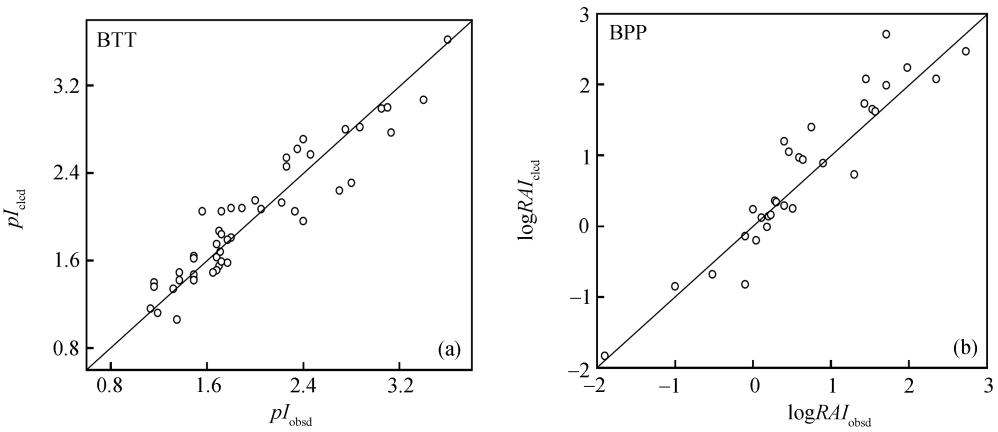

Figure 1 Plot of observed and calculated bioactivity values (log $1/XC_{50}$) of some bioactive oligopeptides (BOP). (a) Plot of the observed and calculated (log$1/T$) values of 48 BTT; (b) Plot of the observed and calculated (log *RAI*) values of 31 BPP.

Table 2 Some QSAM models for several panels of peptides by VMED

No.	Oligopeptides	Data set	Model Mi	n	^{mlr}m	$^{mm}R^2_{cu}$ a)	$^{cv}Q^2_{cu}$ b)	E_{RMS} c)	^{pcr}l	$^{mm}R^2_{cum}$ a)	$^{cv}Q^2_{cum}$ b)	E_{RMS} c)
1	dipeptide, dp, 2p	58ACE inhibitors	M1 e)[39,40]	58	10	0.792	0.677	0.50	2	0.741	0.711	0.50
2	dipeptide, dp, 2p	48BTT thresholds	M1 e)[39,40]	48	10	0.710	0.475	0.35	3	0.648	0.570	0.37
2”	dipeptide, dp, 2p	47BTT thresholds	M1 e)[39,40]	47	10	0.772	0.587	0.33	3	0.734	0.677	0.33
3	pentapeptide, ap, 5p	31BPT agents	M1 e)[39~42]	31	10	0.857	0.578	0.40	2	0.723	0.655	0.55
4	heptapeptide, hp, 7p	24RSK(with d-form)	M1 e)[39~40]	24	10	0.680	0.598	0.75	6	0.642	0.532	0.68
5	nonapeptide, np, 9p	152 CTL epitopes	Vhbe[48,59]	152	10	0.792	0.677	0.51	3	0.712	0.611	0.55

a) R^2_{cu} stands for the cumulative correlation coefficients of molecular modeling in the calibration set (n=58); b) Q^2_{cu} stands for the cumulative correlation coefficients of cross validation in the prediction set (n-1); c) E_{RMS} refers to the rooted mean squares of error; d) n refers to samples, m variables, l latent varibles, nd means not determined; e) Mi stands for QSAM results obtained by SMR-MLR.

of protein epitopes or antigenic determinants is involved in molecular immunology and chemical biology, researches and development of vaccines, protection and control of fatal diseases, and some other important problems[46−60] including design and preparation of vaccines to prevent and control SARS from atypical pneumonia[60], AIV from bird flu. Further approaches are required to really resolve these difficult QSAR/QSAM problems[60−68] by quantitative molecular modeling.

Prof. Cai S X, Li Z Z, Zheng X L, Li S S, Wang G X, Li S L, Chen G P, Li S H, Huang Y, and Li ZY, are thanked for addressing robustness problems and for helpful discussions. Dr. Sun L L, Liang G Z, Liu Z D, Gao J K, Xu Z L, Shu M, Li Y, He R B, Li G, Liu Q F, Peng O, Wu R, Han M, Hu F, Kang J, Hou M, Zhang M J, Chen G H, Li B Y, Nie J Y, Yin Z H, Lan Y K, Qin R H, Wang Y Q, Zeng H, Li B, Qiu L J, Wang J N, Liu Y H, Zhang Y H, Zhou P, Tian F F, Zhang Q X, Peng C Y, Deng J, Sun J Y, Li K, Xu L N, Liao L M, He L, Zhu W P, Yang J, Wu Y Q, Liu Z, Yang C, Zhou Y H, Zou Z H, Jiang X R, Han L, Duan Y C, Wu C Y, Jiang G H, Deng H, Huang P, Fernandez R are acknowledged for supplying some experimental and observation data and for providing technological assistances.

1 Placa J. Human genome—Development of energy on the map. Nature 1986, 321, 371—386
2 Venter J C, Smith H O, Hood L. A new strategy for genome sequencing. Nature 1996, 381, 364—366
3 International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. Nature, 2001, 409: 860—921
4 Chen K -X, Jiang H -L, Ji R -Y. Computer-assisted Drug Design—Principle, Method and Application (in Chinese). Shanghai: Shanghai Science and Techonlogical Publishing House, 2000
5 Martin Y C. Quantitative Drug Design: A Critical Introduction. New York: Marcel Dekker Inc, 1978, Preface Ch7
6 Miyashita Y, Li Z, Sasaki S. Chemical pattern recognition and multivariate analysis for QSAR studies. Trend Anal Chem TrAc, 1993, 12: 50—60, doi: 10.1016/0165-9936(93)87051-X
7 Wang Z X. Assessing the accuracy of protein secondary structure. Nat Struct Biol, 1994, 1: 145—146
8 Wang Z X. How many fold types of protein are there in nature? Proteins: Structure, Function and Genetics, 1996, 26: 186—191
9 Wang Z X. Influence of substrates on *in vitro* dephosphorylation of glycogen phosphorylase a by protein phosphatase-1. Biochem J, 1999, 341: 545—554
10 Wang Z X, Yuan Z. How good is the prediction of protein structural classes by the component-coupled method? Proteins: Structure, Function and Genetics, 2000, 38: 165—175
11 Wang Z X, Wu J W. Autophosphorylation kinetics of protein kinases. Biochem J, 2002, 368: 947—952
12 Wu H, Zheng Y, Wang Z X. Evaluation of the catalytic mechanism of

the p21-activated protein kinase PAK2. Biochemistry, 2003, 42: 1129－1139
13 Wu H, Wang Z X. The mechanism of p21-activated protein kinase 2 autoactivation. J Biol Chem, 2003, 278: 41768－41778
14 Luo Y, Jiang X L, Lai L H. Modeling protein backbone structure based on C_α guiding coordinates. Protein Eng, 1992, 5: 147
15 Luo Y, Lai L H, Xu X J. Defining topological equivalents in protein structures by means of dynamic programming algorithm. Protein Eng, 1993, 6: 373
16 Qu C X, Lai L H, Xu X J. Phyletic relationship of proteins based on structure preference factors. J Mol Evol, 1993, 36: 67
17 Cao W, Liu L, Lai L H, et al. Molecular recognition: monomer of the yeast transcriptional activator GCN4 recognizes its dimer DNA binding target sites specifically. Sci China Ser B-Chem Sci, 2000, 43(5): 466－476
18 Zhang W, Feng J N, Shen B F. Identification of binding epitope of a monoclonal antibody (Z12) against human TNF-α using computer modeling and deletion mutant technique. Sci China Ser C-Life Sci, 2004, 47(3): 279－286
19 Feng J N, Wan T, Wu J J, et al. Epitope prediction based on three-dimensional structure. J Mol Sci, 1999, 15(2): 112－115
20 Wan T, Sun T, Wu J J, et al. The multi-parameter prediction of protein antigenic determinants. Chin J Immunol, 1997, 13(6): 329－333
21 Anfinsen C B, Haber E, Sela M, et al. The kinetics of formation of native ribonuclease during oxidation of the reduced polypeptide chain. Proc Natl Acad Sci USA, 1961, 47: 1309
22 Martin J, Mayhew M, Lattger T, et al. The reaction cycle of GroEL and GroES in chaperonin-assisted protein folding. Nature, 1993, 366: 228
23 Sneath P H A. Relations between chemical structure and biological activity in peptides. J Theor Biol, 1966, 12: 157－195
24 Borea P A, Santo G P, Salvadori S, et al. Opioid peptides. Pharmacological activity and lipophilic character of dermorphin oligopeptides. Farmaco Ed Sci, 1983, 38: 521－526
25 Asao M, Iwamura H, Akamatsu M, et al. Quantitative structure-activity relationships of the bitter thresholds of amino acids, peptides, and their derivatives. J Med Chem, 1987, 30: 1873－1879
26 Fauchere J, Charton M, Kier L B, et al. Amino acid side chain parameters for correlation studies in biology and pharmacology. Int J Pept Protein Res, 1988, 32: 269－278
27 Depriest S A, Mayer D, Naylor C D, et al. 3D-QSAR of angiotensin-converting enzyme and thermolysin inhibitors: A comparison of CoMFA models based on deduced and experimentally determined active site geometries. J Am Chem Soc, 1993, 115: 5372－5384
28 Cocchi M, Johansson E. Amino acids characterization by GRID and multivariate data analysis. Quant Struct Act Relat, 1993, 12: 1－8
29 Charton M. The quantitative description of amino acid, peptide, and protein properties and bioactivities. Prop Phys Org Chem, 1990, 18: 163－284
30 Kidera A, Konishi Y, Oka M, et al. A statistical analysis of the physical properties of the 20 naturally occurring amino acids. J Protein Chem, 1985, 4: 23－55
31 Hellberg S, Sjostrom M, Skagerberg B, et al. Peptide quantitative structure-activity relationships, A multivariate approach. J Med Chem, 1987, 30: 1126－1135
32 Hellberg S, Eriksson L, Jonsson J, et al. Minimum analogue peptide sets (MAPS) for quantitative structure-activity relationships. Int J Pept Protein Res, 1991, 37: 414－424
33 Wold S, Eriksson L, Hellberg S, et al. Principal property values for six non-natural amino acids and their application to a structure-activity relationship for oxytocin peptide analogues. Can J Chem, 1987, 65: 1814－1820
34 Collantes E R, Dunn W J. Amino acid side chain descriptors for quantitative structure activity relationship studies of peptide analogues. J Med Chem, 1995, 38: 2705－2713
35 Zaliani A, Gancia E. MS-WHIM scores for amino acids: A new 3D-description for peptide QSAR and QSPR studies. J Chem Inf Compt Sci, 1999, 39: 525－533
36 Raychaudhury C, Banerjee A, Bag P, et al. Topological shape and size of peptides: Identification of potential allele specific helper T cell antigenic sites. J Chem Inf Comput Sci, 1999, 39: 248－254
37 Liu S S, Cai C Z, Li Z. Approach to estimation and prediction for normal boiling points of alkanes based on a molecular distance-edge vector(*MDE*), lambt. J Chem Inf Comput Sci, 1998, 38(3): 387－394, doi: 10.1021/ci970109z
38 Liu S S, Cai S X, Liu Y, et al. A novel molecular electronegativity-distance vector(MEDV). Acta Chim Sin (in Chinese), 2000, 58(11): 1353－1357
39 Liu S S. Novel molecular electronegativity-distance vector for pharmaceutical characterization and application. PhD Dissertation (in Chinese). Chongqing: Chongqing University, 2001, 05, Ch1-9: 118+19
40 Liu S S. Novel molecular electronegativity-distance vector for organic characterization and application. Selected 100 Excellent PhD Dissertations (in Chinese). Beijing: Higher Education Press, 2005. 07, Ch1-12: 228+17
41 Ling Z. Chemical structural parameterization and chemobiological property quantitation of types of organic compounds. MSc Thesis (in Chinese). Chongqing: Chongqing University, 2000.04, Ch1-7: 127+6
42 Ling Z H, Liu S S, Li Z. Structural parameterization and QSAR study of oligopeptides. Acta Chim Sin (in Chinese), 2001, 59(7): 1004－1008
43 Xu H. Chemical structural parameterization and chemobiological property quantitation of typical organic compounds. MSc Thesis (in Chinese). Chongqing: Chongqing University, 2001, 05, Ch1-9: 118+19
44 Sun L L. Structure expression and function prediction of biologically active compounds. PhD Dissertation (in Chinese). Chongqing: Chongqing University, 2004, 06, Ch1-9: 110+8
45 Sun L L, Zhou Y, Li G R, et al. Molecular eletronegativity-distance vector (MEDV-4):A two-dimensional QSAR method for the estimation and prediction of biological activities of estradiol derivatives. J Molecular Structure (Theochem), 2004, 679: 107－113. doi: 10.1016/j.theochem.204.04.010
46 Xiong Q. Eukaryotic promoter prediction. PhD Dissertation (in Chinese), Chongqing: Chongqing University, 2004,11, Ch1-9: 118+8
47 Xiong Q, Wang Y, Li Z. Eukaryotic promoter recognition using backpropagation neural network. Chin J Biomed Eng, 2004, 13(2): 87－92
48 Mei H. Peptide QSARs. PhD Dissertation (in Chinese), Chongqing:

Chongqing University, 2005, 05, Ch1-9: 129+8
49 Li S Z, Fu B, Wang Y, et al. On structural parameterization and molecular modeling of peptide analogues by molecular electronegativity edge vector (VMEE): Estimation and prediction for biological activity of dipeptides. J Chin Chem Soc, 2001, 48(5): 937－944, doi.wiley.com/10.1002/bip.20296, dx.doi.org/10.1002/bip.20296
50 Zhou P, Tian F F, Zhang M J, et al. Applying generalized hydrophobicity scale of amino acids to quantitative prediction of human leukocyte antigen-A*0201-restricted cytotoxic T lymphocyte epitope. Chin Sci Bull, 2006, 51(12): 1439－1443, doi: 101007/s11434-006-1439-z
51 Zhou P, Zhou Y, Wu S R, et al. A new descriptor of amino acids based on the three-dimensional vector of atomic interaction field. Chin Sci Bull 2006, 51(5): 524－529, doi: 101007/s11434-006-0524-7
52 Rothbard J B, Taylor W R. Sequence pattern common to T cell epitopes. EMBO J, 1988, 7(1): 93－100
53 Willims D B, Ferguson J, Gariepy J, et al. Characterization of the insulin A-chain major immunogenic determinant presented by MHC class II I-Ad molecules. J Immunol, 1993, 151: 3627－3627
54 Sette A S, Buus S, Colon S, et al. I-Ad-binding peptides derived from unrelated protein antigens share a common structural motif. J Immunol, 1988, 141: 45
55 Demotz S, Sette A, Sakaguchi K, et al. Self peptide requirement for class II major histocompatibility complex allorecognition. Proc Natl Acad Sci USA, 1991, 88(19): 8730－8734
56 Marrack P, Kappler J. The T cell receptor. Science, 1987, 238(4830): 1073－1079
57 Adorini L, Sette A, Buus S, et al. Interaction of an immunodominant epitope with Ia molecules in T-cell activation. Proc Natl Acad Sci USA, 1988, 85(14): 5181－5185
58 Ozaki S, Durum SK, Muegge K, et al. Production of T-T hybrids from T cell clones. Direct comparison between cloned T cells and T hybridoma cells derived from them. J Immunol, 1988, 141(1): 71－78
59 Liu S S, Yin C S, Cai S X, et al. QSAR study of steroid benchmark and dipeptides based on MEDV-13. J Chem Inf Comput Sci, 2001, 41(2): 321－329, doi: 10.1021/ci0003350
60 Mei H, Sun L L, Zhou Y, et al. Identification of encoding proteins related to SARS-CoV. Chin Sci Bull, 2004, 49(19): 2037－2040
61 Mei H, Sun L L, Zhou Y, et al. A new set of amino acid descriptors and its application in peptide QSARs. Biopolymers: Pep Sci, 2005, 80(6): 775－786, doi: 10.1002 bip.20296
62 Liang G Z, Li Z L. A new sequence pepresentation (FASGAI) as applied in better specificity elucidation for human immunodeficiency virus type 1 protease. Biopolymers (Pept Sci), 2007, 88(3): 401－412, doi: 10.002/bip 20669
63 Liang G Z, Li Z L. Scores of generalized base properties for quantitative sequence-activity modelings for *E. coli* promoters based on support vector machine. J Mol Graph Model, 2007, 26(1): 269－281, doi: 10.1016/j.jmgm.2006.12.004
64 Liang G Z, Yang S B, Zhou Y, et al. Using scores of amino acid topological descriptors for quantitative sequence-mobility modeling of peptides based on support vector machine. Chin Sci Bull, 2006, 51(22): 2700－2705, doi: 10.1007/s11434-006-2138-5
65 Zhou P, Zeng H, Tian F F, et al. Applying novel molecular electronegativity-interaction vector (MEIV) to QSPR study on collision cross section of singly protonated peptides. QSAR Comb Sci, 2007, 26(1): 117－121, doi: 10.1002/qsar.200510220
66 Zhou P, Mei H, Tian F F, et al. A new two-dimensional approach to quantitative prediction for collision cross-section of more than 110 singly protonated peptides by a novel moecular electronegativity-interaction vector through quantitative structure-spectrometry relationship studies. Frontiers of Chem China, 2007, 2(1): 55－64, doi: 10.1007/s11458-007-0012-x
67 Liu S S, Cai SX, Li Z. Molecular electronegative distance vector (MEDV) related to 15 properties of alkanes, J Chem Inf Comput Sci, 2000, 40(6): 1337－1348, doi: 10.1021/ci0003247
68 Zhou P, Tian F F, Li Z L. Novel molecular electronegativity-interaction vector and its application in quantitative prediction for collision cross-section of singly protonated peptides. Chin J Anal Chem, 2006, 34(6): 778－782, doi: 10.1016/s1872-2040(06)60039-x