Amino Acids (2009) 36:535–554
DOI 10.1007/s00726-008-0116-8

ORIGINAL ARTICLE

In silico quantitative prediction of peptides binding affinity to human MHC molecule: an intuitive quantitative structure–activity relationship approach

F. Tian · L. Yang · F. Lv · Q. Yang · P. Zhou

Received: 20 May 2008 / Accepted: 2 June 2008 / Published online: 25 June 2008

Abstract In this paper, we have handpicked 23 kinds of electronic properties, 37 kinds of steric properties, 54 kinds of hydrophobic properties and 5 kinds of hydrogen bond properties from thousands of amino acid structural and property parameters. Principal component analysis (PCA) was applied on these parameters and thus ten score vectors involving significant nonbonding properties of 20 coded amino acids were yielded, called the divided physicochemical property scores (DPPS) of amino acids. The DPPS descriptor was then used to characterize the structures of 152 HLA-A*0201-restricted CTL epitopes, and significant variables being responsible for the binding affinities were selected by genetic algorithm, and a quantitative structure–activity relationship (QSAR) model by partial least square was established to predict the peptide-HLA-A*0201 molecule interactions. Statistical analysis on the resulted DPPS-based QSAR models were consistent well with experimental exhibits and molecular graphics display. Diversified properties of the different residues in binding peptides may contribute remarkable effect to the interactions between the HLA-A*0201 molecule and its peptide ligands. Particularly, hydrophobicity and hydrogen bond of anchor residues of peptides may have a significant contribution to the interactions. The results showed that DPPS can well represent the structural characteristics of the antigenic peptides and is a promising approach to predict the affinities of peptide binding to HLA-A*0201 in a efficient and intuitive way. We expect that this physical-principle based method can be applied to other protein–peptide interactions as well.

Keywords Quantitative structure–activity relationship · Computer-aided vaccine design · Immunoinformatics · Divided physicochemical property scores of amino acids · HLA-A*0201 · CTL epitope · Genetic algorithm-partial least square regression

F. Tian · F. Lv (✉) · Q. Yang
Research Institute of Surgery, Daping Hospital, Third Military Medical University, Chongqing, China
e-mail: ggootc@163.com

F. Tian · L. Yang
College of Bioengineering, Chongqing University, Chongqing, China

P. Zhou
Department of Chemistry, Zhejiang University, Hangzhou, China

Introduction

Albeit driving much attention, life science is presently known little due to diversity and complexity of biosystems. Emergence of bioinformatics greatly benefits analysis, annotation and induction of a large number of biological data resulted from the experiments and computations. Computational approaches, such as molecular docking (Kitchen et al. 2004), pharmacophore modeling (Chang et al. 2006), protein cleavage site prediction (Chou 1996), protein subcellular location identify (Chou and Shen 2007), enzyme functional class prediction (Shen and Chou 2007a, b), mutation prediction (Xiao et al. 2005), and signal peptide prediction (Shen and Chou 2007a, b), can provide very useful information for drug design in a timely manner. Recently, the term immunoinformatics is proposed, referred to applying bioinformatics methods into immune systems (Rammensee 2003). Because immune system in response to exogenous agents is via some specific fragments of proteins (i.e. epitopes), epitope identification and

prediction are thus becoming the focus of immunoinformatics (Brusic and Flower 2007).

Major histocompatibility complex (MHC), a group of closely linked gene cluster encoding major histocompatibility antigen, is the most complex and polymorphic genes of higher vertebrates. MHC is totally divided into three kinds, of which the type I, including MHC-A, -B and -C, have wide distributions in different kinds of histocytes (Trowsdale et al. 1991; Horton et al. 2004). Mediated by ubiquitin, endogenous antigens (proteins or polypeptides) are degraded into oligepeptide fragments containing 6–30 residues by 20S proteosome. After that, peptide fragments are selectively transported by transporter associated with antigen processing (TAP) into lumen of endoplasmic reticulum, then binding with MHC-I molecules to form MHC–peptide complexes. These complexes are then transferred to cell surface through golgibody and secretory vesicle and subsequently recognized by cytotoxic T lymphocyte (CTL) (or CD8 + T cell), thereby eliciting immune cascades to eventually decompose infected cells (Germain 1994). However, T-cell recognition is not elicited unless antigen segments are bound with MHC molecules (meant MHC restrictions), so antigen peptide–MHC binding potentialities directly relate to the presentation of antigen peptides and activation of CD8 + T cell (Rammensee 1995). CTL epitopes, referring to the antigen peptides that bind with MHC-I molecules to initiate following CD8 + T cell recognition, are typically 9 ± 1 residues long and share a highly conserved properties. Two or more positions of the antigen peptide specifically binding with MHC-I molecules were defined as anchor residues (Falk et al. 1991). Besides anchor residues, some other positions possessing lesser contributions to antigen recognition were defined as the secondary anchor residue (Ruppert et al. 1993). In fact, that a protease-hydrolyzed peptide could be efficiently recognized and presented by MHC-I molecules depend upon not only the anchor residues but also the properties of non-anchor residues. It is believed that the MHC–peptide association involves in many intricate physicochemical effects. Although numerous studies have addressed MHC I–peptide interactions, yet the principles governing antigenic recognition and presentation are not fully understood.

Considering complexity and individual diversity of MHC inheritance, Hagmann pointed out that sought of MHC-binding peptides is the bottleneck of the present vaccine design (Hagmann 2000). Traditionally, short peptide sequences are synthesized in large scale to implement subsequent affinity assay, and in this way, potential active sequences are found. However, considering experimental blindness, it is really a time-consuming and labor-intensive process for developing new vaccines even in assistance of the modern combinatorial chemistry and high-throughout screening technique. In view of that, computer-aided vaccine design (CAVD) was taken emergence in the late 1980s, promising for such a problem (Hagmann 2000; Sette 2000). In the early CAVD stage, Pamer and Hill had ever used successful anchor residue-based simple motif method to predict CTL epitopes of listeria monocytogenes (LM) and human papilloma virus (HPV) (Pamer et al. 1991; Hill et al. 1992). However, Kast et al. (1994) claimed simple motif was infeasible to accurately characterize antigen peptide sequences, with correct prediction rate merely coming up to 30%; later, Kubo et al. (1994) further extended simple motif into non-anchor residues, thereby constructing extended motif which achieved 70% of correct prediction rate. Subsequently, del Guercio et al. (1995) firstly proposed the notion of HLA-supertype and supermotif, and after that, another MHC syperfamilies as HLA-A2 (Altfeld et al. 2001), HLA-A3 (Kawashima et al. 1999a, b) HLA-B7 (Coyle and Gutierrez-Ramos 2001) and HLA-B44 (Sette and Sidney 1998) were continually defined. Recently, research interests have been turned to quantitative predictions of CTL epitopes, with remarkable contributions as of the independent binding of side-chain (IBS) hypothesis proposed by Parker et al. (1994) and the weight coefficient matrix (Sette et al. 1989; Gulukota et al. 1997; Sturniolo et al. 1999). However, restricted to experimental conditions, sequence diversity was insufficient, and even some residue types were missing at some peptide positions, so false positive rate of predictions was high. In addition, some other achievements were accomplished by employing intelligence algorithms as artificial neural network (Honeyman et al. 1998), evolutionary algorithm (Brusic et al. 1998) and hidden Markov model (Udaka et al. 2002) into predictions of CTL epitope-binding affinities. However, these methods were still time consuming, requiring massive experimental data to train the model, thus becoming restricted in applications. In addition, these intricate intelligent algorithm-based models were physiochemically inexplicit, not amenable to interpretation.

Quantitative structure–activity relationship (QSAR) plays a vital role in the modern drug design (Kubinyi 1997). Despite QSAR researches are difficultly performed for biomolecule as peptides and proteins, there are still many successful examples (Zhou et al. 2008). In the field of vaccine design, Doytchinova and Flower (2002) successfully perform a 3D-QSAR study for human leukocyte antigen HLA-A*0201-restricted CTL epitopes by comparative molecular field analysis (CoMFA) and comparative molecular similarity index analysis (CoMSIA), and then the additive model is proposed to analyze each epitope residue in contributions to epitope-binding affinity (Doytchinova et al. 2002); Lin and Guan et al. perform quantitative predictions of binding affinities of HLA-A*0201 restricted

CTL epitopes (Lin et al. 2004; Guan et al. 2005) employ amino acid parameters as isotropic surface area (ISA) by electronic charge index (ECI), *z* scale and physicochemical properties from 2D-QSAR pathway; Recently, Hattotuwagama and Doytchinova et al. discuss MHC–peptide binding potentials in murine and human kinds, respectively, in assistance of iterative self-consistent partial least squares (ISC-PLS) and discriminant analysis partial least squares (DA-PLS), also gaining achievable results (Hattotuwagama et al. 2006; Doytchinova and Flower 2007).

Based upon these previous works, a powerful prediction method targeting to binding affinity of CTL epitopes is attempted in this study, expecting to be potent, simple, quick and interpretable. Such requirements are based upon the following considerations. First, theoretical methods suitable to practical applications should allow high-throughout evaluations of immune activities of unknown sequences, since CTL isomers composed of 20 natural amino acids are extremely huge; for experimental immunologists, theoretical approaches are expected to be physicochemically explicit, thus in favor of guiding experiments. Concomitantly, mathematical and computational methods should not be too intricate and abstract in consideration of practical applications. Therefore, a principal component analysis (PCA)-based amino acid descriptor, focusing on physicochemical properties of CTL epitopes, is devised to characterize sequence structures. To be interpretable, PCA is employed to compress and extract information from 23 electronic properties, 37 steric properties, 54 hydrophobic properties and 5 hydrogen bond properties, respectively, thus yielding a new amino acid descriptor as divided physicochemical property scores (DPPS). Since numerous nonbonding information included in DPPS directly relates to peptide–MHC association, and each component possesses definite physiochemical meanings, DPPS is hopeful to quickly construct potent QSAR models for epitope-binding affinities from a two-dimensional pathway. The resulted model is studied from the aspects of statistics and molecular graphics, providing insight into MHC presentation theory and relevant vaccine design and structural modification to some extent.

Principle and methodology

Divided physicochemical property scores (DPPS) of amino acids

Nonbonding effects, such as electrostatic, van der Waals, hydrophobic interactions and hydrogen bond, play central roles in peptide–MHC interactions (Schueler-Furman 2000; Logean et al. 2001; Doytchinova and Flower 2001; Zhou et al. 2007). As the structure unit in biosystem, properties and arrangements of amino acids directly relate to bioactivity and function of peptides and proteins. Although numerous amino acid properties are presently available, many of them are essentially irrelevant with peptide structures in physiological state (e.g. amino acid parameters in advanced structures of protein and the properties tested in non-aqueous solution, etc.). Considering the CTL epitope-receptor association are completely driven by nonbonding interactions, 23 electronic properties, 37 steric properties, 54 hydrophobic properties and 5 hydrogen bond properties (see Supplementary Material) are manually chosen from hundreds of natural amino acid properties from database (Kawashima et al. 1999a, b; Lu et al. 2007) and literatures (Kidera et al. 1985; Hellberg et al. 1987; Collantes and Dunn 1995; Sandberg et al. 1998; Zhou et al. 2007). These selected parameters mainly include the following information: (1) electronic properties, e.g. net charge, molecular polarity, localized electrostatic effect and polarizability, etc.; (2) steric properties, e.g. normalized van der Waals volume, molecular size, residual side-chain volume, graph shape index and flexibility parameter, etc.; (3) hydrophobic properties as solvation free energy, partition coefficient, retention index, hydrophobic moment, residue bury and solvent accessible surface area, etc.; (4) hydrogen bond properties, including number of hydrogen bond donors/acceptors and hydrogen bond contribution factor, etc. Note that since hydrogen bond parameters are difficultly tested by experiments, reports on single hydrogen bond parameters are lacking, so there are merely 5 hydrogen bond parameters.

First, original variable matrix was subject to autoscaling to eliminate the unit difference between different properties. Then the original variable matrix, including 23 electronic properties, 37 steric properties, 54 hydrophobic properties and 5 hydrogen bond properties, was processed by principal component analysis (PCA), with the top 4, 2, 2 and 2 significant principal component scores accounting for 74.44, 72.72, 73.78 and 77.15% variance of the original information, respectively. For each amino acid, 10 significant principal component scores were yielded by multiplying the scoring coefficient of each principal component by the original variable. When the total of the 10 scores was used as a new amino acid descriptor on behalf of the original data, the information loss was insignificant. Here, these significant principal component score vectors are called the divided physicochemical property scores (DPPS) of amino acids. Amongst, electronic property is characterized by variables V_1, V_2, V_3 and V_4; stetic property by V_5 and V_6; hydrophobic property by V_7 and V_8; and hydrogen bond by V_9 and V_{10}. DPPS values for 20 amino acids are listed in Table 1.

For steric, hydrophobic and hydrogen bond properties in DPSS, only two principal components were used to achieve more than 70% variance of the original variable matrix.

Table 1 Divided physicochemical property scores (DPPS) of amino acids

AAs	Electronic property				Steric property		Hydrophobicity		Hydrogen bond	
	V_1	V_2	V_3	V_4	V_5	V_6	V_7	V_8	V_9	V_{10}
Ala A	−1.02	−2.88	−0.56	0.36	−6.15	−1.68	0.04	−2.51	−1.94	−0.01
Arg R	1.99	4.13	−4.41	−1.02	4.78	3.04	−9.06	6.71	4.41	0.07
Asn N	−2.19	1.86	0.38	−0.13	−2.30	1.41	−5.71	−1.11	1.73	−0.19
Asp D	−6.60	3.32	1.61	0.36	−3.25	1.95	−7.36	0.14	1.24	−0.15
Cys C	0.21	1.12	3.42	−0.68	−2.27	−1.22	3.11	−2.98	−1.70	1.57
Gln Q	−0.47	1.16	−0.57	0.69	0.39	1.93	−5.46	−0.84	1.93	0.85
Glu E	−5.39	0.65	−0.98	1.39	−0.23	2.51	−6.84	−0.68	1.41	1.28
Gly G	−2.86	−5.00	−2.97	0.53	−11.45	1.89	−2.11	−3.99	−2.16	−0.76
His H	0.73	2.68	−0.66	−1.89	1.60	1.13	−1.94	−0.11	0.44	0.15
Ile I	1.91	−3.13	0.01	1.14	2.70	−4.55	8.93	0.18	−1.10	−0.76
Leu L	1.64	−2.57	0.00	1.35	2.62	−2.65	7.72	0.05	−1.03	−1.81
Lys K	2.47	1.54	−4.28	−0.86	2.77	2.06	−6.18	2.05	2.19	−1.65
Met M	1.93	−0.01	1.21	0.99	2.79	−0.56	5.33	−0.87	−0.99	−1.09
Phe F	2.68	0.84	2.22	0.71	5.02	−0.30	8.60	1.13	−1.40	−0.28
Pro P	0.45	−2.89	1.77	−5.81	−3.79	−0.61	0.70	1.21	−1.67	1.79
Ser S	−1.76	−0.19	1.06	−0.69	−5.72	0.14	−4.14	−2.42	−0.13	0.69
Thr T	−0.55	−0.66	0.13	−0.31	−2.76	−1.56	−2.46	−2.12	0.17	0.08
Trp W	3.88	1.78	1.68	2.00	9.31	0.89	7.53	4.27	−0.23	−1.42
Tyr Y	2.10	1.26	1.15	0.91	5.90	0.74	3.71	3.32	0.25	1.33
Val V	0.83	−3.02	−0.22	0.97	0.05	−4.55	5.61	−1.41	−1.44	0.30

While for electronic property, four principal components were needed to achieve an equivalent level. In fact, electronic property has many behaviors, including not only some macroscopic parameters reflecting the total charges (e.g. net charge, positive charge, negative charge, etc.), but also microcosmic parameters focusing on local electronic distributions (e.g. chemical shift, localized electrostatic effect, etc.). Indicating many interaction aspects, these electronic variables contain so much intricate information that at least four principal components are needed to get to valid interpretations. While stetic, hydrophobic and hydrogen bond properties are relatively unified, with parameter information greatly overlapped (meant high colinearity), so the PCA processing is used to deal with original variables. Figure 1 shows the scattering distributions of the top two principal components in the PCA-score space for electronic, steric, hydrophobic and hydrogen bond, respectively, for 20 natural amino acids, with different properties in different marks. Figure 1a shows the electronic property space, and the second principal component space matches well with the polar distribution trends. Polar and charged amino acids are clustered on the top of this figure, while apolar ones are below, with obvious discrimination boundary. In Fig. 1b, amino acids are marked in terms of volumes (mark "□" indicates the bulk > 80 Å^3, and mark "○" denotes the bulk < 80 Å^3. Volumes of natural amino acids are in reference to Bigelow (1967). In this figure, the first principal component space is readily interpretable as amino acid bulk characteristics: on the left, number of heavy atoms on side-chains is all less than 4; while on the right, the number is more than 5. Similarly, Fig. 1c is accountable of hydrophobicity of amino acids by the first principal component of hydrophobic score space, and from left to the right in this figure, hydrophobicity of amino acids are orderly increased, approximately in agreement with polarity distributions of amino acids. Besides, although there are only 5 hydrogen bond parameters for amino acids, Fig. 1d demonstrates these data are generally inclusive of hydrogen bond information of amino acids, and marks "□" and "○" indicate whether or not the amino acid side-chain can form hydrogen bond. Obviously, such a feature is very clear in the PCA-score plot, and validity of the 5 hydrogen bond parameters is thus confirmed.

Dataset

Human major histocompatibility complex, so-called human leukocyte antigen (HLA), plays a crucial role in cellular immunity. Among them, proteins encoded by HLA-A*0201 allele take presence in more than 50% of Caucasians (Peoples et al. 1995), and therefore, virus and

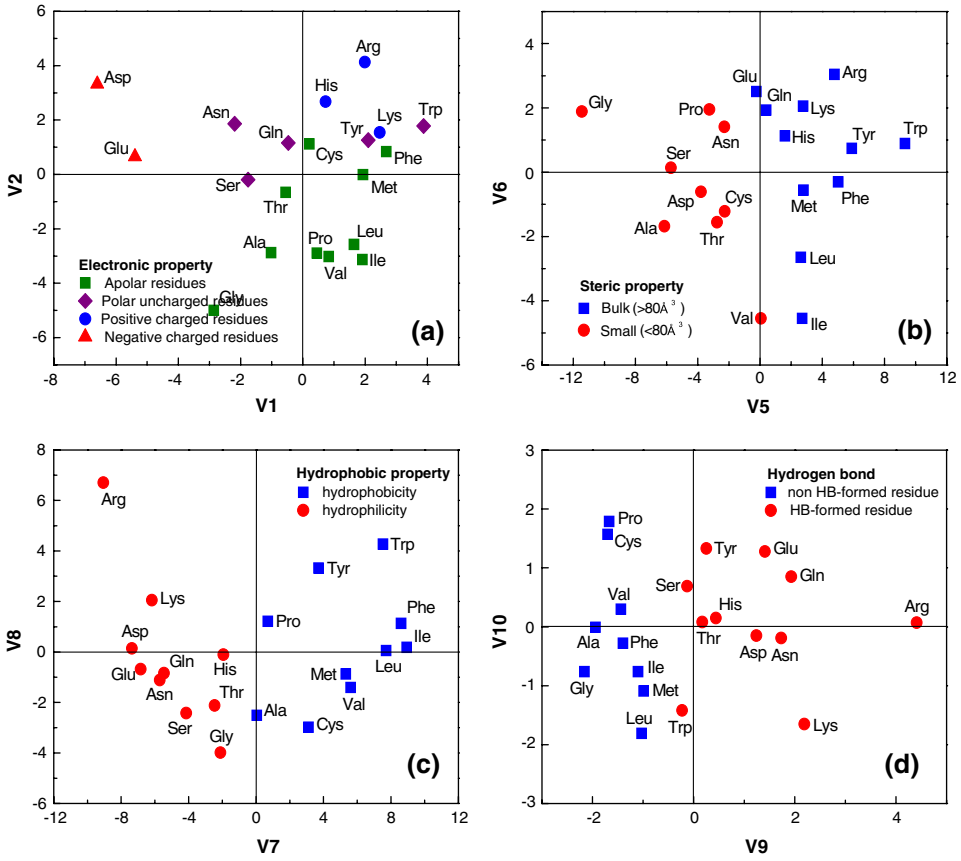

Fig. 1 PCA-score scatters of nonbonding properties of 20 amino acids. **a** Electronic property; **b** steric property; **c** hydrophobic property; and **d** hydrogen bond

tumor antigen presented by HLA-A*0201 are very valuable in vaccine design and immunotherapy. Here, the 152 HLA-A*0201-restricted CTL epitopes are all nonapeptides, with free N- and C- terminus. This group of data was selected from JenPep database (Doytchinova and Flower 2001; Blythe et al. 2002). Amino acid sequences of these antigen peptides (i.e. potential CTL epitopes) and their HLA-A*0201-binding affinities pIC_{50} ($-\log IC_{50}$) are listed in Table 2, including 65 antigen peptides of high affinities ($IC_{50} \leq 50$ nM, $pIC_{50} \geq 7.301$), 59 middle ones (50 nM $< IC_{50} \leq 500$ nM, $7.301 > pIC_{50} \geq 6.301$) and 28 low ones ($IC_{50} > 500$ nM, $pIC_{50} < 6.301$). First, the radio-labeled HBVc18227 (FLPSDYEPSV) CTL epitope of 0.5 nmol/L was served as the standard peptide. Then, test peptides of different doses were incubated 2 h at room temperature together with the standard peptide (FLPSDYFPSV)/HLA-A*0201 complex, and IC_{50} was assayed by the concentration of test peptide replacing 50% of standard peptide in the complex.

Figure 2a reveals the crystal structure of the complex of HLA-A*0201 molecule binding with antigen peptide LLFGYPVYV by X-ray diffraction at a 1.8 Å resolution (PDB entry: 1duz) (Khan et al. 2000), clearly indicating the peptide-binding cleft comprises two α-helixes and one β-sheet, and the antigen peptide embedded in is unfolding. Figure 2b shows the conformation of the antigen peptide extracted from the complex, and all the residues are in *trans*-conformation, thus leading to maximum distance along side-chains while without remarkable distortion. Hence, antigen peptides under binding state are in low-energy conformation, and binding affinity is determined upon the antigen peptide residues interacting with nearby HLA-A*0201 residues. Due to HLA-A*0201 molecule is invariant, different binding affinities are caused by different residue types and positions of antigen peptides.

Golbraikh and Tropsha (2002) claim that cross-validation q^2 has no direct relation with modeling predictabilities, and advocate the importance of the external test set. Therefore, 152 antigen peptides were randomly divided into training/test set as the number of 102/50, and then the QSAR model based on the training set was tested by test set (test samples are numbered 103–152 in Table 2). Two criteria are cooperated for the selection of test set. First, amino acid types at each position in the selection should be present at the same position of training samples; second, the selected binding affinities should be neither the maximum nor the minimum (Doytchinova and Flower 2001).

Table 2 The observed and calculated binding affinities of 152 HLA-A*0201-restricted CTL epitopes

No.	Epitopes	pIC_{50}		No.	Epitopes	pIC_{50}	
		Observed	Calculated			Observed	Calculated
Training set				77	IMDQVPFSV	7.719	7.628
1	VALVGLFVL	5.146	5.602	78	QLFEDNYAL	7.764	7.828
2	VCMTVDSLV	5.146	6.023	79	ALMDKSLHV	7.770	8.187
3	HLESLFTAV	5.301	Outlier	80	YAIDLPVSV	7.796	8.055
4	GTLVALVGL	5.342	5.600	81	FVWLHYYSV	7.824	7.595
5	LLSCLGCKI	5.447	5.955	82	MLGTHTMEV	7.845	7.201
6	LQTTIHDII	5.501	5.169	83	LLFGYPVYV	7.886	8.306
7	TLLVVMGTL	5.580	6.248	84	ILKEPVHGV	7.921	8.021
8	AMFQDPQER	5.740	5.511	85	YLMPGPVTV	7.932	7.757
9	SLHVGTQCA	5.842	6.209	86	WLDQVPFSV	7.939	8.256
10	ALPYWNFAT	5.869	6.219	87	KTWGQYWQV	7.955	Outlier
11	SLNFMGYVI	5.881	5.712	88	ALMPLYACI	8.000	6.862
12	NLQSLTNLL	6.000	6.556	89	YLAPGPVTA	8.032	7.658
13	FVTWHRYHL	6.025	6.270	90	YLYPGPVTV	8.051	7.756
14	WLEPGPVTA	6.082	Outlier	91	LLMGTLGIV	8.097	7.550
15	QVMSLHNLV	6.170	6.853	92	YLWPGPVTV	8.125	8.064
16	DPKVKQWPL	6.176	5.996	93	FLLTRILTI	8.149	8.190
17	ITSQVPFSV	6.196	6.781	94	GLLGWSPQA	8.237	8.144
18	ALAKAAAAI	6.211	6.309	95	ILYQVPFSV	8.310	8.308
19	GLGQVPLIV	6.301	6.978	96	GILTVILGV	8.347	7.853
20	MLDLQPETT	6.335	6.331	97	NMVPFFPPV	8.398	7.768
21	LLSSNLSWL	6.342	6.556	98	ILDQVPFSV	8.481	7.825
22	GLACHQLCA	6.380	6.806	99	YLFPGPVTA	8.495	8.201
23	LIGNESFAL	6.415	6.983	100	YLDQVPFSV	8.638	8.156
24	ALAKAAAAV	6.419	6.715	101	ILFQVPFSV	8.699	8.429
25	LLAVGATKV	6.477	6.189	102	ILWQVPFSV	8.770	8.616
26	KLPQLCTEL	6.484	6.422	Test set			
27	ALAKAAAAL	6.511	6.360	103	LLGCAANWI	5.301	5.567
28	WILRGTSFV	6.556	7.577	104	SAANDPIFV	5.342	5.924
29	IISCTCPTV	6.580	6.819	105	TTAEEAAGI	5.380	5.966
30	FLGGTPVCL	6.623	7.201	106	LTVILGVLL	5.580	5.207
31	ALIHHNTHL	6.623	6.684	107	HLLVGSSGL	5.792	6.121
32	NLSWLSLDV	6.639a	6.389	108	LLVVMGTLV	5.869	6.418
33	YMIMVKCWM	6.663	6.519	109	GIGILTVIL	6.000	5.466
34	VLQAGFFLL	6.682	6.451	110	TVILGVLLL	6.072	5.901
35	GTLGIVCPI	6.714	6.434	111	WTDQVPFSV	6.145	6.805
36	VILGVLLLI	6.785	7.525	112	AIAKAAAAV	6.176	6.593
37	VTWHRYHLL	6.793	6.666	113	VLHSFTDAI	6.380	6.357
38	PLLPIFFCL	6.796	7.524	114	AAAKAAAAV	6.398	5.915
39	TLGIVCPIC	6.815	6.284	115	ILTVILGVL	6.419	6.233
40	CLTSTVQLV	6.832	6.883	116	MLLAVLYCL	6.477	6.947
41	ILLLCLIFL	6.845	6.834	117	AVAKAAAAV	6.495	6.368
42	FLCKQYLNL	6.875	7.574	118	AAGIGILTV	6.581	5.492
43	FAFRDLCIV	6.886	6.904	119	ILDEAYVMA	6.623	7.150
44	FLEPGPVTA	6.898	7.412	120	YLEPGPVTA	6.668	7.343
45	ALAKAAAAA	6.947	7.039	121	LLWFHISCL	6.682	6.961
46	LMAVVLASL	6.954	6.325	122	TLHEYMLDL	6.726	7.253

Table 2 continued

No.	Epitopes	pIC_{50}		No.	Epitopes	pIC_{50}	
		Observed	Calculated			Observed	Calculated
47	YVITTQHWL	6.983	6.735	123	TLDSQVMSL	6.793	7.146
48	LLCLIFLLV	6.996	6.816	124	HLYQGCQVV	6.832	7.025
49	HLAVIGALL	7.000	6.798	125	QLFHLCLII	6.886	6.986
50	ITAQVPFSV	7.020	6.935	126	ITDQVPFSV	6.947	6.874
51	YLEPGPVTL	7.058	6.964	127	ALCRWGLLL	7.000	7.133
52	YTDQVPFSV	7.066	7.205	128	NLGNLNVSI	7.119	6.793
53	NLYVSLLLL	7.114	6.749	129	HLYSHPIIL	7.131	7.601
54	ILHNGAYSL	7.127	7.116	130	ITFQVPFSV	7.179	7.478
55	SIISAVVGI	7.159	7.427	131	FTDQVPFSV	7.212	6.974
56	VVMGTLVAL	7.174	7.175	132	YMLDLQPET	7.310	8.490
57	YLEPGPVTI	7.187	6.913	133	RLMKQDFSV	7.342	7.855
58	GLSRYVARL	7.248	6.767	134	KLHLYSHPI	7.352	6.816
59	LLAQFTSAI	7.301	6.897	135	ITMQVPFSV	7.398	7.358
60	VLLDYQGML	7.328	7.416	136	YMNGTMSQV	7.398	7.127
61	YLEPGPVTV	7.342	7.319	137	KIFGSLAFL	7.478	7.685
62	ILSPFMPLL	7.347	6.789	138	ALVGLFVLL	7.585	7.059
63	YLSPGPVTA	7.383	7.504	139	YLSPGPVTV	7.642	7.180
64	ALAKAAAAM	7.398	6.521	140	GLYSSTVPV	7.699	7.544
65	IIDQVPFSV	7.398	7.904	141	YLYPGPVTA	7.772	8.080
66	SVYDFFVWL	7.444	7.324	142	YLAPGPVTV	7.818	7.334
67	ITWQVPFSV	7.463	7.665	143	VVLGVVFGI	7.845	7.374
68	ITYQVPFSV	7.480	7.357	144	MMWYWGPSL	7.921	7.342
69	GLYSSTVPV	7.481	7.544	145	ILAQVPFSV	7.939	7.886
70	VMGTLVALV	7.553	6.704	146	WLSLLVPFV	8.048	7.405
71	LLLCLIFLL	7.585	7.343	147	FLLSLGIHL	8.053	7.730
72	SLDDYNHLV	7.585	7.401	148	ILMQVPFSV	8.125	8.309
73	VLIQRNPQL	7.644	7.639	149	YLFPGPVTV	8.237	7.877
74	SLYADSPSV	7.658	7.438	150	YLMPGPVTA	8.367	8.081
75	RLLQETELV	7.682	7.731	151	YLWPGPVTA	8.495	8.388
76	ILSQVPFSV	7.699	7.733	152	FLDQVPFSV	8.658	7.925

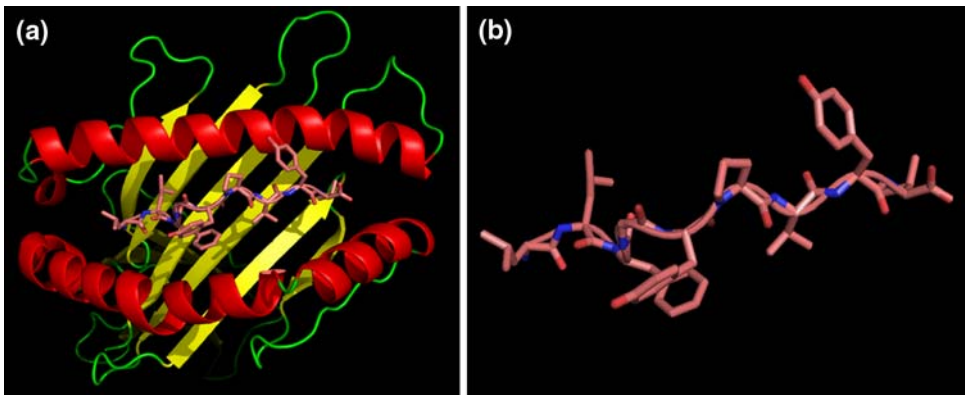

Fig. 2 **a** Crystal structure of LLFGYPVYV–HLA-A*0201 complex solved by X-crystal diffraction (PDB entry 1 duz); **b** conformation of antigen peptide extracted from the complex (this figure was produced using PyMOL, DeLano 2002)

Results and discussion

Modeling for HLA-A*0201-restricted CTL epitopes

For peptide sequences, physicochemical properties of residue at each position were characterized by 10 DPPS descriptors. When employing DPPS to characterize a nonapeptide, totally 90 variables (V_1–V_{90}) were generated. Amongs the 90 variables, V_1–V_{10} indicate 10 DPPS descriptors at position 1, V_{11}–V_{20} are orderly introduced 10 DPPS descriptors at position 2, and so forth. Because different residues and variables (even these variables indicate the same residue) contribute differently to binding affinities, variable selection is required prior to QSAR modeling, expecting to improve modeling qualities and to decrease complexity. Variable selection methods typically include orthogolization (Šoškić 1996), genetic algorithm (GA) (Rogers and Hopfinger 1994) and simulated annealing (SA) (Sutter et al. 1995), etc. Considering that the GA is a potent approach to solve the complex combinatorial optimization problems and has wide applications in large-scale of variable selections (Schefzick and Bradley 2004), and that numerous variables are yielded in DPPS characterization of nonapeptides, generic algorithm-partial least square (GA-PLS) is employed to select significant variables for the QSAR model. Based on a Matlab environment, GA-PLS toolbox as GP-toolbox was developed in our laboratory to fulfill the flexible operations of GA-PLS calculations. GA-PLS parameters settings included: (1) population size, 200; (2) genmax, 500; (3) convergence criteria, 80% of population achieving an agreement or genmax; (4) mutation rate, 1%; (5) hybridization and crossover, 2 points; (6) cross-validation, leave-1/5-out (random grouping); and (7) data pretreatment, autoscaling. Consequently, the optimal variable subset including 40 variables as V_8(Var_1), V_{10}(Var_2), V_{11}(Var_3), V_{14}(Var_4), V_{15}(Var_5), V_{16}(Var_6), V_{17}(Var_7), V_{18}(Var_8), V_{19}(Var_9), V_{20}(Var_10), V_{21}(Var_11), V_{22}(Var_12), V_{24}(Var_13), V_{27}(Var_14), V_{29}(Var_15), V_{31}(Var_16), V_{32}(Var_17), V_{37}(Var_18), V_{39}(Var_19), V_{41}(Var_20), V_{42}(Var_21), V_{43}(Var_22), V_{45}(Var_23), V_{48}(Var_24), V_{52}(Var_25), V_{56}(Var_26), V_{58}(Var_27), V_{61}(Var_28), V_{62}(Var_29), V_{64}(Var_30), V_{68}(Var_31), V_{71}(Var_32), V_{79}(Var_33), V_{81}(Var_34), V_{83}(Var_35), V_{84}(Var_36), V_{87}(Var_37), V_{88}(Var_38), V_{89}(Var_39), V_{90}(Var_40) was generated. Number of principal components is 2, and fitness (referring to root mean square error of cross-validation, RMSCV) is 0.585.

Statistic analysis of QSAR model

Positions of antigen nonapeptides are orderly numbered P1–P9. In Table 3, M1 and M2 denote the PLS and GA-PLS models on the 102 training samples, respectively. Without variable selection, M1 is statistically poor; while M2, subject to variable selection, is greatly improved in both its fitting and predictive powers. So it is suggested that the 9 peptide positions and their nonbonding effects contribute to binding affinities in great difference, and much valueless information is included in the original variable set. For that, insignificant variables are filtered out to improve statistical quality. Classical anchor residue theory deems that P2 and P9 in antigen peptides are the key anchor residues, playing important roles in antigen peptide binding with MHC molecules (Falk et al. 1991). In the GA-PLS-resulted variable subset (including 40 variables), number of variables indicating positions P2 and P9 are the most: 8 and 7, respectively. Figure 3 is the plot of the observed versus calculated binding affinities for 102 training samples, indicating a linear relation, and most samples are closely dispersed along an origin-passed line forming an angle of 45, with only few outliers especially caused by samples #3, #14 and #87; among these three samples, sample #3 (HLESLFTAV) was overestimated. In the investigations of the primary structure, its anchor residues P2 (Leu) and P9 (Val) are hydrophobic, pertaining to classic anchor residues. Besides, experimental value of sample #3 may be slightly low. This can be explained to the following: (1) residue His at P1 of sample #3 may contribute remarkably, and such a case is scare in the whole dataset, thus leading to the statistical model was trained insufficiently; (2) this phenomenon may also caused by experimental error. For sample #14 (WLEPGPVTA), the case is similar to sample #3, i.e. at anchor positions P2 and P9, hydrophobic residues Leu and Ala occur, and moreover, residue Trp at P1 is amphipathic and possesses large volume, possible to exert intricate effects on binding. Sample #87 (KTWGQYWQV) was underestimated. The anchor residue P2 of sample #87 is polar Thr, with adjacent P1 and P3 also occupied by polar residues, which disobey the classical theory "the anchor residues are hydrophobic

Table 3 Statistics of models M1–M3 for the training set

No.	Number of samples (Training/test set)	Outlier	PC	r^2	q^2	RMSEE
M1[a]	102/50	0	1	0.371	0.073	–
M2	102/50	0	2	0.686	0.522	0.480
M3	99/50	3	2	0.767	0.684	0.412

[a] This model is not processed by variable selection

amino acids", thus underestimated by the model. By these analyses, samples #3, #14 and #87 were abnormally predicted due to the particularity of their structures and properties. Removing the three samples, statistical model M3 was subsequently created, and in Fig. 4, all the samples are uniformly distributed, without outliers. In Table 3, M3 was remarkably improved in contrast with M2, especially with modeling stability (q^2) increased from 0.522 (M2) to 0.684 (M3).

To ensure the reliability of M3, statistical diagnosis was further implemented for M3. Figure 5 shows scoring scatters of the 99 samples in M3 at the top two PLS principal component space (sample number here is reedited after removing three outliers), suggesting most samples are falling in the Hotelling T^2 ellipse with 95% of confidence, with only few outliers. Besides, antigen peptides in this figure from the left to right are in increasing distribution in terms of activities, indicating PLS principal component

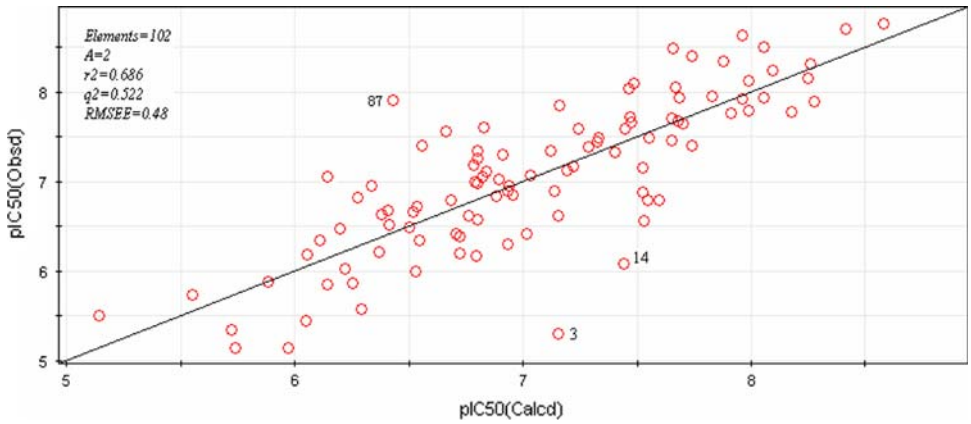

Fig. 3 The observed versus the predicted by model M2 for 102 training samples

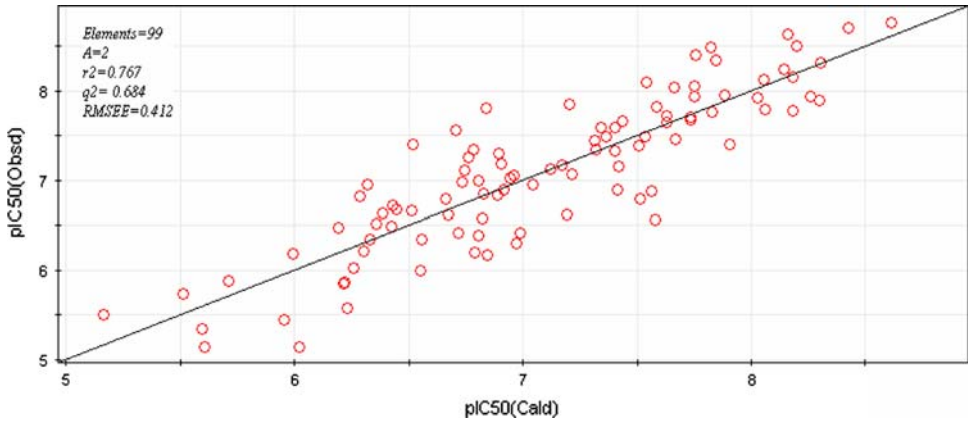

Fig. 4 The observed versus the predicted by model M3 for 99 training samples (removing three outliers)

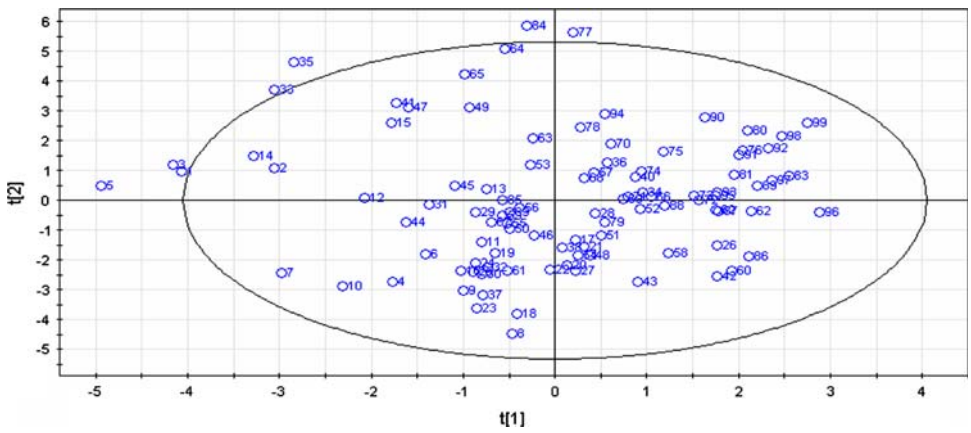

Fig. 5 Scoring scatters at the top PLS principal component space for the 99 samples in M3

spaces match well with structural and property characteristics of peptides. Subsequently, M3 was used to perform predictions on 50 test samples and Fig. 6 shows the predicted versus the observed affinities for test set, revealing correlative coefficient r^2_{pred} and the predicted root mean square error RMSEP are 0.713 and 0.457, respectively, merely generating two big errors (>1 log unit). For test set, Tropsha et al. (2003) suggested there should be a further validation. In their works, statistics referring as external correlation coefficient q^2_{pred}, $r^2_{0,\text{ext}}$, $r\prime^2_{0,\text{ext}}$and predicted slopes k, k' were proposed to evaluate the model, and following criteria were supposed to be satisfied when performing predictions on external samples.

$$\frac{r^2_{\text{pred}} - r^2_{0,\text{ext}}}{r^2_{\text{pred}}} < 0.1 \text{ or } \frac{r^2_{\text{pred}} - r\prime^2_{0,\text{ext}}}{r^2_{\text{pred}}} < 0.1 \quad (1)$$

$$0.85 \le k \le 1.15 \text{ or } 0.85 \le k\prime \le 1.15 \quad (2)$$

On the basis of such criteria, the Tropsha's statistics of M3 was calculated for test set, with results listed in Table 4. Statistical results of M3 satisfy Eqs. 1 and 2, and the predicted r^2_{pred}, $r^2_{0,\text{ext}}$ and $r\prime^2_{0,\text{ext}}$ on test set are approximately equivalent, suggesting that this model has a favorable unbiasedness. Therefore, M3 was deemed to pass through the rigorous statistical diagnosis, possessing excellent stability and predictability.

Comparison of models

The dataset including 152 HLA-A*0201-restricted CTL epitopes used here has been already studied using different methods. Doytchinova and Flower have employed CoMFA and CoMSIA to perform 3D-QSAR analysis for this dataset and found the CoMSIA model was more predictable than the CoMFA model. Therefore, a best model was constructed by removing four outliers in the CoMSIA model (Doytchinova and Flower 2001). Recently, Du et al. (2007) successfully used AABPP approach coupled with iterative double least square technique to study physicochemical properties of amino acids in contribution to the peptide–MHC binding. Table 5 lists the statistics of previous reports and of the DPPS-based model, and by comparison, DPPS model is superior to CoMSIA and AABPP approach. Besides, in comparison with the time-consuming 3D-QSAR (e.g. CoMFA and CoMSIA), DPPS and AABPP approaches that are based upon physicochemical properties of amino acids are more amenable to interpretations, thus feasible to an online prediction.

Analysis of residue positions for HLA-A*0201-restricted CTL epitopes

Figure 7 shows the standardized coefficients of model M3, indicating contributions of the 40 selected variables to the model. Figure 8 exhibits hydrophobic and hydrogen bond interactions for the peptide/HLA-A*0201 molecule association, and this figure was produced by mapping complex crystal structure into two-dimensional plane. The fringe and broken line denote hydrophobic interactions and hydrogen bonds, respectively. In this figure, many positions of the antigen peptide are surrounded by hydrophobic residues, while at the two ends, hydrogen bonds are more intensive. Following, we make a one-by-one discussion for

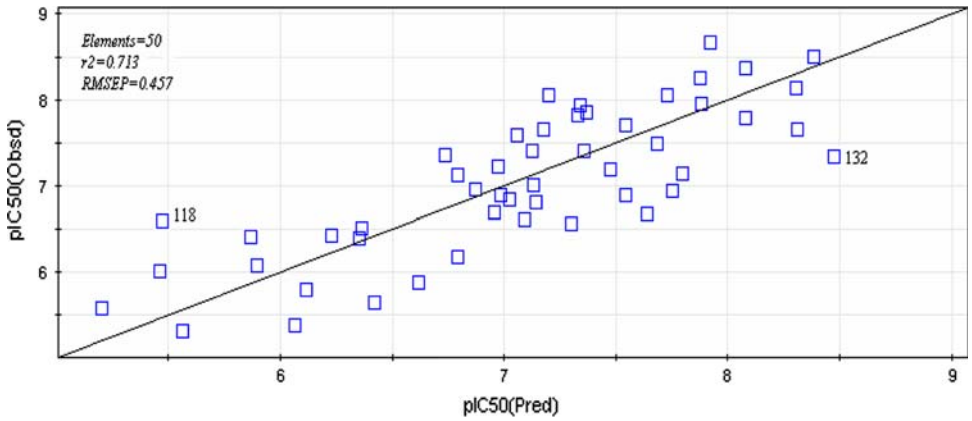

Fig. 6 The predicted versus the observed for 50 test samples by model M3

Table 4 Predicting statistics of model M3 on the test set

Statistics		Tropsha's statistics						
r^2_{pred}	RMSEP	q^2_{ext}	$r^2_{0,\text{ext}}$	$r\prime^2_{0,\text{ext}}$	$\frac{r^2_{\text{pred}} - r^2_{0,\text{ext}}}{r^2_{\text{pred}}}$	$\frac{r^2_{\text{pred}} - r\prime^2_{0,\text{ext}}}{r^2_{\text{pred}}}$	k	k'
0.713	0.457	0.685	0.714	0.712	~0.031	~0.045	0.882	0.867

Table 5 Comparison among different modeling results for 152 HLA-A*0201-restricted CTL epitopes

Method	Training set				Test set	
	Outliers	r^2	q^2	Error	r^2	Error
CoMSIA (Doytchinova and Flower 2001)	4	0.870	0.542	0.563[a]	0.679	–
AABPP (Du et al. 2007)	0	0.514	–	0.594	0.532	0.621
DPPS	3	0.767	0.684	0.412	0.713	0.457

[a] This error is obtained from leave-one-out cross-validation (LOO-CV)

positions P1–P9 on the basis of the statistical model coupled with molecule graphics and some literature reports.

P1, defined as the second anchor residue by Ruppert et al. (1993), is suggested to contribute significantly to binding affinity. In M3, although only two variables were selected at this position (referring to Var_1 and Var_2 in Fig. 7), they represent hydrophobic interactions and hydrogen bonds at P1 and have a significant contributions to the model. Sapper and Bjorkman (1991) indicate that around P1, there are several Tyr residues, forming a complex hydrogen bond network, thus accommodating Phe and Tyr that can form π–π conjugations and hydrogen bonds with receptor. The receptor residue Tyr159, a hydrogen bond donor, interacts with carboxyl oxygen at this position (Madden 1995). Figure 8 illustrates a list of hydrogen bond and hydrophobic interactions occurred at this position. Based upon these analyses, very complex nonbonding interactions are inferred at P1, with the dominance taken by hydrogen bonds and hydrophobic interactions.

P2 is a classical anchor residue, and Falk et al. (1991) point out that HLA-A*0201 molecule forms a hydrophobic pocket at this position, thus attracting strongly hydrophobic amino acids at corresponding position of the antigen peptide. Residue type of P2 is typically restricted to Leu and allows conserved replacements by Val and Ile of which both the side-chains are occupied by hydrocarbonyl group. Around P2 in Fig. 8, there are a great deal of hydrophobic resides forming a few hydrogen bonds. In correspondence, the DPPS model here introduces four nonbonding interaction terms.

P3 also belongs to the secondary anchor residues. Madden and Sarobe suggested that the aromatic and hydrophobic residues, such as Phe and Tyr, are preferred at this position (Madden et al. 1993; Sarobe et al. 1998), are similar to P1. The DPPS model here introduces six variables (referring as Var_10–Var_15 in Fig. 7) indicating four nonbonding effects. Figures 7 and 8 give consistent conclusions that P3 is partially surrounded by hydrophobic residues with its backbone forming two hydrogen bonds, suggesting the hydrophobic π–π stacking and hydrogen bonds are the main nonbonding interactions at P3.

P4 directly contacts with polar Arg65 and Lys66, thereby forming a small hydrophilic region. Kirksey et al. (1999) found that a hydrogen bond between P1 and P4 could be mediated by a water molecule in some cases, in favor of the peptide/HLA-A*0201 association and following recognition by T-cell receptor (TCR). Moreover, in the PLS coefficient plot (Fig. 7), negative charges at P4 are indicated to enhance the binding affinity (Var_16).

P5, a less important position, is far away from receptor (there is only one receptor residue Gln155 (Zhou et al. 2007)) and has few contributions. P5 is insignificant in the PLS model, and in Fig. 8, P5 is revealed to scarcely

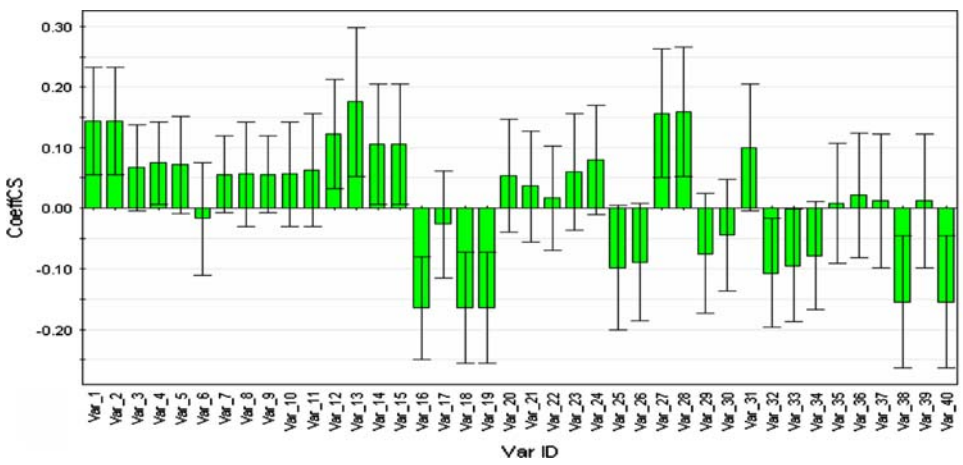

Fig. 7 Standardized coefficients of model M3

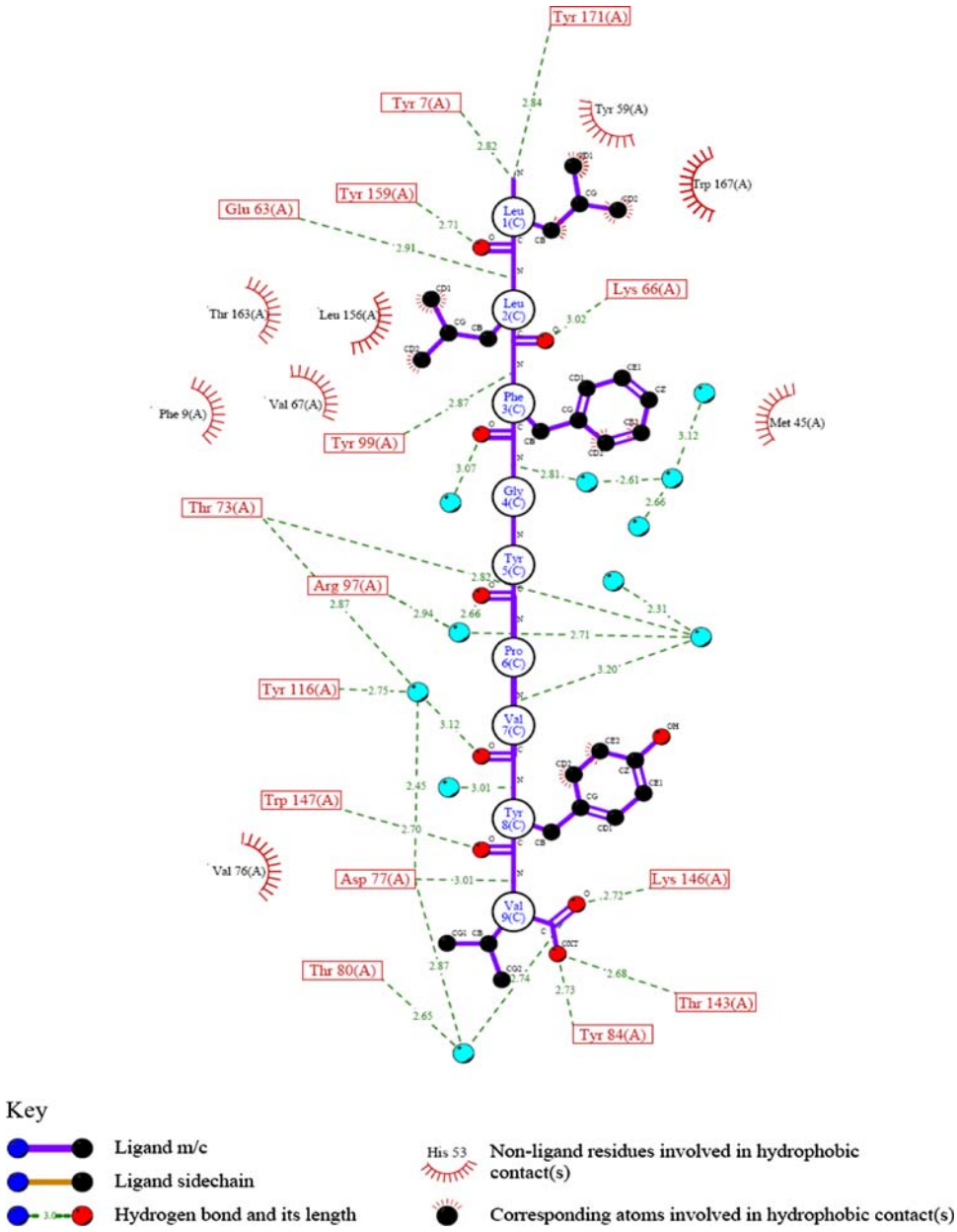

Fig. 8 A schematic diagram of the distributions of hydrophobicity and hydrogen bonds for HLA-A*0201–peptide complex (PDB entry 1 duz, produced using LIGPLOT v4.0, Wallace et al. 1995)

interact with surrounding environments. So P5 is deemed as a non-anchor residue.

Similar to P5, P6 is differed by that P6 is in a polar pocket formed by receptor residues His70, Thr73, His74 and Arg97 (Sapper and Bjorkman 1991), thus affected by electrostatic potential to some extent. Such a case is reflected in M3, e.g. Var_25 indicating electronic property has remarkable contributions. Moreover, steric term Var_27 was also introduced in M3, suggesting the van der Waals interactions also has some contributions at P6. Besides, a hydrogen bond formed between P6 and Arg97 mediated by a water molecule is given in Fig. 8.

P7, although served as the secondary anchor residue, has fewer contributions to binding in comparison with P1 and P3, so the effects of P7 are more insignificant. Sapper et al. suggested that most regions around P7 are overlapped by hydrophobic residues, and meanwhile one of its side is occupied by strongly polar Arg97 (Sapper and Bjorkman 1991), so P7 is amphipathic, suitable to occur with small hydrophobic side-chains of those residues (e.g. Val and Ala). In the PLS coefficients plot (Fig. 7), electronic and hydrophobic terms at P7 have a significant contribution, which is consistent with the case in Fig. 8 that such a position is in the juncture of positive and negative potential.

Carboxyl oxygen at the backbone of P8 forms a stable hydrogen bond with Trp147 (Madden 1995), with other nonbonding interactions not intensive at this position, so the model merely introduces a hydrogen bond term (Var_33) at P8; and P8 is deemed to be insignificant, due to its loose contact with receptors.

P9, a classical anchor residues at the C-terminal of antigen peptides, is usually occupied by Leu and Val, meanwhile allowing for conserved replacement by some similar hydrophobic resides as Ile and Ala (Falk et al. 1991). P9 is in a negative potential region, but due to the occurrence of protonated Lys146 (Madden 1995), its top is partially exposed to positive potential. In the M3, many kinds of nonbonding interactions are introduced at P9 (referring as Var_34–Var_40 in Fig. 7), with hydrophobic interactions and hydrogen bonds the most intensive (indicated by Var_38 and Var_40 in Fig. 7). We believe the hydrogen bonds of P9 are essential to the binding specificity, and hydrophobicities ensure strong binding affinity.

Conclusions

Computer-aided vaccine design (CAVD) is a newly emerging field. CAVD involves in multiple subjects as immunology, molecular biology, drug design, bioinformatics, and computational science. However immunology, one of its fundamental subjects, is presently an archetypal experimental science (Hagmann 2000), and CAVD is essentially in its early stage, with relevant theories and algorithms immature. Epitope identification and prediction, as a research branch of CAVD, are pivotal in developing new vaccines and performing immunotherapy, thus driving considerable interests. However, many epitope predicting methods are based upon time-consuming bioinformatics and machine learning algorithms. In this study, physicochemical property scores (DPPS) of amino acids, involving hydrophobicity, steric properties, electronic properties and hydrogen bond properties, were proposed to characterize peptide structures that bind with the MHC molecule. Parameters responsible for the binding affinity were selected by genetic algorithm, and a quantitative structure–affinity relationship (QSAR) model was established by partial least square (PLS). The results showed that different properties of the residues in nonapeptide (P1P2P3P4P5P6P7P8P9) may contribute distinctly to the binding. Particularly, hydrophobicity and hydrogen bond of anchor residues are more significant during the association process.

Acknowledgments This work was supported by National Project 863 Fund (grant number 2006AA02Z312) and the National Natural Science Fund (grant number 30371339 and 30571748). We thank Prof. Zhiliang Li for commenting on this manuscript.

Appendix

Table 6 The names of 119 physicochemical properties for amino acids

No.	Physicochemical property
	Electronic property (No. 1–23)
1	alpha-NH chemical shifts (Bundi-Wuthrich 1979)
2	A parameter of charge transfer capability (Charton-Charton 1983)
3	A parameter of charge transfer donor capability (Charton-Charton 1983)
4	Localized electrical effect (Fauchere et al. 1988)
5	Positive charge (Fauchere et al. 1988)
6	Negative charge (Fauchere et al. 1988)
7	Polarity (Grantham 1974)
8	Net charge (Klein et al. 1984)
9	Mean polarity (Radzicka-Wolfenden 1988)
10	a-CH chemical shifts (Andersen et al. 1992)
11	Electron-ion interaction potential (Veljkovic et al. 1985)
12	Polar requirement (Woese 1973)
13	Polarity (Zimmerman et al. 1968)
14	Isoelectric point (Zimmerman et al. 1968)
15	a-CH chemical shifts (Bundi-Wuthrich 1979)
16	Polarizability parameter (Charton-Charton 1982)
17	Dependence of partition coefficient on ionic strength (Zaslavsky et al. 1982)
18	N.m.r. chemical shift of alpha-carbon (Fauchere et al. 1988)
19	pK-N (Fasman 1976)
20	pK-a(RCOOH) (Fauchere et al. 1988)
21	pK-C (Fasman 1976)
22	Principal property value z3 (Wold et al. 1987)
23	Side-chain electronic charge index (Collantes and Dunn 1995)
	Steric Property (No. 24–60)
24	Flexibility parameter for two rigid neighbors (Karplus-Schulz 1985)
25	Residue volume (Bigelow 1967)
26	Average volume of buried residue (Chothia 1975)
27	Apparent partial specific volume (Bull-Breese 1974)
28	The number of atoms in the side-chain labeled 1 + 1 (Charton-Charton 1983)
29	The number of atoms in the side-chain labeled 2 + 1 (Charton-Charton 1983)
30	The number of atoms in the side-chain labeled 3 + 1 (Charton-Charton 1983)
31	The number of bonds in the longest chain (Charton-Charton 1983)
32	Partial specific volume (Cohn-Edsall 1943)
33	Size (Dawson 1972)
34	Molecular weight (Fasman 1976)
35	Graph shape index (Fauchere et al. 1988)
36	Normalized van der Waals volume (Fauchere et al. 1988)
37	STERIMOL length of the side-chain (Fauchere et al. 1988)
38	STERIMOL minimum width of the side-chain (Fauchere et al. 1988)
39	STERIMOL maximum width of the side-chain (Fauchere et al. 1988)

Table 6 continued

No.	Physicochemical property
40	Residue volume (Goldsack-Chalifoux 1973)
41	Volume (Grantham 1974)
42	Side-chain volume (Krigbaum-Komoriya 1979)
43	Refractivity (McMeekin et al. 1964)
44	Distance between Ca and centroid of side-chain (Levitt 1976)
45	Radius of gyration of side-chain (Levitt 1976)
46	van der Waals parameter R0 (Levitt 1976)
47	van der Waals parameter e (Levitt 1976)
48	Bulkiness (Zimmerman et al. 1968)
49	Steric parameter (Charton 1981)
50	Smoothed upsilon steric parameter (Fauchere et al. 1988)
51	Side-chain angle theta(AAR) (Levitt 1976)
52	Side-chain torsion angle phi(AAAR) (Levitt 1976)
53	Residue accessible surface area in tripeptide (Chothia 1976)
54	Relative mutability (Dayhoff et al. 1978a)
55	Optimized propensity to form reverse turn (Oobatake et al. 1985)
56	Accessible surface area (Radzicka-Wolfenden 1988)
57	Average flexibility indices (Bhaskaran-Ponnuswamy 1988)
58	Flexibility parameter for no rigid neighbors (Karplus-Schulz 1985)
59	Flexibility parameter for one rigid neighbor (Karplus-Schulz 1985)
60	Residue Side-Chain Volume (Zhou et al. 2006)
	Hydrophobic property (No. 61–114)
61	Hydrophobicity index (Argos et al. 1982)
62	Retention coefficient in TFA (Browne et al. 1982)
63	Retention coefficient in HFBA (Browne et al. 1982)
64	Free energy of solution in water, kcal/mole (Charton-Charton 1982)
65	Normalized hydrophobicity scales for alpha-proteins (Cid et al. 1992)
66	Normalized hydrophobicity scales for beta-proteins (Cid et al. 1992)
67	Normalized hydrophobicity scales for alpha + beta-proteins (Cid et al. 1992)
68	Normalized hydrophobicity scales for alpha/beta-proteins (Cid et al. 1992)
69	Normalized average hydrophobicity scales (Cid et al. 1992)
70	Consensus normalized hydrophobicity scale (Eisenberg 1984)
71	Solvation free energy (Eisenberg-McLachlan 1986)
72	Atom-based hydrophobic moment (Eisenberg-McLachlan 1986)
73	Direction of hydrophobic moment (Eisenberg-McLachlan 1986)
74	Hydrophobic parameter p (Fauchere-Pliska 1983)
75	Hydrophobicity factor (Goldsack-Chalifoux 1973)
76	Hydration number (Hopfinger 1971)
77	Hydrophilicity value (Hopp-Woods 1981)
78	Hydrophobicity (Jones 1975)
79	Fraction of site occupied by water (Krigbaum-Komoriya 1979)
80	Hydropathy index (Kyte-Doolittle 1982)
81	Transfer free energy, CHP/water (Lawson et al. 1984)

Table 6 continued

No.	Physicochemical property
82	Hydrophobic parameter (Levitt 1976)
83	Retention coefficient in HPLC, pH7.4 (Meek 1980)
84	Retention coefficient in HPLC, pH2.1 (Meek 1980)
85	Retention coefficient in NaClO4 (Meek-Rossetti 1981)
86	Retention coefficient in NaH2PO4 (Meek-Rossetti 1981)
87	Transfer energy, organic solvent/water (Nozaki-Tanford 1971)
88	Optimized transfer energy parameter (Oobatake et al. 1985)
89	HPLC parameter (Parker et al. 1986)
90	Partition coefficient (Pliska et al. 1981)
91	Surrounding hydrophobicity in folded form (Ponnuswamy et al. 1980)
92	Average gain in surrounding hydrophobicity (Ponnuswamy et al. 1980)
93	Average gain ratio in surrounding hydrophobicity (Ponnuswamy et al. 1980)
94	Surrounding hydrophobicity in alpha-helix (Ponnuswamy et al. 1980)
95	Surrounding hydrophobicity in beta-sheet (Ponnuswamy et al. 1980)
96	Surrounding hydrophobicity in turn (Ponnuswamy et al. 1980)
97	Hydrophobicity (Prabhakaran 1990)
98	Transfer free energy from chx to wat (Radzicka-Wolfenden 1988)
99	Transfer free energy from oct to wat (Radzicka-Wolfenden 1988)
100	Energy transfer from out to in(95%buried) (Radzicka-Wolfenden 1988)
101	Hydration free energy (Robson-Osguthorpe 1979)
102	Side-chain hydropathy, uncorrected for solvation (Roseman 1988)
103	Side-chain hydropathy, corrected for solvation (Roseman 1988)
104	Transfer free energy (Simon 1976)
105	Optimal matching hydrophobicity (Sweet-Eisenberg 1983)
106	Transfer free energy to lipophilic phase (von Heijne-Blomberg 1979)
107	Free energy change of a(Ri) to a(Rh) (Wertz-Scheraga 1978)
108	Free energy change of e(i) to a(Rh) (Wertz-Scheraga 1978)
109	Hydration potential (Wolfenden et al. 1981)
110	Unfolding Gibbs energy in water, pH7.0 (Yutani et al. 1987)
111	Unfolding Gibbs energy in water, pH9.0 (Yutani et al. 1987)
112	Activation Gibbs energy of unfolding, pH7.0 (Yutani et al. 1987)
113	Activation Gibbs energy of unfolding, pH9.0 (Yutani et al. 1987)
114	Hydrophobicity (Zimmerman et al. 1968)
	Hydrogen Bond Property (No. 115–119)
115	Number of hydrogen bond donors (Fauchere et al. 1988)
116	Number of hydrogen bond acceptors (Fauchere et al. 1988)
117	Information measure for extended without H-bond (Robson-Suzuki 1976)
118	Hydrogen bond donor factors (Zhou et al. 2006)
119	Hydrogen bond acceptor factors (Zhou et al. 2006)

Table 7 The values of 119 physicochemical properties for amino acids

No.	ala	arg	asn	asp	cys	gln	glu	gly	his	ile	leu	lys	met	phe	pro	ser	thr	trp	tyr	val
1	8.249	8.274	8.747	8.410	8.312	8.411	8.368	8.391	8.415	8.195	8.423	8.408	8.418	8.228	0.000	8.380	8.236	8.094	8.183	8.436
2	0.000	0.000	1.000	1.000	0.000	0.000	1.000	1.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
3	0.000	1.000	1.000	0.000	1.000	1.000	0.000	0.000	1.000	0.000	0.000	1.000	1.000	1.000	0.000	0.000	0.000	1.000	1.000	0.000
4	−0.01	0.040	0.060	0.150	0.120	0.050	0.070	0.000	0.080	−0.01	−0.01	0.000	0.040	0.030	0.000	0.110	0.040	0.000	0.030	0.01
5	0.000	1.000	0.000	0.000	0.000	0.000	0.000	0.000	1.000	0.000	0.000	1.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
6	0.000	0.000	0.000	1.000	0.000	0.000	1.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
7	8.100	10.500	11.600	13.000	5.500	10.500	12.300	9.000	10.400	5.200	4.900	11.300	5.700	5.200	8.000	9.200	8.600	5.400	6.200	5.900
8	0.000	1.000	0.000	−1.000	0.000	0.000	−1.000	0.000	0.000	0.000	0.000	1.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
9	−0.060	−0.840	−0.480	−0.800	1.360	−0.730	−0.770	−0.410	0.490	1.310	1.210	−1.180	1.270	1.270	0.000	−0.500	−0.270	0.880	0.330	1.090
10	0.037	0.096	0.004	0.126	0.083	0.076	0.006	0.005	0.024	0.000	0.000	0.037	0.082	0.095	0.020	0.083	0.094	0.055	0.052	0.006
11	7.000	9.100	10.000	13.000	5.500	8.600	12.500	7.900	8.400	4.900	4.900	10.100	5.300	5.000	6.600	7.500	6.600	5.300	5.700	5.600
12	0.000	52.000	3.380	49.700	1.480	3.530	49.900	0.000	51.600	0.130	0.130	49.500	1.430	0.350	1.580	1.670	1.660	2.100	1.610	0.130
13	6.000	10.760	5.410	2.770	5.050	5.650	3.220	5.970	7.590	6.020	5.980	9.740	5.740	5.480	6.300	5.680	5.660	5.890	5.660	5.960
14	4.350	4.380	4.750	4.760	4.650	4.370	4.290	3.970	4.630	3.950	4.170	4.360	4.520	4.660	4.440	4.500	4.350	4.700	4.600	3.950
15	4.349	4.396	4.755	4.765	4.686	4.373	4.295	3.972	4.630	4.224	4.385	4.358	4.513	4.663	4.471	4.498	4.346	4.702	4.604	4.184
16	0.046	0.291	0.134	0.105	0.128	0.180	0.151	0.000	0.230	0.186	0.186	0.219	0.221	0.290	0.131	0.062	0.108	0.409	0.298	0.140
17	−0.152	−0.089	−0.203	−0.355	0.000	−0.181	−0.411	−0.190	0.000	−0.086	−0.102	−0.062	−0.107	0.001	−0.181	−0.203	−0.170	0.275	0.000	−0.125
18	7.300	11.100	8.000	9.200	14.400	10.600	11.400	0.000	10.200	16.100	10.100	10.900	10.400	13.900	17.800	13.100	16.700	13.200	13.900	17.200
19	9.690	8.990	8.800	9.600	8.350	9.130	9.670	9.780	9.170	9.680	9.600	9.180	9.210	9.180	10.640	9.210	9.100	9.440	9.110	9.620
20	4.760	4.300	3.640	5.690	3.670	4.540	5.480	3.770	2.840	4.810	4.790	4.270	4.250	4.310	0.000	3.830	3.870	4.750	4.300	4.860
21	2.340	1.820	2.020	1.880	1.920	2.170	2.100	2.350	1.820	2.360	2.360	2.160	2.280	2.160	1.950	2.190	2.090	2.430	2.200	2.320
22	0.090	−3.440	0.840	2.360	4.130	−1.140	−0.070	0.300	1.110	−1.030	−0.980	−3.140	−0.410	0.450	2.230	0.570	−1.400	0.850	0.01	−1.290
23	0.05	1.69	1.31	1.25	0.15	1.36	1.31	0.02	0.56	0.09	0.10	0.53	0.34	0.14	0.16	0.56	0.65	1.08	0.72	0.07
24	0.892	0.901	0.930	0.932	0.925	0.885	0.933	0.923	0.894	0.872	0.921	1.057	0.804	0.914	0.932	0.923	0.934	0.803	0.837	0.913
25	52.600	109.100	75.700	68.400	68.300	89.700	84.700	36.300	91.900	102.000	102.000	105.100	97.700	113.900	73.600	54.900	71.200	135.400	116.200	85.100
26	91.500	202.000	135.200	124.500	117.700	161.100	155.100	66.400	167.300	168.800	167.900	171.300	170.800	203.400	129.300	99.100	122.100	237.600	203.600	141.700
27	0.691	0.728	0.596	0.558	0.624	0.649	0.632	0.592	0.646	0.809	0.842	0.767	0.709	0.756	0.730	0.594	0.655	0.743	0.743	0.777
28	0.000	1.000	1.000	1.000	1.000	1.000	1.000	0.000	1.000	2.000	1.000	1.000	1.000	1.000	0.000	1.000	2.000	1.000	1.000	2.000
29	0.000	1.000	1.000	1.000	0.000	1.000	1.000	0.000	1.000	1.000	2.000	1.000	1.000	1.000	0.000	0.000	0.000	1.000	1.000	0.000
30	0.000	1.000	0.000	0.000	0.000	1.000	1.000	0.000	1.000	0.000	0.000	1.000	1.000	1.000	0.000	0.000	0.000	1.500	1.000	0.000
31	0.000	5.000	2.000	2.000	1.000	3.000	3.000	0.000	3.000	2.000	2.000	4.000	3.000	4.000	0.000	1.000	1.000	5.000	5.000	1.000
32	0.750	0.700	0.610	0.600	0.610	0.670	0.660	0.640	0.670	0.900	0.900	0.820	0.750	0.770	0.760	0.680	0.700	0.740	0.710	0.860
33	2.500	7.500	5.000	2.500	3.000	6.000	5.000	0.500	6.000	5.500	5.500	7.000	6.000	6.500	5.500	3.000	5.000	7.000	7.000	5.000
34	89.090	174.200	132.120	133.100	121.150	146.150	147.130	75.070	155.160	131.170	131.170	146.190	149.210	165.190	115.130	105.090	119.120	204.240	181.190	117.150
35	1.280	2.340	1.600	1.600	1.770	1.560	1.560	0.000	2.990	4.190	2.590	1.890	2.350	2.940	2.670	1.310	3.030	3.210	2.940	3.670
36	1.000	6.130	2.950	2.780	2.430	3.950	3.780	0.000	4.660	4.000	4.000	4.770	4.430	5.890	2.720	1.600	2.600	8.080	6.470	3.000
37	2.870	7.820	4.580	4.740	4.470	6.110	5.970	2.060	5.230	4.920	4.920	6.890	6.360	4.620	4.110	3.970	4.110	7.680	4.730	4.110
38	1.520	1.520	1.520	1.520	1.520	1.520	1.520	1.000	1.520	1.900	1.520	1.520	1.520	1.520	1.520	1.520	1.730	1.520	1.520	1.900

Table 7 continued

No.	ala	arg	asn	asp	cys	gln	glu	gly	his	ile	leu	lys	met	phe	pro	ser	thr	trp	tyr	val
39	2.040	6.240	4.370	3.780	3.410	3.530	3.310	1.000	5.660	3.490	4.450	4.870	4.800	6.020	4.310	2.700	3.170	5.900	6.720	3.170
40	88.300	181.200	125.100	110.800	112.400	148.700	140.500	60.000	152.600	168.500	168.500	175.600	162.200	189.000	122.200	88.700	118.200	227.000	193.000	141.400
41	31.000	124.000	56.000	54.000	55.000	85.000	83.000	3.000	96.000	111.000	111.000	119.000	105.000	132.000	32.500	32.000	61.000	170.000	136.000	84.000
42	27.500	105.000	58.700	40.000	44.600	80.700	62.000	0.000	79.000	93.500	93.500	100.000	94.100	115.500	41.900	29.300	51.300	145.500	117.300	71.500
43	4.340	26.660	13.280	12.000	35.770	17.560	17.260	0.000	21.810	19.060	18.780	21.290	21.640	29.400	10.930	6.350	11.010	42.530	31.530	13.920
44	0.770	3.720	1.980	1.990	1.380	2.580	2.630	0.000	2.760	1.830	2.080	2.940	2.340	2.970	1.420	1.280	1.430	3.580	3.360	1.490
45	0.770	2.380	1.450	1.430	1.220	1.750	1.770	0.580	1.780	1.560	1.540	2.080	1.800	1.900	1.250	1.080	1.240	2.210	2.130	1.290
46	5.200	6.000	5.000	5.000	6.100	6.000	6.000	4.200	6.000	7.000	7.000	6.000	6.800	7.100	6.200	4.900	5.000	7.600	7.100	6.400
47	0.025	0.200	0.100	0.100	0.100	0.100	0.100	0.025	0.100	0.190	0.190	0.200	0.190	0.390	0.170	0.025	0.100	0.560	0.390	0.150
48	11.500	14.280	12.820	11.680	13.460	14.450	13.570	3.400	13.690	21.400	21.400	15.710	16.250	19.800	17.430	9.470	15.770	21.670	18.030	21.570
49	0.520	0.680	0.760	0.760	0.620	0.680	0.680	0.000	0.700	1.020	0.980	0.680	0.780	0.700	0.360	0.530	0.500	0.700	0.700	0.760
50	0.530	0.690	0.580	0.590	0.660	0.710	0.720	0.000	0.640	0.960	0.920	0.780	0.770	0.710	0.000	0.550	0.630	0.840	0.710	0.890
51	121.900	121.400	117.500	121.200	113.700	118.000	118.200	0.000	118.200	118.900	118.100	122.000	113.100	118.200	81.900	117.900	117.100	118.400	110.000	121.700
52	243.200	206.600	207.100	215.000	209.400	205.400	213.600	300.000	219.900	217.900	205.600	210.900	204.000	203.700	237.400	232.000	226.700	203.700	195.600	220.300
53	115.000	225.000	160.000	150.000	135.000	180.000	190.000	75.000	195.000	175.000	170.000	200.000	185.000	210.000	145.000	115.000	140.000	255.000	230.000	155.000
54	100.000	65.000	134.000	106.000	20.000	93.000	102.000	49.000	66.000	96.000	40.000	56.000	94.000	41.000	56.000	120.000	97.000	18.000	41.000	74.000
55	1.340	0.950	2.490	3.320	1.070	1.490	2.200	2.070	1.270	0.660	0.540	0.610	0.700	0.800	2.120	0.940	1.090	−4.650	−0.170	1.320
56	93.700	250.400	146.300	142.600	135.200	177.700	182.900	52.600	188.100	182.200	173.700	215.200	197.600	228.600	0.000	109.500	142.100	271.600	239.900	157.200
57	7.000	9.100	10.000	13.000	5.500	8.600	12.500	7.900	8.400	4.900	4.900	10.100	5.300	5.000	6.600	7.500	6.600	5.300	5.700	5.600
58	1.041	1.038	1.117	1.033	0.960	1.165	1.094	1.142	0.982	1.002	0.967	1.093	0.947	0.930	1.055	1.169	1.073	0.925	0.961	0.982
59	0.946	1.028	1.006	1.089	0.878	1.025	1.036	1.042	0.952	0.892	0.961	1.082	0.862	0.912	1.085	1.048	1.051	0.917	0.930	0.927
60	13.7	77.3	32.7	30	25	42.7	40.2	3.5	45.1	44.4	44.4	61.5	45	56.1	30.7	18.3	28.5	74.8	59.1	34.1
61	0.61	0.60	0.06	0.46	1.07	0.00	0.47	0.07	0.61	2.22	1.53	1.15	1.18	2.02	1.95	0.05	0.05	2.65	1.88	1.32
62	7.30	−3.60	−5.70	−2.90	−9.20	−0.30	−7.10	−1.20	−2.10	6.60	20.00	−3.70	5.60	19.20	5.10	−4.10	0.80	16.30	5.90	3.50
63	3.90	3.20	−2.80	−2.80	−14.30	1.80	−7.50	−2.30	2.00	11.00	15.00	−2.50	4.10	14.70	5.60	−3.50	1.10	17.80	3.80	2.10
64	−0.37	−1.03	0.00	2.06	4.53	0.73	1.77	−0.52	0.00	0.79	1.07	0.00	0.66	1.06	−2.24	−0.52	0.00	1.60	4.91	0.40
65	−0.45	−0.24	−0.20	−1.52	0.79	−0.99	−0.80	−1.00	1.07	0.76	1.29	−0.36	1.37	1.48	−0.12	−0.98	−0.70	1.38	1.49	1.26
66	−0.08	−0.09	−0.70	−0.71	0.76	−0.40	−1.31	−0.84	0.43	1.39	1.24	−0.09	1.27	1.53	−0.01	−0.93	−0.59	2.25	1.53	1.09
67	0.36	−0.52	−0.90	−1.09	0.70	−1.05	−0.83	−0.82	0.16	2.17	1.18	−0.56	1.21	1.01	−0.06	−0.60	−1.20	1.31	1.05	1.21
68	0.17	−0.70	−0.90	−1.05	1.24	−1.20	−1.19	−0.57	−0.25	2.06	0.96	−0.62	0.60	1.29	−0.21	−0.83	−0.62	1.51	0.66	1.21
69	0.02	−0.42	−0.77	−1.04	0.77	−1.10	−1.14	−0.80	0.26	1.81	1.14	−0.41	1.00	1.35	−0.09	−0.97	−0.77	1.71	1.11	1.13
70	0.25	−1.76	−0.64	−0.72	0.04	−0.69	−0.62	0.16	−0.40	0.73	0.53	−1.10	0.26	0.61	−0.07	−0.26	−0.18	0.37	0.02	0.54
71	0.67	−2.10	−0.60	−1.20	0.38	−0.22	−0.76	0.00	0.64	1.90	1.90	−0.57	2.40	2.30	1.20	0.01	0.52	2.60	1.60	1.50
72	0.00	10.00	1.30	1.90	0.17	1.90	3.00	0.00	0.99	1.20	1.00	5.70	1.90	1.10	0.18	0.73	1.50	1.60	1.80	0.48
73	0.00	−0.96	−0.86	−0.98	0.76	−1.00	−0.89	0.00	−0.75	0.99	0.89	−0.99	0.94	0.92	0.22	−0.67	0.09	0.67	−0.93	0.84
74	0.31	−1.01	−0.60	−0.77	1.54	−0.22	−0.64	0.00	0.13	1.80	1.70	−0.99	1.23	1.79	0.72	−0.04	0.26	2.25	0.96	1.22
75	0.75	0.75	0.69	0.00	1.00	0.59	0.00	0.00	0.00	2.95	2.40	1.50	1.30	2.65	2.60	0.00	0.45	3.00	2.85	1.70
76	1.00	2.30	2.20	6.50	0.10	2.10	6.20	1.10	2.80	0.80	0.80	5.30	0.70	1.40	0.90	1.70	1.50	1.90	2.10	0.90

Table 7 continued

No.	ala	arg	asn	asp	cys	gln	glu	gly	his	ile	leu	lys	met	phe	pro	ser	thr	trp	tyr	val
77	−0.50	3.00	0.20	3.00	−1.00	0.20	3.00	0.00	−0.50	−1.80	−1.80	3.00	−1.30	−2.50	0.00	0.30	−0.40	−3.40	−2.30	−1.50
78	0.87	0.85	0.09	0.66	1.52	0.00	0.67	0.10	0.87	3.15	2.17	1.64	1.67	2.87	2.77	0.07	0.07	3.77	2.67	1.87
79	4.32	6.55	6.24	6.04	1.73	6.13	6.17	6.09	5.66	2.31	3.93	7.92	2.44	2.59	7.19	5.37	5.16	2.78	3.58	3.31
80	1.80	−4.50	−3.50	−3.50	2.50	−3.50	−3.50	−0.40	−3.20	4.50	3.80	−3.90	1.90	2.80	−1.60	−0.80	−0.70	−0.90	−1.30	4.20
81	−0.48	−0.06	−0.87	−0.75	−0.32	−0.32	−0.71	0.00	−0.51	0.81	1.02	−0.09	0.81	1.03	2.03	0.05	−0.35	0.66	1.24	0.56
82	−0.50	3.00	0.20	2.50	−1.00	0.20	2.50	0.00	−0.50	−1.80	−1.80	3.00	−1.30	−2.50	−1.40	0.30	−0.40	−3.40	−2.30	−1.50
83	0.50	0.80	0.80	−8.20	−6.80	−4.80	−16.90	0.00	−3.50	13.90	8.80	0.10	4.80	13.20	6.10	1.20	2.70	14.90	6.10	2.70
84	−0.10	−4.50	−1.60	−2.80	−2.20	−2.50	−7.50	−0.50	0.80	11.80	10.00	−3.20	7.10	13.90	8.00	−3.70	1.50	18.10	8.20	3.30
85	1.10	−0.40	−4.20	−1.60	7.10	−2.90	0.70	−0.20	−0.70	8.50	11.00	−1.90	5.40	13.40	4.40	−3.20	−1.70	17.10	7.40	5.90
86	1.00	−2.00	−3.00	−0.50	4.60	−2.00	1.10	0.20	−2.20	7.00	9.60	−3.00	4.00	12.60	3.10	−2.90	−0.60	15.10	6.70	4.60
87	0.50	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.50	1.80	1.80	0.00	1.30	2.50	0.00	0.00	0.40	3.40	2.30	1.50
88	0.46	−1.54	1.31	−0.33	0.20	−1.12	0.48	0.64	−1.31	3.28	0.43	−1.71	0.15	0.52	−0.58	−0.83	−1.52	1.25	−2.21	0.54
89	2.10	4.20	7.00	10.00	1.40	6.00	7.80	5.70	2.10	−8.00	−9.20	5.70	−4.20	−9.20	2.10	6.50	5.20	−10.00	−1.90	−3.70
90	−2.89	−3.30	−3.41	−3.38	−2.49	−3.15	−2.94	−3.25	−2.84	−1.72	−1.61	−3.31	−1.84	−1.63	−2.50	−3.30	−2.91	−1.75	−2.42	−2.08
91	12.28	11.49	11.00	10.97	14.93	11.28	11.19	12.01	12.84	14.77	14.10	10.80	14.33	13.43	11.19	11.26	11.65	12.95	13.29	15.07
92	7.62	6.81	6.17	6.18	10.93	6.67	6.38	7.31	7.85	9.99	9.37	5.72	9.83	8.99	6.64	6.93	7.08	8.41	8.53	10.38
93	2.63	2.45	2.27	2.29	3.36	2.45	2.31	2.55	2.57	3.08	2.98	2.12	3.18	3.02	2.46	2.60	2.55	2.85	2.79	3.21
94	13.65	11.28	12.24	10.98	14.49	11.30	12.55	15.36	11.59	14.63	14.01	11.96	13.40	14.08	11.51	11.26	13.00	12.06	12.64	12.88
95	14.60	13.24	11.79	13.78	15.90	12.02	13.59	14.18	15.35	14.10	16.49	13.28	16.23	14.18	14.10	13.36	14.50	13.90	14.76	16.30
96	10.67	11.05	10.85	10.21	14.15	11.71	11.71	10.95	12.07	12.95	13.07	9.93	15.00	13.27	10.62	11.18	10.53	11.41	11.52	13.86
97	−6.70	51.50	20.10	38.50	−8.40	17.20	34.30	−4.20	12.60	−13.00	−11.70	36.80	−14.20	−15.50	0.80	−2.50	−5.00	−7.90	2.90	−10.90
98	1.81	−14.92	−6.64	−8.72	1.28	−5.54	−6.81	0.94	−4.66	4.92	4.92	−5.55	2.35	2.98	0.00	−3.40	−2.57	2.33	−0.14	4.04
99	0.52	−1.32	−0.01	0.00	0.00	−0.07	−0.79	0.00	0.95	2.04	1.76	0.08	1.32	2.09	0.00	0.04	0.27	2.51	1.63	1.18
100	−0.29	−2.71	−1.18	−1.02	0.00	−1.53	−0.90	−0.34	−0.94	0.24	−0.12	−2.05	−0.24	0.00	0.00	−0.75	−0.71	−0.59	−1.02	0.09
101	−1.00	0.30	−0.70	−1.20	2.10	−0.10	−0.70	0.30	1.10	4.00	2.00	−0.90	1.80	2.80	0.40	−1.20	−0.50	3.00	2.10	1.40
102	−0.67	12.10	7.23	8.72	−0.34	6.39	7.35	0.00	3.82	−3.02	−3.02	6.13	−1.30	−3.24	−1.75	4.35	3.86	−2.86	0.98	−2.18
103	−0.67	3.89	2.27	1.57	−2.00	2.12	1.78	0.00	1.09	−3.02	−3.02	2.46	−1.67	−3.24	−1.75	0.10	−0.42	−2.86	0.98	−2.18
104	0.73	0.73	−0.01	0.54	0.70	−0.10	0.55	0.00	1.10	2.97	2.49	1.50	1.30	2.65	2.60	0.04	0.44	3.00	2.97	1.69
105	−0.40	−0.59	−0.92	−1.31	0.17	−0.91	−1.22	−0.67	−0.64	1.25	1.22	−0.67	1.02	1.92	−0.49	−0.55	−0.28	0.50	1.67	0.91
106	−12.04	39.23	4.25	23.22	3.95	2.16	16.81	−7.85	6.28	−18.32	−17.79	9.71	−8.86	−21.98	5.82	−1.54	−4.15	−16.19	−1.51	−16.22
107	0.15	−0.37	0.69	−0.22	−0.19	−0.06	0.14	0.36	−0.25	0.02	0.06	−0.16	0.11	1.18	0.11	0.13	0.28	−0.12	0.19	−0.08
108	−0.07	−0.40	−0.57	−0.80	0.17	−0.26	−0.63	0.27	−0.49	0.06	−0.17	−0.45	0.03	0.40	−0.47	−0.11	0.09	−0.61	−0.61	−0.11
109	1.94	−19.92	−9.68	−10.95	−1.24	−9.38	−10.20	2.39	−10.27	2.15	2.28	−9.52	−1.48	−0.76	−3.68	−5.06	−4.88	−5.88	−6.11	1.99
110	8.50	0.00	8.20	8.50	11.00	6.30	8.80	7.10	10.10	16.80	15.00	7.90	13.30	11.20	8.20	7.40	8.80	9.90	8.80	12.00
111	6.80	0.00	6.20	7.00	8.30	8.50	4.90	6.40	9.20	10.00	12.20	7.50	8.40	8.30	6.90	8.00	7.00	5.70	6.80	9.40
112	18.08	0.00	17.47	17.36	18.17	17.93	18.16	18.24	18.49	18.62	18.60	17.96	18.11	17.30	18.16	17.57	17.54	17.19	17.99	18.30
113	18.56	0.00	18.24	17.94	17.84	18.51	17.97	18.57	18.64	19.21	19.01	18.36	18.49	17.95	18.77	18.06	17.71	16.87	18.23	18.98
114	0.83	0.83	0.09	0.64	1.48	0.00	0.65	0.10	1.10	3.07	2.52	1.60	1.40	2.75	2.70	0.14	0.54	0.31	2.97	1.79

Table 7 continued

No.	ala	arg	asn	asp	cys	gln	glu	gly	his	ile	leu	lys	met	phe	pro	ser	thr	trp	tyr	val
115	0.00	4.00	2.00	1.00	0.00	2.00	1.00	0.00	1.00	0.00	0.00	2.00	0.00	0.00	0.00	1.00	1.00	1.00	1.00	0.00
116	0.00	3.00	3.00	4.00	0.00	3.00	4.00	0.00	1.00	0.00	0.00	1.00	0.00	0.00	0.00	2.00	2.00	0.00	2.00	0.00
117	0.00	1.10	−2.00	−2.60	5.40	2.40	3.10	−3.40	0.80	−0.10	−3.70	−3.10	−2.10	0.70	7.40	1.30	0.00	−3.40	4.80	2.70
118	−3.27	−12.27	−6.00	−5.82	−3.26	−8.18	−8.00	−1.09	−6.27	−9.81	−9.81	−12.54	−7.63	−7.63	−6.54	−4.46	−6.64	−7.36	−7.95	−7.63
119	0.00	5.90	3.52	2.80	1.04	3.52	2.80	0.00	3.32	0.00	0.00	4.24	1.04	0.00	0.00	1.40	1.40	1.66	1.40	0.00

References

Altfeld MA, Livingston B, Reshamwala N, Nguyen PT, Addo MM, Shea A, Newman M, Fikes J, Sidney J, Wentworth P, Chesnut R, Eldridge RL, Rosenberg ES, Robbins GK, Brander C, Sax PE, Boswell S, Flynn T, Buchbinder S, Goulder PJR, Walker BD, Sette A, Kalams SA (2001) Identification of novel HLA-A2-restricted human immunodeficiency virus type 1-specific cytotoxic T-lymphocyte epitopes predicted by the HLA-A2 supertype peptide-binding motif. J Virol 75:1301–1311

Bigelow CC (1967) On the average hydrophobicity of proteins and the relation between it and protein structure. J Theor Biol 16:187–211

Blythe MJ, Doytchinova IA, Flower DR (2002) JenPep: a database of quantitative functional peptide data for immunology. Bioinformatics 18:434–439

Brusic V, Flower DR (2007) Bioinformatics tools for identifying T-cell epitopes. Drug Discov Today Biosilico 2:18–23

Brusic V, Rudy G, Honeyman G, Hammer J, Harrison L (1998) Prediction of MHC class II-binding peptides using an evolutionary algorithm and artificial neural network. Bioinformatics 14:121–130

Chang C, Ekins S, Bahadduri P, Swaan PW (2006) Pharmacophore-based discovery of ligands for drug transporters. Adv Drug Deliv Rev 58:1431–1450

Chou KC (1996) Review: prediction of HIV protease cleavage sites in proteins. Anal Biochem 233:1–14

Chou KC, Shen HB (2007) Review: recent progresses in protein subcellular location prediction. Anal Biochem 370:1–16

Collantes ER, Dunn WJ (1995) Amino acid side chain descriptors for quantitative structure activity relationship studies of peptide analogues. J Med Chem 38:2705–2713

Coyle AJ, Gutierrez-Ramos JC (2001) The expanding B7 superfamily: increasing complexity in costimulatory signals regulating T cell function. Nat Immunol 2:203–209

del Guercio MF, Sidney J, Hermanson G, Perez C, Grey HM, Kubo RT, Sette A (1995) Binding of a peptide antigen to multiple HLA alleles allows definition of an A2-like supertype. J Immunol 154:685–693

DeLano WL (2002) The PyMOL molecular graphics system. DeLano Scientific, San Carlos

Doytchinova IA, Flower DR (2001) Toward the quantitative prediction of T-cell epitopes: CoMFA and CoMSIA studies of peptides with affinity for the class I MHC molecule HLA-A*0201. J Med Chem 44:3572–3581

Doytchinova IA, Flower DR (2002) Physicochemical explanation of peptide binding to HLA-A*0201 major histocompatibility complex: a three-dimensional quantitative structure–activity relationship-study. Proteins 48:505–518

Doytchinova IA, Flower DR (2007) Predicting class I major histocompatibility complex (MHC) binders using multivariate statistics: comparison of discriminant analysis and multiple linear regression. J Chem Inf Model 47:234–238

Doytchinova IA, Blythe MJ, Flower DR (2002) Additive method for the prediction of protein–peptide binding affinity: application to the MHC class I molecule HLA-A*0201. J Proteome Res 1:263–272

Du QS, Wei YT, Pang ZW, Chou KC, Huang RB (2007) Predicting the affinity of epitope-peptides with class I MHC molecule HLA-A*0201: an application of amino acid-based peptide prediction. Protein Eng Des Sel 20:417–423

Falk K, Rötzschke O, Stefanovic S, Jung G, Rammensee HG (1991) Allele specific motifs revealed by sequencing of self-peptides eluted from MHC molecules. Nature 351:290–296

Germain RN (1994) MHC-dependent antigen processing and peptide presentation: providing ligands for T lymphocyte activation. Cell 76:287–299

Golbraikh A, Tropsha A (2002) Beware of q2!. J Mol Graphics Mod 20:269–276

Guan P, Doytchinova IA, Walshe VA, Borrow P, Flower DR (2005) Analysis of peptide-protein binding using amino acid descriptors: prediction and experimental verification for HLA-A*0201. J Med Chem 48:7418–7425

Gulukota K, Sidney J, Sette A, DeLisi C (1997) Two complementary methods for predicting peptides binding major histocompatibility complex molecules. J Mol Biol 267:1258–1267

Hagmann M (2000) Computers aid vaccine design. Science 290:80–82

Hattotuwagama CK, Toseland CP, Guan P, Taylor DJ, Hemsley SL, Doytchinova IA, Flower DR (2006) Toward prediction of class II mouse major histocompatibility complex peptide binding affinity: in silico bioinformatic evaluation using partial least squares, a robust multivariate statistical technique. J Chem Inf Model 46:1491–1502

Hellberg S, Sjostrom M, Skagerberg B, Wold S (1987) Peptide quantitative structure–activity relationships, a multivariate approach. J Med Chem 30:1126–1135

Hill AV, Elvin J, Willis AC, Aidoo M, Allsopp CE, Gotch FM, Gao XM, Takiguchi M, Greenwood BM, Townsend AR (1992) Molecular analysis of the association of HLA-B53 and resistance to severe malaria. Nature 360:434–439

Honeyman MC, Brusic V, Stone NL, Harrison LC (1998) Neural network-based prediction of candidate T-cell epitopes. Nat Biotechnol 16:966–969

Horton R, Wilming L, Rand V, Lovering RC, Bruford EA, Khodiyar VK, Lush MJ, Povey S, Talbot CC Jr, Wright MW, Wain HM, Trowsdale J, Ziegler A, Beck S (2004) Gene map of the extended human MHC. Nat Rev Genet 5:889–899

Kast WM, Brandt RM, Sidney J, Drijfhout JW, Kubo RT, Grey HM, Melief CJ, Sette A (1994) Role of HLA-A motifs in identification of potential CTL epitopes in human papillomavirus type 16 E6 and E7 proteins. J Immunol 152:3904–3912

Kitchen DB, Decornez H, Furr JR, Bajorath J (2004) Docking and scoring in virtual screening for drug discovery: Methods and applications. Nat Rev 3:935–949

Kubo RT, Sette A, Grey HM, Appella E, Sakaguchi K, Zhu NZ, Arnott D, Sherman N, Shabanowitz J, Michel H (1994) Definition of specific peptide motifs for four major HLA-A alleles. J Immunol 152:3913–3925

Kawashima I, Ogata H, Kanehisa M (1999a) AAindex: amino acid index database. Nucleic Acids Res 27:368–369

Kawashima I, Tsai V, Southwood S, Takesako K, Sette A, Celis E (1999b) Identification of HLA-A3-restricted cytotoxic T lymphocyte epitopes from carcinoembryonic antigen and HER–2/neu by primary in vitro immunization with peptide-pulsed dendritic cells. Cancer Res 59:431–435

Kidera A, Konishi Y, Oka M, Ooi T, Scheraga HA (1985) A statistical analysis of the physical properties of the 20 naturally occurring amino acids. J Protein Chem 4:23–55

Kirksey TJ, Pogue-Caley RR, Frelinger JA, Collins EJ (1999) The structural basis for the increased immunogenicity of two HIV-reverse transcriptase peptide variant/class I major histocompatibility complexes. J Biol Chem 274:37259–37264

Khan AR, Baker BM, Ghosh P, Biddison WE, Wiley DC (2000) The structure and stability of an HLA-A*0201/octameric Tax peptide complex with an empty conserved peptide-N-terminal binding site. J Immunol 164:6398–6405

Kubinyi H (1997) QSAR and 3D-QSAR in drug design. Drug Discov Today 2:457–467

Lin Z, Wu Y, Zhu B, Ni B, Wang L (2004) Toward the quantitative prediction of T-Cell. epitopes: QSAR studies on peptides having affinity with the class I MHC molecular HLA-A*0201. J Comput Boil 11:683–694

Logean A, Sette A, Rognan D (2001) Customized versus universal scoring functions: application to class I MHC–peptide binding free energy predictions. Bioorg Med Chem Lett 11:675–679

Lu Y, Bulka B, desJardins M, Freeland SJ (2007) Amino acid quantitative structure property relationship database: a web-based platform for quantitative investigations of amino acids. Protein Eng Des Sel 20:347–351

Madden DR (1995) The three-dimensional structure of peptide–MHC complexes. Annu Rev Immunol 13:587–622

Madden DR, Garboczi DN, Wiley DC (1993) The antigenic identity of peptide/MHC complexes, a comparison of the conformations of five viral peptides presented by HLA-A2. Cell 75:693–708

Pamer EG, Harty JT, Bevan MJ (1991) Precise prediction of a dominant class I MHC-restricted epitope of Listeria monocytogenes. Nature 353:852–855

Parker KC, Bednarek MA, Coligan JE (1994) Scheme for ranking potential HLA-A2 binding peptides based on independent binding of individual peptide side-chain. J Immunol 152:163–175

Peoples GE, Goedegebuure PS, Smith R, Linehan DC, Yoshino I, Eberlein TY (1995) Breast and ovarian cancer-specific cytotoxic T lymphocytes recognize the same HER2/neu-derived peptide. Proc Natl Acad Sci USA 92:432–436

Rammensee HG (1995) Chemistry of peptides associated with MHC class 1 and class II molecules. Curr Opin Immunol 7:85–96

Rammensee HG (2003) Immunoinformatics: bioinformatic strategies for better understanding of immune function. Novartis Found Symp 254:1–2

Rogers D, Hopfinger AJ (1994) Application of genetic function approximation to quantitative-structure activity relationships and quantitative structure–property relationships. J Chem Inf Comput Sci 34:854–866

Ruppert J, Sidney J, Celis E, Kubo RT, Grey HM, Sette A (1993) Prominent role of secondary anchor residues in peptide binding to HLA-A*0201 molecules. Cell 74:929–937

Sandberg M, Eriksson L, Jonsson J, Wold S (1998) New chemical descriptors for the design of biologically active peptides. A multivariate charaterrization of 87 amino acids. J Med Chem 41:2481–2491

Sapper MA, Bjorkman PJ (1991) Refined structure of the human histocompatibility antigen HLA-A2 at 2.6Å resolution. J Mol Biol 219:277–319

Sarobe P, Pendleton CD, Akatsuka TD, Engelhard VH, Feinstone SM, Berzofsky JA (1998) Enhanced in vitro potency and in vivo immunogenicity of a CTL epitope from hepatitis C virus core protein following amino acid replacement at secondary HLA-A 2.1 binding positions. J Clin Invest 102:1239–1248

Schefzick S, Bradley M (2004) Comparison of commercially available genetic algorithms: GAs as variable selection tool. J Comput Aided Mol Des 18:511–521

Schueler-Furman O (2000) Structure-based prediction of binding peptides to MHC class I molecules: application to a broad range of MHC alleles. Protein Sci 9:1838–1846

Sette A (2000) Tools of the trade in vaccine design. Science 290:2074–2075

Sette A, Buus S, Appella E, Smith JA, Chesnut R, Miles C, Colon SM, Grey HM (1989) Prediction of major histocompatibility complex binding regions of protein antigens by sequence pattern analysis. Proc Natl Acad Sci USA 86:3296–3300

Sette A, Sidney J (1998) HLA supertypes and supermotifs: a functional perspective on HLA polymorphism. Curr Opin Immunol 10:478–482

Shen HB, Chou KC (2007a) EzyPred: a top–down approach for predicting enzyme functional classes and subclasses. Biochem Biophys Res Comm 364:53–59

Shen HB, Chou KC (2007b) Signal–3L: a 3-layer approach for predicting signal peptide. Biochem Biophys Res Comm 363:297–303

Šoškić M (1996) Link between orthogonal and standard multiple linear regression models. J Chem Inf Comput Sci 36:829–832

Sturniolo T, Bono E, Ding J, Raddrizzani L, Tuereci O, Sahin U, Braxenthaler M, Gallazzi F, Protti MP, Sinigaglia F, Hammer J (1999) Generation of tissue-specific and promiscuous HLA ligand databases using DNA microarrays and virtual HLA class II matrices. Nat Biotechnol 17:555–561

Sutter JM, Dixon SL, Jurs PC (1995) Automated descriptor selection for quantitative structure–activity relationships using generalized simulated annealing. J Chem Inf Comput Sci 35:77–84

Tropsha A, Gramatica P, Gombar VK (2003) The importance of being earnest: validation is the absolute essential for successful application and interpretation of QSPR models. QSAR Comb Sci 22:69–77

Trowsdale J, Ragoussis J, Campbell RD (1991) Map of the human MHC. Immunol Today 12:443–446

Udaka K, Mamitsuka H, Nakaseko Y, Abe N (2002) Empirical evaluation of a dynamic experiment design method for prediction of MHC class I-binding peptides. J Immunol 169:5744–5753

Wallace AC, Laskowski RA, Thornton JM (1995) LIGPLOT: a program to generate schematic diagrams of protein–ligand interactions. Protein Engin 8:127–134

Xiao X, Shao S, Ding Y, Huang Z, Chen X, Chou KC (2005) An application of gene comparative image for predicting the effect on replication ratio by HBV virus gene missense mutation. J Theor Biol 235:555–565

Zhou P, Tian F, Zhang M, Li Z (2006) Applying generalized hydrophobicity scale of amino acids to quantitative prediction of human leukocyte antigen-A*0201-restricted cytotoxic T lymphocyte epitope. Chin Sci Bull 51:1439–1443

Zhou P, Tian F, Li Z (2007) A structure-based, quantitative structure–activity relationship approach for predicting HLA-A*0201-restricted cytotoxic T lymphocyte epitopes. Chem Biol Drug Des 69:56–67

Zhou P, Tian F, Wu Y, Li Z, Shang Z. (2008) Quantitative sequence-activity model (QSAM): Applying QSAR strategy to model and predict bioactivity and function of peptides, proteins and nucleic acids. Curr Comput Aided Drug Des (in press)