

Toward Prediction of Binding Affinities Between the MHC Protein and Its Peptide Ligands Using Quantitative Structure-Affinity Relationship Approach

Feifei Tian^{1,2}, Fenglin Lv^{1,*}, Peng Zhou³, Qinwu Yang¹ and Abraham F. Jalbout⁴

¹Research Institute of Surgery, Daping Hospital, Third Military Medical University, Chongqing, China; ²College of Bioengineering, Chongqing University, Chongqing, China; ³Department of Chemistry, Zhejiang University, Hangzhou, China; ⁴Department of Chemistry, The University of Arizona, Tucson, AZ, USA

Abstract: It is important but challenging to determine the binding specificity of MHC-peptide interactions accurately and to predict their binding affinity quantitatively. In this paper, we discuss the application of an effective amino acid descriptor to model and predict the binding affinities between the MHC protein and its peptide ligands. This amino acid descriptor was derived from 23 electronic properties, 37 steric properties, 54 hydrophobic properties and 5 hydrogen bond properties of coded amino acids using principal component analysis (PCA), called the divided physicochemical property scores (DPPS). The DPPS descriptor was used to characterize a set of mouse MHC (H-2K^K) binding peptides, and genetic algorithm-partial least square (GA-PLS) models were then constructed. In analyses, these models were statistically consistent with previous reports and molecular graphics exhibition. Hydrophobic interactions and hydrogen bonds were important to antigen recognition and presentation, especially exerting effects on anchor residues of peptides.

Keywords: Quantitative structure-affinity relationship (QSAR), divided physicochemical property scores (DPPS), immunoinformatics, MHC, peptide.

INTRODUCTION

With the advent of bioinformatics, powerful tools are in emergence for collection, maintenance and processing of biological data. Computational approaches, such as structural bioinformatics [1-4], molecular docking [5-9], pharmacophore modeling [10,11], QSAR [12,13], protein cleavage site prediction [14-17], protein subcellular location prediction [18-23], membrane protein type prediction [24], enzyme functional class prediction [25], mutation prediction [26,27], and signal peptide prediction [28,29], can provide very useful information for drug design in a timely manner. In immune system, the astounding diversity of components (e.g. immunoglobulins, lymphocyte receptors, or cytokines) together with the complexity of the regulatory pathways and network-type interactions makes immunology a combinatorial science. Currently available data represent only a tiny fraction of possible situations and data continues to accrue at an exponential rate. Computational analysis has therefore become an essential element of immunology research with a main role of immunoinformatics being the management and analysis of immunological data [30].

The products of the major histocompatibility complex (MHC) play a fundamental role in regulating immune responses. There are two classes of MHC proteins: MHC I and MHC II. The MHC I protein is encoded by different genetic loci and extremely polymorphic. The main processing pathway for peptide ligands of MHC I involves degradation of

proteins by the proteasome, followed by transport of the products by the transporter associated with antigen processing (TAP) to the endoplasmic reticulum (ER), where peptides are bound to MHC I proteins, and then presented on the cell surface by MHCs. Peptides binding to MHC I usually are 8–12 amino acids long and require free N- and C-terminal. In addition to a specific size, a combination of two main anchor residues is required. These anchors have been described as Leu at position 2 and Leu or Val at the C-terminal end [31]. The presence of anchors is necessary, but not sufficient, for high-affinity binding. Prominent roles for several other positions, so-called secondary anchor residues, have also been demonstrated [32].

Identification of MHC binding peptide sequences was believed to be the bottleneck of present vaccine development [33] and the peptide library experiments were often used to serve this purpose. Although the peptide library has a limited coverage of all possible peptides, it is too time-consuming and expensive to synthesize all potential peptides found in infected cells. Therefore, it is important to develop computational methods to derive such information. Some rigorous computational approaches were developed to calculate the binding free energy between the MHC and peptides to analyze the interaction modes [34,35]. However, accurate computation of binding free energy is quite trivial. Besides, it is quite time-consuming on protein-peptide systems [36]. Quantitative structure-affinity relationship (QSAR) provides a practical tool for exploring the MHC-peptides interactions. Previously, Doytchinova *et al.* [37,38] used comparative molecular field analysis (CoMFA) and comparative molecular similarity indices analysis (CoMSIA) to analyze the binding affinities of different MHC proteins and peptides. Lin

Address correspondence to this author at the Research Institute of Surgery, Daping Hospital, Third Military Medical University, Chongqing 40042, China; Tel: ++86-23-68757413; Fax: ++86-23-68757413; E-mail: lufenglin001@yahoo.com.cn

[39] and Guan [40] *et al.* employed the ISA-ECI index, z-scale and physicochemical properties from two dimensional QSAR pathway to quantitatively predict the interactions between the MHC and its peptide ligands. Recently, we have discussed the interaction modes using structure-based QSAR and nonbonding interaction matrix, and to obtain models with good prediction ability [41].

Reasonable representation is essential to QSAR study of peptides. This work used an effective structural representation approach as divided physicochemical property scores (DPPS) to characterize the structures of a set of mouse MHC (H-2K^k) binding peptides. A quantitative structure-affinity relationship (QSAR) model with low computation complexity was developed by genetic algorithm-partial least square (GA-PLS) to predict the binding affinities and acquire some information concerning the interactions between the mouse MHC (H-2K^k) protein and its peptide ligands. Further, key features closely related to interactions were investigated, that provides instructive reference for designing and synthesizing peptide fragments with high binding affinities.

METHODS

It is well known that nonbonding factors play significant effects in biosystem, including molecular recognition in protein/drug, antibody/antigen and enzyme/substrate complexes. Although there were numerous literature reports on amino acid properties, many of them are essentially irrelevant with our researches (e.g. amino acid parameters in advanced structures of protein and properties tested in non-aqueous solution, etc.). Nonbonding interactions of receptor/ligand have been mainly expressed as electrostatic, van der Waals, hydrophobic and hydrogen bond interactions, while other factors (e.g. charge transfers and salt bridges, etc.) are regarded as special forms. A total 119 physicochemical parameters of 20 coded amino acids were collected from databases and literatures [42-47] and divided into four groups according to different categories of properties: 23 electronic properties, 37 steric properties, 54 hydrophobic properties and 5 hydrogen bond properties respectively. These properties possess straightforward and physicochemical information such as electrostatic effect, bulky property, hydrophobicity, hydrogen bond contribution factor, etc. Then, these four categories of property matrices were analyzed by principal component analysis (PCA) separately. For the matrices of electronic, steric, hydrophobic and hydrogen bond properties, the first 4, 2, 2 and 2 principal components accounted for 74.44%, 72.72%, 73.78% and 77.15% variance of original data matrices, respectively. That is to say, the most information in the four original matrices can be replaced by the first 4, 2, 2 and 2 principal component scores, respectively. Hence, the electronic, steric, hydrophobic and hydrogen bond properties of 20 amino acids can be expressed by the 10 principal component scores with less information lost [48]. The 10 score vectors were called the divided physicochemical property scores (DPPS). For amino acids, D1-D4 are related to electronic properties, D5 and D6 to steric properties, D7 and D8 to hydrophobic properties and D9 and D10 to hydrogen bond properties (Table 1). Each residue in a sequence is described by ten DPPS descriptors according to varied amino acid positions. Accordingly, the structural fea-

tures of a sequence with n residues were characterized by the concatenation of $10 \times n$ DPPS vectors.

For the DPPS descriptors, steric, hydrophobic and hydrogen bond property only use the top 2 principal components accounting for more than 70% variance of the original matrix, whereas the electronic property requires 4 principal components to achieve an equivalent level. In analysis, the molecular polarity is very complex, including both the macroscopic properties as net charge and electrostatic potential, and microscopic properties as electronic structure and chemical shift. Fig. (1) illustrates the distributions of 20 amino acids in the top 2 principal component space of electronic, steric, hydrophobic and hydrogen bond properties, respectively. Fig. (1a) shows the space of electronic property, and the second PCA space favorably characterizes polarity features of the amino acids. In Fig. (1b), amino acids are labeled in terms of their volumes. In this Fig., the first PCA space is interpretable as amino acid bulks. Similarly in Fig. (1c), the hydrophobicity of amino acids is successfully reflected by the first principal component; from left to the right of this figure, hydrophobicity of amino acids is increasing, which is 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 basically inclusive of information on amino acids forming hydrogen bond, and marks "□" and "○" indicate whether or not this amino acid side-chain can form hydrogen bond. Obviously, amino acid properties are well discriminated in these four PCA scoring plots, confirming the validity of the DPPS descriptors.

PEPTIDE DATASET

154 MHC I (H2-K^k allele)-restricted CTL epitopes (binding peptides) with free C- and N-terminal were compiled from JenPep database [49,50]. Table 2 summarizes their amino acid sequences and binding affinities pIC₅₀, respectively, including 121 high affinities (IC₅₀ ≤ 50nM, pIC₅₀ ≥ 7.301), 17 moderate ones (50nM < IC₅₀ ≤ 500nM, 7.301 > pIC₅₀ ≥ 6.301), and 16 low ones (IC₅₀ > 500nM, pIC₅₀ < 6.301). The binding affinities IC₅₀ we used here were assayed quantitatively by the following: the complex of a radiolabeled standard peptide/MHC I and experimental peptides of different doses are incubated given time at room temperature to determine the IC₅₀.

As shown in Fig. (2a), the three-dimensional crystal structures of H2-K^k-peptide at a 2.5Å resolution by X-ray diffraction (PDB ID: 1zt1) [51] is revealed, and 2α-helices and 1β-sheet make up the peptide-binding cleft of the H2-K^k protein, with the embedded peptide unfolding. Fig. (2b) is the conformation of the antigen peptide extracted from the complex, and all the residues are in *trans*-conformation, thus leading to maximum distance along the side-chain but posing no remarkable distortion. Hence, peptides under binding state are in low-energy conformation, influenced insignificantly by MHC protein, and the interaction between peptide residues and nearby MHC residues is deemed to be decisive to the binding.

It is generally accepted that the internal validation of the QSAR model based upon the training set is sufficient to establish its predictive power. However, Golbraikh *et al.* [52]

Table 1. Divided Physicochemical Property Scores (DPPS)

AAs	Electronic Property				Steric Property		Hydrophobicity		Hydrogen Bond	
	D ₁	D ₂	D ₃	D ₄	D ₅	D ₆	D ₇	D ₈	D ₉	D 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

suggest that there is no direct relation between leave-one-out (LOO) cross validated q^2 on the training set and the correlation coefficient r_{pred}^2 on the test set. Therefore, 154 antigen peptides are randomly divided into training/test set as 104/50, and then the model constructed by the training set is tested by test set (test samples are numbered 105-154 in Table 2). Two criteria are cooperated to select test peptides. First, amino acid kinds at each peptide position should be in presence at the same position of training samples; Secondly, binding affinities of peptides should be neither the maximum nor the minimum [37].

RESULTS AND DISCUSSION

QSAR Modeling

Each amino acid is characterized by 10 DPPS descriptors. So for a peptide comprising m amino acids, $m \times 10$ variables are generated. For instance, a dipeptide can be characterized by 2×10 variables. When employing DPPS to characterize an octapeptide, totally 80 variables (D_1 — D_{80}) are generated in correspondence. Amongst, D_1 — D_{10} indicate 10 DPPS descriptors at position 1, D_{11} — D_{20} are orderly-existed 10 DPPS descriptors at position 2, and so forth. However,

not all the variables are relevant with biological activity, and the unrelated variables are supposed to be omitted to promote modeling stability and predictability. Here, variable selection is implemented by genetic algorithm-partial least square (GA-PLS), which is a sophisticated hybrid approach that combines the powerful optimization method GA with the robust statistical method PLS [54]. In the GA-PLS, the chromosome indicates the selection scheme of variable subset, and the fitness is evaluated by the internal prediction ability of the resulted PLS model. Based upon Matlab environment, GA-PLS toolbox (GP-toolbox for short) was developed by our laboratory, enabling flexible operations of GA-PLS calculations and thus suitable for complex statistical modeling and data processing. GA-PLS parameters settings included: population size, 200; genmax, 300; convergence criteria, 80% of population achieving an agreement or genmax; mutation rate, 0.5%; hybridization and crossover, 2 points; cross-validation, leave-1/5-out (random grouping); and data pretreatment, autoscaling. Consequently, an optimal variable subset is yielded, composed of 34 variables as V_1, V_3, V_5, V_7, V_8, V_9, V_10, V_12, V_14, V_17, V_18, V_19, V_25, V_26, V_29, V_39, V_41, V_45, V_46, V_52, V_55, V_56, V_59, V_62, V_65, V_66, V_69, V_70, V_71,

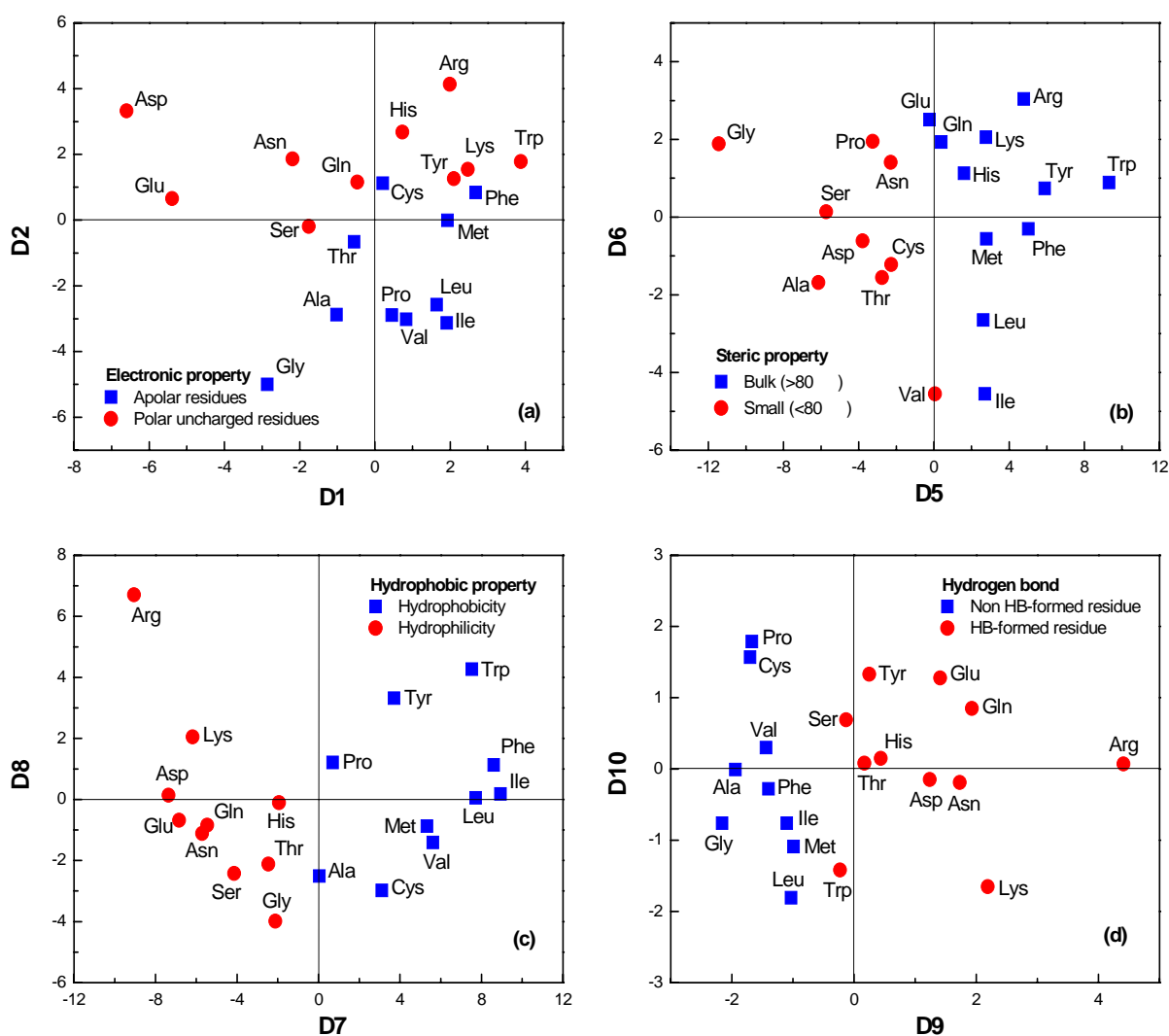


Figure 1. Distributions of amino acid in the top two PCA spaces. (a) Electronic property; (b) Steric property; (c) Hydrophobic property; (d) Hydrogen bond.

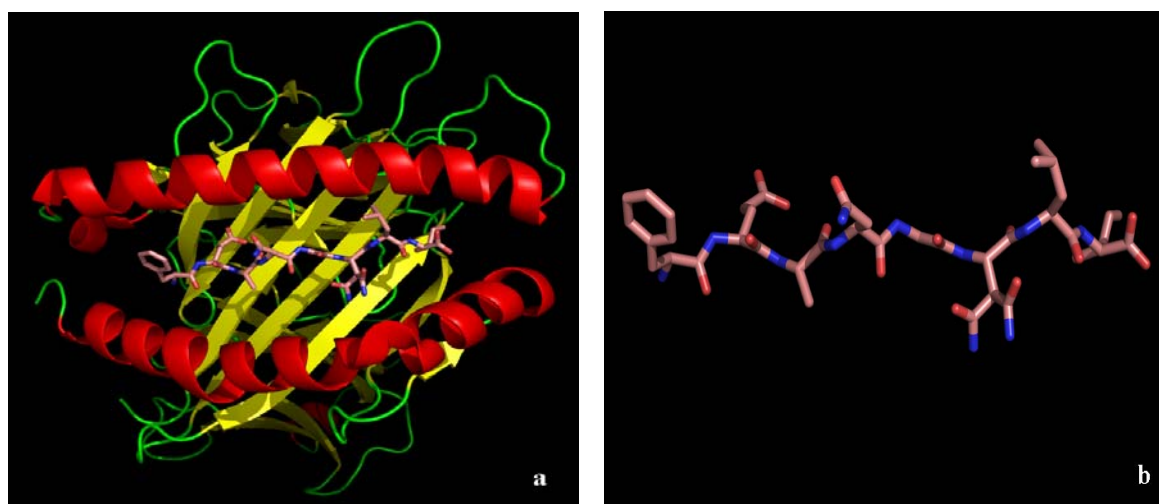


Figure 2. (a) Crystal structure of mouse MHC (H-2K^k)-peptide complex solved by X-ray (PDB ID: 1zt1); (b) Conformation of antigen peptide extracted from the complex (produced using PyMOL [53]).

Table 2. The Primary Sequences and Binding Affinities of MHC Binding Peptides

No.	Epitopes	pIC ₅₀		No.	Epitopes	pIC ₅₀	
		Obsd	Cald			Obsd	Cald
Training set				78	TESTGNLI	7.535	7.549
1	GESTGNLI	7.665	7.817	79	VESTGNLI	7.545	7.539
2	HESTGNLI	7.607	7.593	80	WESTGNLI	7.740	outlier
3	IESTGNLI	7.715	7.678	81	YESTGNLI	7.740	8.009
4	KESTGNLI	7.308	7.454	82	DGLGGKLV	7.959	outlier
5	LESTGNLI	7.716	7.721	83	FAFPGELL	7.022	7.055
6	MESTGNLI	7.716	7.831	84	FAFWAFVV	7.523	7.717
7	NESTGNLI	7.736	7.821	85	FLHPSMPV	7.149	7.094
8	PESTGNLI	7.426	7.283	86	HAIHGLLV	7.319	7.231
9	QESTGNLI	7.727	7.558	87	LEILNGEI	7.921	7.903
10	RESTGNLI	7.544	outlier	88	FESTGNYI	7.793	7.747
11	SESTGNLI	7.641	7.650	89	FESTGPLI	8.302	8.139
12	FEFTGNLN	8.000	7.995	90	FESTGQLI	7.920	7.948
13	FEGTGNLN	8.265	8.019	91	FESTGRLI	8.222	8.111
14	FEHTGNLN	7.982	8.269	92	FESTGSLI	7.992	7.865
15	FEITGNLN	8.197	8.102	93	FESTGTLI	7.922	8.052
16	FEKTGNLN	7.904	7.623	94	FESTGVLI	8.023	8.128
17	FELTGNLN	8.343	8.062	95	FESTGWLI	7.872	8.079
18	FEMTGNLN	8.222	8.233	96	FESTGYLI	8.215	8.082
19	FENTGNLN	8.224	7.977	97	FESTHNLI	7.836	7.734
20	FEPTGNLN	8.043	7.853	98	FESTINLI	7.887	8.088
21	FEQTGNLN	8.217	7.990	99	FESTKNLI	7.304	7.417
22	FERTGNLN	8.300	8.270	100	FESTLNLI	7.898	8.017
23	FESAGNLI	8.031	8.085	101	FESTMNLI	7.888	7.905
24	FESDGNLI	7.890	7.845	102	FESTNNLI	7.748	7.584
25	FESEGNLI	7.972	7.835	103	FESTPNLI	8.141	7.959
26	FESFGNLI	8.085	7.721	104	FESTQNLI	7.819	7.617
27	FESGGNLI	7.985	8.140	Test set			
28	FESHGNLI	8.248	8.046	105	FESTRNLI	7.679	7.241
29	FESIGNLI	8.239	8.194	106	FESTSNLI	7.821	7.530
30	FESKGNLI	7.978	8.033	107	FESTTNLI	7.821	7.734
31	FESLGNLI	8.403	8.223	108	FESTVNLI	7.912	8.276
32	FESMGNLI	8.040	8.009	109	FESTWNLI	7.832	7.874
33	FESNGNLI	7.880	7.884	110	FESTYNLI	7.460	6.904
34	FESPGNLI	8.042	8.085	111	FESVGNLI	8.230	7.724
35	FESQGNLI	8.094	7.992	112	FESWGNLI	7.989	7.805

(Table 2) contd....

No.	Epitopes	pIC ₅₀		No.	Epitopes	pIC ₅₀	
		Obsd	Cald			Obsd	Cald
36	FESRGNLI	8.095	8.122	113	FESYGNLI	8.099	8.016
37	FESSGNLI	8.046	7.875	114	FETTGNLN	8.232	8.502
38	FESTANLI	7.994	7.877	115	FEVTGNLN	8.223	8.376
39	FESTDNLI	7.743	7.504	116	FEWTGNLN	8.225	7.903
40	FESTENLI	7.583	7.555	117	FEYTGNLN	8.176	8.374
41	FESTFNLI	7.895	7.934	118	FFSTGNLI	5.421	5.934
42	FESTGALI	7.964	8.008	119	FGSTGNLI	7.846	7.840
43	FESTGDLI	7.683	7.610	120	FHSTGNLI	5.122	6.358
44	FESTGELI	7.593	7.786	121	FISTGNLI	6.329	6.543
45	FESTGFLI	8.267	8.148	122	FKSTGNLI	5.026	5.710
46	FESTGGLI	7.946	7.926	123	FLSTGNLI	7.088	6.827
47	FESTGHLI	7.997	8.068	124	FMSTGNLI	6.863	6.598
48	FESTGILI	8.098	8.228	125	FNSTGNLI	6.244	6.702
49	FESTGKLI	7.927	8.222	126	FPSTGNLI	8.113	7.416
50	FESTGLLI	8.079	8.187	127	FQSTGNLI	7.013	6.550
51	FESTGMLI	7.979	8.115	128	FRSTGNLI	4.192	5.388
52	FESTGNAI	7.602	7.780	129	FSSTGNLI	7.718	7.364
53	FESTGNDI	7.290	7.500	130	FTSTGNLI	7.547	7.106
54	FESTGNEI	7.541	7.364	131	FVSTGNLI	7.216	7.359
55	FESTGNFI	8.044	7.905	132	FWSTGNLI	5.325	4.757
56	FESTGNGI	7.209	7.043	133	FYSTGNLI	5.592	5.532
57	FESTGNHI	7.742	7.430	134	AESKSVII	6.648	6.538
58	FESTGNII	7.551	7.889	135	NEKSFKDI	6.910	7.186
59	FESTGNKI	7.159	7.236	136	QTFVVGCI	6.796	6.873
60	FESTGNLA	7.455	7.576	137	AESTGNLI	7.624	7.950
61	FESTGNLD	5.010	5.298	138	DESTGNLI	7.712	7.620
62	FESTGNLE	4.707	7.283	139	EESTGNLI	7.732	7.681
63	FESTGNLFI	7.848	7.576	140	FASTGNLI	7.429	8.168
64	FESTGNLGI	6.051	6.300	141	FDSTGNLI	7.814	7.659
65	FESTGNLHI	6.000	outlier	142	FEATGNLN	8.178	8.465
66	FESTGNLI	8.046	7.869	143	FEDTGNLN	8.199	7.976
67	FESTGNLKI	5.010	5.744	144	FEETGNLN	8.028	7.565
68	FESTGNLLI	7.737	7.797	145	FESTGNLY	6.010	6.657
69	FESTGNLMI	7.212	7.007	146	FESTGNMI	7.612	7.428
70	FESTGNLNI	7.000	6.893	147	FESTGNNI	7.521	8.026
71	FESTGNLPI	5.919	6.265	148	FESTGNPI	7.410	7.661

(Table 2) contd....

No.	Epitopes	pIC ₅₀		No.	Epitopes	pIC ₅₀	
		Obsd	Cald			Obsd	Cald
72	FESTGNLQ	5.687	6.196	149	FESTGNQI	7.612	7.980
73	FESTGNLR	5.232	4.507	150	FESTGNRI	8.004	7.380
74	FESTGNLS	7.525	7.951	151	FESTGNSI	7.612	7.128
75	FESTGNLT	7.293	7.089	152	FESTGNTI	7.652	7.034
76	FESTGNLV	7.626	7.756	153	FESTGNVI	7.421	7.398
77	FESTGNLW	7.293	6.985	154	FESTGNWI	7.974	7.629

V_72, V_75, V_76, V_78, V_79. Number of principal components is 4, and fitness (referring to root mean square error of cross-validation, RMSCV) is 0.454.

Statistical Analysis

The two anchor residues occupy the sites of the hydrophobic pocket of MHC-peptide binding cleft that are normally referred as position P₂ and P₉ and MHC binding octapeptides are thus presented as P₂–P₉. First, a PLS model without variable selection is directly constructed, thus resulting model 1 (M1). In the M1, the cross-validation q_{LOO}^2 and q_{LGO}^2 on the training set as well as fitting correlation coefficient r^2 are 0.446, 0.413 and 0.668, respectively, showing poor internal fitting ability and stability, especially with its q^2 below 0.5. While in the GA-PLS model (M2), original 80 DPPS descriptors are mostly excluded from the model during the process of variable selection, yielding an optimal variable subset composed of 34 DPPS descriptors. In contrast with M1, M2 is slightly improved in its fitting ability on the training set, and greatly advanced in stability, with the cross validation q_{LOO}^2 and q_{LGO}^2 achieving 0.653 and 0.621 (Table 3). However in analysis of the calculated results of M2, large errors are indicated in the training set in spite of a favorable statistical quality. In investigations, sequence structures of these abnormal samples are somewhat special. Peptide WESTGNLI is overestimated by the model. In analysis, the anchor residues P2 (Trp) and P9 (Ile) of WESTGNLI are hydrophobic, pertaining to classic anchor residues, which is thus deemed to be highly active by our model. From another aspect, the experimental value may be a little low. Besides, Trp at P2 possibly exerts significant effects on the binding, but at P2, occurrence of Trp is really scarce in the training set, thus leading to the statistical model insufficiently trained. In contrast, peptides RESTGNLI, DGLGGKLV and FESTGNLH are underestimated by the model. The anchor residues P2 and P9 of these peptides are occupied by polar Arg, Asp and His, contrary to the theory that anchor residues are hydrophobic amino acids. Removing these four samples, model M3 is rebuilt. As shown in Fig. (3), all training samples are dispersed along an origin-passed line forming an angle of 45°, without outliers.

To validate modeling reliability, statistical diagnosis was further performed on the M3. In Fig. (5), loading contributions of the DPPS descriptors at positions 2 and 9 are above

0.3 at PC1 and 0.2 at PC2, indicating significant effects of nonbonding interactions on anchor residues. In addition, the DPPS descriptors indicating hydrophobicity and hydrogen bond have large loading contributions. In the PLS scoring scatter (Fig. (6)), 10 outliers are out of the Hotelling T2 ellipse confidence interval, which shows that the high-dimensional characteristics of these 10 peptides are significantly different from others. Then, M3 was employed to perform predictions on 50 test samples, and Fig Fig. (4) delineates the predicted versus the observed for test set, revealing correlative coefficient r_{pred}^2 and root mean square error RMSEP are 0.739 and 0.454, respectively. Besides the conventional predictive correlation r_{pred}^2 and root mean square error of prediction RMSEP, statistics proposed by Tropsha *et al.* [55] were meanwhile used for the test set. Corresponding criteria for a QSAR model to perform high predictive power are suggested as follows:

$$q_{ext}^2 = 1 - \frac{\sum_{i=1}^{n_{ext}} (y_i^{obsd} - y_i^{pred})^2}{\sum_{i=1}^{n_{ext}} (y_i^{obsd} - \bar{y}_{tr})^2} \quad (1)$$

$$\frac{r_{pred}^2 - r_{0,ext}^2}{r_{pred}^2} < 0.1 \text{ or } \frac{r_{pred}^2 - r_{0,ext}'^2}{r_{pred}^2} < 0.1 \quad (2)$$

$$0.85 \leq k \leq 1.15 \text{ or } 0.85 \leq k' \leq 1.15 \quad (3)$$

where q_{ext}^2 (external q^2) is external correlation coefficient indicating unbiased prediction on the test set; $r_{0,ext}^2$ and $r_{0,ext}'^2$ are the coefficients of determination for the regression through origin (predicted versus observed activities $r_{0,ext}^2$, and observed versus predicted activities $r_{0,ext}'^2$), and k together with k' are the slopes of the origin-passed regression line. In Table 3, the Tropsha's statistics of the test set indicate that all the models match the criteria of eq.3, but the M1 fails eq.2, which confirms that the unbiasedness of models M2 and M3 is better than the M1.

Analysis of Residue Positions

Fig. (7) shows the standardized coefficients of the 34 variables in M3, indicating the introduced DPPS descriptors are mainly on behalf of anchor residue positions (P2 and P9) [56] and secondary anchor residue positions (P3 and P7) [32]. At positions P2 and P9, 7 and 6 DPPS descriptors were

Table 3. Statistics of Models M1–M3

No.	Size ^a	Outliers	NPC ^b	r^2	q_{LOO}^2 ^c	q_{LGO}^2 ^d	RMSEE	r_{pred}^2	RMSEP	Tropsha's Statistics				
										q_{ext}^2	$r_{0,ext}^2$	$r_{0,ext}^{\prime 2}$	k	k'
M1	104/50	0	2	0.668	0.446	0.413	0.418	0.524	0.581	0.526	0.495	0.464	0.867	1.102
M2	104/50	0	4	0.789	0.653	0.621	0.334	0.712	0.466	0.723	0.686	0.663	0.915	1.005
M3	100/50	4	4	0.796	0.688	0.669	0.323	0.739	0.454	0.742	0.722	0.708	0.996	0.998

^a The two numbers separated by slashes denote the numbers of compounds in training set and in test set, respectively.
^b NPC: number of principal components.
^c q_{LOO}^2 : leave-one-out cross-validated correlation coefficient.
^d q_{LGO}^2 : leave-group-out cross-validated correlation coefficient.

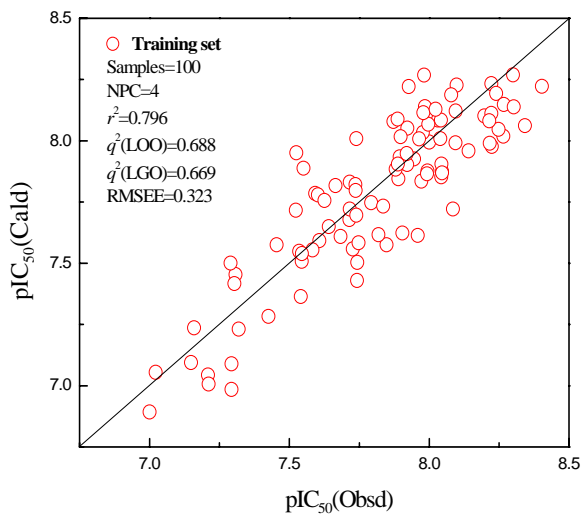


Figure 3. The observed versus the calculated affinities for training set samples by M3.

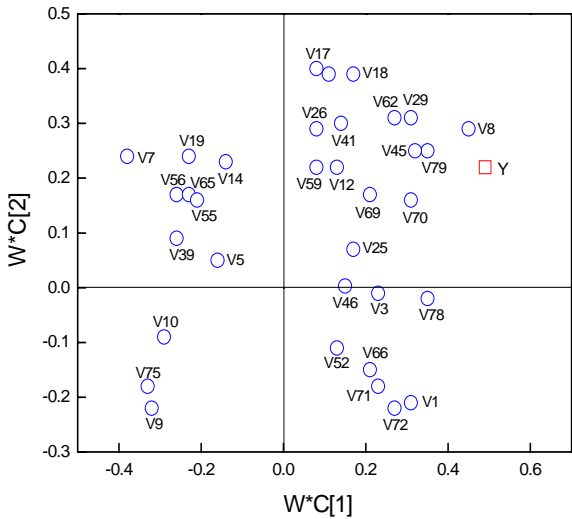


Figure 5. Loading contributions of 34 DPPS descriptors.

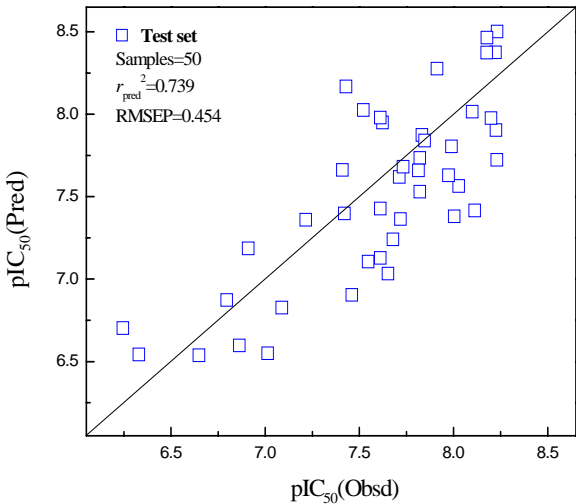


Figure 4. The predicted versus observed for 50 test samples by M3.

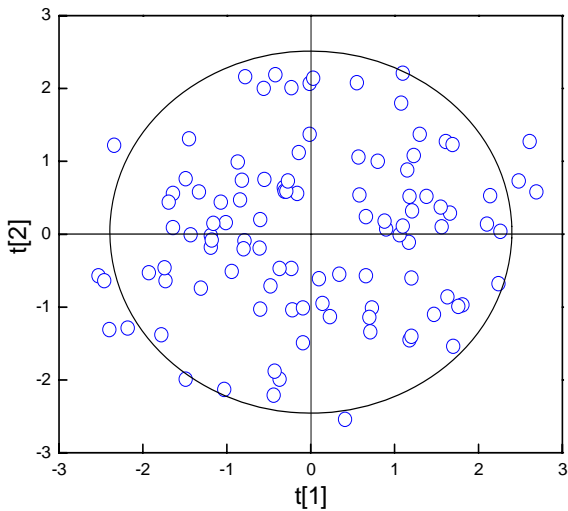


Figure 6. Scoring scatters in the top two PLS principal component spaces.

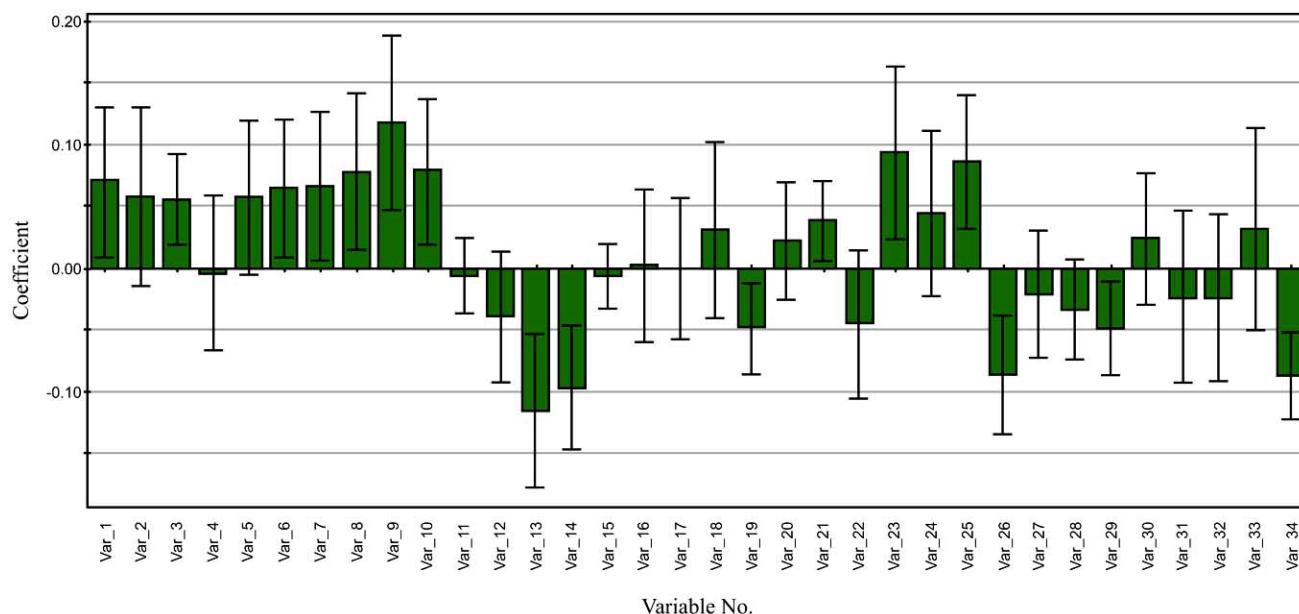


Figure 7. Contributions of standardized coefficients in the M3.

selected, respectively, and the standardized coefficients also confirm the significance of these variables. Fig. (8) schematically illustrates the hydrophobic interactions and hydrogen bonds between the MHC protein and a peptide ligand, wherein the fringed and broken lines indicate hydrophobic interaction and hydrogen bond, respectively. In this figure, many peptide positions are encircled with hydrophobic residues, whereas at the two ends are many hydrogen bonds. The following presents discussions on positions P2–P9 in details.

Positions P2 and P9 of MHC-restricted CTL epitopes, pertaining to the anchor residue, directly determine the peptide binding. Falk *et al.* [56] point out that in correspondence with positions P2 and P9, two hydrophobic pockets are formed at the peptide-binding cleft, attracting strongly hydrophobic amino acids, e.g. Leu at P2 and Val/Leu at P9. Fig. (8) also clearly reveals intensive hydrogen bonds and hydrophobic interactions between the MHC protein and these two positions. In Fig. (7), the conclusion is consistent, i.e. variables indicating hydrogen bond and hydrophobic terms at positions P2 and P9 are all introduced with large contributions.

Positions P3 and P7 are defined as secondary anchor residue by Ruppert *et al.* [32], with their significance secondary to P2 and P9. Fig. (8) illustrates several hydrogen bonds at P3, and Sarobe *et al.* [57] suggest residues bearing hydrophobic aromatic rings (e.g. phenylalanine and tyrosine) are preferred here. In correspondence, hydrophobicity and hydrogen bond are introduced at this position. Around P7, many hydrophobic residues are in presence, but a strongly polar arginine (Arg97) is meanwhile taken place, so P7 exhibits amphipathicity and residues with small hydrophobic side chains are preferred (e.g. Val and Ala). Moreover, electrostatic interaction at P7 is indicated, referred to be caused by the salt bridge between the MHC protein and P7.

For positions P4, P5, P6 and P8 less important than anchor residues, they exert effects on the MHC-peptide binding

to some extent [58]. In correspondence, few DPPS descriptors are introduced at these positions with low contributions.

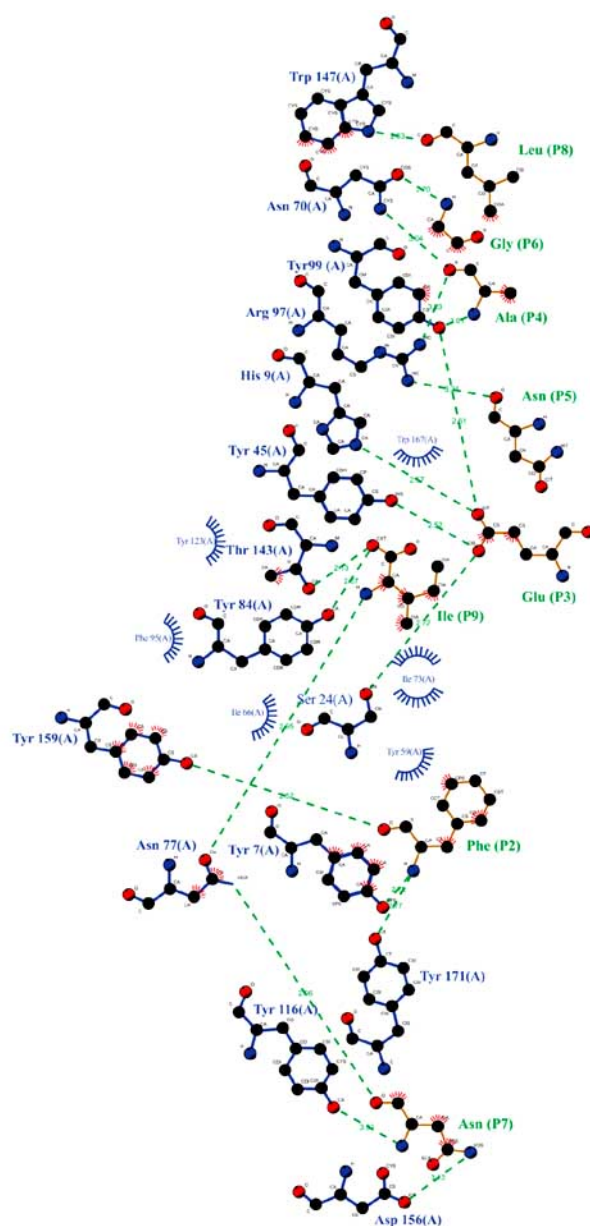
CONCLUSIONS

As a fundamental subject, immunology is presently in an original “wet” state [33], with related theories and algorithms immature. Epitope identification and prediction, an important aspect of immunoinformatics, is pivotal to developing new vaccines and performing immunotherapy, thus driven considerable interests. However, many an epitope predicting methods proposed are based upon intricate bioinformatics and machine learning algorithms, and researches are focused on statistical analysis of resulting accuracies, neglecting explanation at the molecular level. Currently, quick, valid and physicochemically interpretable prediction methods are in lack. In view of that, by PCA technique, 4 kinds of nonbonding effects mostly influencing bioactivity of antigen peptides are extracted to characterize sequence structures, and on that basis, GA-PLS is successfully employed to construct quantitatively predicting model for 154 MHC I-restricted CTL epitopes. Consequently, the following points are concluded.

(i) From the points of nonbonding interaction types, hydrophobic interactions and hydrogen bonds are the most important to the peptides binding, secondarily as electrostatic interaction and the least by steric interaction.

(ii) From the points of peptide positions, anchor residues P2 and P9 have direct relations with the binding, secondary anchor residues P3 and P7 also exert important effects, and non-anchor residues P4, P5, P6 and P8, in spite of less contributions than anchor residues, also have some influences on antigen peptides binding with MHC I proteins.

(iii) From the points of distributions of nonbonding interactions, hydrophobic interactions and hydrogen bond exert large effects on most antigen peptide positions, electrostatic interaction is prominent in the middle region and van der Waals interaction is insignificant.



Key






- | | | | |
|---|------------------------------|---|--|
|  | Residues of first surface |  | Residues involved in hydrophobic contact(s) |
|  | Residues of second surface |  | Corresponding atoms involved in hydrophobic contact(s) |
|  | Hydrogen bond and its length | | |

Figure 8. A schematic representation of the hydrophobic interactions and hydrogen bonds for MHC-peptide complex (PDB ID 1zt1, produced using LIGPLOT [59]).

ACKNOWLEDGEMENTS

This work was supported by National Natural Science Fund (grant number 30371339 and 30571748) and the National Project 863 Fund (grant number 2006AA02Z312).

REFERENCES

- [1] Chou, K.C. (2004) *Curr. Med. Chem.*, 11, 2105.
- [2] Chou, K.C. (2004) *Biochem. Biophys. Res. Comm.*, 316, 636.
- [3] Chou, K.C. (2004) *J. Proteome Res.*, 3, 1284.
- [4] Chou, K.C., Wei, D.Q. and Zhong, W.Z. (2003) *Biochem. Biophys. Res. Comm.*, 308, 148.
- [5] Zhang, R., Wei, D.Q., Du, Q.S. and Chou, K.C. (2006) *Med. Chem.*, 2, 309.
- [6] Gao, W.N., Wei, D.Q., Li, Y., Gao, H., Xu, W.R., Li, A.X. and Chou, K.C. (2007) *Med. Chem.*, 3, 221.
- [7] Zheng, H., Wei, D.Q., Zhang, R., Wang, C., Wei, H. and Chou, K.C. (2007) *Med. Chem.*, 3, 488.
- [8] Li, Y., Wei, D.Q., Gao, W.N., Gao, H., Liu, B.N., Huang, C.J., Xu, W.R., Liu, D.K., Chen, H.F. and Chou K.C. (2007) *Med. Chem.*, 3, 576.

- [9] Wang, J.F., Wei, D.Q., Chen, C., Li, Y. and Chou, K.C. (2008) *Protein Pept. Lett.*, 15, 27.
- [10] Sirois, S., Wei, D.Q., Du, Q.S. and Chou, K.C. (2004) *J. Chem. Inf. Comput. Sci.*, 44, 1111.
- [11] Chou, K.C., Wei, D.Q., Du, Q.S., Sirois, S. and Zhong, W.Z. (2006) *Curr. Med. Chem.*, 13, 3263.
- [12] Du, Q.S., Mezey, P.G. and Chou, K.C. (2005) *J. Comput. Chem.*, 26, 461.
- [13] Du, Q.S., Huang, R.B., Wei, Y.T., Du, L.Q. and Chou, K.C. (2008) *J. Comput. Chem.*, 29, 211.
- [14] Chou, K.C. (1993) *J. Biol. Chem.*, 268, 16938.
- [15] Chou, K.C. (1996) *Anal. Biochem.*, 233, 1.
- [16] Du, Q.S., Wang, S.Q., Jiang, Z.Q., Gao, W.N., Li, Y.D., Wei, D.Q. and Chou, K.C. (2005) *Med. Chem.*, 1, 209.
- [17] Shen, H.B. and Chou, K.C. (2008) *Anal. Biochem.*, 375, 388.
- [18] Chou, K.C. and Shen, H.B. (2006) *J. Proteome Res.*, 5, 1888.
- [19] Chou, K.C. and Shen, H.B. (2006) *Biochem. Biophys. Res. Comm.*, 347, 150.
- [20] Chou, K.C. and Shen, H.B. (2007) *J. Proteome Res.*, 6, 1728.
- [21] Shen, H.B. and Chou, K.C. (2007) *Biochem. Biophys. Res. Commun.*, 355, 1006.
- [22] Jin, Y., Niu, B., Feng, K.Y., Lu, W.C., Cai, Y.D. and Li, G.Z. (2008) *Protein Pept. Lett.*, 15, 286.
- [23] Chou, K.C. and Shen, H.B. (2008) *Nat. Prot.*, 3, 153.
- [24] Chou, K.C. and Shen, H.B. (2007) *Biochem. Biophys. Res. Comm.*, 360, 339.
- [25] Shen, H.B. and Chou, K.C. (2007) *Biochem. Biophys. Res. Comm.*, 364, 53.
- [26] Xiao, X., Shao, S., Ding, Y., Huang, Z., Chen, X. and Chou, K.C. (2005) *J. Theor. Biol.*, 235, 555.
- [27] Wu, G. and Yan, S. (2008) *Protein Pept. Lett.*, 15, 144.
- [28] Shen, H.B. and Chou, K.C. (2007) *Biochem. Biophys. Res. Comm.*, 363, 297.
- [29] Chou, K.C. and Shen, H.B. (2007) *Biochem. Biophys. Res. Comm.*, 357, 633.
- [30] Brusic, V. and Petrovsky, N. (2003) *Novartis Foundation symposium*, 254, 3.
- [31] Shastri, N., Schwab, S. and Serwold, T. (2002) *Annu. Rev. Immunol.*, 20, 463.
- [32] Ruppert, J., Sidney, J., Celis, E., Kubo, R.T., Grey, H.M. and Sette, A. (1993) *Cell*, 74, 929.
- [33] Hagmann, M. (2000) *Science*, 290, 80.
- [34] Froloff, N., Windemuth, A. and Honig, B. (1997) *Protein Sci.*, 6, 1293.
- [35] Wan, S., Coveney, P.V. and Flower, D.R. (2005) *J. Immunol.*, 175, 1715.
- [36] Hou, T., McLaughlin, W., Lu, B., Chen, K. and Wang, W. (2006) *J. Proteome Res.*, 5, 32.
- [37] Doytchinova, I.A. and Flower, D.R. (2001) *J. Med. Chem.*, 44, 3572.
- [38] Doytchinova, I.A. and Flower, D.R. (2002) *Proteins*, 48, 505.
- [39] Lin, Z., Wu, Y., Zhu, B., Ni, B. and Wang, L. (2004) *J. Comput. Boil.*, 11, 683.
- [40] Guan, P., Doytchinova, I.A., Walshe, V.A., Borrow, P. and Flower, D.R. (2005) *J. Med. Chem.*, 48, 7418.
- [41] Zhou, P., Tian, F. and Li, Z. (2007) *Chem. Biol. Drug Des.*, 69, 56.
- [42] Kawashima, S., Ogata, H. and Kanehisa, M. (1999) *Nucleic Acids Res.*, 27, 368.
- [43] Lu, Y., Bulka, B., des Jardins, M. and Freeland, S.J. (2007) *Protein Eng. Des. Sel.*, 20, 347.
- [44] Kidera, A., Konishi, Y., Oka, M., Ooi, T. and Scheraga, H.A. (1985) *J. Protein Chem.*, 4, 23.
- [45] Hellberg, S., Sjostrom, M., Skagerberg, B. and Wold, S. (1987) *J. Med. Chem.*, 30, 1126.
- [46] Sandberg, M., Eriksson, L., Jonsson, J. and Wold, S. (1998) *J. Med. Chem.*, 41, 2481.
- [47] Zhou, P., Li, Z., Tian, F. and Zhang, M. (2006) *Acta Chimica Sinica*, 64, 2065.
- [48] Tian, F., Lv, F., Li, Y., Yang, Q. and Zhou, P. (2008) *Amino Acids*, 35, 418.
- [49] Blythe, M.J., Doytchinova, I.A. and Flower, D.R. (2002) *Bioinformatics*, 18, 434.
- [50] Hattotuwigama, C.K., Doytchinova, I.A. and Flower, D.R. (2005) *J. Chem. Inf. Model.*, 45, 1415.
- [51] Kellenberger, A., Roussel, and B. Malissen. (2005) *J. Immunol.*, 175, 3819.
- [52] Golbraikh, A. and Tropsha, A. (2002) *J. Mol. Graphics Mod.*, 20, 269.
- [53] DeLano, W.L. (2002) *The PyMOL molecular graphics system*, DeLano Scientific, San Carlos, CA, USA.
- [54] Tian, F., Zhou, P., Lv, F., Song, R. and Li, Z. (2007) *J. Pept. Sci.*, 13, 549.
- [55] Tropsha, A., Gramatica, P. and Gombar, V.K. (2003) *QSAR Comb. Sci.*, 22, 69.
- [56] Falk, K., Rotzschke, O., Stefanovic, S., Jung, G. and Rammensee, H.G. (1991) *Nature*, 351, 290.
- [57] Sarobe, P., Pendleton, C.D., Akatsuka, T.D., Engelhard, V.H., Feinstone, S.M. and Berzofsky, J.A. (1998) *J. Clin. Invest.*, 102, 1239.
- [58] Kubo, R.T., Sette, A., Grey, H.M., Appella, E., Sakaguchi, K., Zhu, N.Z., Arnott, D., Sherman, N., Shabanowitz, J. and Michel, H. (1994) *J. Immunol.*, 152, 3913.
- [59] Wallace, A.C., Laskowski, R.A. and Thornton, J.M. (1995) *Protein Eng.*, 8, 127.