Peptide Reagent Design Based on Physical and Chemical Properties of Amino Acid Residues

QI-SHI DU,[1,2,3] RI-BO HUANG,[1,4] YU-TUO WEI,[1] CHENG-HUA WANG,[1] KUO-CHEN CHOU[3]
[1]*Guangxi Key Laboratory of Subtropical Bioresource Conservation and Utilization,*
Guangxi University, Nanning, Guangxi, 530004, China
[2]*Institute of Bioinformatics and Drug Discovery, Tianjin Normal University,*
Tianjin, 300074, China
[3]*Gordon Life Science Institute, San Diego, California 92130, USA*
[4]*Guangxi Academy of Sciences, 98 Daling Road, Nanning, Guangxi 530004, China*

Received 23 January 2007; Revised 18 February 2007; Accepted 19 February 2007
DOI 10.1002/jcc.20732
Published online 20 April 2007 in Wiley InterScience (www.interscience.wiley.com).

Abstract: It has tremendous values for both drug discovery and basic research to develop a solid bioinformatical tool for guiding peptide reagent design. Based on the physical and chemical properties of amino acids, a new strategy for peptide reagent design, the so-called AABPD (amino acid based-peptide design), is proposed. The peptide samples in a training dataset are described by a series of HMLP (heuristic molecular lipophilicity potential) parameters and other physicochemical properties of amino acid residues that form a three-dimensional data matrix where each component is defined by three indexes: the first index refers to the peptide samples, the second to the amino acid positions, and the third to the amino acid parameters. The binding free energy between a peptide ligand and its protein receptor is calculated by a linear free energy equation through the physicochemical parameters, resulting in a set of simultaneous linear equations between the bioactivity of the peptides and the physicochemical properties of amino acids. An iterative double least square technique is developed for the solution of the three-dimensional simultaneous linear equation set to determine the amino acid position coefficients of peptide sequence and the physicochemical parameter coefficients of amino acid residues alternately. The two sets of coefficients thus obtained are used for predicting the bioactivity of other query peptide reagents. Two calculation examples, the peptide substrate specificity of the SARS coronavirus 3C-like proteinase and the affinity prediction for epitope-peptides with Class I MHC molecules are studied by using the peptide reagent design strategy.

© 2007 Wiley Periodicals, Inc. J Comput Chem 28: 2043–2050, 2007

Key words: peptides; amino acids; drug design; QSAR; SARS; class I MHC

Introduction

The elucidation of the human genome has revitalized the interest in using proteins or pieces of proteins (peptides) for the treatment of presently incurable diseases.[1–3] Human diseases for which the peptide-based drug therapies are applied include osteoporosis (calcitonin), diabetes (insulin), prostate cancer and endometriosis (gonadotropin-releasing hormone), acromegaly and ulcers (somatostatin), as well as hypothyroidism (thyrotropin-releasing hormone, TRH).[1,4] The brain, as a major control center, is an important target for many pharmaceutical drugs, such as opioid peptide analogues designed for the treatment of pain, neuropeptides and growth peptides for the treatment of neurological disorders. One of the recently most remarkable examples of peptide drugs is the HIV fusion inhibitor T-20, which is a 36-peptide derived from the structure of gp41 and practically used in clinical treatment for AIDS patients.[5] HIV entry inhibitors[6] are the first antiretroviral drugs in widespread clinical evaluation to target HIV replication at sites other than reverse transcriptase or protease as focused previously (see, e.g., Refs. 7–14). The HIV fusion inhibitors have received the fast-track designation by the United States Food and Drug Administration because of the resistance of HIV to the drugs targeting on reverse transcriptase and protease. The successful implication

Correspondence to: Q. S. Du; e-mail: duqishi@yahoo.com or K. C. Chou; e-mail: kchou@san.rr.com

Contract/grant sponsor: Chinese National Basic Research Program; contract/grant number: 2004CB719606

Contract/grant sponsors: Chinese National Science Foundation (NSFC), Tianjin Commission of Science and Technology

of diverse peptides with special physiological activity in medicinal treatment has also increased the interest in exploiting peptides, peptide analogues, and peptide mimetics as therapeutic drugs.[15–20]

Peptide reagent design may be one of the most challenging research topics in life science owing to the extremely large amount of possible arrangements. For example, for an octapeptide sample, the number of possible sequence arrangements[13,21] are $20^8 = 2.56 \times 10^{10}$; for designing a peptide regent of 36 amino acids, the number of possible amino acid sequence arrangement would be $20^{36} = 6.87 \times 10^{46}$. This is an astronomical number! Therefore, it is vitally important to develop a computational method to guide us for efficiently designing peptide reagents. The present study was initiated in an attempt to develop a rational peptide reagent design method on the basis of the physical and chemical properties of amino acids, particularly their heuristic molecular lipophilicity potential (HMLP) parameters.[22–24] Unlike the purely statistical techniques in drug design that only pursue the statistical results without physical and chemical rationale, the HMLP approach has a solid theoretical foundation. Moreover, the prediction equations not only can indicate the bioactivity of a query peptide reagent but also can reveal the physical and chemical features for each of the subsites along the peptide sequence.

Theory and Method

The calculation of interaction free energy between two peptides, or between a peptide and a protein, is a fundamental problem in protein science and medicinal chemistry. In this study, the peptide–protein interaction is simplified as the interaction between a peptide reagent P_i and a target peptide P_0, which is a segment of a protein receptor, or a set of amino acid residues at its active sites. The main idea of peptide–protein interaction is schematically illustrated in Figure 1, where panel A shows the peptide fusion inhibitors interacting with a part of the HIV-1 gp41 envelope glycoprotein, and panel B shows the amino acid residues of peptide substrates interacting with the active sites of SARS coronavirus main protease.

The binding free energy ΔG_i^0 between peptide P_i and its target peptide P_0 is considered as the summation of interactions from all residue pairs of two peptides at the corresponding sequence positions; i.e.,

$$\Delta G_i^0 = \sum_{\alpha=1}^{M} b_\alpha \Delta g_{i,\alpha} \tag{1}$$

where $\Delta g_{i,\alpha}$ is the interaction free energy of residue pair at position α of peptide P_i and target peptide P_0, and M is the total number of residue pairs involved. The binding free energy $\Delta g_{i,\alpha}$ from different residue pairs may have different weights to the total free energy ΔG_i^0 due to their different microenvironments and different roles in bioactivity. We use a set of sensitive coefficients $\{b_\alpha\}$ to describe their microenvironments of residues. The binding free energy $\Delta g_{i,\alpha}$ of residue pair α is described by a series of physical and chemical properties of amino acids. In this study we use the heuristic molecular lipophilicity potential (HMLP) parameters of amino acids through the following linear free energy equation[22–24]:

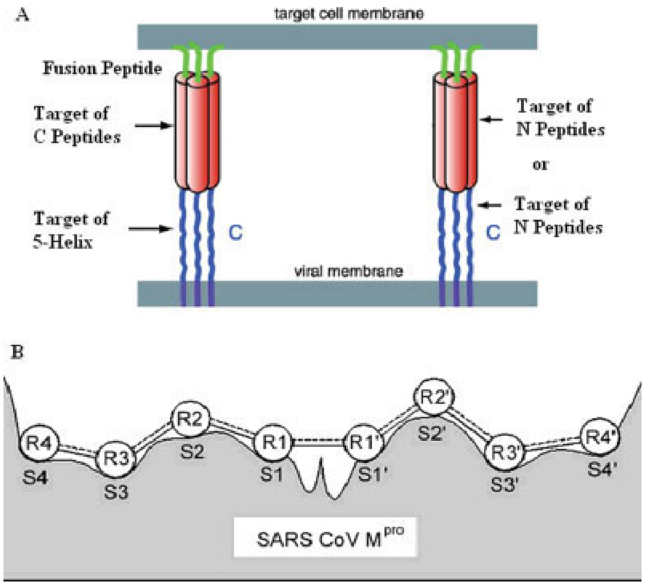

Figure 1. Protein and peptide targets of peptide reagents. (A) C peptides derived from the C-region of gp41, such as C34, bind to the N-region in its trimeric coiled-coil state; the proteins N_{CCG}-gp41 and 5-helix, which expose either the complete or a portion of the N-region trimeric coiled-coil in a stable form, bind to the C-region. N peptides may target a vulnerable C-helix region of gp41. Alternatively, the N peptides could intercalate with the N helices of gp41 to form a heterotrimeric coiled-coil and interfere with the coiled-coil formation of gp41. (B) Interaction between peptide substrate and SARS coronavirus main protease. The amino acid residues of peptide substrates interact with the active sites of SARS coronavirus main protease. Figure 1B was adapted from Figure 2 of Chou (Ref. 13, with permission from Academic Press). [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

$$\Delta g_{i,\alpha} = \frac{H_{0,\alpha} H_{i,\alpha}}{(R_{0,i,\alpha})^\gamma} + \frac{L_{0,\alpha} L_{i,\alpha}}{(R_{0,i,\alpha})^\gamma} + \frac{S_{0,\alpha}^{\mathrm{H}} S_{i,\alpha}^{\mathrm{H}}}{(R_{0,i,\alpha})^\gamma} + \frac{S_{0,\alpha}^{\mathrm{L}} S_{i,\alpha}^{\mathrm{L}}}{(R_{0,i,\alpha})^\gamma} + \cdots \tag{2}$$

where $R_{0,i,\alpha}$ is the distance between peptide P_i and peptide P_0 at residue position α, and γ is an exponent. In eq. (2), $H_{i,\alpha}$ and $L_{i,\alpha}$ are respectively the hydrophilic parameter and lipophilic parameter of residue α in peptide P_i; $S_{i,\alpha}^{\mathrm{H}}$ and $S_{i,\alpha}^{\mathrm{L}}$ are the surface areas of residue α with hydrophilic potential and lipophilic potential, respectively. Similarly, $H_{0,\alpha}$, $L_{0,\alpha}$, $S_{0,\alpha}^{\mathrm{H}}$, and $S_{0,\alpha}^{\mathrm{L}}$ are the HMLP parameters of target peptide P_0. The HMLP parameters of amino acids were derived from the quantum chemical electrostatic potential with the ability to reflect the hydrophilic and lipophilic interactions of amino acid residues.[25–27] The first term in eq. (2) describes the hydrophilic interactions, including the hydrogen bond interaction, ionic interaction, and dipole interaction; the second term describes the lipophilic interaction between amino acid residues; while the third and fourth terms describe the free energies from the solvation and dissolvation of amino acid residues, respectively. In eq. (2) we can add other terms that can affect the interaction or recognition between peptide ligand and protein receptor.

Actually, peptide P_0 is a virtual target peptide, and we do not need to know its real chemical structure. If the HMLP parameters $H_{0,\alpha}$, $L_{0,\alpha}$, $S^{H}_{0,\alpha}$, and $S^{L}_{0,\alpha}$ of the virtual target peptide P_0 and the distance $R_{0,i,\alpha}$ are assumed to be constants, eq. (2) will be reduced to a linear equation of the HMLP parameters of peptide P_i; i.e.,

$$\Delta g_{i,\alpha} = a_1 H_{i,\alpha} + a_2 L_{i,\alpha} + a_3 S^{H}_{i,\alpha} + a_4 S^{L}_{i,\alpha} + \cdots\cdots = \sum_{l=1}^{L} a_\alpha v_{i,\alpha,l} \tag{3}$$

where $v_{i,\alpha,l}$ denotes the l-th HMLP parameter of peptide P_i of amino acid residue at peptide position α. The role of the coefficients $\{a_l\}$ is the same as in traditional 2D-QSAR. Inserting the $\Delta g_{i,\alpha}$ of eq. (3) into eq. (1) and transferring the binding free energy ΔG^0_i to bioactivity $pk_i = -\log k_i = \Delta G^0_i$ of peptide P_i, we obtain the following simultaneous linear equations,

$$\sum_{\alpha=1}^{M} b_\alpha \left(\sum_{l=1}^{L} a_l v_{i,\alpha,l} \right) = pk_i \qquad (i = 1, 2, \ldots\ldots, N) \tag{4}$$

where N is the number of peptide samples, M is the number of amino acid residues in the peptides, and L is the number of physicochemical parameters of amino acid residues. In a training set of peptide reagents, the physicochemical parameters form a three-dimensional data matrix, $V_{N\times M\times L}$, as shown in Figure 2.

Although the transformation from eq. (1) to eq. (4) is not a rigorous theoretical derivation, it can be used to explain the physical implication of the linear free energy equation and the functions of HMLP parameters and some theoretical considerations in our model. Like all other QSAR approaches, the linear free energy equation is not unique. However, we can refine the binding free energy by utilizing other linear free energy equations and including more parameters.

In eq. (4) there are two sets of coefficients: $\{a_l\}$ are sensitive coefficients of the physicochemical parameters, and $\{b_\alpha\}$ are the sensitive coefficients of amino acid residue positions in peptides. An iterative double least square (IDLS) technique was developed to determine the values of the coefficient sets $\{a_l\}$ and $\{b_\alpha\}$ alternately by solving the three-dimensional simultaneous linear equations. By using a set of initial values of coefficients $\{a^{(0)}_l\}$, the three-dimensional data matrix $V_{N\times M\times L}$ is reduced to a two-dimensional data matrix $D^{(1)}_{N\times M}$ with the elements given by

$$d^{(1)}_{i,\alpha} = \sum_{l=1}^{L} a^{(0)}_l v_{i,\alpha,l} \tag{5}$$

Thus, the set of three-dimensional simultaneous linear equations [eq. (4)] is reduced to a set of two-dimensional equations, i.e.,

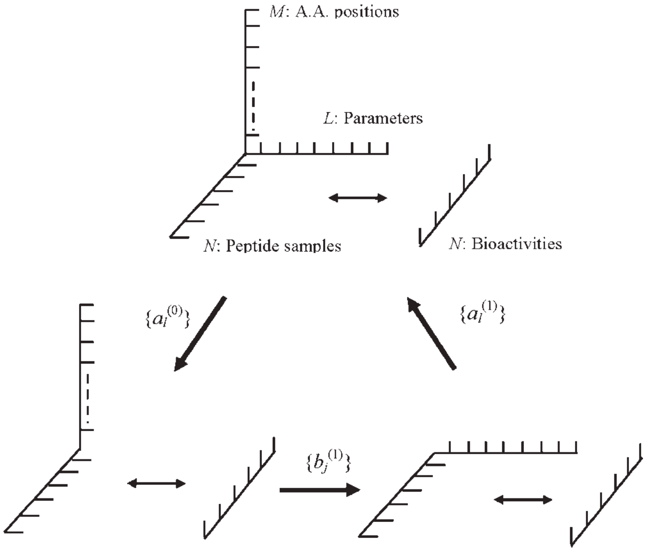

Figure 2. Schemetic illustration of the three-dimensional data matrix and iterative double least square (IDLS) solution. N is the number of peptide samples, M is the number of amino acid residues in peptide, and L is the number of physicochemical parameters of amino acid residues.

$$\sum_{\alpha=1}^{M} d^{(1)}_{i,\alpha} b^{(1)}_\alpha = pk_i \quad (i = 1, 2, \ldots\ldots, N) \tag{6}$$

The above equation set can be solved by using the least square approach, yielding the first solutions of the sensitive coefficients $\{b^{(1)}_a\}$. Then the values of $\{b^{(1)}_a\}$ are used to reduce the three-dimensional data matrix $V_{N\times M\times L}$ to a two-dimensional data matrix $T^{(1)}_{N\times L}$ with the elements given by

$$t^{(1)}_{i,l} = \sum_{\alpha=1}^{M} b^{(1)}_\alpha v_{i,\alpha,l} \tag{7}$$

Similarly, the set of three-dimensional simultaneous linear equations [eq. (4)] is reduced to a set of two-dimensional equations by eq. (7), as given by

$$\sum_{l=1}^{L} a^{(1)}_l t^{(1)}_{i,l} = pk_i \qquad (i = 1, 2, \ldots\ldots, N) \tag{8}$$

The above equation can be solved by using the least square approach, leading to the solution of sensitive coefficients $\{a^{(1)}_l\}$. Then the values of $\{a^{(1)}_l\}$ are used for the new solutions of the sensitive coefficients $\{b^{(2)}_\alpha\}$ of amino acid residue positions. The above procedure is performed iteratively for n steps, i.e., until reaching the converged solutions as denoted by $\{a^{(n)}_l\}$ and $\{b^{(n)}_\alpha\}$. Now, the values of $\{a^{(n)}_l\}$ and $\{b^{(n)}_\alpha\}$ can be used to predict the bioactivities $pk^{(\mathrm{pred})}_i$ of the i-th peptide reagent through the following equation:

Table 1. HMLP Parameters[a] and Secondary Structural Potencies of 20 Amino Acids Side Chains.

A.A.	S^L (Å^2)	S^H (Å^2)	L	H	[b]P_α	[b]P_β	[b]P_c
Leu (L)	84.5476	0.0000	1.2906	0.0000	1.21	1.30	0.68
Ile (I)	88.6055	0.0000	1.1046	0.0000	1.08	1.60	0.66
Val (V)	77.8108	0.0000	0.5324	0.0000	1.06	1.70	0.62
Phe (F)	105.7054	11.2472	0.4412	−0.1195	1.13	1.38	0.71
Met (M)	70.3631	23.2299	1.0768	−0.3068	1.45	1.05	0.58
Trp (W)	133.6980	14.8820	0.8364	−0.4310	1.08	1.37	0.75
Ala (A)	34.7760	0.0000	0.1744	0.0000	1.42	0.83	0.70
Cys (C)	23.5563	30.4540	0.2479	−0.2402	0.70	1.19	1.18
Gly (G)	3.7616	0.0000	0.0208	0.0000	0.57	0.75	1.50
Tyr (Y)	80.9646	42.7160	0.4534	−0.5896	0.69	1.47	1.06
Thr (T)	46.7285	16.0490	1.4265	−0.4369	0.83	1.19	1.07
Ser (S)	26.0681	15.9613	0.2346	−0.6040	0.77	0.75	1.32
His (H)	82.1701	13.8631	0.8124	−0.7766	1.00	0.87	1.06
Gln (Q)	70.0876	17.8662	1.0036	−0.7211	1.11	1.10	0.86
Lys (K)	97.7144	8.0786	1.4600	−0.6229	1.16	0.74	0.98
Asn (N)	50.5075	17.7804	0.6396	−0.7211	0.67	0.89	1.35
Glu (E)	57.1582	25.5726	1.0315	−0.9298	1.51	0.37	0.84
Asp (D)	37.4173	25.2736	0.6058	−0.9298	1.01	0.54	1.20
Arg (R)	90.8008	35.3095	1.2424	−1.4797	0.98	0.93	1.04
Pro (P)	69.2297	0.0000	0.3226	0.0000	0.57	0.55	1.59

[a] Notations used: S^L, surface area of side chain with lipophilic potential; S^H, surface area of side chain with hydrophilic potential; L, lipophilic index of side chains; H, hydrophilic index of side chains; P_α, α potency index; P_β, β potency index; P_c, coil potency index.

$$pk_i^{\mathrm{pred}} = \sum_{\alpha=1}^{M} b_\alpha^{(n)} \Delta g_{i,\alpha} = \sum_{\alpha=1}^{M} b_\alpha^{(n)} \left(\sum_{l=1}^{L} a_l^{(n)} v_{i,\alpha,l} \right) \qquad (9)$$

where the term $b_\alpha^{(n)} \Delta g_{i,\alpha}$ is the contribution of amino acid α of the i-th peptide reagent to the bioactivity. The convergence criterion for the iterative procedure is given by the following equation

$$\left|Q^{(n+1)} - Q^{(n)}\right| = \left| \sqrt{\frac{1}{N}\sum_{l=1}^{N}\left(pk_i^{expt} - pk_i^{(n+1)-pred}\right)^2} - \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(pk_i^{expt} - pk_i^{(n)-pred}\right)^2} \right| \le \varepsilon \quad (10^{-6}) \qquad (10)$$

where $Q^{(n)}$ represents difference between the square root of the summation of squared differences between experimental bioactivities and the predicted bioactivities in the nth step, and $Q^{(n+1)}$ that in the $(n + 1)$th step.

Calculations and Results

In this section two calculation examples are used to prove the predictive power of AABPD. The HMLP parameters of the 20 natural amino acid residues are taken from our previous work[22] and listed in Table1 with the secondary structural potency indices. One of the merits of the HMLP approach is that it can provide a lipophilic index and a hydrophilic index for every amino acid side chain, describing its lipophilic moiety and hydrophilic moiety, respectively. In the first example the 22 octapeptides in training set is taken from Ref. 28, which was originally used in the substrate specificity study for the SARS coronavirus 3C-like proteinase.[29,30] In Table2, the notation ↓ indicates the cleavage site in octapeptides. From ↓ to left side, the amino acid residues are numbered as R1, R2, R3, R4, and R5; from ↓ to right side, the amino acid residues are numbered as R1′, R2′, and R3′. The first octapeptide S12 (SAVLQ↓SGF-CONH2) is the parent of other octapeptides in the training set. It is a naturally cleavable peptide of 3C-like proteinase, derived from polyprotein pp1ab of SARS coronavirus, covering the first cleavage site of pp1ab.[31–34] Other octapeptides are mutated from octapeptide S12, in which one or more residues are replaced. In Table 2 the mutated residues are indicated with bold letters. The names of octapeptides used in this study are the same as in Ref. 28.

The iterative double least square (IDLS) technique described in the second Section is used in the peptide substrate specificity study for the SARS coronavirus 3C-like proteinase based on the experimental data listed in Table 2. The initial values of sensitive coefficients of physicochemical parameters $\{a_l^{(0)}\}$ are assigned to be 1.0. In the first calculation only four HMLP parameters are used, which yields the correlation coefficient R = 0.9375 and average predictive residue Q = ±0.4804. Q is the average square root of the summation of squared differences between predicted bioactivities and experimental bioactivities. Its definition can be found in eq. (10). In the second calculation total of 7 parameters are used, which gives better results than the first calculation (R = 0.9705 and Q = ±0.2305). The other

Table 2. Experimental and Predicted Peptide Substrate Cleavage Activities of SARS Coronavirus 3C-like Proteinase.

Name[a]	Sequence[b]	$(K_{cat}/K_m)_{rel}$[a]	$\log(K_{cat}/K_m)_{expt}$[a]	$\log(K_{cat}/K_m)_{calc}$[c]	$\log(K_{cat}/K_m)_{pred}$[d]
S12	SAVLQ↓SGF	1.0000	0.0000	−0.0338	−0.0396
P5L	**L**AVLQ↓SGF	3.9000	1.3610	1.7060	1.9394
P5T	**T**AVLQ↓SGF	3.6600	1.2975	1.3252	1.3306
P5V	**V**AVLQ↓SGF	3.5900	1.2782	1.0332	0.9280
P5A	**A**AVLQ↓SGF	3.3300	1.2030	0.6025	0.3333
P4L	S**L**VLQ↓SGF	0.1500	−1.8971	−0.8152	0.4788
P4T	S**T**VLQ↓SGF	1.4700	0.3853	−0.5455	−0.9877
P4V	S**V**VLQ↓SGF	2.4400	0.8920	−0.3388	−0.5486
P3L	SA**L**LQ↓SGF	0.8700	−0.1393	0.5149	1.0315
P3T	SA**T**LQ↓SGF	1.1900	0.1740	0.2043	0.2272
P3A	SA**A**LQ↓SGF	0.1900	−1.6607	−0.3850	0.3295
P3K	SA**K**LQ↓SGF	2.6800	0.9858	−0.0289	0.1694
P1′A	SAVLQ↓**A**GF	2.0400	0.7130	0.4130	0.1084
P1′G	SAVLQ↓**G**GF	0.8300	−0.1863	0.2582	Outlier
P2M	SAV**M**Q↓SGF	0.2080	−1.5702	−1.4607	−1.3500
P2F	SAV**F**Q↓SGF	0.0460	−3.0791	−3.5378	−4.5758
P2V	SAV**V**Q↓SGF	0.0056	−5.1850	−3.6989	−2.9612
P2A	SAV**A**Q↓SGF	0.0046	−5.3817	−6.0456	−5.5328
S21	**TV**VLQ↓SGF	3.9800	1.3813	1.0201	0.9197
S22	**TVT**LQ↓SGF	2.6200	0.9632	1.2582	1.4161
S23	**VVT**LQ↓SGF	2.4300	0.8879	0.9663	0.9465
S24	**TVK**LQ↓AGF	4.3100	1.4609	1.4717	1.5935

[a] The names of peptides are taken from Ref. 28. S12 is the parent peptide in the training set. All other peptides are mutated from it.

[b] The notation ↓ indicates the cleavage site of SARS coronavirus 3C-like proteinase.[13]

[c] Calculated bioactivities of 22 octapeptides after 10 iterative calculations.

[d] Predicted bioactivities of 21 octapeptides in Jackknife test.

three parameters are the secondary structural potency indices of amino acid residues. The curves of correlation coefficients R is shown in Figure 3, where R_a is for the iterated coefficients $\{a_l^{(n)}\}$ and R_b is for the iterated coefficients $\{b_\alpha^{(n)}\}$. The average residue Q between the predicted bioactivities and the experimental bioactivities of octapeptides are shown in Figure 4, where Q_a is for $\{a_l^{(n)}\}$ and Q_b is for $\{b_\alpha^{(n)}\}$. It has been observed that, after 5–6 iterations, the iterative result is converged smoothly. The converged sensitive coefficient sets $\{a_l^{(n)}\}$ and $\{b_\alpha^{(n)}\}$, as well as

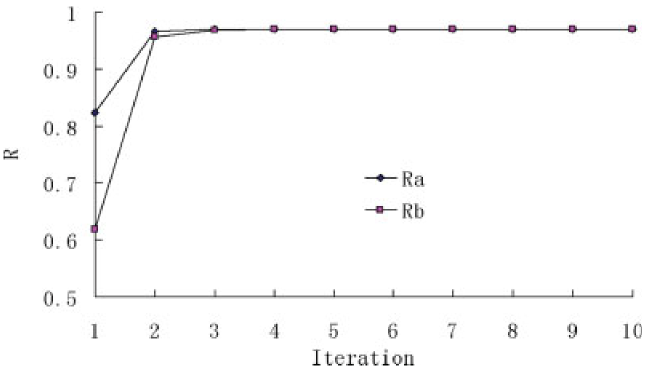

Figure 3. The correlation coefficients between experimental and predicted bioactivities. R_a is for $\{a_l^{(n)}\}$ iteration and R_b is for $\{b_j^{(n)}\}$ iteration. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

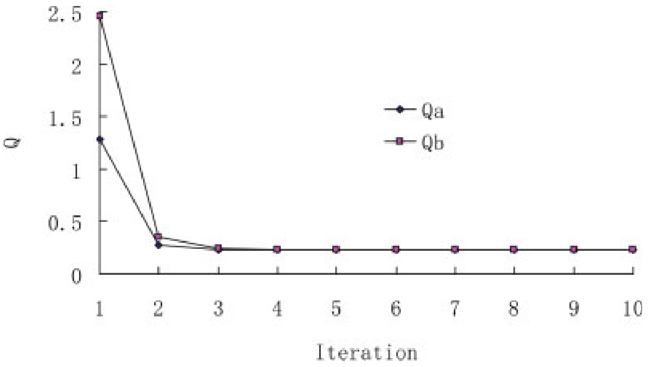

Figure 4. The value of Q between predicted bioactivities and experimental bioactivities of octapeptides. Q_a is for $\{a_l^{(n)}\}$ iteration and Q_b is for $\{b_j^{(n)}\}$ iteration. The Q is the average square root of the summation of squared differences between predicted bioactivities and experimental bioactivities (cf. Eq. 10). [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

Table 3. The Converged Sensitive Coefficient Sets $\{a_l^{(10)}\}$, $\{b_a^{(10)}\}$, Correlation Coefficient *R*, and Average Residue *Q* of 22 Octapeptides in Training Calculation and in Jackknife Test.

4 Parameters	7 Parameters
$a_1(S^L) = 0.16205$	$a_1(S^L) = 0.64288$
$a_2(S^H) = 0.94055$	$a_2(S^H) = -0.60573$
$a_3(L) = 45.11523$	$a_3(L) = 31.18381$
$a_4(H) = 78.24573$	$a_4(H) = 7.90862$
–	$a_5(P_\alpha) = 50.47151$
–	$a_6(P_\beta) = -26.15246$
–	$a_7(P_c) = -3.59095$
–	–
R	*R*
Q	Q
Jackknife Test	
R	–
Q	–
$b_1(R_5) = 0.01925$	$b_1(R_5) = 0.01875$
$b_2(R_4) = -0.01321$	$b_2(R_4) = -0.04513$
$b_3(R_3) = 0.01682$	$b_3(R_3) = 0.00510$
$b_4(R_2) = 0.10315$	$b_4(R_2) = 0.11854$
$b_5(R_1) = -0.75093$	$b_5(R_1) = -0.10948$
$b_6(R'_1) = 0.01418$	$b_6(R_1') = 0.00863$
$b_7(R'_2) = 26.69763$	$b_7(R_2') = -0.26087$
$b_8(R'_3) = -0.79655$	$b_8(R_3') = 0.05023$
0.9375	0.9705
±0.4804	±0.2305
0.8472	–
±0.6375	–

the corresponding *R* and *Q* for the octapeptides in the training dataset are given in Table 3.

The two sets of coefficients $\{a_l^{(n)}\}$ and $\{b_\alpha^{(n)}\}$ and eq. (4) can then be used for the bioactivity prediction of new peptide reagents. Jackknife test[35] is performed for evaluating the predictive power of AABPD approach. The predicted bioactivities of the 22 octapeptides obtained in the Jackknife test are listed in Table 2 with correlation coefficient $R = 0.8472$ and predictive residue $Q = \pm 0.7375$. In the Jackknife test we found that the octapeptide P1′G (SAVLQ↓GGF) is an outlier with an very high predicted bioactivity. This may be caused by the small training dataset with limited diversity. In octapeptide P1′G the residue serine (S) in position R1′ of the parent octapeptide S12 (SAVLQ↓SGF) is replaced by glycine (G). On this position only two mutations are made. In the Jacknife test for P1′G in training

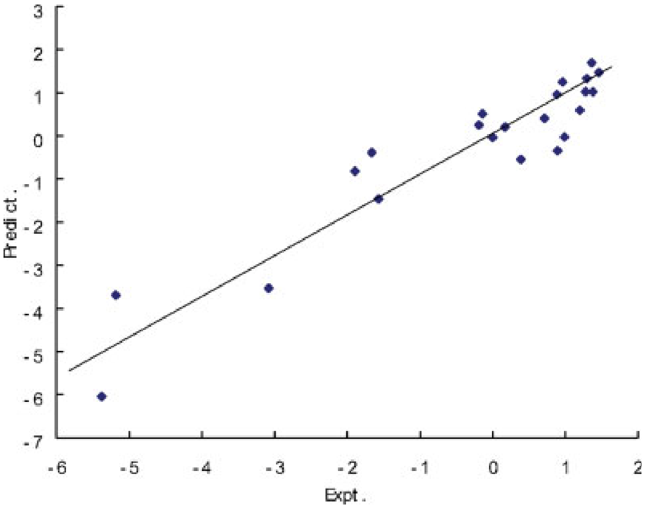

Figure 5. The experimental and predicted bioactivities of octapeptides in training set of 22 octapeptides after 10 iterations. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

Table 4. Experimental and Predicted Bioactivities of 40 Epitope-Peptides.

Name	Sequence	Exp. pIC_{50}[a]	Calc. pIC_5[b]	Pred. pIC_{50}[c]
P01	WLEPGPVTA	6.082	6.341	6.582
P02	ITSQVPFSV	6.196	6.495	6.568
P03	FLEPGPVTA	6.898	7.163	7.465
P04	ITAQVPFSV	7.020	6.799	6.726
P05	YLEPGPVTL	7.058	7.364	7.498
P06	YTDQVPFSV	7.066	7.498	7.758
P07	YLEPGPVTI	7.187	7.375	7.443
P08	YLEPGPVTV	7.342	7.453	7.461
P09	YLSPGPVTA	7.383	7.305	7.287
P10	IIDQVPFSV	7.398	7.667	7.707
P11	ITWQVPFSV	7.463	7.362	7.348
P12	ITYQVPFSV	7.480	7.302	7.234
P13	ILSQVPFSV	7.699	7.797	7.815
P14	IMDQVPFSV	7.719	7.717	7.717
P15	YLMPGPVTV	7.932	8.056	8.074
P16	WLDQVPFSV	7.939	7.574	7.285
P17	YLAPGPVTA	8.032	7.608	7.526
P18	YLYPGPVTV	8.051	7.998	7.975
P19	YLWPGPVTV	8.125	8.058	8.090
P20	ILYQVPFSV	8.310	8.603	8.648
P21	ILDQVPFSV	8.481	7.827	7.713
P22	YLFPGPVTA	8.495	8.131	8.069
P23	YLDQVPFSV	8.638	8.799	8.892
P24	ILFQVPFSV	8.699	8.623	8.605
P25	ILWQVPFSV	8.770	8.663	8.649
P26	WTDQVPFSV	6.145	6.273	6.381
P27	YLEPGPVTA	6.668	7.567	7.748
P28	ITDQVPFSV	6.947	6.526	6.643
P29	ITFQVPFSV	7.179	7.321	7.344
P30	FTDQVPFSV	7.212	7.094	7.028
P31	ITMQVPFSV	7.398	7.360	7.357
P32	YLSPGPVTV	7.642	7.191	7.124
P33	YLYPGPVTA	7.772	8.111	8.147
P34	YLAPGPVTV	7.818	7.495	7.466
P35	ILAQVPFSV	7.939	8.100	8.119
P36	ILMQVPFSV	8.125	8.661	8.749
P37	YLFPGPVTV	8.237	8.017	7.981
P38	YLMPGPVTA	8.367	8.169	8.135
P39	YLWPGPVTA	8.495	8.171	8.112
P40	FLDQVPFSV	8.658	8.395	8.191

[a] The experimental bioactivities pIC_{50} are taken from Ref. 38.
[a] Calculated bioactivities pIC_{50} of 9-peptides in training set.
[b] Predicted bioactivities pIC_{50} of 9-peptides in Jackknife test.

set there is only one mutated octapeptide P1′A on the position R1′, no sufficient information on this position. The comparison between experimental and predictive bioactivities of the 21 octapeptides are shown in Figure 5.

Peptide-based vaccines, in which small peptides derived from target proteins (epitopes) are used to provoke an immune reaction, have attracted considerable attention as a potential means both of treating infectious diseases and promoting the destruction of cancerous cells by a patient's own immune system.[36,37] With the availability of large sequence databases, computer aided design of peptide-based vaccines has emerged as a promising approach for screening among billions of possible immune-active peptides to find those likely to provoke an immune response to a particular cell type. The second example is the affinity prediction for epitope-peptides with class I MHC (major histocompatibility complex) molecules. Forty samples of 9-peptides are used in the training set, which are taken form Ref. 38. The 9-peptide sequences, the experimental bioactivities pIC_{50}, and the predicted bioactivities in Jackknife test are shown in Table4. In the training calculation we get the correlation coefficient $R = 0.8988$ and average residue $Q = \pm 0.6246$, and in the Jackknife test we get the correlation coefficient $R = 0.8285$ and predictive average residue $Q = \pm 0.7949$.

Discussion

The theoretical model of AABPD is built upon the structure features of peptides with clear physical implications. A merit of the peptide reagent design method developed in this paper is that the binding free energy between peptide reagent P_i and the target peptide P_0 is described by the physicochemical parameters of amino acids at every sequence site. It is through eq. (4) that has made it possible to not only predict the bioactivities of new peptide reagents, but also describe the physical and chemical features of an amino acid at every sequence position. This is very helpful for designing peptide reagents, peptide analogues, as well as peptide mimetics and modified peptides for drug development. In the traditional QSAR only one set of predictive coefficients $\{a_l\}$ is used, that is for the physical parameters in linear free energy equation. However, in the AABPD model two sets of predictive coefficients $\{a_l\}$ and $\{b_\alpha\}$ are used for physical parameters and for the position of residues in peptide, respectively. Two least square procedures are performed for $\{a_l\}$ and $\{b_\alpha\}$ alternately and iteratively. In this way the predictive residue Q decreases and the correlation coefficient R increases step by step. In the first calculation example the first correlation coefficient $R^{(0)} = 0.6175$ is the results of traditional QSAR method and the converged correlation coefficient $R^{(10)} = 0.9705$ is the improved result with the AABPD method. Usually after 5–6 iteration steps, the iterative procedure converged smoothly for two sensitive coefficient sets $\{a_l\}$ and $\{b_\alpha\}$. However, in the case the diversification of peptides is very poor we may get very bad solution, because of the matrix singularity problem. To build a good training set with proper diversification, two criteria have to be satisfied. First, the range of binding affinities in the test set should not exceed the range of affinities in the training set. Second, each amino acid at each position in the test set should also be present at that position in the training set in different peptides.

The iterative double least square (IDLS) technique has been used in more than 10 peptide systems. In all these cases we got converged results. So far we have not observed any unconverged example. In the IDLS technique general inverse matrix method is applied for the least square solution of simultaneous linear equation set. This is a very robust method. Only in the case the diversification of peptide training set is very poor, we may get a very bad solution because of the matrix singularity problem.

Conclusion

The linear free energy equation plays an important role in the QSAR study. The physicochemical parameters used in the linear free energy equation should describe the binding free energy from all aspects and include all factors that affect the interaction and recognition between ligand and receptor. Generally speaking, more physical parameters are used in linear free energy equation, better results may be achieved. However, sometimes additional parameters do not work harmonically and may conflict each other. For rational drug design, chemists developed many good parameters for QSAR study. However, the physicochemical parameters of amino acid residues for peptide design are not so ready. For further improvement of AABPD, better linear free energy equations and more optimized physicochemical parameters are needed.

Acknowledgments

The authors like to expresses their gratitude to the two anonymous reviewers, whose constructive comments were very helpful for improving the presentation of this paper.

References

1. Adessi, C.; Soto, C. Curr Med Chem 2002, 9, 963.
2. Venkatesan, N.; Kim, B. H. Curr Med Chem 2002, 9, 2243.
3. Chou, K. C. Curr Med Chem 2004, 11, 2105.
4. Sawyer, T. K. In ACS Professional Reference Book; Reid, R. E., Ed.; ACS Publisher: New York, 2000; Vol. 17, pp. 387–422.
5. Tremblay, C. L.; Kollmann, C.; Giguel, F.; Chou, T. C.; Hirsch, M. S. J Acquir Immune Defic Syndr 2000, 25, 99.
6. Nagashima, K. A.; Thompson, D. A.; Rosenfield, S. I.; Maddon, P. J.; Dragic, T.; Olson, W. C. J Infect Dis 2001, 183, 1121.
7. Althaus, I. W.; Chou, J. J.; Gonzales, A. J.; Diebel, M. R.; Chou, K. C.; Kezdy, F. J.; Romero, D. L.; Aristoff, P. A.; Tarpley, W. G.; Reusser, F. J Biol Chem 1993, 268, 6119.
8. Althaus, I. W.; Gonzales, A. J.; Chou, J. J.; Diebel, M. R.; Chou, K. C.; Kezdy, F. J.; Romero, D. L.; Aristoff, P. A.; Tarpley, W. G.; Reusser, F. J Biol Chem 1993, 268, 14875.
9. Althaus, I. W.; Chou, J. J.; Gonzales, A. J.; Diebel, M. R.; Chou, K. C.; Kezdy, F. J.; Romero, D. L.; Aristoff, P. A.; Tarpley, W. G.; Reusser, F. Biochemistry 1993, 32, 6548.
10. Althaus, I. W.; Chou, J. J.; Gonzales, A. J.; Diebel, M. R.; Chou, K. C.; Kezdy, F. J.; Romero, D. L.; Aristoff, P. A.; Tarpley, W. G.; Reusser, F. Experientia 1994, 50, 23.

11. Althaus, I. W.; Chou, J. J.; Gonzales, A. J.; Diebel, M. R.; Chou, K. C.; Kezdy, F. J.; Romero, D. L.; Thomas, R. C.; Aristoff, P. A.; Tarpley, W. G.; Reusser, F. Biochem Pharmacol 1994, 47, 2017.
12. Althaus, I. W.; Chou, K. C.; Franks, K. M.; Diebel, M. R.; Kezdy, F. J.; Romero, D. L.; Thomas, R. C.; Aristoff, P. A.; Tarpley, W. G.; Reusser, F. Biochem Pharmacol 1996, 51, 743.
13. Chou, K. C. Anal Biochem 1996, 233, 1.
14. Chou, K. C.; Kezdy, F. J.; Reusser, F. Anal Biochem 1994, 221, 217.
15. Bewley, C. A.; Louis, J. M.; Ghirlando, R.; Clore, G. M. J Biol Chem 2002, 277, 14238.
16. Eckert, D. M.; Kim, P. S. Proc Natl Acad Sci USA 2001, 98, 11187.
17. Root, M. J.; Kay, M. S.; Kim, P. S. Science 2001, 291, 884.
18. Doms, R. W.; Moore, J. P. J Cell Biol 2000, 151, F9.
19. Doms, R. W. Aids 2001, 15 Suppl 1, S34.
20. Chen, J.; Skehel, J. J.; Wiley, D. C. Proc Natl Acad Sci USA 1999, 96, 8967.
21. Chou, K. C. J. Biol. Chem. 1993, 268, 16938–16948.
22. Du, Q. S.; Li, D. P.; He, W. Z.; Chou, K. C. J Comput Chem 2006, 27, 685.
23. Du, Q. S.; Mezey, P. G.; Chou, K. C. J Comput Chem 2005, 26, 461.
24. Du, Q. S.; Liu, P. J.; Mezey, P. G. J Chem Inf Model 2005, 45, 347.
25. Du, Q.; Mezey, P. G. J Comput Aided Mol Des 1998, 12, 451.
26. Du, Q.; Arteca, G. A.; Mezey, P. G. J Comput Aided Mol Des 1997, 11, 503.
27. Du, Q.; Arteca, G. A. J Comput Aided Mol Des 1996, 10, 133.
28. Fan, K.; Ma, L.; Han, X.; Liang, H.; Wei, P.; Liu, Y.; Lai, L. Biochem Biophys Res Commun 2005, 329, 934.
29. Anand, K.; Ziebuhr, J.; Wadhwani, P.; Mesters, J. R.; Hilgenfeld, R. Science 2003, 300, 1763.
30. Yang, H.; Yang, M.; Ding, Y.; Liu, Y.; Lou, Z.; Zhou, Z.; Sun, L.; Mo, L.; Ye, S.; Pang, H.; Gao, G. F.; Anand, K.; Bartlam, M.; Hilgenfeld, R.; Rao, Z. Proc Natl Acad Sci USA 2003, 100, 13190.
31. Chou, K. C.; Wei, D. Q.; Zhong, W. Z. Biochem Biophys Res Comm 2003, 308, 148.
32. Du, Q. S.; Wang, S. Q.; Wei, D. Q.; Zhu, Y.; Guo, H.; Sirois, S.; Chou, K. C. Peptides 2004, 25, 1857.
33. Du, Q. S.; Wang, S.; Wei, D. Q.; Sirois, S.; Chou, K. C. Anal Biochem 2005, 337, 262.
34. Du, Q. S.; Wang, S. Q.; Jiang, Z. Q.; Gao, W. N.; Li, Y. D.; Wei, D. Q.; Chou, K. C. Med Chem 2005, 1, 209.
35. Chou, K. C.; Zhang, C. T. Crit Rev Biochem Mol Biol 1995, 30, 275.
36. Garcia, K. C. Immunol Rev 1999, 172, 73.
37. Jardetzky, T. S.; Lane, W. S.; Robinson, R. A.; Madden, D. R.; Wiley, D. C. Nature 1991, 353, 326.
38. Parkhurst, M. R.; Salgaller, M. L.; Southwood, S.; Robbins, P. F.; Sette, A.; Rosenberg, S. A.; Kawakami, Y. J Immunol 1996, 157, 2539.