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 CHOU3
1Guangxi Key Laboratory of Subtropical Bioresource Conservation and Utilization, Guangxi University, Nanning, Guangxi, 530004, China
2Institute of Bioinformatics and Drug Discovery, Tianjin Normal University, Tianjin, 300074, China
3Gordon Life Science Institute, San Diego, California 92130, USA
4Guangxi 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 protease 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 (thryrotropin-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, neuroptides 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.2 HIV entry inhibitors6 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

© 2007 Wiley Periodicals, Inc.
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 are \(20^8 = 2.56 \times 10^{10}\), for designing a peptide reagent of 36 amino acids, the number of possible amino acid sequence arrangement would be \(20^{36} = 6.87 \times 10^{56}\). 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, a computational method to guide us for efficiently designing peptide reagents, 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 \(\Delta G_i^0\) 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.,

\[
\Delta G_i^0 = \sum_{\alpha=1}^{M} b_{\alpha} \Delta \xi_{i,\alpha}
\]

(1)

where \(\Delta \xi_{i,\alpha}\) is the interaction free energy of residue pair at position \(\alpha\) of peptide \(P_i\) and target peptide \(P_0\) and \(M\) is the total number of residue pairs involved. The binding free energy \(\Delta G_i^0\) from different residue pairs may have different weights to the total free energy \(\Delta G_i^0\) 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 \(\Delta \xi_{i,\alpha}\) of residue pair \(\alpha\) 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 equation22–24:

\[
\Delta \xi_{i,\alpha} = H_{i,\alpha} H_{0,\alpha} \left(\frac{1}{R_{i,\alpha}}\right)^{\gamma} + L_{i,\alpha} L_{0,\alpha} \left(\frac{1}{R_{i,\alpha}}\right)^{\gamma} + S_{i,\alpha} S_{0,\alpha} \left(\frac{1}{R_{i,\alpha}}\right)^{\gamma} + \ldots
\]

(2)

where \(R_{i,\alpha}\) is the distance between peptide \(P_i\) and peptide \(P_0\) at residue position \(\alpha\), and \(\gamma\) is an exponent. In eq. (2), \(H_{i,\alpha}\) and \(L_{i,\alpha}\) are respectively the hydrophilic parameter and lipophilic parameter of residue \(\alpha\) in peptide \(P_i\). \(S_{i,\alpha}\) and \(S_{0,\alpha}\) are the surface areas of residue \(\alpha\) with hydrophilic potential and lipophilic potential, respectively. Similarly, \(H_{0,\alpha}\), \(L_{0,\alpha}\), \(S_{0,\alpha}^H\), and \(S_{0,\alpha}^L\) 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 dissolution 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 \( H_{0,\alpha}, L_{0,\alpha}, S_{0,\alpha} \), and \( S_{0,\alpha}^T \) of the virtual target peptide \( P_0 \) and the distance \( R_{0,\alpha} \) are assumed to be constants, eq. (2) will be reduced to a linear equation of the HMLP parameters of peptide \( P_i \); i.e.,

\[
\Delta G_{i,\alpha} = a_1 H_{i,\alpha} + a_2 L_{i,\alpha} + a_3 S_{i,\alpha}^1 + a_4 S_{i,\alpha}^2 + \cdots = \sum_{l=1}^L a_0 v_{i,\alpha l} 
\]

where \( v_{i,\alpha l} \) denotes the \( l \)-th HMLP parameter of peptide \( P_i \) of amino acid residue at peptide position \( \alpha \). The role of the coefficients \( \{a_i\} \) 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 \( \Delta G_i \) to bioactivity \( \log p_{ki} = -\Delta G_i \) of peptide \( P_i \), we obtain the following simultaneous linear equations,

\[
\sum_{\alpha=1}^M b_{\alpha} \left( \sum_{l=1}^L a_0 v_{i,\alpha l} \right) = p_{ki} \quad (i = 1, 2, \ldots, N) \tag{4}
\]

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, \( V_{N,M,L} \) 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: \( \{a_i\} \) 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 \( \{a_i\} \) and \( \{b_{\alpha}\} \) alternately by solving the three-dimensional simultaneous linear equations. By using a set of initial values of coefficients \( \{a_i^{(0)}\} \), the three-dimensional data matrix \( V_{N,M,L} \) is reduced to a two-dimensional data matrix \( D_{N,M,L} \) with the elements given by

\[
d_{i,\alpha}^{(1)} = \sum_{l=1}^L a_0 v_{i,\alpha l} \tag{5}
\]

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

\[
\sum_{\alpha=1}^M b_{\alpha} d_{i,\alpha}^{(1)} = p_{ki} \quad (i = 1, 2, \ldots, N) \tag{6}
\]

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

\[
t_{i,\alpha}^{(1)} = \sum_{\alpha=1}^M b_{\alpha}^{(1)} v_{i,\alpha l} \tag{7}
\]

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

\[
\sum_{l=1}^L a_1 v_{i,\alpha l} = p_{ki} \quad (i = 1, 2, \ldots, N) \tag{8}
\]

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

\[
M: A.A. \text{positions} \\
L: \text{Parameters} \\
N: \text{Peptide samples} \\
N': \text{Bioactivities}
\]

\[
\begin{align*}
\{a_i^{(n)}\} \\
\{b_{\alpha}^{(n)}\}
\end{align*}
\]

\[
\begin{align*}
\{a_i^{(n)}\} \\
\{b_{\alpha}^{(n)}\}
\end{align*}
\]

\[
\begin{align*}
\{a_i^{(n)}\} \\
\{b_{\alpha}^{(n)}\}
\end{align*}
\]

Figure 2. Schematic 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.
where the term $b_{i}^{a}(\Delta G_{i,a})$ is the contribution of amino acid $a$ of the $i$-th peptide reagent to the bioactivity. The convergence criterion for the iterative procedure is given by the following equation

$$
|Q^{(n+1)} - Q^{(n)}| = \frac{1}{N} \sum_{i=1}^{N} \left( pk_{i}^{\text{exp}} - pk_{i}^{(n+1)-\text{pred}} \right)^{2} - \frac{1}{N} \sum_{i=1}^{N} \left( pk_{i}^{\text{exp}} - pk_{i}^{(n)\text{pred}} \right)^{2} \leq \varepsilon \quad (10)
$$

where $Q^{(n)}$ represents the difference between the square root of the summation of squared differences between experimental bioactivities and the predicted bioactivities in the $n$th 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 and listed in Table 1 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 protease. In Table 2, the notation $\downarrow$ indicates the cleavage site in octapeptides. From $\downarrow$ to left side, the amino acid residues are numbered as R1, R2, R3, R4, and R5; from $\downarrow$ to right side, the amino acid residues are numbered as R1', R2', and R3'. The first octapeptide S12 (SAVLQSGF-CONH2) is the parent of other octapeptides in the training set. It is a naturally cleavable peptide of 3C-like protease, derived from polyprotein pp1ab. 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 protease based on the experimental data listed in Table 2. The initial values of sensitive coefficients of physicochemical parameters $\{a_{i}^{(0)}\}$ 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 $Q = \pm 0.4804$. $Q$ 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 $Q = \pm 0.2305$). The other

---

**Table 1.** HMLP Parameters and Secondary Structural Potencies of 20 Amino Acids Side Chains.

| A.A. | $S_{L}^{i}$ (Å$^2$) | $S_{H}^{i}$ (Å$^2$) | $L$ | $H$ | $\beta_{p_{\alpha}}$ | $\beta_{p_{\beta}}$ | $\beta_{p_{\gamma}}$ |
|------|-----------------|-----------------|-----|-----|----------------|----------------|----------------|
| 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 | 1.0768 | 0.0000 | 1.06 | 1.70 | 0.62 |
| Phe (F) | 105.7054 | 11.2472 | 0.4412 | 0.0000 | 0.95 | 1.07 | 0.62 |
| Met (M) | 70.3631 | 0.0000 | 1.0768 | 0.0000 | 1.06 | 1.70 | 0.62 |
| Trp (W) | 133.6980 | 14.8820 | 0.8364 | 0.0000 | 0.95 | 1.07 | 0.62 |
| Ala (A) | 34.7760 | 0.0000 | 0.1744 | 0.0000 | 1.42 | 0.83 | 0.70 |
| Cys (C) | 23.5563 | 30.4540 | 0.2479 | 0.0000 | 1.06 | 1.70 | 0.62 |
| Gly (G) | 80.9646 | 42.7160 | 1.0036 | 0.0000 | 0.69 | 1.47 | 1.06 |
| Asp (D) | 37.4173 | 25.2736 | 0.6058 | 0.0000 | 0.57 | 0.75 | 1.50 |
| Arg (R) | 90.8008 | 35.3095 | 1.2424 | 0.0000 | 1.45 | 1.05 | 0.58 |
| Glu (E) | 57.1582 | 25.5726 | 1.0315 | 0.0000 | 1.05 | 1.70 | 0.62 |
| Ser (S) | 26.0681 | 15.9613 | 0.2346 | 0.0000 | 0.97 | 0.75 | 1.32 |
| His (H) | 82.1701 | 13.8631 | 0.8124 | 0.0000 | 0.77 | 1.47 | 1.06 |
| Lys (K) | 97.7144 | 8.0786 | 1.4600 | 0.0000 | 1.16 | 0.74 | 0.78 |
| Asn (N) | 50.5075 | 17.7804 | 0.6396 | 0.0000 | 0.82 | 1.10 | 0.86 |
| Thr (T) | 46.7285 | 16.0490 | 1.4265 | 0.0000 | 1.00 | 0.87 | 1.06 |
| Pro (P) | 69.2297 | 0.0000 | 0.3226 | 0.0000 | 0.57 | 0.55 | 1.59 |

*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; $\beta_{p_{\alpha}}$, $\alpha$ potency index; $\beta_{p_{\beta}}$, $\beta$ potency index; $\beta_{p_{\gamma}}$, coil potency index.
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 $\{a^{(n)}_i\}$ and $R_b$ is for the iterated coefficients $\{b^{(n)}_i\}$. The average residue $Q$ between the predicted bioactivities and the experimental bioactivities of octapeptides are shown in Figure 4, where $Q_a$

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

| Name$^a$ | Sequence$^b$ | $(K_{cat}/K_m)_{rel}^c$ | log$(K_{cat}/K_m)_{exp}^c$ | log$(K_{cat}/K_m)_{calc}^d$ | log$(K_{cat}/K_m)_{pred}^d$ |
|----------|--------------|------------------------|---------------------------|---------------------------|---------------------------|
| S12      | SAVLQ|SGF      | 1.0000                  | 0.0000                    | $-0.0338$                 | $-0.0396$                |
| P5L      | LAVLQ|SGF      | 3.9000                  | 1.3610                    | $1.7060$                  | $1.9394$                |
| P5T      | TAVLQ|SGF      | 3.6600                  | 1.2975                    | $1.3252$                  | $1.3306$                |
| P5V      | VAVLQ|SGF      | 3.5900                  | 1.2782                    | $1.0332$                  | $0.9280$                |
| P5A      | AAVLQ|SGF      | 3.3300                  | 1.2030                    | $0.6025$                  | $0.3333$                |
| P4L      | SLVVLQ|SGF | 0.1500                  | $-1.8971$                 | $-0.8152$                 | $0.4788$                |
| P4T      | STVLQ|SGF      | 1.4700                  | 0.3853                    | $-0.5455$                 | $-0.9877$                |
| P4V      | SVVLQ|SGF      | 2.4400                  | 0.8920                    | $-0.3388$                 | $-0.5486$                |
| P3L      | SALLQ|SGF      | 0.8700                  | $-1.1393$                 | $0.5149$                  | $1.0315$                |
| P3T      | SATLQ|SGF      | 1.1900                  | 0.1740                    | $0.2043$                  | $0.2272$                |
| P3A      | SAALQ|SGF      | 0.1900                  | $-1.6607$                 | $-0.3850$                 | $0.3295$                |
| P3K      | SAKLQ|SGF      | 2.6800                  | 0.9858                    | $-0.0289$                 | $0.1694$                |
| P'1A     | SAVLQ|AGF      | 2.0400                  | 0.7130                    | $0.4130$                  | $0.1084$                |
| P'1G     | SAVLQ|GGF      | 0.8300                  | $-0.1863$                 | $0.2582$                  | Outlier                |
| P2M      | SAVMQ|SGF      | 0.2080                  | $-1.5702$                 | $-1.4607$                 | $-1.3500$                |
| P2F      | SAVAQ|SGF      | 0.0460                  | $-3.0791$                 | $-3.5378$                 | $-4.5758$                |
| P2V      | SAVVQ|SGF      | 0.0056                  | $-5.1850$                 | $-3.6989$                 | $-2.9612$                |
| P2A      | SAVAQ|SGF      | 0.0046                  | $-5.3817$                 | $-6.0456$                 | $-5.5238$                |
| S21      | TVVLQ|SGF      | 3.9800                  | 1.3813                    | 1.0201                    | 0.1917                  |
| S22      | TVTLQ|SGF      | 2.6200                  | 0.9632                    | 1.2582                    | 1.4161                  |
| S23      | VVTLQ|SGF      | 2.4300                  | 0.8879                    | 0.9663                    | 0.9465                  |
| S24      | TVKLQ|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 \(\text{;}\) 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.

Figure 3. The correlation coefficients between experimental and predicted bioactivities. $R_a$ is for $\{a^{(n)}_i\}$ iteration and $R_b$ is for $\{b^{(n)}_i\}$ 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 $\{a^{(n)}_i\}$ iteration and $Q_b$ is for $\{b^{(n)}_i\}$ 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 \( \{a_{(20)}^1\}, \{b_{(20)}^1\} \)
Correlation Coefficient \( R \), and Average Residue \( Q \) of 22 Octapeptides in Training Calculation and in Jackknife Test.

| Parameters | 4 Parameters | 7 Parameters |
|------------|--------------|--------------|
| \( a_1(S^1) \) | 0.16205 | \( a_1(S^1) \) | 0.64288 |
| \( a_2(S^1) \) | 0.94055 | \( a_2(S^1) \) | -0.60573 |
| \( a_3(S^1) \) | 45.1152 | \( a_3(S^1) \) | 31.1838 |
| \( a_4(H) \) | 78.24573 | \( a_4(H) \) | 7.90862 |
| \( \ldots \) | \( \ldots \) | \( \ldots \) | \( \ldots \) |
| \( R \) | \( R \) | \( R \) | \( Q \) |
| \( Q \) | \( Q \) | \( Q \) | \( Q \) |
| \( b_1(R_1) \) | 0.001925 | \( b_1(R_4) \) | 0.01875 |
| \( b_2(R_1) \) | -0.001321 | \( b_2(R_4) \) | -0.04513 |
| \( b_3(R_1) \) | 0.001682 | \( b_3(R_4) \) | 0.00510 |
| \( b_4(R_1) \) | 0.10315 | \( b_4(R_2) \) | 0.11854 |
| \( b_5(R_1) \) | -0.75093 | \( b_5(R_4) \) | -0.10948 |
| \( b_6(R_1) \) | 0.001418 | \( b_6(R_4) \) | 0.00863 |
| \( b_7(R_1) \) | 26.69763 | \( b_7(R_4) \) | -0.26087 |
| \( b_8(R_1) \) | -0.79655 | \( b_8(R_4) \) | 0.05023 |
| 0.9375 | 0.9705 | \( \pm 0.4804 \) | \( \pm 0.2305 \) |
| \( 0.8472 \) | \( 0.8472 \) | \( \pm 0.6375 \) | \( \pm 0.6375 \) |

The corresponding \( R \) and \( Q \) for the octapeptides in the training dataset are given in Table 3.

The two sets of coefficients \( \{a_{(20)}^1\} \) and \( \{b_{(20)}^1\} \) 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 (SAVLQGGF) 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 (SAVLQSGF) is replaced by glycine (G). On this position only two mutations are made. In the Jackknife test for P1'G in training reagents.

![Figure 5](https://www.interscience.wiley.com)

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.]
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 immuneactive 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 from Ref. 38. The 9-peptide sequences, the experimental bioactivities pIC50, and the predicted bioactivities in Jackknife test are shown in Table 4. 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 \( P \) 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 \( \{a_i\} \) is used, that is for the physical parameters in linear free energy equation. However, in the AABPD model two sets of predictive coefficients \( \{a_i\} \) and \( \{b_n\} \) are used for physical parameters and for the position of residues in peptide, respectively. Two least square procedures are performed for \( \{a_i\} \) and \( \{b_n\} \) 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 \( \{a_i\} \) and \( \{b_n\} \). 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 express 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.; Gigué, 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. J Biol Chem 1993, 268, 6119.
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 Commun 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.