J. Mol. Model. **1999**, *5*, 189 – 195

Journal of Molecular Modeling

FULL PAPER

© Springer-Verlag 1999

Calculating Partition Coefficients of Peptides by the Addition Method

Peng Tao, Renxiao Wang, and Luhua Lai

Institute of Physical Chemistry, Peking University, Beijing 100871, P.R.China. Tel: +86-10-62756833;
Fax: +86-10-62751725. E-mail:lai@ipc.pku.edu.cn

Received: 16 July 1999/ Accepted: 20 September 1999/ Published: 19 October 1999

Abstract A new addition method is described in this study for calculating the partition coefficients of peptides. Log*P* and log*D* values of peptides are calculated by summing the contributions of the component amino acids. The final models are derived from a multivariate linear regression analysis of 219 peptides with known experimental data. The standard errors in a leave-one-out cross-validation are 0.23 and 0.24 log units for the log*P* and log*D* values, respectively. The predictive ability of the model is tested by an extra set of ten peptides, and the self-consistency of the model is further demonstrated by a new validation procedure called the evolution test. The parameters obtained in regression could be used as hydrophobicity scales for amino acids. The application of such hydrophobicity scales has also been discussed.

Keywords Partition coefficient, Peptide, Addition method, Fragment method, log*P*, log*D*, Evolution test

Introduction

Since the pioneering work of Hansch and Fujita [1], the logarithm of l-octanol/water partition coefficient (log*P*) has been successfully introduced into quantitative structure-activity relationship (QSAR) studies. It is widely used as a parameter to measure hydrophobicity in studying many biochemical, pharmacological, and environmental processes [2]. This has prompted extensive work on predicting log*P* values based solely on chemical structure [3-8].

Although they work well for common organic compounds, most of the current methods fail to calculate log*P* values of peptides accurately. Peptides and their analogs are often regarded as a very important class of potential therapeutic reagents [9]. Modeling the hydrophobicity properties of peptides is undoubtedly meaningful for drug design and discovery. Furthermore, establishing a set of hydrophobicity scales for amino acids will aid studies of three-dimensional protein structure and may provide insights into processes such as protein folding and binding [10,11].

Several hydrophobicity scales have been proposed for amino acids or peptides [12,13]. Steinmetz employed comparative molecular field analysis (CoMFA). [14] to analyze the octanol-water distribution coefficient of free and blocked amino acids [15]. Buchwald and Bordor used a size-based model to predict the octanol-water partition of non-zwitterionic peptides and other organic compounds [16]. Sotomatsu-Niwa and Ogino analyzed the experimentally determined log*P* values of free and blocked di- and tripeptides statistically to derive hydrophobicity parameters for amino acids [17]. Maybe the most convincing approach put forward so far is from Akamatsu et al. [18-21]. They have carefully measured the partition coefficients of a wide

Correspondence to: L. Lai

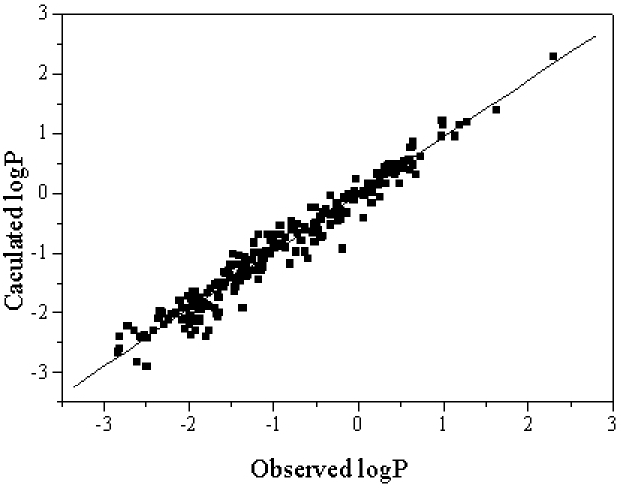


Figure 1 *Correlation between the experimental and calculated log*P *values of 219 peptides given by the residue addition model*

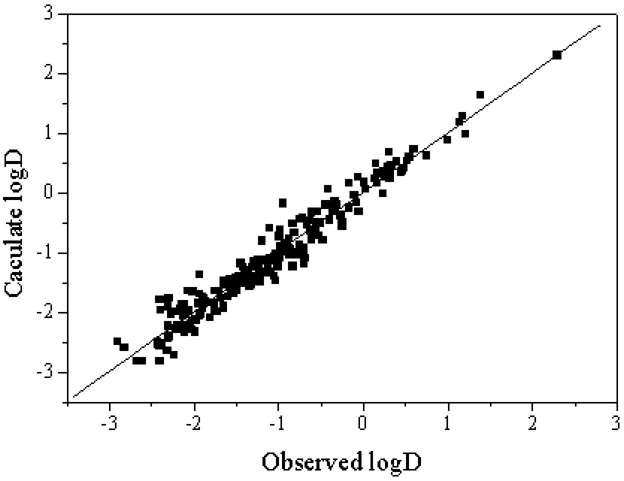


Figure 2 *Correlation between the experimental and calculated log*D *values of 216 peptides given by the residue addition model*

variety of peptides under controlled experimental conditions. After studying these data with linear regression analysis, they have obtained different regression models for different kinds of peptides. Various parameters are used in their models (i.e. structural effects, β-turn formation corrections, N- and C-terminal effects, etc.). A good correlation was observed between the experimental and calculated log*P* values.

In this paper, we describe new addition methods for calculating log*P* and log*D* values of peptides. Using the same data set as Akamatsu, we have derived much simpler regression models, which give comparable results. Furthermore, we have demonstrated that log*P* and log*D* values of peptides can be calculated reliably either by the amino acid addition or fragment addition method. The hydrophobicity parameters we have obtained are applicable to QSAR and protein modeling studies.

Method and computation results

Training set

We use 219 peptides with known 1-octanol/water partition coefficients as our training set (see Supplementary Material). These peptides range from di- to pentapeptides, including N-acetyl-peptide amides (usually referred as "blocked" peptides) and peptides with free N- and C-terminus (usually referred as "unblocked" or "free" peptides). The experimentally determined partition coefficients, log*P*, and distribution coefficients, log*D*, of these peptides are all taken from the literature [16-21]. Most of the data come from Akamatsu's work.

Residue addition model

We assume that the partition coefficient of a peptide can be calculated by summing the contribution of each component amino acid (we will refer to "component amino acid" as "residue" in the following text). Based on the regression analysis of the training set, we have obtained an equation for calculating log*P* values of peptides:

$$\log P = \sum_n a_n R_n^P + bB_P + uU_P \tag{1}$$

Here, a_n is the occurrence of the *n*th kind of residue; R^P_n is the log*P* contribution of the *n*th kind of residue; and *b* and *u* are indicator variables to account for different forms of peptides. When the sample compound is a blocked peptide, *b* is set to 1 and *u* is set to 0; while if the sample compound is an unblocked peptide, *b* is set to 0 and *u* is set to 1. B_p and U_P are the corrections for log*P* values of blocked and unblocked peptides, respectively. The regression analysis yielded: *n* = 219, r = 0.978, s = 0.21, and F = 189.8. The correlation between the observed and calculated log*P* values, given by equation 1, is illustrated in Figure 1. The slope and intercept of the fitted line are 0.96 and –0.04, respectively. The regression analysis results, i.e. *R*, *B*, and *U*, are listed in Table 1.

We have also obtained a similar equation for calculating log*D* values of peptides:

$$\log D = \sum_n a_n R_n^D + bB_D + uU_D \tag{2}$$

The meaning of the parameters in this equation is similar to the corresponding variable in equation 1. However, in equation 2, R^D_n, B_D, and U_D are concerned with log*D* values rather than log*P* values. The regression results of this equation are *n* = 216, r = 0.974, s = 0.21, F = 155.8. The correlation between the observed and calculated log*D* values given by equa-

Table 1 *Hydrophobicity contributions of 21 natural amino acids*

Amino acid	log*P* Contribution [a]		log*D* Contribution	
Ala	-0.27	(+/-0.06)	-0.27	(+/-0.07)
Arg	-0.79	(+/-0.21)	-1.65	(+/-0.22)
Asn	-0.98	(+/-0.22)	-0.98	(+/-0.22)
Asp	-0.28	(+/-0.21)	-2.06	(+/-0.22)
Cys	0.83	(+/-0.33)	0.82	(+/-0.34)
Gln	-1.00	(+/-0.21)	-1.00	(+/-0.22)
Glu	-0.34	(+/-0.21)	-2.19	(+/-0.22)
Gly	-0.22	(+/-0.06)	-0.22	(+/-0.06)
His	-0.31	(+/-0.19)	-0.44	(+/-0.20)
Ile	0.70	(+/-0.06)	0.69	(+/-0.06)
Leu	0.80	(+/-0.06)	0.80	(+/-0.06)
Lys	0.17	(+/-0.19)	-2.27	(+/-0.20)
Met	0.51	(+/-0.14)	0.51	(+/-0.14)
Phe	1.16	(+/-0.06)	1.16	(+/-0.06)
Pro	0.15	(+/-0.13)	0.15	(+/-0.13)
Ser	-0.45	(+/-0.15)	-0.45	(+/-0.15)
Thr	-0.26	(+/-0.14)	-0.26	(+/-0.14)
Trp	1.46	(+/-0.11)	1.46	(+/-0.11)
Tyr	0.55	(+/-0.09)	0.55	(+/-0.09)
Val	0.32	(+/-0.06)	0.32	(+/-0.06)
Orn	-0.29	(+/-0.21)	-2.17	(+/-0.27)
Blocked [b]	-1.19	(+/-0.12)	-1.18	(+/-0.12)
Unblocked [c]	-3.25	(+/-0.15)	-3.25	(+/-0.15)

[a] The values in brackets are 95% confidence interval
[b] For N-acetyl-peptide amides
[c] For free peptides

tion 2 is illustrated in Figure 2. The slope and intercept of the fitted line are 1.00 and 0.00, respectively. The regression analysis results, i.e. *R*, *B*, and *U*, are also listed in Table 1.

To test the predictive ability of the above two regression models, we have used equations 1 and 2 to perform leave-one-out cross-validation on the training set. The results obtained are q=0.974, s=0.23 for equation 1, and q=0.969, s=0.24 for equation 2, respectively.

Fragment addition model

The fragment addition method has been widely used in calculating log*P* values for common organic compounds [6,7,22,23]. We dissected the structures of the amino acids into 18 elementary chemical fragments (see Table 2). Either blocked or unblocked peptides can be represented as the assembly of these fragments. On the assumption that the log*D* value of a peptide can be calculated by summing the contribution of each component fragment, we have obtained the following equation by the multivariate regression analysis of the training set:

$$\log D = \sum_n a_n F_n \; + u U_D \qquad (3)$$

Here, F_n is the log*D* contribution of the *n*th kind of fragment; a_n is the occurrence of the *n*th kind of fragment; and *u* is an indicator variable. If the sample compound is an unblocked peptide, *u* is set to 1; otherwise it is set to 0. U_D is the correction for log*D* values of unblocked peptides. The regression analysis of the training set by using equation 3 yielded *n* = 216, r = 0.967, s = 0.24, and F = 150.8. The correlation between the observed and calculated log*D* values given by equation 3 is illustrated in Figure 3. The slope and intercept of the fitted line in Figure 3 are 0.90 and –0.10, respectively. The parameters derived for each fragment are listed in Table 2.

Test set

We used ten tetra- and pentapeptides as a test set (see Table 3). The experimental log*D* values of these peptides are taken from the literature [24]. We calculated the log*D* values for these peptides with the residue addition model, i.e. equation 2. The predictive correlation coefficient (r) is 0.929, and the standard deviation (s) is 0.47 log units. The correlation between the observed and calculated log*D* values of the test set is illustrated in Figure 4.

The aqueous buffer used in the log*D* measurement experiment for the test set (i.e. Hank’s balanced salt solution modified to contain 25 mM glucose and 10 mM HEPES, pH 7.35, 37°C)[24], is not the same as that used for the training set (i.e. 0.1M aqueous sodium phosphate plus phosphoric acid pH 7, ionic strength 0.1, 25°C) [20]. This is probably the reason why the standard deviation in calculating the test set is a little larger than the one in calculating the training set. However, considering that the average error in partition experiments is about 0.4 log units, the deviation of 0.47 log units is still acceptable. This shows that our model is robust

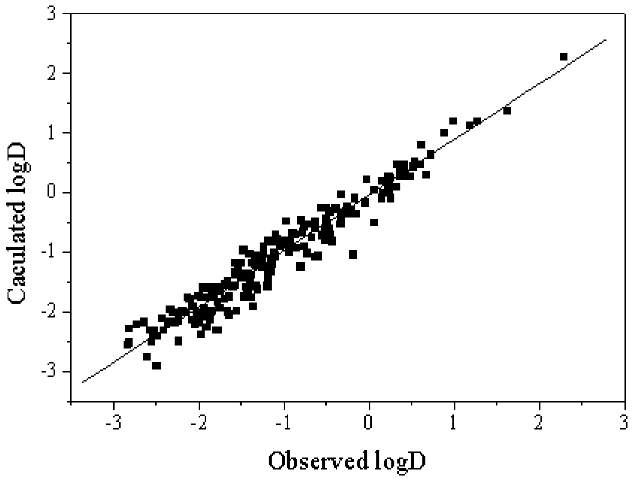


Figure 3 *Correlation between the experimental and calculated log*D *values of 216 peptides given by the fragment addition model*

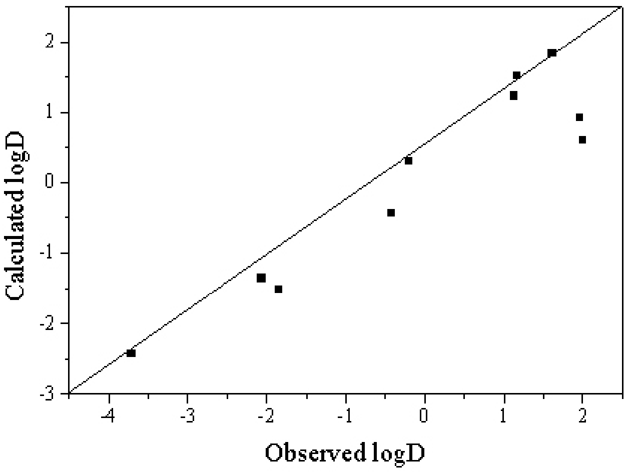


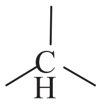
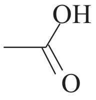
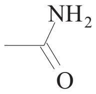
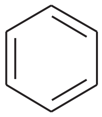
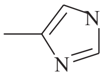
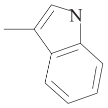
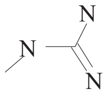
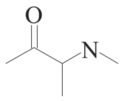
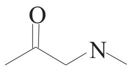
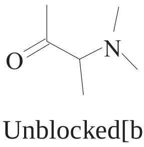
Figure 4 *The correlation between the experimental and calculated log*D *values of 10 peptides in the test set*

for predicting partition coefficients of peptides smaller than hexapeptides.

The program

Based on the final model, we have written a computer program, PLOGP, which can calculate log*P* and log*D* values of a given peptide and the MLP of protein with known 3D structure. This program is written in C language. Its source code and a detailed description are available in the Supplementary Material.

Table 2 *Hydrophobicity contributions of 18 fragments*

Fragments	Contribution	Confidence interval[a]
$-CH_3$	0.051	(+/-)0.140
$-CH_2-$	0.388	(+/-)0.064
	0.577	(+/-)0.188
-OH (in Ser)	-0.498	(+/-)0.192
-OH (in Thr)	-0.565	(+/-)0.214
-OH (in Tyr)	-0.604	(+/-)0.097
	-2.344	(+/-)0.214
	-1.244	(+/-)0.103
$-NH_2$	-3.306	(+/-)0.283
-S-	-0.007	(+/-)0.204
-SH	0.759	(+/-)0.360
	1.087	(+/-)0.141
	-0.506	(+/-)0.236
	1.401	(+/-)0.160
	-2.456	(+/-)0.306
	-0.323	(+/-)0.168
	-0.227	(+/-)0.069
	-0.986	(+/-)0.250
Unblocked[b]	-3.235	(+/-)0.167

[a] The confidence level is 95%.
[b] For free peptides.

Table 3 *Test set*

Peptide	Obs. log*D* [a]	Calc. log*D* [b]
Ac-Tyr-Pro-Ile-Asp-Val-N	-1.85	-1.52
Ac-Tyr-Pro-Gly-Asp-Val-N	-3.71	-2.44
Ac-Tyr-Pro-Ile-Asn-Val-N	-0.42	-0.44
Ac-Tyr-Pro-Gly-Asn-Val-N	-2.06	-1.36
Ac-Tyr-Pro-Ile-Ile-Val-N	1.13	1.23
Ac-Tyr-Pro-Gly-Ile-Val-N	-0.20	0.31
Ac-Phe-Pro-Ile-Ile-Val-N	1.61	1.84
Ac-Phe-Pro-Gly-Ile-Val-N	1.96	0.92
Ac-Phe-Pro-Ile-Ile-N	1.17	1.52
Ac-Phe-Pro-Gly-Ile-N	2.00	0.60

*[a] Observed log*D *value, cited from Ref. [24].*
*[b] Calculated log*D *value, given by equation 2.*

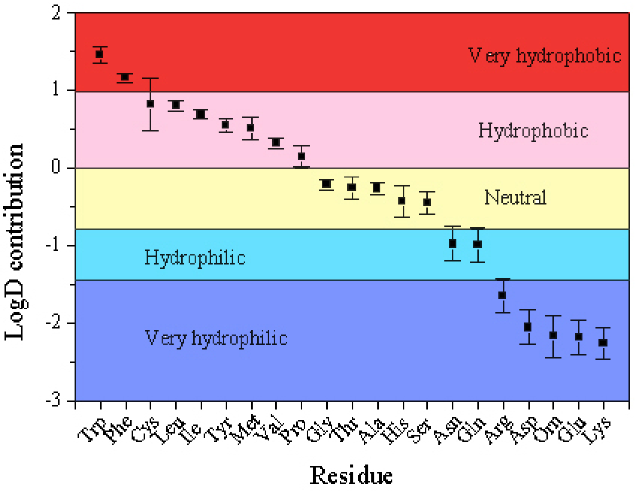


Figure 5 *The subdivision of 21 natural amino acids according to their log*D *contribution*

Discussion

Residue addition model

While reproducing the experimental log*P* and log*D* values satisfactorily, the residue addition model is rather simple and straightforward. By using this model, the regression analysis of the training set gives an "eigenvalue" for each kind of amino acid. This value represents the contribution of a specific amino acid to the partition coefficient and therefore can be regarded as its hydrophobicity scale. According to these hydrophobicity scales, we can roughly divide the 21 kinds of amino acids into five groups (see Figure 5). It is not surprising that those amino acids with aromatic side chains, e.g. Trp and Phe, are "very hydrophobic", while those amino acids with ionizable side chains, e.g. Asp, Glu, Orn, Lys, and Arg, are "very hydrophilic". However, it is noticeable that Ala, Gly, and Pro are not as hydrophobic as they are usually considered to be (these amino acid residues are typically treated as hydrophobic residues in approaches such as protein structure modeling). Furthermore, Thr and Ser are only slightly hydrophilic in spite of the existence of a hydroxy group in the side chain of each amino acid. All of these amino acids have relatively short side chains that cannot extend into the solvent. Perhaps that is why the hydrophobicity properties of these amino acids differ from the conventional concept.

Fragment addition model

Compared to the residue addition model, the fragment addition model provides further information on the distribution of hydrophobicity properties upon the whole molecule. For example, Lys is considered to be very hydrophilic (its contribution to the log*D* value is –2.27 according to the residue addition model). By using the fragment addition model, it can be seen very clearly that the significantly negative log*D* value of Lys comes mainly from the amino group (see Table 3). This is also true for Orn.

Using the hydrophobicity contribution of each fragment, we can calculate the Molecular Lipophilicity Potential (MLP) for peptides or even protein molecules. The MLP can provide a 3D profile to illustrate the spatial distribution of hydrophobicity properties around a molecule and is widely applied to QSAR and molecular docking approaches. The MLP value at a certain grid point *i* around the molecule is usually calculated as:

$$H_i = \sum_k \frac{h_k}{1 + r_{ik}} \tag{4}$$

where H_i is the hydrophobic potential at the *i*th grid point, h_k is the hydrophobicity contribution of the *k*th fragment in the molecule, and r_{ik} is the distance between the *i*th grid point and the geometric center of the *k*th fragment in the molecule (a minimum cutoff distance of 5 Å is imposed to avoid artificially large values of H_i).

As an example, we have calculated the MLP profile of the HIV-1 protease enzyme (PDB entry 1aaq), and show its MLP contour lines at level 1.2 in red (Figure 6). The HIV-1 protease enzyme has five rigid domains: two flap domains, two core domains and one terminal domain [25]. As indicated by red contour lines at a high MLP level (1.2), the space between the two core domains is the largest hydrophobic area in the whole molecule. Perhaps this means that hydrophobicity plays an important role in the stability of the dimer.

Evolution test

The two models we have described above, both the residue addition model and the fragment addition model, are empirical methods. They are derived from the regression analysis of a training set. Therefore, the final model is inevitably dependent on the training set. It is commonly believed that an ideal training set should contain adequate samples to guarantee the reliability of the final regression model. However, no consensus on the actual size of an ideal training set has been reached so far. If a training set contains N samples and the model to be studied includes M terms, it is generally accepted that N/M should be larger than 3 or 5 at the very minimum. This is too rough of an estimate to be a convincing standard. To resolve this problem, we have put forward a stepwise procedure, the evolution test, to investigate the relationship between the size of the training set and the predictive ability of the regression model.

We performed an evolution test for equation 2. The evolution procedure began by selecting a subset from the training set. As mentioned above, the training set contains a total of 219 samples. We randomly selected 50 samples from the training set to form a subset. Then we performed the regression analysis on this subset with equation 2 and therefore derived a regression model. Using this model, we calculated the log*D* values of the samples in the test set. We recorded the correlation coefficient (*q*). and the standard deviation (*s*) of the regression fitting of the subset. We also recorded the correlation coefficient ($q_{\_pred}$). and the standard deviation ($s_{\_pred}$). of the calculated and observed log*D* values of samples in the test set. Here $q_{\_pred}$ is defined as equation 5, representing the predictive ability of the regression model. To minimize the coincidence in such analysis, the whole process, (i.e. selection, regression, and prediction), was repeated 20 times and the average values of $q_{\_pred}$, q , $s_{\_pred}$ and s were recorded. Then the evolution procedure moved to the next step by increasing the size of the subset to 70 and performing all the analysis, again. Then the size of the subset increased to 90, 110, …, until the size of the training set itself was reached. Figure 7a shows the trends of q and $q_{\_pred}$ throughout the entire procedure, and Figure 7b shows the trends of s and $s_{\_pred}$, respectively.

$$q = \sqrt{1 - \frac{\sum (\log D_{pred} - \log D_{obs})^2}{\sum (\log D_{obs} - \log D_{mean})^2}} \qquad (5)$$

For a good stable model, we assume that the predictive ability will increase steadily with the increase in the size of the subset. The results of the evolution test on our model confirm this assumption. We also note that, when the size of the training set is larger than 100, the predictive ability of our model rises stably but slightly. This indicates that a training set containing 100 to 200 samples is sufficient for "training" the models we have proposed. Using an even larger training set is unnecessary.

Conclusion

In this study, we have demonstrated that partition coefficients, log*P* and log*D*, of oligo-peptides can be calculated reliably

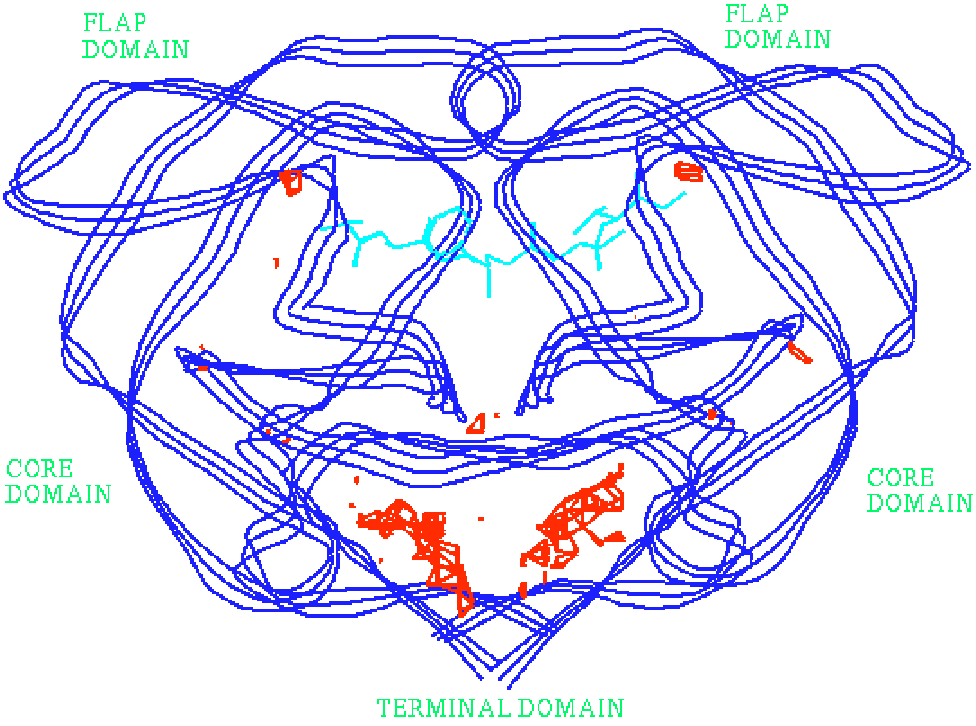


Figure 6 *MLP contours at level 1.2 of HIV-1 protease enzyme (PDB entry 1aaq). Five domains of HIV-1 protease are indicated by green notes. Each part of the figure is shaded: the contour lines are shaded red; the ribbon that represents the backbone of the protein is shaded blue; and the ligand molecule is shaded cyan.*

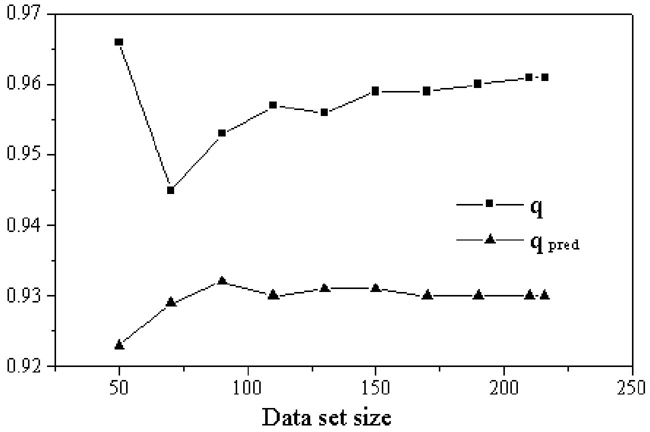


Figure 7a *The correlation coefficients found for the evolution test*

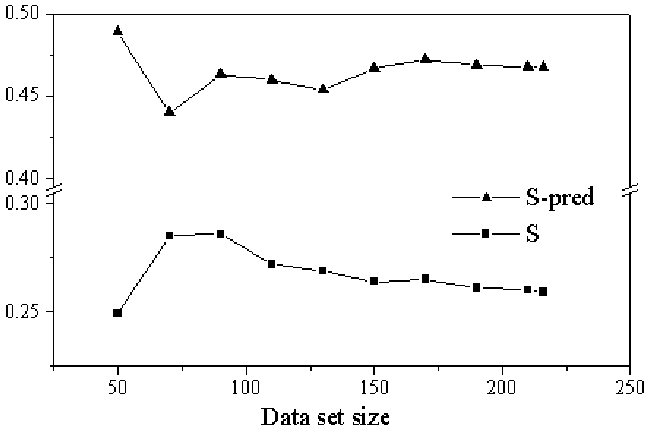


Figure 7b *The standard deviations found for the evolution test*

by either the residue addition model or the fragment addition model. Our addition models are derived from regression analysis of 219 peptides and validated by an extra test set. We have also introduced a new stepwise procedure, the evolution test, to test the self-consistency of the regression model. The hydrophobicity scales obtained by our models for 21 kinds of natural amino acids are valuable for QSAR studies and protein structure modeling.

Supplementary material available statement The training set, and the source codes of PLOGP (in C), are available from the authors.

References

1. Hansch, C.; Fujita, T. *J.Am.Chem.Soc.* **1964**, *86*, 1616-1626.
2. Fujita, T.; Iwamura, H. *Top.Curr.Chem.* **1983**, *114*, 119-157.
3. Leo, A. *Chem.Rev.* **1993**, *93*, 1281-1306.
4. Leo, A.; Hansch, C.; Elkins, D. *Chem.Rev.* **1971**, *71*, 525-616.
5. Chou, J.T.; Jurs, P.C. *J.Chem.Inf.Comput.Sci.* **1979**, *19*, 172-178.
6. Suzuki, T.; Kudo, Y. *J.Comput.-Aided Mol.Des.* **1990**, *4*, 155-198.
7. Klopman, G.; Li, J.-Y.; Wang, S. *J.Chem.Inf.Comput.Sci.* **1994**, *34*, 752-781.
8. Renxiao, W.; Ying, F.; Luhua, L. *J.Chem.Inf.Comput.Sci.* **1997**, *37*, 615-621.
9. Claassen, V.*Trends in Drug Research*; Elsevier Science: Amsterdam: 1990; pp 73-108.
10. Kauzman, W. *Adv.Protein Chem.* **1959**, *14*, 1-63.
11. Kyte, J.; Doolittle, R.F. *J.Mol.Biol.* **1982**, *157*, 105-132.
12. Fauchere, J.L.; Charton, M.; Kier, L.B.; Verloop, A.; Pliska, V. *Int.J.Pept.Protein Res.* **1988**, *32*, 269-278.
13. Abraham, D.J.; Leo, A. *J.Proteins Struct.Funct.Genet.* **1987**, *2*, 130-152.
14. Cramer, R.D.; Patterson, D.E.; Bunce, J.D. *J.Am.Chem.Soc.* **1988**, *110*, 5959-5967.
15. Steinmetz, W.E. *Quant.Struct.-Act.Relat.* **1995**, *14*, 19-23.
16. Buchwald, P.; Bodor, N. *Proteins: Struct.Funct.Genet.* **1998**, *30*, 86-99.
17. Sotomatsu-Niwa, T.; Ogino, A. *J.Mol.Struct.(Theochem).* **1997**, *392*, 43-54.
18. Akamatsu, M.; Yoshida, Y.; Nakamura, H.; Asao, M.; Iwamura, H.; Fujita, T. *Quant.Struct.-Act.Relat.* **1989**, *8*, 195-203.
19. Akamatsu, M.; Okutani, S.; Nakao, K.; Hong, N.J.; Fujita, T. *Quant.Struct.-Act.Relat.* **1990**, *9*, 189-194.
20. Akamatsu, M.; Fujita, T. *J.Pharm.Sci.* **1992**, *81*, 164-174.
21. Akamatsu, M.; Katayama, T.; Kishimoto, D.; Kurokawa, Y.; Shibata, H.; Ueno, T.; Fujita, T. *J.Pharm.Sci.* **1994**, *83*, 1026-1033.
22. Rekker, R.F.*The Hydrophobic Fragment Constant*; Elsevier: New York: 1977.
23. Moriguchi, I.; Hirono, S.; Liu, Q.; Nakagome, I.; Matsushita, Y. *Chem.Pharm.Bull.* **1992**, *40*, 127-130.
24. Sorensen, M.; Steenberg, B.; Knipp, G.T.; Wang, W.; Steffansen, B.; Frokjaer, S.; Borchardt, R.T. *Pharm.Res.* **1997**, *14*, 1341-1348.
25. Rose, R.B.; Craik, C.S.; Stroud, R.M. *Biochemistry* **1998**, *37*, 2607-2621.