I.J. Engineering and Manufacturing **2011, 1, 47-56**
Published Online February 2011 in MECS (http://www.mecs-press.net)
DOI: 10.5815/ijem.2011.01.08

Available online at http://www.mecs-press.net/ijem

A Simple, Yet Rapid and Effective Method for LogP Prediction of Dipeptides Based on Theoretical Descriptors (HMLP)

Jiajian Yin [a,*], Yong Liu [a]

[a] *Department of chemistry, College of life and science, Sichuan Agricultural University, Yaan, P.R.China,625014*

Abstract

The hydrophobicity of peptide is an important factor that affects the dissolution behavior of proteins and peptides, also affect the physical and chemical properties. In this study, each amino acid side chain was characterized using three structure parameters (heuristic molecular lipophilicity potential, HMLP). The HMLP parameters, total surface area(S), lipophilic indices (L), and hydrophilic indices (H) of amino acid side chains are derived from theoretical computation. Based on HMLP descriptors, QSAR models of the logP were constructed for blocked and unblocked dipeptides by multiple linear regression (MLR) and support vector regression (SVR). All the results showed that the logP relates to the total surface area(S) and hydrophilic indices (H), and the prediction results of SVR are better than that of MLR. The prediction results are in agreement with the experimental values. The result shows HMLP parameters (S,L,H) could preferably describe the structure features of the peptides responsible for their octanol to water partition behavior.

Index Terms: HMLP descriptors; peptides; logP; QSAR; support vector regression

1. Introduction

Peptide analogues with high activity, high selectivity and small side effects widely exist in nature. Bioactive peptides is a very active research area in recent years, its application is related to biology, medicine, chemistry and other disciplines, shows a bright future. While the hydrophobicity (logP) of peptides is a very important parameter for reasonable design of bioactive peptide and discovery of peptide drugs. Therefore, it is important meaningful to study the relationships between peptide logP and its structure by theory tool both in academic significance and application value. The literatures [1-5] built a linear model for calculation of logP

* Corresponding author.
E-mail address: yinjiajian_100@163.com

by the hydrophobic parameter π and index variable of amino acids. Tao Peng et al. [6] had been constructed the multivariate linear regression logP prediction model of peptides using the contribution addition method of amino acid side chain. Gulyaeva N. et al. [7] estimated the relative hydrophobicity of the peptide with the equivalent additon principles of the number of methylene units. Darren R. F. et al. [8,9] studied comparatively for prediction procedures of peptides logP values with seven methods based on peptide molecules fragments addition and one method based on the whole property of peptide molecules, and constructed multiple linear regression and partial least squares regression prediction models. The literatures utilized parameters of amino acid structural information characterization to discuss the peptides logP, and some beneficial exploring work was made by them, while these methods have shown certain degree of success in peptide logP prediction.

However, in these documents, the calculation method to the structural characteristics of peptides is too complicated and time-consuming, and the prediction results were not satisfactory by linear model methods. In addition, the quantitative description for peptide structure characteristics is difficult and key issue. Therefore, in this paper, we try to introduce heuristic molecular lipophilicity potential (HMLP) parameters [10] in order to characterize the peptide structural information, and to build new descriptors to represent the whole peptide chain structure information with direct substitution. The support vector regression (SVR) prediction model of dipeptides logP value has successfully constructed. The results also indicate that SVR can be used as an alternative powerful modeling tool for peptide QSAR studies, and give one advice (LOPO) about evaluating the importance of parameter in SVR model. Moreover, one view was pointed out that the logP value of the peptides relates mainly to its molecular surface area and hydrophilicity index. This will be provided with certain guidance meaning to design and exploit peptide analogues.

2. Methods

2.1. Amino acid descriptors and structure information representation of peptide

Amino acid descriptors HMLP (S, L, H) [10] were selected in the article and the code is 20 natural amino acid (AA) single letter symbols. See Table 1. The HMLP parameters, total surface area(S), lipophilic indices (L), and hydrophilic indices (H) of amino acid side chains are derived from lipophilicity potential, respectively. The HMLP has clear physical and chemical meaning and provides useful lipophilic and hydrophilic parameters for the studies of proteins and peptides [10]. The HMLP parameters were studied in the affinity prediction for epitope-peptides with Class I MHC molecules [11], and bioactivities prediction of proteins and peptides [12]. For a set of peptide analogues, the chemical structure can now be quantified by describing each varied amino acid position with three HMLP descriptors. So the chemical structure of a dipeptides, for example, can be described by 2×3 variables. Thus, a set of peptide analogues varied in n positions can be described by $n \times 3$ descriptors, namely, S, L, and H. The amino acid at the amino terminus was designed as n_1, and its properties were described as n_1S, n_1L, and n_1H; the amino acid adjacent to the amino terminus was designed as n_2, and its properties were described as n_2S, n_2L, and n_2H, etc.

Table 1 Components HMLP descriptor of 20 natural occurring amino acids

AA	**S/nm^2**	**L**	**H**	**AA**	**S/nm^2**	**L**	**H**
Ala(A)	0.3478	0.1744	0.0000	Leu(L)	0.8455	1.2906	0.0000
Arg(R)	1.2611	1.2424	-1.4797	Lys(K)	1.0579	1.46	-0.6229
Asn(N)	0.6829	0.6396	-0.7211	Met(M)	0.9359	1.0768	-0.3068
Asp(D)	0.6269	0.6058	-0.9298	Phe(F)	1.1695	0.4412	-0.1195
Cys(C)	0.5401	0.2479	-0.2402	Pro(P)	0.6923	0.3226	0.0000
Gln(Q)	0.8795	1.0036	-0.7211	Ser(S)	0.4203	0.2346	-0.604
Glu(E)	0.8273	1.0315	-0.9298	Thr(T)	0.6278	1.4265	-0.4369
Gly(G)	0.0376	0.0208	0.0000	Trp(W)	1.4858	0.8364	-0.431
His(H)	0.9603	0.8124	-0.7766	Tyr(Y)	1.2368	0.4534	-0.5896
Ile(I)	0.8861	1.1046	0.0000	Val(V)	0.7781	0.5324	0.0000

2.2. Data processing

In this article, multiple linear regression (MLR) and support vector regression (SVR) were used to build linear and nonlinear models of the peptides QSAR by leave-one-out cross validation (LOOCV).

1) Multiple linear regression

First, multiple linear regression method was used to detect relationship between peptides logP and structural descriptors HMLP of amino acids, which aims to build a linear model, namely,

$$\log P = b_0 + \sum b_{ij} S + \sum b_{ij} L + \sum b_{ij} H \tag{1}$$

Where S_i, L_i, and H_i refer to the ith amino acid residue, and the b_{ij} refers to the coefficients which will be given by the multiple linear regression analysis of the entire data set.

2) Support vector regression

In this paper, the support vector regression (SVR) was used to build the nonlinear model for the prediction of the peptides logP. Originally, SVM is developed for pattern recognition problems. And now, with the introduction of ε-insensitive loss function, SVM has been extended to solve nonlinear regression estimation, function approximation and time series prediction. The most attractive characteristics of SVM are the absence of local minima, the sparseness of the solution, and the use of the kernel-induced feature spaces. LogP of peptides prediction problem has been looked upon a complicated non-linear function relation approximation solution between logP value and impact factors, so we attempt to construct peptides logP prediction model by SVR. The algorithm of SVR has been shown in the literature [13,14].

3) Validation and evaluation of model

Multiple linear regression and support vector regression (SVR) method were used for all QSAR analysis. In order to find the optimum QSAR model, prohibit the over-fitting of the model and so as to have the best prediction performance, the LOOCV of the whole dataset is performed. The goodness of the model was accessed by the following statistical parameters: the prediction root mean square error (RMSE), and finally by the modeling standard and cross-validated correlation coefficient R^2 and $Q^2_{(CV)}$, respectively.

$$RMSE = \sqrt{\frac{\sum_{1}^{n} (\log P_{\exp} - \log P_{pre})^2}{n}}$$

$$R^2 = 1 - \sum_{i=1}^{n} (\log P_{\exp} - \log P_{cal})^2 / \sum_{i=1}^{n} (\log P_{\exp} - \log P_{avl})^2$$

$$Q^2_{(CV)} = 1 - \sum_{i=1}^{n} (\log P_{\exp} - \log P_{pre})^2 / \sum_{i=1}^{n} (\log P_{\exp} - \log P_{ave})^2 \tag{2}$$

Where n is the number of peptides in the training set, $logP_{exp}$ is experiment value ,$logP_{pre}$ is predict value by LOOCV, $logP_{cal}$ is calculation value by MLR and $logP_{ave}$ is average value of $logP_{exp}$. Generally, without a high R^2, it is impossible to obtain a high $Q^2_{(CV)}$. A $Q^2_{(CV)}>0.5$ is regarded as good model and a $Q^2_{(CV)}>0.9$ as excellent, and difference between R^2 and $Q^2_{(CV)}$ ought not to exceed 0.3[15].

In addition, For the sake of examining infection of each amino acid property VS logP value in the peptides at the SVR model, the importance of each parameter or property is estimated by the value of the model RMSE obtained by using leave-one-parameter-out(LOPO) approach($RMSE_{lopo}$) subtract that($RMSE_{lopo}$) of original whole parameter based on estab- lishing optimal model in turn, namely:

$$\Delta RMSE = RMSE_{lopo} - RMSE_{whole} \quad (3)$$

Where, if $\Delta RMSE$ is positive value, indicating that this parameter or properties to logP of peptides have major impact, and the greater the positive, shows the bigger the implication, so the parameter preserved; if $\Delta RMSE$ is negative value, indicating that this parameter or property to logP of peptides have no importance and account for the greater the negative, the smaller the implication, so the parameter canceled. Thus the expected amino acid properties in each position are evaluated according to their importance to the logP of peptides.

4) Experiment condition

The MLR algorithm is implemented by using MATLAB6.5 program, while the SVR algorithm is implemented by using LIBSVM-MAT program[16] in the article.

3. Results and Discussions

3.1. LogP for blocked dipeptides

The sequences and the logP value of the studied peptides were collected from the literature [1-5], which can be seen in Table 2.

Table 2 The sequence and logP of blocked DIpeptides

NO.	peptide	Exp.LogP	NO.	peptide	Exp.LogP	NO.	peptide	Exp.LogP
1	GV	-1.33	13	VA	-1.14	25	NI	-1.43
2	AV	-1.13	14	YV	-0.2	26	NF	-1.14
3	LV	0.26	15	YL	0.32	27	LN	-1.30
4	IV	0.16	16	YF	0.54	28	IN	-1.41
5	GF	-0.56	17	WV	0.73	29	QV	-1.85
6	VV	-0.32	18	MV	-0.28	30	QL	-1.32
7	FV	0.43	19	MF	0.42	31	QF	-1.14
8	AL	-0.54	20	SV	-1.53	32	FQ	-1.03
9	AA	-2.00	21	SF	-0.79	33	VQ	-1.82
10	GL	-0.78	22	TV	-1.25	34	KL	-0.26
11	LI	0.68	23	TI	-0.86	35	KF	0.12
12	FG	-0.50	24	NV	-1.85	36	FK	0.14

1) MLR method was used to perform QSAR analysis.

The results of MLR analysis was showed in Table 3. After LOOCV, it yield a cross-validated correlation coefficient $Q^2_{(CV)}$ of 0.781, and RMSE of 0.375, and after MLR fitting analysis, it yielded a standard multiple correlation coefficient R^2 of 0.868. Comparing the regression coefficients of each descriptor in column 2 in

Table 3, it is found that contributions of the first, the third, the fourth and the sixth descriptors are more than those of the second and the fifth descriptor. Actually, the first and the fourth descriptor refer to the total surface area(S), the third and the sixth descriptor refers to hydrophilic indices (H). So S and H parameter of the side chain residual are closely related to the logP of blocked dipeptide.

Table 3 The value of regression coefficient and statistics of various linear models

Statistics	**Blocked dipeptides**	**Unblocked dipeptides**
b0	-3.286	-3.89
b11	2.058	1.313
b12	-0.081	0.045
b13	2.165	0.259
b21	2.26	1.197
b22	0.148	0.072
b23	2.336	1.145
n	36	46
$Q_{(cv)}^2$	0.781	0.659
RMSE	0.375	0.387
R^2	0.868	0.755

2) SVR method was used to perform QSAR analysis

The results of SVR (radical basis kernel function, C=200, γ =0.08, ε=0.0001, loss=0.001) analysis was showed in Fig.1 and Fig.2. After LOOCV, it yield a cross-validated correlation coefficient Q2(CV) of 0.970, and RMSE of 0.140, and after SVR fitting analysis, it yielded a standard multiple correlation coefficient R2 of 0.993. The importance of each parameter or property in the SVR-LOOCV model was showed in Fig.2.

How found that the impact factors of peptide logP by QSAR model, which is a very key question. For the sake of examining infection of each amino acid property VS logP in the peptides by SVR model, the importance of each parameter or property is estimated by the $\Delta RMSE$ value in turn. For dipeptides, the very important impact of the n_1S, the n_1H, the n_2S, and the n_2H can be seen in the model, because of RMSE increasing of the model while one of them having been taken out, and hardly any infection of the n_1L, and the n_2L were shown in Fig.2, the n_1S, the n_1H, the n_2S, and the n_2H are positively related to the logP value. Looking at the $\Delta RMSE$ value, it is evident that position n1, which corresponds to the amino terminus for a dipeptide, is more important than position n_2. For both positions, amino acid residue with S as well as H side chains are preferred, which display the logP correlate to total surface area(S) and hydrophilic indices (H), and slightly relates to lipophilic indices (L). Moreover, this also demonstrates that the importance of selected parameters by using SVR and MLR are the same. But the results of SVR is better than that of MLR, it also shows the non-linear relationship between logP and HMLP parameters.

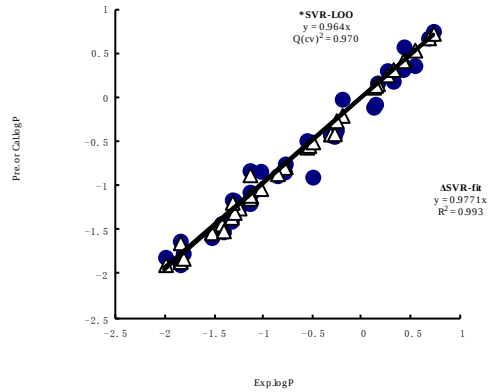

Fig.1 Plot of prediction or calculation and experiment activities of 36 blocked dipeptides (Δ-SVR-fit, *-SVR-LOO)

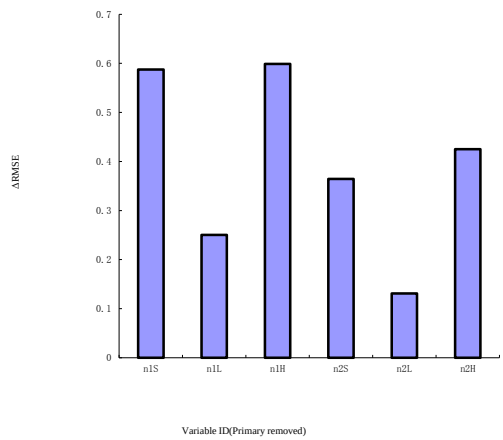

Fig.2 Plot of $\Delta RMSE$ and removed amino acids property position of a set 36 blocked dipeptides by SVR-LOOCV

3.2. LogP for unblocked dipeptides

The sequences and the logP value of the studied peptides were collected from the literature [1-5], which can be seen in Table 4.

Table 4 The sequence and logP of unblocked peptides

NO.	peptide	Exp.LogP	NO.	peptide	Exp.LogP	NO.	peptide	Exp.LogP
1	FL	-1.17	17	YL	-1.75	33	AL	-2.46
2	LF	-1.15	18	VY	-2.52	34	AW	-2.21
3	FF	-0.85	19	FY	-1.68	35	FG	-2.31
4	LL	-1.46	20	YY	-1.87	36	FS	-2.59
5	LV	-2.05	21	LM	-1.87	37	GF	-2.3
6	VL	-2.07	22	ML	-1.84	38	GG	-2.92
7	AI	-2.6	23	MV	-2.53	39	GV	-2.98
8	LI	-1.64	24	FM	-1.59	40	GW	-2.17
9	VV	-2.82	25	SL	-2.49	41	LH	-2.74
10	II	-1.82	26	PF	-2.07	42	SF	-2.54
11	WW	-0.27	27	PL	-2.41	43	VG	-2.74
12	WF	-0.47	28	PI	-2.56	44	WG	-1.98
13	WA	-1.98	29	FP	-1.36	45	WS	-2.2
14	WL	-0.73	30	LP	-1.76	46	WL	-0.68
15	WY	-1.13	31	IP	-1.79			
16	LY	-1.94	32	AF	-2.21			

1) MLR method was used to perform QSAR analysis

The results of MLR analysis was showed in Table 3. After LOOCV, it yield a cross-validated correlation coefficient $Q^2_{(CV)}$ of 0.659, and RMSE of 0.387, and after MLR fitting analysis, it yielded a standard multiple correlation coefficient R^2 of 0.755. Comparing the regression coefficients of each descriptor in column 3 in Table 3, it is found that contributions of the first, the fourth and the sixth descriptors are more than those of the second, the third and the fifth descriptor. In fact, the first and the fourth descriptor refer to the total surface area(S), the sixth descriptor refers to hydrophilic indices (H). So S and H parameter of the side chain residual are closely related to the logP of unblocked dipeptide.

2) SVR method was used to perform QSAR analysis

The results of SVR (radical basis kernel function, C=50, γ =0.16, ε=0.0001, loss=0.001) analysis was showed in Fig.3 and Fig.4. After LOOCV, it yield a cross-validated correlation coefficient $Q^2_{(CV)}$ of 0.928, and RMSE of 0.176, and after SVR fitting analysis, it yielded a standard multiple correlation coefficient R^2 of 0.975. The importance of each parameter or property in the SVR-LOOCV model was showed in Fig.4.

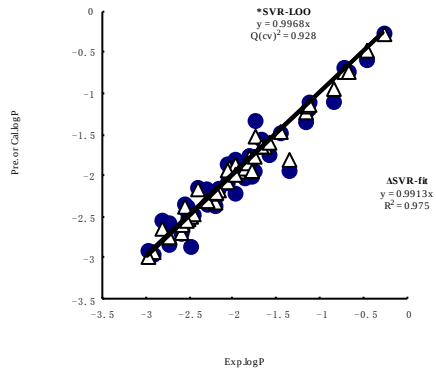

Fig.3 Plot of prediction or calculation and experiment activities of 46 unblocked tripeptides (Δ-SVR-fit, *-SVR-LOO)

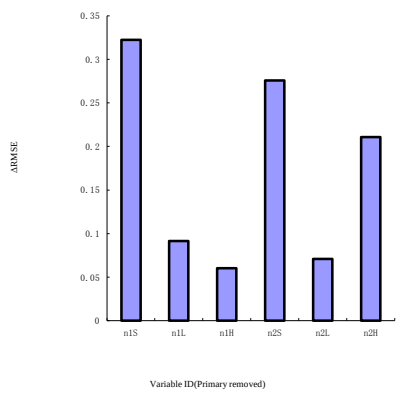

Fig.4 Plot of $\Delta RMSE$ and removed amino acids property position of a set 46 unblocked tripeptides by SVR-LOOCV

For dipeptides, the very important impact of the n_1S, the n_2S, and the n_2H can be seen in the model, because of $\Delta RMSE$ increasing of the model while one of them having been taken out, and hardly any infection of the n_1L, the n_1H and the n_2L were shown in Fig.4, the n_1S, the n_2S, and the n_2H are positively related to the logP value. Looking at the $\Delta RMSE$ value, it is evident that logP relates mainly to S in position n_1, while relates mainly to S and H in position n_2. For both positions, amino acid residue with S as well as H side chains are preferred, which display the logP correlate to total surface area(S) or hydrophilic indices (H), and slightly relates to lipophilic indices (L). Moreover, this also demonstrates that the importance of selected parameters by using SVR and MLR are the same. But the results of SVR is better than that of MLR, it also shows the non-linear relationship between logP and HMLP parameters.

4. Conclusions

In this paper, a series of heuristic molecular lipophilicity potential (HMLP) parameters (S, L, H) were introduced to describe Structural characteristics of amino acid side chain. All the results showed that the logP relates to the total surface area(S) and hydrophilic indices (H), and the prediction results of SVR are better than that of MLR. In a word, this paper provided a simple and effective method for predicting the logP of peptide and some insight into what structural features are related to the logP of peptides. Based on amino acid descriptors HMLP (S, L, H), we construct a new dipeptides QSAR model by SVR. Moreover, it also offered an idea about nonlinear relation between logP of peptides and their structural descriptors (HMLP). In addition, the HMLP descriptors will be useful in structure characterization and activity prediction of biological molecules, and will become a group of general parameters for QSAR analyses on polypeptides and proteins.

Acknowledgements

This work was supported by by the “211 Engineering Double Support Plan” Foundation of Sichuan Agricultural University (Yaan, China).

References

[1] M.Akamatsu, Y.Yoshida, H.Nakamura, et al., “Hydrophobicity of di- and tripeptides having unionizable side chains and correlation with substituent and structural parameters”, Quant. Struct.-Act. Relat., 1989, 8, pp.195-203.
[2] M.Akamatsu, S.Okutani, K.Nakao, et al., “Hydrophobicity of N-acetyl-di- and tripeptide amides having unionizable side chains and correlation with substituent and structural parameters”, Quant. Struct.-Act. Relat., 1990,9, pp.189-194.
[3] M.Akamatsu, and T.Fujita, “Quantitative analyses of hydrophobicity of di- to pentapeptides having un-ionizable side chains with substituent and structural parameters”, J.Pharm. Sci., 1992, 81(2), pp.164-174.
[4] M.Akamatsu, T.Katayama, D.Kishimoto, et al., “Quantitative analyses of the structure-hydrophobicity relationship for N-acetyl di- and tripeptide amides”, J.Pharm. Sci., 1994, 83(7), pp.1026-1033.
[5] S.N.Tomoko, O.Akio, “Evaluation of the hydrophobic parameters of the amino acid side chains of peptides and their application in QSAR and conformational studies”, J.Mol. Struct. -Theochem, 1997, 392, pp.43-54.
[6] P.Tao, R.X.Wang, L.H.Lai, “Calculating Partition Coefficients of Peptides by the Addition method”, J.Mol. Model., 1999,5(10), pp.189-195.
[7] N.Gulyaeva, A.Zaslavsky, A.Chait, et al., “Relative hydrophobicity of di- to hexapeptides as measured by aqueous two-phase partitioning”, J. Pept. Res., 2003, 61(3), pp.129-139.
[8] J. T.Sarah, K. H.Channa, D. H.John, et al., “On the hydrophobicity of peptides: Comparing empirical predictions of peptide log P values”, Bioinformation, 2006, 1(7), pp. 237-241.
[9] K. H.Channa and R. F.Darren, “Empirical prediction of peptide octanol-water partition coefficients”, Bioinformation, 2006, 1(7), pp. 257-259.
[10] Q.S., Du, D.P. Li, W.Z. He, et al., “Heuristic molecular lipophilicity potential (HMLP): Lipophilicity and hydrophilicity of amino acid side chains”, J. Comput. Chem., 2006, 27(6), pp. 685-692.
[11] Q.S. Du, R.B. Huang, Y.T.Wei, et al., “Peptide reagent design based on physical and chemical properties of amino acid residues”, J. Comput. Chem., 2007, 28(12), pp. 2043-2050.
[12] Huang, R.B., Du, Q.S., Wei, Y.T., et al., Physics and chemistry-driven artificial neural network for predicting bioactivity of peptides and proteins and their design.. Journal of Theoretical Biology, 2009, 256(3):428-435

[13] J.S.Alex, S.Bernhrd, "A tutorial on support vector regression", Stat. Comput., 2004, 14, pp.199-222.
[14] N.V. Vapnik, "Statistical learning theory", Beijing: Publishing House of Electronics Industry, 2004.
[15] L.Eriksson, E.Johansson, N.Kettaneh-Wold, et al., "Multi- and megavariate data analysis: principle and application ', Umea, Sweden: Umetrics AB. 2001.
[16] R.E.Fan, P.H.Chen, C.J.Lin, "Working set selection using the second order information for training SVM", J. Mach. Learn. Res., 2005, 6, pp.1889-1918 or http://www.csie.ntu.edu.tw/~cjlin/libsvm/.