Partial Least Squares Regression; PLS

Because PLS enables extraction of latent variables correlating with a response variable, it is desirable that all irrelevant descriptors are excluded to calculate latent variables from relevant ones with a response variable. Therefore, backward and forward selection procedures were applied. The q^2 value in LOO-CV was used as an index of both variable selection procedures. In the forward selection procedure, three variables, logP(o/w) and both dummy variables with respect to aromatic amines and aliphatic amines were used for the initial variables. The number of latent variables was selected on the basis of the lowest root mean squared error of prediction (RMSEP) by LOO-CV.

Results and Discussions

MLR analysis

The results from four MLR analysis models are shown in Table 3. Firstly, 29 variables were selected by the backward selection procedure ("MLR_00" in Table 2), but the q^2 value calculated by LOO-CV was negative, indicating that it was an inappropriate model for predicting, although AIC had been minimized and the adjusted coefficient of determination, R_a^2 was high. Since too many variables were selected, over-fitting occurred in this model. Based on the significance tests for the regression coefficients of the 29 explanatory variables, 26, 24 and 11 variables were selected according to the significance levels 0.1, 0.05 and 0.01, respectively.

Table 3. Results of models by multiple linear regression analysis

Variable set [*N* *]	R_a^2	q^2 (LOO-CV)	Correct [%]	Outlier [%]
MLR_00 [29]	0.610	-0.544	187 [70.6]	8 [3.0]
MLR_01 [26]	0.600	0.448	189 [71.3]	8 [3.0]
MLR_02 [24]	0.594	0.404	194 [73.2]	8 [3.0]
MLR_03 [11]	0.555	0.521	192 [72.5]	6 [2.2]
logP(o/w)	0.432	0.422	173 [65.3]	7 [2.6]

*: number of variables

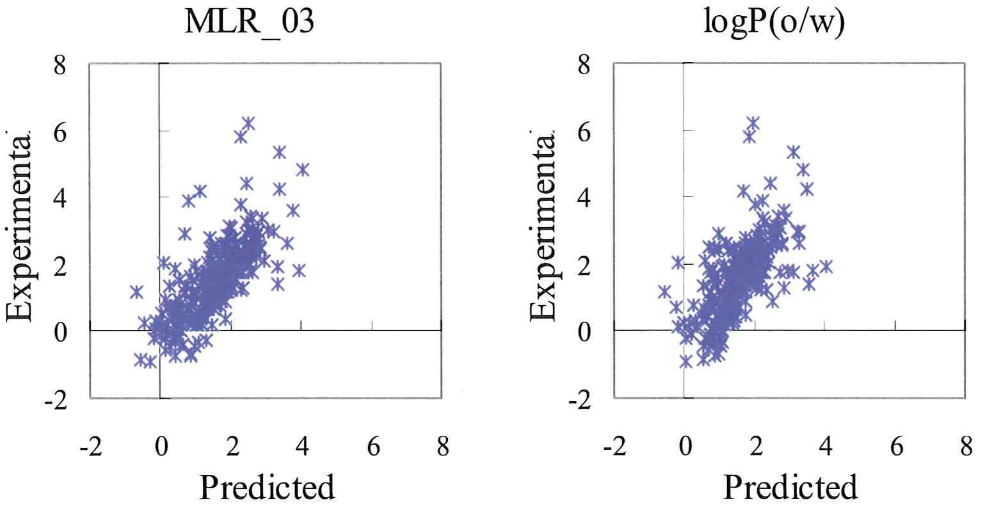
R_a^2: adjusted coefficient of determination

Three models were then calculated using the newly selected variables ("MLR_01", "MLR_02" and "MLR_03" in Table 3). The "MLR_03" prediction model was the most appropriate model for predicting the toxicity based on the q^2 values. The plots of experimental and predicted values are shown in Figure 1.

In order to evaluate the number of outliers in these models, the samples were classified based on their prediction errors. More specifically, samples having prediction errors less than 0.7 were classified into "Correct" and those having prediction errors greater than 2 were classified into "Outlier". In the "MLR_02" regression model, the number of compounds that were classified into "Correct" was greater than that of the other models, but eight compounds were classified into "Outlier". The prediction error of one compound was actually six or more. In contrast, in the "MLR_03" model, the number of compounds that were classified into "Outlier" was six and the prediction error of the above compound (prediction error > 6) was improved to 0.27, although the number of compounds classified into "Correct" decreased. Because q^2 value is the highest, "MLR_03" might be excellent model in the examination done this time. The "MLR_03" regression formula is shown below as Equation (1). The correlation coefficients between explanatory variables of "MLR_03" are shown in Table 4.

$$y = 2.304 + 0.377\ (\pm 0.166)\mathit{aroma} + 0.380\ (\pm 0.037)\mathit{logP(o/w)} - 0.131\ (\pm 0.040)\mathit{dipole} + 0.267\ (\pm 0.079)\mathit{HOMO} - 0.152\ (\pm 0.060)\mathit{LUMO} + 0.070\ (\pm 0.037)\mathit{E\_ele} - 0.00012\ (\pm 0.00003)\mathit{pmi} - 0.0141(\pm 0.0025)[\mathit{ASA+}] + 0.0014\ (\pm 0.0003)[\mathit{CASA-}] + 3.01\ (\pm 0.887)[\mathit{FASA+}] + 0.480\ (\pm 0.100)\mathit{std\_dim}1 \quad (1)$$

Since logP(o/w), regarded as an important factor for predicting toxicity, was also the important variable in the "MLR_03" model, variable selections in the present study were considered to be one of the appropriate procedures. The possibility that is the item that relates from the selection also of "HOMO" and "LUMO" that shows the energy status of the compound to the reactiveness of the compound is thought. FASA+ is a similar descriptor to ASA+ but the regression coefficients of both variables had inverse effects on the model. Therefore, to evaluate the multicolinearity, we excluded one of the variables from the dataset and regenerated the two models. As a result, the q^2 values calculated from both models were lower than that of the former model (data is not shown). Thus, it is likely that both descriptors are necessary to predict toxicity. Because the dataset in the present study comprises compounds with various structures, complicated relationships may result.



(a) (b)

Figure 1. Scatter plots of experimental and predicted values in MLR.
(a): MLR_03; variable selection by p-value ≤ 0.01 in significant tests for regression coefficients.
(b): logP(o/w); single linear regression analysis by logP(o/w).

Table 4. Intercorrelation matrix for selected variables (MLR_03)

	aromatic	logP	dipole	HOMO	LUMO	E_ele	pmi	ASA+	CASA-	FASA+	std_dim1
aromatic	1										
logP	-0.137	1									
dipole	-0.038	-0.215	1								
HOMO	0.551	0.187	-0.286	1							
LUMO	0.131	-0.13	-0.294	0.026	1						
E_ele	-0.293	0.072	-0.179	-0.119	0.038	1					
pmi	-0.131	0.676	0.134	0.026	-0.194	-0.071	1				
ASA+	0.352	0.023	-0.058	0.553	-0.2	-0.056	0.133	1			
CASA-	0.115	0.169	0.235	0.142	-0.665	-0.074	0.479	0.526	1		
FASA+	0.416	-0.278	-0.065	0.489	-0.148	-0.062	-0.262	0.881	0.349	1	
std_dim1	-0.04	0.519	0.109	0.089	0.028	-0.144	0.821	0.281	0.229	-0.075	1

PLS analysis

The results of three PLS analysis models are shown in Table 5. Some selected variables were different in the backward and forward selection procedures, although both procedures resulted in the same number of selected variables. We selected the number of latent variables bases on the lowest RMSEP. The transitions of RMSEPs and the scatter plots between experimental and predicted values are shown in Figure 2.

The q^2 values in the models with variable selection procedures were higher than that of the model without a variable selection procedure. It is possible that the unnecessary variables influenced the calculations of the latent variables and subsequent predictions. "Outlier" was also small, better predict was able to be done to "PLSR_baskward". The variables selected by both variable selection procedures are shown in Table 6, and there were some variables which were selected in both procedures. The standardized partial regression coefficients of these common variables are shown in Figure 3.

The coefficients of these common variables had identical signs (positive or negative) between both models, although there were slight differences between their absolute values. Therefore, it was thought that the significances of the descriptors that were selected in both models for predicting toxicity were higher than the other descriptors.

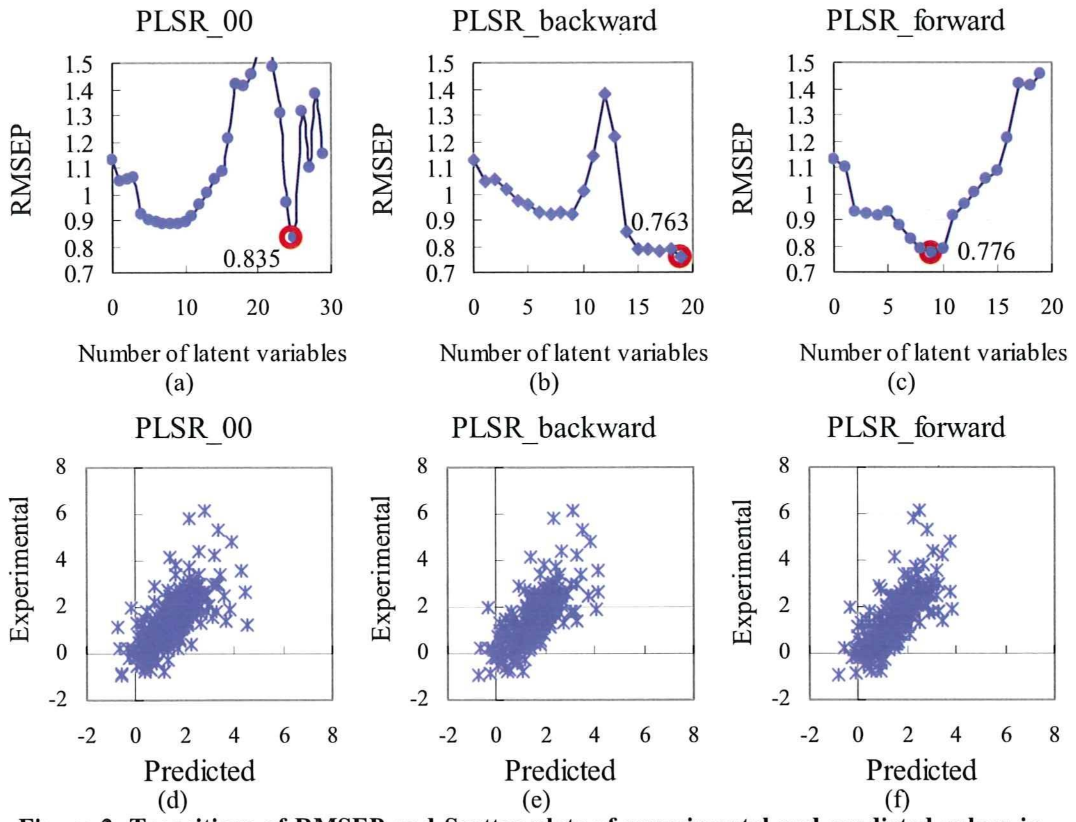


Figure 2. Transitions of RMSEP and Scatter plots of experimental and predicted values in PLSR.

(a), (b), (c): transitions of RMSEP in PLSR without a variable selection, with backward and forward selection procedures.

(d), (e), (f): scatter plots of experimental and predicted values in PLSR without a variable selection, with backward and forward selection procedures.

Table 5. Results of models by partial least squares regression analysis

Variable set [N *^{1}]	N_lv*^{2}	q^2	R^2	Correct [%]	Outlier [%]
PLSR_00 [45]	25	0.448	0.468	185 [70.0]	8 [3.0]
PLSR_backward*^{3} [19]	19	0.539	0.543	186 [70.2]	5 [1.9]
PLSR_forward*^{4} [19]	9	0.523	0.526	187 [70.6]	7 [2.6]
logP(o/w)*^{5}	-	0.422	0.432	173 [65.3]	7 [2.6]

*^{1}: number of variables

*^{2}: number of latent variables

*^{3}: backward selection procedure

*^{4}: forward selection procedure

*^{5}: this is same result as presented in Table 2

Table 6. List of variables in two models with variable selection procedures

common*^{1}	PLSR_backward*^{2}	PLSR_forward*^{3}
logP(o/w)	E_ele	aliphatic
aromatic	E	E_ele
HOMO	E_stb	E_nb
LUMO	E_str	ASA_H
dipole	E_strain	FASA-
HF	E_tor	FCASA+
pmi	CASA-	FCASA-
E_vdw	DCASA	dens
std_dim3	std_dim1	glob
	vol	VSA

*^{1}: variables selected in both models

*^{2}: variables selected in only backward selection procedure

*^{3}: variables selected in only forward selection procedure

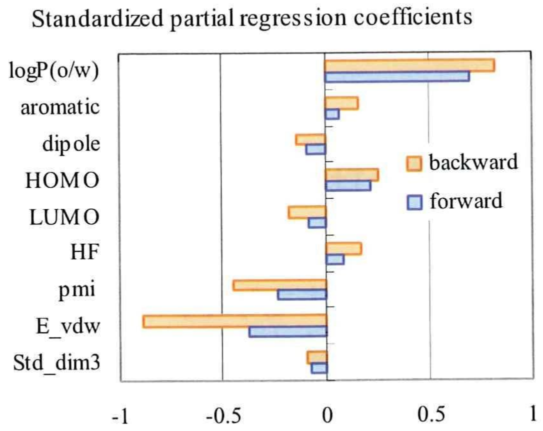


Figure 3. Standardized partial regression coefficients of common variables selected in both backward and forward selection procedures.

Conclusion

In the present study, both MLR and PLS analyses were conducted using 3D descriptors and logP(o/w), and some 3D descriptors were found to be important for predicting toxicity. Predicton accuracy of the generated model was adequate and improved compared to that of the model using only logP(o/w). It seems that MLR is useful when thinking about mechanical analysis and a structural improvement by the variable for the predict model. Moreover, it is thought that PLS is suitable for valuing the prediction accuracy. However, the value of a 3D descriptor easily changes when the steric structure changes, therefore, careful attention must be paid to the possibility that predictions using 3D descriptors may worsen due to structural changes of the compound.

In future studies, prediction accuracy may be improved by adding 2D descriptors to the 3D descriptors used in the present study. Moreover, re-examination of not only structure optimization, which affects the values of 3D descriptors, but also variable selection procedures for excluding unnecessary descriptors are important for improving prediction accuracy.

References and Notes

[1] Act on the Evaluation of Chemical Substances and Regulation of Their Manufacture, etc. (Act No. 117 of October 16, 1973), http://www.env.go.jp/en/laws/chemi/cscl/CSCL_law.pdf

[2] Results of Eco-toxicity tests of chemicals conducted by Ministry of the Environment in Japan (- 2005), http://www.env.go.jp/chemi/sesaku/02e.pdf

[3] OECD Quantitative Structure-Activity Relationships [(Q)SARs] Project, http://www.oecd.org/document/23/0,3343,en_2649_34379_33957015_1_1_1_1,00.html

[4] Structure of the Guidance on Information Requirements and Chemical Safety Assessment Chapter R.6 QSARs and grouping of chemicals, http://guidance.echa.europa.eu/docs/guidance_document/information_requirements_r6_en.pdf?vers=20_08_08

[5] P. Reuschenbach, M. Sulvani, M. Dammann, D. Warnecke, and T. Knacker, *Chemosphere*, **71**, 1986-1995 (2008).

[6] J. C. Faucon, R. Bureau, J. Faisant, F. Briens, and S. Rault, *Chemosphere*, **44**, 407-422 (2001).

[7] Ecological Structure Activity Relationships (ECOSAR), http://www.epa.gov/oppt/newchems/tools/21ecosar.htm

[8] TIssue MEtabolism Simulator (TIMES), (LMC, Bourgas, Bulgaria), http://oasis-lmc.org/

[9] G. H. Lu, C. Wang, and X. L. Guo, *Biomed. Environ. Sci.* **21**, 193-196 (2008).

[10] GUIDELINE ON THE ENVIRONMENTAL RISK ASSESSMENT OF MEDICINAL PRODUCTS FOR HUMAN USE, http://www.emea.europa.eu/pdfs/human/swp/444700en.pdf

[11] Development and application of QSAR models for predicting the results of Fish Acute Toxicity Test and Daphnia sp. Acute Immobilization Test (in Japanese), http://www.env.go.jp/council/05hoken/y051-69/mat01-1.pdf, mat01-2.pdf, mat01-3.pdf, mat01-4.pdf

[12] R Development Core Team, R: A language and environment for statistical computing, R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0, http://www.R-project.org

[13] K. Hasegawa, K. Funatsu, in Chemoinformatics and Advanced Machine Learning Perspectives: Complex Computational Methods and Collaborative Techniques, Medical Information Science Reference, ISBN-13: 978-1615209118.

[14] A. W. Hughes, M. L. King, *J. Statist. Plann. Infer.* **115**, 397-411 (2003).

3 次元構造記述子を用いた環境毒性予測

日髙　伸之介[a]，白石　寛明[b]，大眉　佳大[a]，山﨑　広之[a]，岡本　晃典[a]，川下　理日人[a,c]，安永　照雄[c]，高木　達也[a,c,d*]

[a] 大阪大学大学院薬学研究科，〒565-0871　大阪府吹田市山田丘1-6
[b] 国立環境研究所，〒305-8506　茨城県つくば市小野川16-2
[c] 大阪大学微生物病研究所，〒565-0871　大阪府吹田市山田丘3-1
[d] 大阪大学感染症国際研究拠点タイ感染症共同研究センター，〒565-0871　大阪府吹田市山田丘3-1

我々の周りの環境中には膨大な数の化学物質が存在しており，社会的にも必要不可欠なものとなっている．しかしながら，ヒトの健康と地球環境に対して深刻な影響を与えるような危険な化学物質も存在している．日本では 1995 年から OECD における高生産量 (HPV) 化学物質の有害性評価プログラムに貢献するために，化学物質を対象とした生態影響試験が実施されている．しかし，未だに約 500 種しか試験されておらず，全ての化合物を確認することは非常に困難である．そこで，毒性試験の代替法としてコンピュータ技術を活用した定量的構造活性相関 (QSAR) 解析法を用いることで，化学物質の環境毒性，特性や物性を予測することが出来ると期待されている．本研究では，環境省の実施したミジンコ急性遊泳阻害能試験結果と3 次元構造記述子を用いた，多様な構造の化合物の生態毒性を予測する QSAR モデルの構築・改良を試みた．本研究で構築した QSAR モデルでは，従来の n- オクタノール / 水分配係数 (logP(o/w)) だけを使ったモデルに比べ，予測精度を向上させることができた．

キーワード: 環境生態毒性，SAR，3 次元記述子，MLR，PLS，予測モデル

* *ttakagi@phs.osaka-u.ac.jp*