EXCLI Journal 2006;5:150-163 – ISSN 1611-2156
Received: 1. September 2006, accepted: 12. October 2006, published: 16. October 2006

Original article:

Prediction of selectivity index of pentachlorophenol-imprinted polymers

Chanin Nantasenamat[1], Tanawut Tantimongcolwat[1], Thanakorn Naenna[2], Chartchalerm Isarankura-Na-Ayudhya[1], Virapong Prachayasittikul[1*]

[1]Department of Clinical Microbiology, Faculty of Medical Technology, Mahidol University, Bangkok 10700, Thailand. [2]Department of Industrial Engineering, Faculty of Engineering, Mahidol University, Nakhon Pathom 73170, Thailand. *Corresponding author. Telephone: 662-849-6318, Fax: 662-849-6330, E-mail: mtvpr@mahidol.ac.th

ABSTRACT

A data set comprising of the selectivity index of pentachlorophenol-imprinted polymers against 53 pentachlorophenol and related compounds was obtained from the excellent work of Baggiani et al. Molecular descriptors of the phenol compounds were calculated with E-DRAGON to obtain a total of 1,666 descriptors spanning 20 categories of molecular properties. Multivariate analysis of the data set was performed using multiple linear regression, partial least squares regression, and principal component regression. Partial least squares regression was found to deliver an excellent predictive model and was chosen for further investigation. The descriptor dimension was reduced by the combined use of partial least squares and Unsupervised Forward Selection algorithm. The obtained Quantitative Structure-Property Relationship (QSPR) model based on the smaller subset of the molecular descriptors displayed substantial gain in predictive ability when compared to the model of Baggiani et al. Such QSPR model can help in the computational design of MIPs with predefined selectivity toward template molecules of interest.

Keywords: selectivity, pentachlorophenol, molecularly imprinted polymer, partial least squares regression, QSPR

INTRODUCTION

Molecular imprinting is a technique that enables the production of artificial receptors that is tailor-made to any given target molecule of interest. Upon removal of the template, receptors possessing specific and selective recognition are formed within the macromolecular matrix of the molecularly imprinted polymer (MIP). Some of the advantages of MIPs over biological receptors are their ease of preparation and durability (Bruggemann, 2002; Haupt, 2003). Successful applications of MIPs have been demonstrated as separation media (Martin-Esteban, 2001; Turiel and Martin-Esteban, 2004; Takeuchi and Haginaka, 1999; Tamayo and Martin-Esteban, 2005; Machtejevas et al., 2004; Spegel et al., 2003; Watabe et al., 2006), enzyme mimetics (Piacham et al., 2003), artificial receptors (Hsieh et al., 2006; Chianella et al., 2002; Ramstrom et al., 1996), recognition elements of biosensors (Piacham et al., 2005), synthetic receptors for drug assays (Vlatakis et al., 1993; Sellergren and Andersson, 2000; Piletsky et al., 2000; Ansell and Mosbach, 1998; Ye et al., 2002) and serologic tests (Tai et al., 2006), and nanofactories for synthesis of enzyme inhibitors (Ye et al., 2001; Mosbach et al., 2001).

Phenolic compounds are commonly used as raw material for petrochemical, pharmaceutical, plastic, and pesticide industry (Ahlborg and Thunberg, 1980; Exon, 1984). Common consumer products made of phenol include detergents, plastic packagings, polycarbonate plastic coatings of compact discs, aspirins and other pharmaceuticals. In fact, phenol ranked among the top 50 chemicals produced in the United States (1998). The adverse effects of phenols have been summarized by the U.S. Environmental Protection Agency (Bruce et al., 1987). A variety of methods have been attempted for the detection and removal of phenols such as whole-cell based biosensors (Shaw et al., 1999; Sinclair et al., 1999; Weitz et al., 2002), phytoremediation (Santos de Araujo et al., 2002; Harvey et al., 2002; Agostini et al., 2003), enzymatic detoxification (Wang et al., 2004; Bollag et al., 1988; Wright and Nicell, 1999; Buchanan and Nicell, 1997), photochemical degradation (Catalkaya et al., 2003; Bali et al., 2003), Fenton reaction (Kavitha and Palanivelu, 2004; Detomaso et al., 2003), and degradation by acoustic cavitation (Gogate et al., 2004).

Alternatively, detection and removal of phenolic compounds may be achieved by molecular imprinting. This endeavor has been realized using bisphenol A (Sanbe et al., 2003; Sanbe and Haginaka, 2003), nitrophenols (Huang et al., 2003; Caro et al., 2003; Caro et al., 2002; Masque et al., 2000), and chlorophenols (Caro et al., 2003; Baggiani et al., 2004) as templates for preparation of molecularly imprinted polymers possessing selectivity and specificity toward the compounds. Baggiani et al. prepared a pentachlorophenol-imprinted polymer and explored its selectivity against a library of 52 phenolic compounds comprising of chloro-, alkyl-, aryl-, methoxy-, and polyphenols. In their study, quantitative structure-retention relationship was constructed and modeled by principal component regression using molecular descriptors derived from quantum chemical calculations.

We have previously proposed the feasibility of using molecular descriptors, which were derived from molecular charge densities of template and functional monomer molecules, with artificial neural networks for prediction of the imprinting factors of MIPs (Nantasenamat et al., 2005a). Artificial neural networks were demonstrated to be a suitable modeling method for biological and chemical systems in our previous studies (Nantasenamat et al., 2005a; Nantasenamat et al., 2005b; Nantasenamat et al., 2006). In the present investigation, we further the development of a robust quantitative structure-property relationship (QSPR) model for the prediction of selectivity index using an extensive library of molecular descriptors provided by E-DRAGON. The molecular descriptors comprising of 20 categorical blocks provided a thorough physicochemical representation of the phenol compounds. The mass number of descriptors was reduced sequentially via confidence interval filter or regression coefficients and multi-collinear variable removal. Partial least squares regression was demonstrated to be superior to principal component regression and was chosen as the method of choice for this investigation. The final subset of variables showed good predictive ability in modeling the selectivity index of the pentachlorophenol-imprinted polymers toward related phenols.

MATERIALS AND METHODS

Data set

The data set used in this study was taken from the work of Baggiani et al. In their study, Baggiani and co-workers prepared a pentachlorophenol-imprinted polymer using 4-vinylpyridine as functional monomer, ethylene glycol dimethacrylate as cross-linker, and methanol-water (3/1, v/v) as porogen. Chromatographic evaluation of the HPLC-packed polymer was performed against 52 PCP-related phenols comprising of 22 chloro-, 21 akly-, 4 aryl-, 3 methoxy-, and 6 polyphenols. The molecular recognition properties was evaluated from the selectivity index as calculated from the retention factors of non-imprinted polymer and imprinted polymer from the following equation:

$$SI = \frac{k_{NIP}}{k_{MIP}} \qquad (1)$$

Table 1: Summary of molecular descriptors calculated from E-DRAGON.

No.	Type of Descriptor	No. of Descriptors
1	Constitutional descriptors	48
2	Topological descriptors	119
3	Walk and path counts	47
4	Connectivity indices	33
5	Information indices	47
6	2D autocorrelations	96
7	Edge adjacency indices	107
8	Burden eigenvalue descriptors	64
9	Topological charge indices	21
10	Eigenvalue-based indices	44
11	Randic molecular profiles	41
12	Geometrical descriptors	74
13	RDF descriptors	150
14	3D-MoRSE descriptors	160
15	WHIM descriptors	99
16	GETAWAY descriptors	197
17	Functional group counts	154
18	Atom-centred fragments	120
19	Charge descriptors	14
20	Molecular properties	29

where *SI* is the selectivity index, k_{NIP} is the retention factor of non-imprinted polymers, and k_{MIP} is the retention factor of imprinted polymers. Therefore, a data set of 53 phenol compounds comprising of pentachlorophenol and 52 related phenols was obtained for this investigation.

Molecular descriptors

The chemical structures of the 53 phenol compounds were drawn with MarvinSketch (ChemAxon, Budapest, Hungary) and exported as SMILES (Weininger, 1988) notation. Next, phenol compounds represented by SMILES format was used as input for calculation of 1,666 molecular descriptors with the online software, E-DRAGON (Tetko et al., 2005; VCCLAB). The software converted the molecules from SMILES notation to 3-dimensional structures using the algorithm derived from CORINA (Gasteiger et al., 1990; Sadowski et al., 1994; Sadowski and Gasteiger, 1993). The molecular descriptors comprising of 20 descriptor blocks is shown in Table 1. The definition and description of these molecular descriptors was described by Todeschini et al. (Todeschini et al., 2000).

Data Pre-processing

The molecular descriptors were standardized to mean of zero and standard deviation of one with the following equation:

$$x_{ij}^{stn} = \frac{x_{ij} - \bar{x}_j}{\sum_{i=1}^{N}(x_{ij} - \bar{x}_j)^2 \big/ N} \qquad (2)$$

where x_{ij}^{stn} is the standardized value, x_{ij} is the value of each sample, $\bar{x}_j$ is the mean of each descriptor, and N is the sample size of the data set.

Multivariate analysis

Three multivariate analysis methods comprising of multiple linear regression (MLR), partial least squares (PLS) regression, and principal component regression (PCR) were used to model the *SI* property of 53 phenols. All multivariate analysis was performed with The Unscrambler 9.6 software package (Camo Process AS, Norway) as previously described in our previous study (Nantasenamat et al., 2006). The phenol compounds represented by 1,666 molecular descriptors were used as independent variables while *SI* was used as dependent variable. The descriptor matrix comprising of several hundred variables were reduced to a few latent variables called Principal Components (PC). The PCs are orthogonal and are therefore not redundant since the PCs are perpendicular to one another (Esbensen, 2004). The optimal number of PCs was determined according to the method of Haaland and Thomas from a plot of PC against MSE using LOO-CV. Mean squared error (MSE) was calculated according to the following equation:

$$MSE = \frac{\sum_{i=1}^{n} (p_i - a_i)^2}{n} \qquad (3)$$

where p_i is the predicted output, a_i is the actual output, and n is the number of compounds presented in the data set.

Reduction of descriptors

Although PLS and PCR are able to handle data sets with multi-collinear variables, the rather large size of the variables is undesirable since it takes longer to calculate as well as not revealing crucial information, particularly the contributions of the variables in modeling the *SI* property. Constant variables were removed from the data set as they provide no useful information. Next, the data set was subjected to standardization according to equation 2. PLS regression was performed using PLS1 algorithm. The regression coefficients derived from PLS regression was filtered by retaining those located outside the defined confidence interval, which was calculated according to the following equation:

$$CI = \bar{x} \pm (z \times s) \qquad (4)$$

where CI is the confidence interval, $\bar{x}$ is the mean, z is the standard score of the level of confidence, and s is the standard deviation. The level of confidence at 75, 80, 90, 95, 99, and 99.9 % were used for variable reduction. Descriptors found within the defined confidence interval were removed while variables outside the defined confidence interval were retained.

The second phase of variable reduction utilized the Unsupervised Forward Selection (UFS) program (Whitley et al., 2000), to further remove redundant variables while still maintaining the core information of the data set. The UFS algorithm was described in our previous study (Nantasenamat et al., 2005a) and by Whitley et al. (Whitley et al., 2000).

Generation of training and testing sets

Training and testing sets were generated according to the leave-one-out cross-validation (LOO-CV) method (Nantasenamat et al., 2006; Witten and Frank, 2000). Briefly, one sample of the data set was left out as the testing set and the rest was used as the training set. This procedure was performed reiteratively until all samples of the data set were given the chance to be used as testing sets.

RESULTS AND DISCUSSION

Structural Considerations

Factors governing the selectivity of MIPs were extensively reviewed by Spivak (Spivak, 2004; Spivak and Campbell, 2001). In the investigation by Baggiani et al., the core structure of the library compounds was based on phenol. However, each of the compounds differs in the substituent groups that they bear, which may consequently change the structural and electronic properties of the molecules. As a result, this affects the molecular recognition properties of the compounds toward the imprinted polymer. The molecular descriptors produced by E-DRAGON provide a thorough representation of the phenolic compounds investigated in this study. Thus, the observed property differences among the different phenolic compounds can be attributed to their structural deviations and this is well accounted for by the molecular descriptors. The ability to model the selectivity index provides useful insights on the theoretical design of novel artificial receptors specific for pentachlorophenol and related

Table 2: Initial comparison of quantum chemical and E-DRAGON descriptors.

Correlation coefficient	Quantum chemical descriptors		E-DRAGON descriptors	
	PLS1	PCR	PLS1	PCR
$r_{Training}$ [a]	0.8571	0.8516	0.9447	0.8250
$r_{Testing}$ [b]	0.8331	0.8333	0.8441	0.7869

[a] Training set correlation coefficient
[b] Testing set correlation coefficient

compounds based on molecular imprinting. Furthermore, the QSPR model could be used to help control the degree of polymer selectivity toward pentachlorophenol and related phenols. Therefore, MIPs with predefined selectivity toward template molecule of interest could be realized.

Initial Comparison of Molecular Descriptors

Assessment of the initial performance of unprocessed data set prior to optimization of the number of molecular descriptors was performed and results are presented in Table 2. It was observed that PLS and PCR performed at similar level of performance when using quantum chemical descriptors. On the other hand, PLS yield better results than PCR when E-DRAGON descriptors were used. Both types of molecular descriptors shows similar level of performance as indicated from the testing set correlation coefficient in excess of 0.83 for quantum chemical descriptors modeled by PLS and PCR, and for E-DRAGON descriptors modeled by PLS.

Reduction of Molecular Descriptors and Prediction of Selectivity Index

The intial number of molecular descriptors derived from E-DRAGON amounted to 1,666 variables. They were scaled to mean of zero and unit variance by standardization using equation 2. The Unscrambler software detected that 436 were constant variables and was automatically removed to yield a reduced set of 1,230 descriptors. Of the two PC-based regression methods, PLS was found to perform better than PCR as observed from the greater correlation coefficent values. PLS had training set correlation coefficient ($r_{Training}$) and testing set correlation coefficient ($r_{Testing}$) of 0.9447 and 0.8441, respectively, while PCR obtained $r_{Training}$ = 0.8250 and $r_{Testing}$ = 0.7869. Therefore, PLS was chosen as the optimal PC-based regression method for further investigations.

The variables were filtered according to equation 4 based on the confidence interval of regression coefficients derived from PLS. Briefly, those situating inside the defined confidence interval were removed as they were considered to be redundant variables, whereas those located outside the defined confidence interval were retained for further processing. For example, at the 90% confidence interval 1,113 variables were found located outside the confidence interval, thus warranting their removal from the data set. This generated a reduced subset of 117

Table 3: Summary of variable selection as a function of the level of confidence.

CI (%)	z-score	N_{CI}	$r_{CI}^{Training}$	$r_{CI}^{Testing}$	N_{CI+UFS}	$r_{CI+UFS}^{Training}$	$r_{CI+UFS}^{Testing}$
75.0	1.15	235	0.9580	0.8937	43	0.9377	0.8245
80.0	1.28	195	0.9472	0.8866	43	0.9484	0.8579
90.0	1.65	117	0.9408	0.8837	40	0.9525	**0.8913**
95.0	1.96	72	0.9111	0.8523	34	0.9205	0.8684
99.0	2.58	38	0.8999	0.8492	23	0.9029	0.8629
99.9	3.30	2	0.3782	0.1918	—	—	—

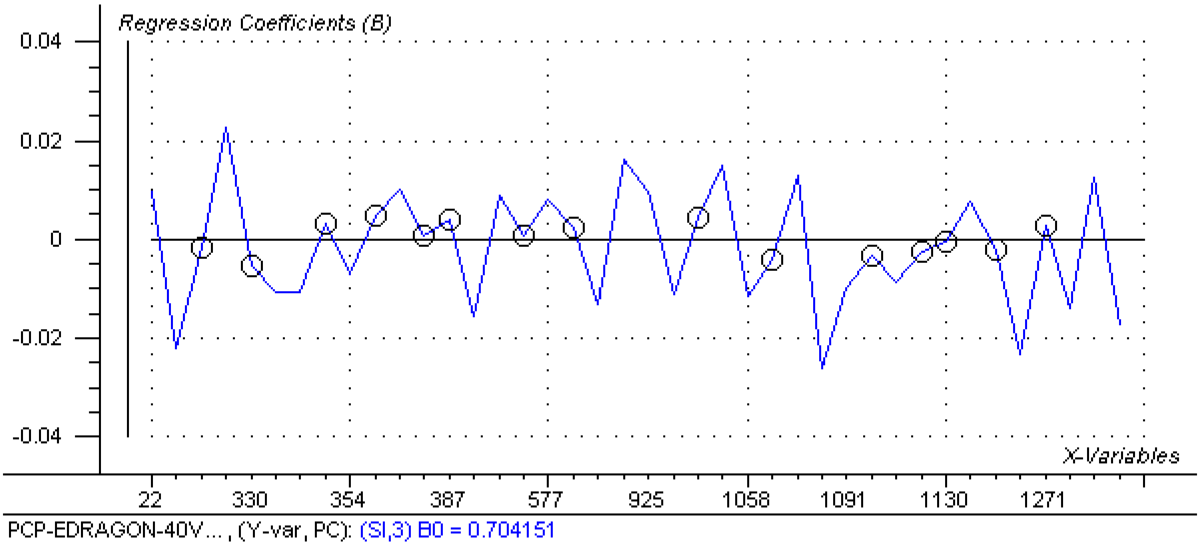

Figure 1: Plot of the PLS regression coefficients as a function of the variables. Descriptors marked with an empty circle were subjected to removal as they possess low regression coefficient values.

variables. Second phase of filtering with UFS were performed to remove redundant multi-collinear variables. The 117 variables were further reduced to a subset of 40 variables. The same procedures were performed for the other five level of confidence and are summarized in Table 3. The 40 variables were comprised of a mixture of steric and electronic descriptors as shown in Table 4. This subset of variables gave rather good predictive ability as indicated by the r_{Training} and r_{Testing} values of 0.9525 and 0.8913, respectively.

A third phase of variable filtering was performed by removing descriptors possessing regression coefficient near the origin axis as observed from Figure 1. This led to the removal of 15 additional variables (Table 4), further condensing the variables to a subset of 25. The eliminated variables were made up of a combination of steric and electronic descriptors comprising mostly of 2D autocorrelation indices, WHIM, and GETAWAY descriptors. The reduced subset

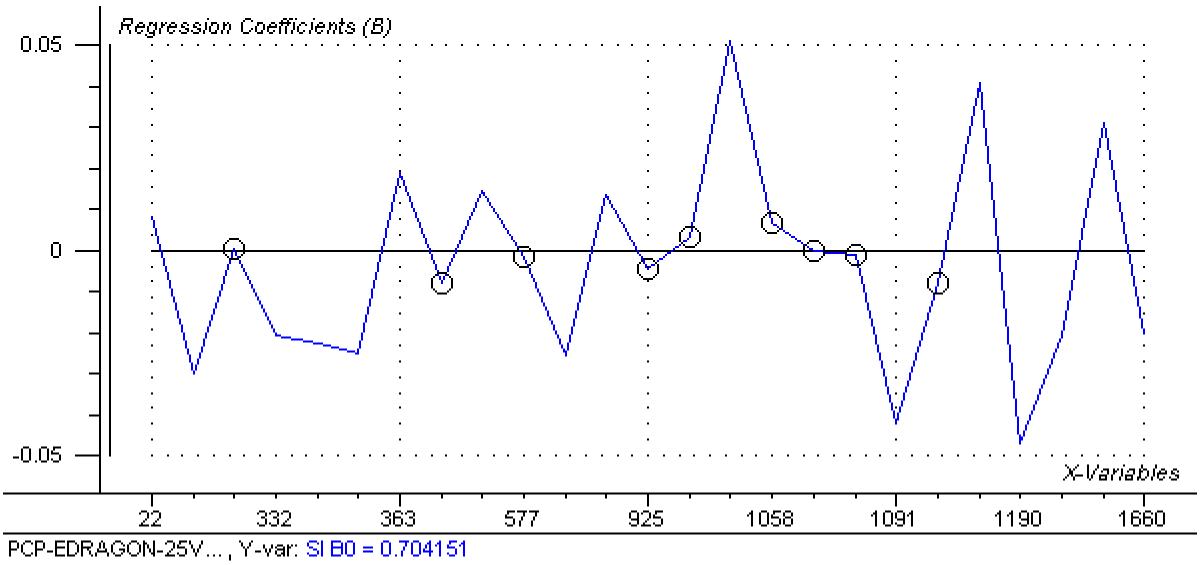

Figure 2: Plot of the MLR regression coefficients as a function of the variables. Descriptors marked with an empty circle were subjected to removal as they possess low regression coefficient values.

increased the predictive power slightly as observed from the $r_{Training}$ and $r_{Testing}$ values of 0.9531 and 0.9054, respectively.

The set of 25 variables was then modeled by MLR but was shown to give lower predictive power than the PLS model as indicated from the correlation coefficients of $r_{Training}$ = 0.9681 and $r_{Testing}$ = 0.8772. A plot of the regression coefficients prompted further variable reduction by removing descriptors having low regression coefficients as shown in Figure 2. This resulted in the elimination of 9 variables to give an optimal set of 16 descriptors (Table 4). The 9 variables were made up of topological descriptor, edge adjacency indices, topological charge indices, 3D-MoRSE descriptors, and WHIM descriptors. Results indicated that the final phase of variable reduction on the MLR model boosted the predictive power significantly to $r_{Training}$ = 0.9657 and $r_{Testing}$ = 0.9332. A plot of the predicted *SI* versus the experimental *SI* as modeled by MLR is shown in Figure 3A.

For the benefit of comparison, the same set of 16 descriptors was then modeled by PLS (Figure 3B). It was observed that the predictive performance of PLS with $r_{Training}$ = 0.9636 and $r_{Testing}$ = 0.9380 was slightly higher than MLR but the superiority in performance was not significant. The superiority of PLS over MLR on the set of 25 variables can possibly be explained by the non-linear nature of the descriptor matrix. Since the PLS approach is capable of handling non-linear data well, it outperforms the MLR approach. The removal of 9 variables transformed the descriptor matrix to a linear form, which boosted the performance of MLR by 0.056 from the testing set correlation coefficient of $r_{Testing}$ = 0.8772 to $r_{Testing}$ = 0.9332. The linearity of the data was confirmed further by the comparable level of performance of PLS to the linear MLR method as observed from the testing set correlation coefficient of $r_{Testing}$ = 0.9380 and $r_{Testing}$ = 0.9332, respectively. It should be noted that when the data is of linear form, the non-linear approach is not necessary and so reverts to the simplified linear approach.

Final Assessment of Molecular Descriptors

Since the number of molecular descriptors for E-DRAGON has been optimized as outlined above, the next step is to assess the performance of the new QSPR model in relation to the previously reported model (Baggiani et al., 2004). A summary of the results is displayed in Table 5. The data set using quantum chemical descriptors was pre-processed according to Baggiani et al. (Baggiani et al., 2004) by removing 8 variables to obtain a “minimum dimensionality model” that gave the same level of performance as the model using all descriptors. Furthermore, Baggiani et al. removed 3 outliers (sample number 48, 49, and 52) from their model, thus for the benefit of comparison, in this investigation the outliers were removed from the data sets. Results indicate that the quantum chemical descriptors gave similar level of performance for both PLS and PCR methods, whereas PLS was the better performing approach when E-DRAGON descriptors were used. The removal of 3 outliers did not exert a significant influence on the predictive performance of the PLS model as indicated from the slight drop in performance from $r_{Testing}$ = 0.9380 to $r_{Testing}$ = 0.9294 for the data set with all data samples intact and the data set omitting 3 outliers, respectively.

CONCLUSION

In summary, we have demonstrated the feasibility of using molecular descriptors derived from E-DRAGON in modeling the selectivity index of pentachlorophenol-imprinted polymer. The variable reduction method used in this study starts with the reduction of the variable dimension from 1,666 to 117 descriptors using the confidence interval approach. This is followed by further removal of multi-collinear variables with the UFS algorithm from 117 to 40 descriptors. Moreover, 15 additional descriptors were removed by filtering off descriptors bearing low PLS regression coefficient to give a set

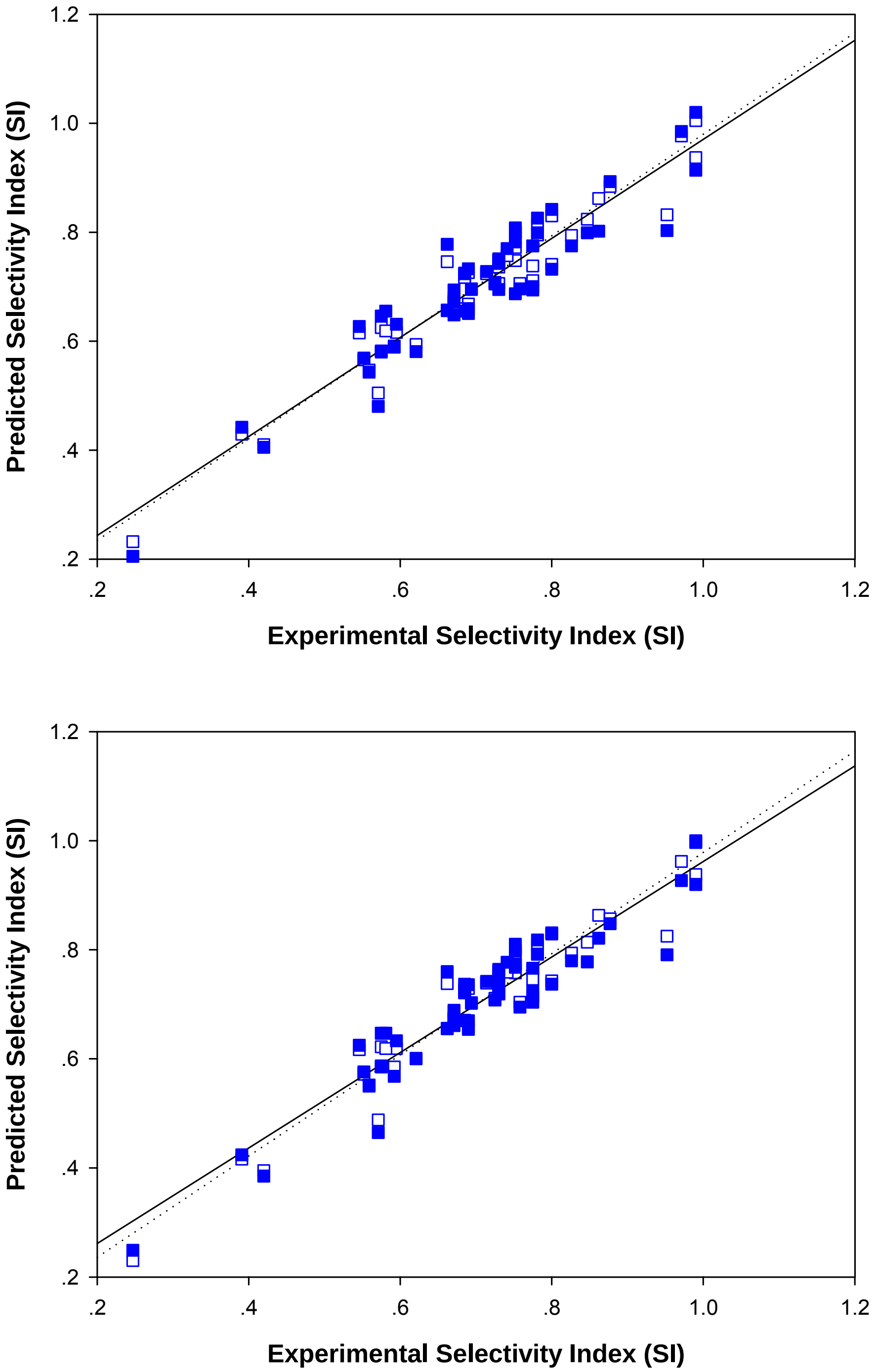

Figure 3: Plot of the predicted *SI* as a function of the experimental *SI* for the training set (□; regression line is represented as dotted line) and testing set (■; regression line is represented as solid line).

of 25 variables. A final phase of variable reduction was performed on the set of 25 variables modeled by MLR using the same criteria as the previous step by removing variables with low MLR regression coefficients. Our results indicated that PLS regression and MLR reliably predicted the *SI* of the phenolic compounds as observed from the correlation coefficient of 0.9380 and 0.9332, respectively. The QSPR model investigated in this study are valuable for predicting the selectivity index of a library of related compounds and provides theoretical guidance for molecular design as observed from the retention property as it is influenced by its substituents.

Acknowledgements: Financial support from the Thailand Research Fund through the Royal Golden Jubilee Ph.D. Program to C.N. under V.P. supervision is gratefully acknowledged. We also acknowledge partial financial support from Thailand Toray Science Foundation (TTSF) and a grant from the annual budget of Mahidol University (B.E.2546-2550).

Table 4: Reduced subset of 40 descriptors obtained after variable reduction with UFS.

Descriptor	Description	Type of descriptor
nDB	number of double bonds	constitutional descriptors
MAXDP	maximal electrotopological positive variation	topological descriptors
ICR [a]	radial centric information index	topological descriptors
T(O..O) [b]	sum of topological distances between O..O	topological descriptors
MATS4m [a]	Moran autocorrelation - lag 4 / weighted by atomic masses	2D autocorrelation indices
MATS6m	Moran autocorrelation - lag 6 / weighted by atomic masses	2D autocorrelation indices
MATS3v	Moran autocorrelation - lag 3 / weighted by atomic van der Waals volumes	2D autocorrelation indices
MATS4e [a]	Moran autocorrelation - lag 4 / weighted by atomic Sanderson electronegativities	2D autocorrelation indices
MATS4p	Moran autocorrelation - lag 4 / weighted by atomic polarizabilities	2D autocorrelation indices
GATS4m [a]	Geary autocorrelation - lag 4 / weighted by atomic masses	2D autocorrelation indices
GATS5m	Geary autocorrelation - lag 5 / weighted by atomic masses	2D autocorrelation indices
GATS4v [a]	Geary autocorrelation - lag 4 / weighted by atomic van der Waals volumes	2D autocorrelation indices
GATS5p [a]	Geary autocorrelation - lag 5 / weighted by atomic polarizabilities	2D autocorrelation indices
EEig08d [b]	Eigenvalue 08 from edge adj. matrix weighted by dipole moments	edge adjacency indices
EEig12r	Eigenvalue 12 from edge adj. matrix weighted by resonance integrals	edge adjacency indices
JGI1 [a]	mean topological charge index of order1	topological charge indices
JGI6 [b]	mean topological charge index of order6	topological charge indices
DISPm [a]	d COMMA2 value / weighted by atomic masses	geometrical descriptors
RDF050m	Radial Distribution Function - 5.0 / weighted by atomic masses	RDF descriptors
RDF035v	Radial Distribution Function - 3.5 / weighted by atomic van der Waals volumes	RDF descriptors
Mor02m [b]	3D-MoRSE - signal 02 / weighted by atomic masses	3D-MoRSE descriptors
Mor22m [b]	3D-MoRSE - signal 22 / weighted by atomic masses	3D-MoRSE descriptors
Mor23m [a]	3D-MoRSE - signal 23 / weighted by atomic masses	3D-MoRSE descriptors
Mor29e	3D-MoRSE - signal 29 / weighted by atomic Sanderson electronegativities	3D-MoRSE descriptors
G2u [b]	2st component symmetry directional WHIM index / unweighted	WHIM descriptors
G2m [a]	2st component symmetry directional WHIM index / weighted by atomic masses	WHIM descriptors

G3m [b]	3st component symmetry directional WHIM index / weighted by atomic masses	WHIM descriptors
E1v [b]	1st component accessibility directional WHIM index / weighted by atomic van der Waals volumes	WHIM descriptors
G2e	2st component symmetry directional WHIM index / weighted by atomic Sanderson electronegativities	WHIM descriptors
G2p [a]	2st component symmetry directional WHIM index / weighted by atomic polarizabilities	WHIM descriptors
G3p [b]	3st component symmetry directional WHIM index / weighted by atomic polarizabilities	WHIM descriptors
E2p [a]	2nd component accessibility directional WHIM index / weighted by atomic polarizabilities	WHIM descriptors
Gu [a]	G total symmetry index / unweighted	WHIM descriptors
H6u	H autocorrelation of lag 6 / unweighted	GETAWAY descriptors
HATS5u [a]	leverage-weighted autocorrelation of lag 5 / unweighted	GETAWAY descriptors
HATS5m	leverage-weighted autocorrelation of lag 5 / weighted by atomic masses	GETAWAY descriptors
R5u+ [a]	R maximal autocorrelation of lag 5 / unweighted	GETAWAY descriptors
R2e+	R maximal autocorrelation of lag 2 / weighted by atomic Sanderson electronegativities	GETAWAY descriptors
R7p+	R maximal autocorrelation of lag 7 / weighted by atomic polarizabilities	GETAWAY descriptors
Infective-80	Ghose-Viswanadhan-Wendoloski antiinfective-like index at 80%	molecular properties

[a] Set of 15 variables with low PLS regression coefficients were criteria for removal.
[b] Set of 9 variables low MLR regression coefficients were criteria for removal.

Table 5: Final assessment of quantum chemical and E-DRAGON descriptors.

Correlation coefficient	Quantum chemical descriptors [c]		E-DRAGON descriptors [d]	
	PLS1	PCR	PLS1	PCR
$r_{Training}$ [a]	0.8696	0.8687	0.9604	0.9364
$r_{Testing}$ [b]	0.8429	0.8402	0.9294	0.8813

Both data sets were subjected to removal of 3 outliers reported by Baggiani et al. (sample no. 48, 49, and 52).
[a] Training set correlation coefficient
[b] Testing set correlation coefficient
[c] 7 descriptors derived from the "minimum dimensionality model" of ref. 48
[d] 16 descriptors obtained from a series of variable filter and reduction

REFERENCES

Agency for Toxic Substances and Disease Registry (ATSDR). Toxicological profile for phenol. Atlanta, GA: U.S. Department of Health and Human Services, Public Health Service. 1998.

Agostini E, Coniglio MS, Milrad SR, Tigier HA, Giulietti AM. Phytoremediation of 2,4-dichlorophenol by Brassica napus hairy root cultures. *Biotechnol Appl Biochem*. 2003;37:139-44.

Ahlborg UG, Thunberg TM. Chlorinated phenols: occurrence, toxicity, metabolism, and environmental impact. *Crit Rev Toxicol*. 1980;7:1-35.

Ansell RJ, Mosbach K. Magnetic molecularly imprinted polymer beads for drug radioligand binding assay. *Analyst*. 1998;123:1611-6.

Baggiani C, Anfossi L, Giovannoli C, Tozzi C. Multivariate analysis of the selectivity for a pentachlorophenol-imprinted polymer. *J Chromatogr B*. 2004;804:31-41.

Bali U, Catalkaya EC, Sengul F. Photochemical degradation and mineralization of phenol: a comparative study. *J Environ Sci Health A*. 2003;38:2259-75.

Bollag JM, Shuttleworth KL, Anderson DH. Laccase-mediated detoxification of phenolic compounds. *Appl Environ Microbiol*. 1988;54:3086-91.

Bruce RM, Santodonato J, Neal MW. Summary review of the health effects associated with phenol. *Toxicol Ind Health*. 1987;3:535-68.

Bruggemann O. Molecularly imprinted materials: receptors more durable than nature can provide. *Adv Biochem Eng Biotechnol*. 2002;76:127-63.

Buchanan ID, Nicell JA. Model development for horseradish peroxidase catalyzed removal of aqueous phenol. *Biotechnol Bioeng*. 1997;54:251-61.

Caro E, Masque N, Marce RM, Borrull F, Cormack PA, Sherrington DC. Non-covalent and semi-covalent molecularly imprinted polymers for selective on-line solid-phase extraction of 4-nitrophenol from water samples. *J Chromatogr A*. 2002;963:169-78.

Caro E, Marce RM, Cormack PA, Sherrington DC, Borrull F. On-line solid-phase extraction with molecularly imprinted polymers to selectively extract substituted 4-chlorophenols and 4-nitrophenol from water. *J Chromatogr A*. 2003;995:233-8.

Catalkaya EC, Bali U, Sengul F. Photochemical degradation and mineralization of 4-chlorophenol. *Environ Sci Pollut Res Int*. 2003;10:113-20.

Chianella I, Lotierzo M, Piletsky SA, Tothill IE, Chen B, Karim K, Turner AP. Rational design of a polymer specific for microcystin-LR using a computational approach. *Anal Chem*. 2002;74:1288-93.

Detomaso A, Lopez A, Lovecchio G, Mascolo G, Curci R. Practical applications of the Fenton reaction to the removal of chlorinated aromatic pollutants. Oxidative degradation of 2,4-dichlorophenol. *Environ Sci Pollut Res Int*. 2003;10:379-84.

Esbensen KH. Multivariate Data Analysis: in practice. Fifth ed. Oslo: CAMO Process AS. 2004.

Exon JH. A review of chlorinated phenols. *Vet Hum Toxicol*. 1984;26:508-20.

Gasteiger J, Rudolph C, Sadowski J. Automatic Generation of 3D Atomic Coordinates for Organic Molecules. *Tetrahedron Comp Method*. 1990;3:537-47.

Gogate PR, Mujumdar S, Thampi J, Wilhelm AM, Pandit AB. Destruction of phenol using sonochemical reactors: scale up aspects and comparison of novel configuration with conventional reactors. *Sep Purif Technol*. 2004;34:25-34.

Harvey PJ, Campanella BF, Castro PM, Harms H, Lichtfouse E, Schaffner AR, Smrcek S, Werck-Reichhart D. Phytoremediation of polyaromatic hydrocarbons, anilines and phenols. *Environ Sci Pollut Res Int*. 2002;9:29-47.

Haupt K. Imprinted polymers-tailor-made mimics of antibodies and receptors. *Chem Commun*. 2003;21:171-8.

Hsieh RY, Tsai HA, Syu MJ. Designing a molecularly imprinted polymer as an artificial receptor for the specific recognition of creatinine in serums. *Biomaterials*. 2006;27:2083-9.

Huang X, Kong L, Li X, Zheng C, Zou H. Molecular imprinting of nitrophenol and hydroxybenzoic acid isomers: effect of molecular structure and acidity on imprinting. *J Mol Recognit*. 2003;16:406-11.

Kavitha V, Palanivelu K. The role of ferrous ion in Fenton and photo-Fenton processes for the degradation of phenol. *Chemosphere*. 2004;55:1235-43.

Machtejevas E, Sellergren B, Martynaitis V, Owens PK, Maruska A. Screening of oxazepine indole enantiomers by means of high performance liquid chromatography with imprinted polymer stationary phase. *J Sep Sci*. 2004;27:547-51.

Martin-Esteban A. Molecularly imprinted polymers: new molecular recognition materials for selective solid-phase extraction of organic compounds. *Fresenius J Anal Chem*. 2001;370:795-802.

Masque N, Marce RM, Borrull F, Cormack PA, Sherrington DC. Synthesis and evaluation of a molecularly imprinted polymer for selective on-line solid-phase extraction of 4-nitrophenol from environmental water. *Anal Chem*. 2000;72:4122-6.

Mosbach K, Yu Y, Andersch J, Ye L. Generation of new enzyme inhibitors using imprinted binding sites: the anti-idiotypic approach, a step toward the next generation of molecular imprinting. *J Am Chem Soc*. 2001;123:12420-1.

Nantasenamat C, Naenna T, Isarankura Na Ayudhya C, Prachayasittikul V. Quantitative prediction of imprinting factor of molecularly imprinted polymers by artificial neural network. *J Comput Aid Mol Des*. 2005a;19:509-24.

Nantasenamat C, Naenna T, Isarankura-Na-Ayudhya C, Prachayasittikul V. Recognition of DNA Splice Junction via Machine Learning Approaches. *EXCLI Journal*. 2005b;4:114-29.

Nantasenamat C, Isarankura-Na-Ayudhya C, Tansila N, Naenna T, Prachayasittikul V. Prediction of GFP spectral properties using artificial neural network. *J Comput Chem*. 2006;In press.

Piacham T, Isarankura Na Ayudhya C, Prachayasittikul V, Bülow L, Ye L. A polymer supported manganese catalyst useful as a superoxide dismutase mimic. *Chem Commun*. 2003;(11):1254-5.

Piacham T, Josell A, Arwin H, Prachayasittikul V, Ye L. Molecularly imprinted polymer thin films on quartz crystal microbalance using a surface bound photo-radical initiator. *Anal Chim Acta*. 2005;536:191-6.

Piletsky SA, Piletska EV, Chen B, Karim K, Weston D, Barrett G, Lowe P, Turner AP. Chemical grafting of molecularly imprinted homopolymers to the surface of microplates. Application of artificial adrenergic receptor in enzyme-linked assay for beta-agonists determination. *Anal Chem*. 2000;72:4381-5.

Ramstrom O, Ye L, Mosbach K. Artificial antibodies to corticosteroids prepared by molecular imprinting. *Chem Biol*. 1996;3:471-7.

Sadowski J, Gasteiger J. From atoms and bonds to three-dimensional atomic coordinates: automatic model builders. *Chem Rev*. 1993;93:2567-81.

Sadowski J, Gasteiger J, Klebe G. Comparison of Automatic Three-Dimensional Model Builders Using 639 X-ray Structures. *J Chem Inf Comput Sci*. 1994;34:1000-8.

Sanbe H, Hosoya K, Haginaka J. Preparation of uniformly sized molecularly imprinted polymers for phenolic compounds and their application to the assay of bisphenol A in river water. *Anal Sci*. 2003;19:715-9.

Sanbe H, Haginaka J. Uniformly sized molecularly imprinted polymers for bisphenol A and beta-estradiol: retention and molecular recognition properties in hydro-organic mobile phases. *J Pharm Biomed Anal*. 2003;30:1835-44.

Santos de Araujo B, Charlwood BV, Pletsch M. Tolerance and metabolism of phenol and chloroderivatives by hairy root cultures of Daucus carota L. *Environ Pollut*. 2002;117:329-35.

Sellergren B, Andersson LI. Application of imprinted synthetic polymers in binding assay development. *Methods*. 2000;22:92-106.

Shaw LJ, Beaton Y, Glover LA, Killham K, Meharg AA. Development and characterization of a lux-modified 2,4-dichlorophenol-degrading Burkholderia sp. RASC. *Environ Microbiol*. 1999;1:393-9.

Sinclair GM, Paton GI, Meharg AA, Killham K. Lux-biosensor assessment of pH effects on microbial sorption and toxicity of chlorophenols. *FEMS Microbiol Lett*. 1999;174:273-8.

Weitz HJ, Campbell CD, Killham K. Development of a novel, bioluminescence-based, fungal bioassay for toxicity testing. *Environ Microbiol*. 2002;4:422-9.

Spegel P, Schweitz L, Nilsson S. Molecularly imprinted polymers in capillary electrochromatography: recent developments and future trends. *Electrophoresis*. 2003;24:3892-9.

Spivak DA, Campbell J. Systematic study of steric and spatial contributions to molecular recognition by non-covalent imprinted polymers. *Analyst*. 2001;126:793-7.

Spivak DA. Selectivity in Molecularly Imprinted Matrices. In: Yan M, Ramström O, editors. Molecularly Imprinted Materials: Science and Technology. New York: Marcer Dekker/CRC Press. 2004. p. 395-417.

Tai DF, Lin CY, Wu TZ, Huang JH, Shu PY. Artificial receptors in serologic tests for the early diagnosis of dengue virus infection. *Clin Chem*. 2006;52:1486-91.

Takeuchi T, Haginaka J. Separation and sensing based on molecular recognition using molecularly imprinted polymers. *J Chromatogr B*. 1999;728:1-20.

Tamayo FG, Martin-Esteban A. Selective high performance liquid chromatography imprinted-stationary phases for the screening of phenylurea herbicides in vegetable samples. *J Chromatogr A*. 2005;1098:116-22.

Tetko IV, Gasteiger J, Todeschini R, Mauri A, Livingstone D, Ertl P, Palyulin VA, Radchenko EV, Zefirov NS, Makarenko AS, Tanchuk VY, Prokopenko VV. Virtual computational chemistry laboratory - design and description. *J Comput Aid Mol Des*. 2005;19:453-63.

Todeschini R, Consonni V, Mannhold R, Kubinyi H, Timmerman H. Handbook of Molecular Descriptors. Weinheim: Wiley-VCH. 2000.

Turiel E, Martin-Esteban A. Molecularly imprinted polymers: towards highly selective stationary phases in liquid chromatography and capillary electrophoresis. *Anal Bioanal Chem*. 2004;378:1876-86.

VCCLAB. Virtual Computational Chemistry Laboratory, http://www.vcclab.org, 2005.

Vlatakis G, Andersson LI, Muller R, Mosbach K. Drug assay using antibody mimics made by molecular imprinting. *Nature*. 1993;361:645-7.

Wang GD, Li QJ, Luo B, Chen XY. Ex planta phytoremediation of trichlorophenol and phenolic allelochemicals via an engineered secretory laccase. *Nat Biotechnol*. 2004;22:893-7.

Watabe Y, Kubo T, Nishikawa T, Fujita T, Kaya K, Hosoya K. Fully automated liquid chromatography-mass spectrometry determination of 17beta-estradiol in river water. *J Chromatogr A*. 2006;1120:252-9.

Weininger D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J Chem Inf Comput Sci*. 1988;28:31-6.

Whitley DC, Ford MG, Livingstone DJ. Unsupervised Forward Selection: A Method for Eliminating Redundant Variables. *J Chem Inf Model*. 2000;40:1160-8.

Witten IH, Frank E. Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations. San Francisco: Morgan Kaufmann Publishers. 2000.

Wright H, Nicell JA. Characterization of soybean peroxidase for the treatment of aqueous phenols. *Bioresource Technol*. 1999;70:69-79.

Ye L, Yu Y, Mosbach K. Towards the development of molecularly imprinted artificial receptors for the screening of estrogenic chemicals. *Analyst*. 2001;126:760-5.

Ye L, Surugiu I, Haupt K. Scintillation proximity assay using molecularly imprinted microspheres. *Anal Chem*. 2002;74:959-64.