Asian Journal of Chemistry Vol. 19, No. 6 (2007), 4512-4522

Application of High-Order MCI Method to Predict Aqueous Solubility of Aliphatic Alcohols

YING-LONG WANG†, YANG-DONG HU*, GUO-DONG LI and QIN-SHENG WEI
College of Chemistry and Chemical Engineering, Ocean University of China
Qingdao 266003, P.R. China
Tel:(86)(134)55231781; E-mail: chem_ouc@yahoo.com.cn

Correlations for estimation of the aqueous solubility (ln S) of aliphatic alcohols are proposed. The high-order MCI (molecular connectivity index) based quantitative structure-property relationship (QSPR) models are obtained by stepwise regression and support vector regression. On the basis of the data set of 50 aliphatic alcohols, the optimal linear model obtained by stepwise regression has a correlation coefficient of 0.990 and an average absolute error of 0.149 ln units and is comparable with the existing models. The optimal nonlinear model obtained by support vector regression has a correlation coefficient of 0.996 and an average absolute error of 0.116 ln units and is better than the existing models. The new models are predictive and easy to apply for it requires only connectivity indices in the calculations and does not require any experimental physicochemical properties in the calculation.

Key Words: QSPR, High-order molecular connectivity index, Support vector regression, Aqueous solubility, Aliphatic alcohols.

INTRODUCTION

Modeling of aqueous solubility of organic chemicals from molecular structure is a significant activity for aqueous solubility is particularly important physicochemical property of organic chemicals and at the same time is a property which is difficult to measure experimentally. Reliable computational methods to predict aqueous solubility have been the focus of relative researches in recent years. Comprehensive reviews on developed water solubility computational methods, main issues that affect the applicability of different techniques and aspects of emerging scientific understanding that may lead to breakthroughs in the computational modeling of aqueous solubility have been reported recently[1-3].

There are a large number of reports relating to the estimation of this highly important property, which can be divided into two main groups based on model parameters. The first approach[4-8] is to build models from more

†College of Chemical Engineering, Qingdao University of Science and Technology, Qingdao 266042, P.R. China.

easily measured physicochemical properties, such as melting point, boiling point, molar volume, partition coefficient, chromatographic retention time, *etc*. The other method is based on the information from the molecules of the organic chemicals which can be further divided into two classes, one is group contributions method[9-12] and the other is QSPR approach[13-37].

As for the model method, there are also two main groups, one is linear method and the other is nonlinear method. Representatives of the two methods are multiple linear regression and neural networks[38]. Recently, a novel type of modeling method, support vector regression, has been adopted as a popular method in QSPR research.

The molecular connectivity indices which proposed about 30 years ago, have been successfully used in the calculations of various physiochemical properties of organic chemicals especially in the applications to computational molecular design studies recently[39-43]. In the previous works, correlations of aqueous solubility using molecular connectivity indices and other descriptors have been studied and demonstrated the possibility of molecular connectivity indices in modeling aqueous solubility. Zhong and Hu[29] have developed a nonlinear model with R^2 of 0.885 to predict the aqueous solubility of organic compounds, with three molecular connectivity indices involved. Wang and Hu[35,37] proposed the MCI based linear models to predict aqueous solubility of different chemicals.

In this study, we first use high-order molecular connectivity indices to develop linear model that relate the structures of a group of 50 aliphatic alcohols to their aqueous solubility and obtain the new model with the same accuracy comparing with the existing models, then use the five molecular descriptors to build the nonlinear model by the support vector regression and obtain the new model with the higher accuracy.

EXPERIMENTAL

50 Aliphatic alcohols that has been studied by Amic[22] and Zakarya[28] are adopted as data set and listed in Table-3. This data set consists of different types of structures: aliphatic (linear and branched), primary, secondary and tertiary carbon. The water solubility used is expressed in ln S and their values ranges from -8.2208 for decanol up to 0.0953 for 1-butanol.

Methods

High-order molecular connectivity index and stepwise regression: The simple and valence connectivity indices defined by earlier workers[44-50] that used in this work can be expressed in the following equation:

$$^{m}\chi_{k}^{q} = \sum_{j=1}^{n_m} \prod_{i=1}^{m+1} \left(\delta_{i}^{q}\right)_{j}^{-0.5} \tag{1}$$

where m is the order of the connectivity index, k denotes a continuous path type of fragment, which is divided into paths (p), clusters (c), *etc.,* q denotes connectivity index is simple or valence or other types; n_m is the number of the relevant paths; δ_i^q is the connectivity index.

In this work, for each chemical, the values of the connectivity indices up to third order are calculated using the vertex adjacency matrix. The simple connectivity index (δ) the valence connectivity index (δ^v) used in this study are summarized in Table-1.

TABLE-1
CONNECTIVITY INDEX VALUES OF GROUPS USED IN THIS WORK

Group	δ	δ^v	Group	δ	δ^v
$-CH_3$	1	1	$=CH_2$	1	2
$-CH_2-$	2	2	=CH-	2	3
>CH-	3	3	=C<	3	4
>C<	4	4	-OH	1	5

The detailed equations used in this work for the simple and valence molecular connectivity indices for zeroth, first, second and third orders are listed as follows:

$$^0\chi = \sum_{\text{vertices}} \frac{1}{\sqrt{\delta_i}} \quad (2)$$

$$^0\chi^v = \sum_{\text{vertices}} \frac{1}{\sqrt{\delta_i^v}} \quad (3)$$

$$^1\chi = \sum_{\text{edges}} \frac{1}{\sqrt{\delta_i \delta_j}} \quad (4)$$

$$^1\chi^v = \sum_{\text{edges}} \frac{1}{\sqrt{\delta_i^v \delta_j^v}} \quad (5)$$

$$^2\chi = \sum_{\text{triplets}} \frac{1}{\sqrt{\delta_i \delta_j \delta_k}} \quad (6)$$

$$^2\chi^v = \sum_{\text{triplets}} \frac{1}{\sqrt{\delta_i^v \delta_j^v \delta_k^v}} \quad (7)$$

$$^3\chi_p = \sum_{\text{quaternion-path}} \frac{1}{\sqrt{\delta_i \delta_j \delta_k \delta_l}} \quad (8)$$

$$^{3}\chi_{p}^{v} = \sum_{\text{quaternion-path}} \frac{1}{\sqrt{\delta_i^v \delta_j^v \delta_k^v \delta_l^v}} \tag{9}$$

$$^{3}\chi_{c} = \sum_{\text{quaternion-cluster}} \frac{1}{\sqrt{\delta_i \delta_j \delta_k \delta_l}} \tag{10}$$

$$^{3}\chi_{c}^{v} = \sum_{\text{quaternion-cluster}} \frac{1}{\sqrt{\delta_i^v \delta_j^v \delta_k^v \delta_l^v}} \tag{11}$$

After the calculation of 10 molecular connectivity indices, stepwise regression using MATLAB Statistics Toolbox are used in choosing the variables and fitting the experimental data of the data set.

The average absolute error (AAE) and the root-mean-square error (RMSE) were calculated as the following to compare with the existing model.

The AAE was calculated as

$$\text{AAE} = \frac{\sum \left| \log S_{cal} - \log S_{exp} \right|}{N} \tag{12}$$

The RMSE was calculated as

$$\text{RMSE} = \sqrt{\frac{\sum \left(\log S_{cal} - \log S_{exp} \right)^2}{N}} \tag{13}$$

where N is the number of compounds.

Support vector regression: The support vector machine, which introduced by Vapnik[51,52], has gained popularity both in pattern recognition and in QSPR/QSAR in recent years for its outstanding features and attractive principle of structure risk minimization. The principle of support vector regression is to map the input data *x* into a higher-dimensional feature space F and then to do regression between the target output z and the transformed *x*. The detailed description of support vector regression can be found in monographs of Vapnik[51,52].

Given a data set, $D = \{(x_1, z_1), (x_2, z_2) \cdots, (x_l, z_l)\} \subset R^n \times R$, where $x_i \in R^n$ is an input and $z_i \in R$ is a target output, the standard form of $\varepsilon - \text{SVR}$ can be expressed in the following form:

$$\min_{w,b,\xi,\xi^*} \frac{1}{2} w^T w + C \sum_{i=1}^{l} \xi_i + C \sum_{i=1}^{l} \xi_i^* \tag{14}$$

$$\text{s.t.} \quad b - z_i + w^T \phi(x_i) \le \varepsilon + \xi_i$$

$$z_i - w^T \phi(x_i) - b \le \varepsilon + \xi_i^*$$

$$\xi_i, \xi_i^* \geq 0, i = 1,2,L,1$$

The dual form of function (14) which has the following form is solved to get the Lagrange multiplier α, α^*:

$$\min_{\alpha,\alpha^*} \frac{1}{2}(\alpha - \alpha^*)^T Q(\alpha - \alpha^*) + \varepsilon \sum_{i=1}^{1} (\alpha_i + \alpha_i^*) + \sum_{i=1}^{1} z_i (\alpha_i - \alpha_i^*) \quad (15)$$

$$\text{s.t.} \ \sum_{i=1}^{1} (\alpha_i - \alpha_i^*) = 0,$$

$$0 \leq \alpha_i, \alpha_i^* \leq C, \ i = 1,2,L,1$$

where $Q_{ij} = K(x_i, x_j) = \phi(x_i)^T \phi(x_j)$ is the kernel function that satisfies Mercer's condition.

Thus the decision function can be obtained with the following form:

$$f(x) = \sum_{i=1}^{1} (\alpha_i - \alpha_i^*) K(x_i, x) + b \quad (16)$$

where b can be determined according to Karush-Kuhn-Tucker conditions as the following form

$$b = y_j - \varepsilon - \sum_{i=1}^{1} (\alpha_i - \alpha_i^*) K(x_j, x_i) \quad (17)$$

$$\alpha_j \in (0, C)$$

In this work, the Gaussian radical basis function is used as the kernel in regression.

RESULTS AND DISCUSSION

Amic[22] performed the structure-water solubility modeling of aliphatic alcohols using the weighted path numbers and Zakarya[28] modeled the structure-water solubility of aliphatic alcohols using the multifunctional autocorrelation method. They both got the satisfying results. Linear predictive QSPR models that based on high-order molecular connectivity indices are proposed in this work to correlate the aqueous solubility of 50 aliphatic alcohols and the results are comparable with the existing models and the nonlinear model that developed by support vector regression has the better statistical results than the existing models.

Stepwise regression was used to develop the linear model for the prediction of aqueous solubility using the molecular connectivity index. The coefficients of the best correlation model for aqueous solubility of the 50 aliphatic alcohols used in this study are shown in Table-2 and eqn. 14. The best linear model contains 5 indices with different meanings.

TABLE-2
THE BEST CORRELATION MODEL OF lnS FOR 50 COMPOUNDS

n	Descriptor	Coefficient	t-test
0	Intercept	8.1229	–
1	$^1\chi$	-7.5933	-6.5015
2	$^1\chi^v$	5.2500	4.8608
3	$^2\chi^v$	-0.7004	-3.1738
4	$^3\chi_p^v$	0.4424	3.5451
5	$^3\chi_c$	0.4964	2.2567

The $^1\chi$ and $^1\chi^v$ that reflect the size of the molecule are the most important descriptors, as can be seen by their high t-test values. This conclusion is in agreement with the existing models. The other descriptors $^2\chi^v$, $^3\chi_p^v$ and $^3\chi_c$ that reflect the contribution of clusters in a molecule to aqueous solubility are also important in describing the aqueous solubility of aliphatic alcohols. This demonstrates again that higher-order connectivity indices contain a large mount of information about the molecule, the larger-scale structural features (such as branching), to name a one.

The linear model obtained is as the following general correlation:

$$\ln S = 8.1229 - 7.5933\,{}^1\chi + 5.2500\,{}^1\chi^v - 0.7004\,{}^2\chi^v + 0.4424\,{}^3\chi_p^v + 0.4964\,{}^3\chi_c \quad (18)$$

$R^2 = 0.990$, $F = 839.5$, $n = 50$

The results calculated with eqn. 18 are shown in Table-3 and the scatter plot is shown in Fig. 1. The AAE for our linear model is 0.149 and the RMSE is 0.048 indicating that the new model has comparable accuracy to the existing models.

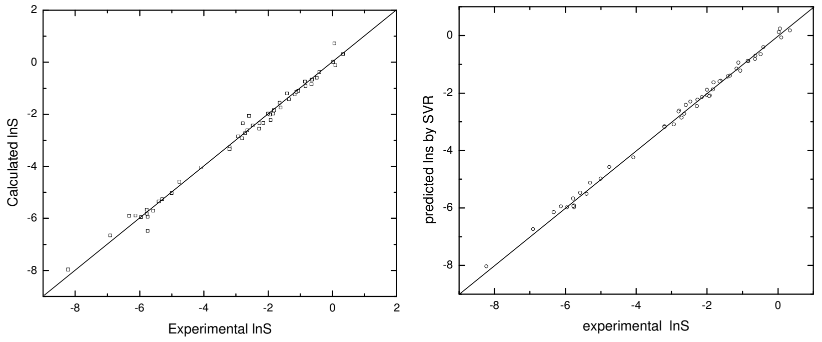

Fig. 1. Scatter plot of the calculated *vs.* the experimental values of ln S. –Linear, o–Non-linear

TABLE-3
CALCULATED AND EXPERIMENTAL VALUES OF THE MOLAR AQUEOUS SOLUBILITY FOR 50 ALIPHATIC ALCOHOLS

Compd. no.	Name	Experimental	Linear	Non-linear
1	2-Methyl-1-propanol	0.0227	0.0112	0.11469
2	2-Butanol	0.0658	0.7110	0.23675
3	3-Methyl-1-butanol	-1.1680	-1.2404	-1.15769
4	2-Methyl-1-butanol	-1.0584	-1.1079	-1.22727
5	2-Pentanol	-0.6349	-0.6760	-0.70465
6	3-Pentanol	-0.4861	-0.6026	-0.65680
7	3-Methyl-2-butanol	-0.4050	-0.3773	-0.40515
8	2-Methyl-2-butanol	0.3386	0.3117	0.17080
9	3-Hexanol	-1.8326	-1.9741	-1.87723
10	3-Methyl-3-pentanol	-0.8301	-0.9129	-0.89567
11	2-Methyl-2-pentanol	-1.1178	-1.1391	-0.94822
12	2-Methyl-3-pentanol	-1.6094	-1.7524	-1.57359
13	3-Methyl-2-pentanol	-1.6399	-1.5632	-1.60137
14	2,3-Dimethyl-2-butanol	-0.8510	-0.7460	-0.88252
15	3,3-Dimethyl-1-butanol	-2.5903	-2.0707	-2.41682
16	3,3-Dimethyl-2-butanol	-1.4106	-1.2049	-1.42845
17	4-Methyl-1-pentanol	-2.2828	-2.5565	-2.45789
18	4-Methyl-2-pentanol	-1.8140	-1.8491	-1.63931
19	2-Ethyl-1-butanol	-2.7871	-2.3502	-2.61375
20	2-Methyl-2-hexanol	-2.4734	-2.4344	-2.30304
21	3-Methyl-3-hexanol	-2.2634	-2.3504	-2.23631
22	3-Ethyl-3-pentanol	-1.9173	-2.2232	-2.08819
23	2,3-Dimethyl-2-pentanol	-2.0025	-1.9776	-2.08024
24	2,3-Dimethyl-3-pentanol	-1.9379	-2.0143	-2.11339
25	2,4-Dimethyl-2-pentanol	-2.1456	-2.3438	-2.14529
26	2,4-Dimethyl-3-pentanol	-2.8018	-2.9344	-2.65358
27	2,2-Dimethyl-3-pentanol	-2.6437	-2.6225	-2.72695
28	3-Heptanol	-3.1942	-3.2744	-3.16549
29	4-Heptanol	-3.1966	-3.3456	-3.18457
30	2,2,3-Trimethyl-3-pentanol	-2.9318	-2.8586	-3.09995
31	2-Octanol	-4.7560	-4.5936	-4.58207
32	2-Ethyl-1-hexanol	-4.9967	-5.0357	-4.98713
33	2-Nonanol	-6.3200	-5.9022	-6.15126
34	3-Nonanol	-6.1193	-5.8917	-5.95364
35	4-Nonanol	-5.9522	-5.9545	-5.98424
36	5-Nonanol	-5.7446	-5.9461	-5.98424

Compd. no.	Name	Experimental	Linear	Non-linear
37	2,6-Dimethyl-4-heptanol	-5.7764	-5.6775	-5.67899
38	3,5-Dimethyl-4-heptanol	-5.2983	-5.2698	-5.12628
39	2,2-Diethyl-1-pentanol	-5.5728	-5.7158	-5.48074
40	7-Methyl-1-octanol	-5.7446	-6.4825	-5.91937
41	3,5,5-Trimethyl-1-hexanol	-5.7699	-5.8187	-5.94183
42	1-Butanol	0.0953	-0.1146	-0.07249
43	1-Pentanol	-1.3471	-1.4233	-1.39511
44	2,2-Dimethyl-1-propanol	-0.6463	-0.8495	-0.81820
45	1-Hexanol	-2.7181	-2.7320	-2.85564
46	2-Hexanol	-1.9951	-1.9762	-1.89283
47	1-Heptanol	-4.0745	-4.0406	-4.24789
48	1-Octanol	-5.4015	-5.3493	-5.51666
49	1-Nonanol	-6.9078	-6.6580	-6.73726
50	1-Decanol	-8.2208	-7.9667	-8.03454

After the building of the linear model, $\varepsilon - \mathrm{SVR}$ is adopted to develop a nonlinear model based on the same five molecular connectivity indices. The parameters that affect the performances of SVR in this data set are sequentially optimized to get the best model. The parameters include capacity parameter C, ε of ε-insensitive loss function and the parameter γ of Gaussian function and their optimal values is found as 4000, 0.02 and 0.3. The detailed selection is shown in Figs. 2-4.

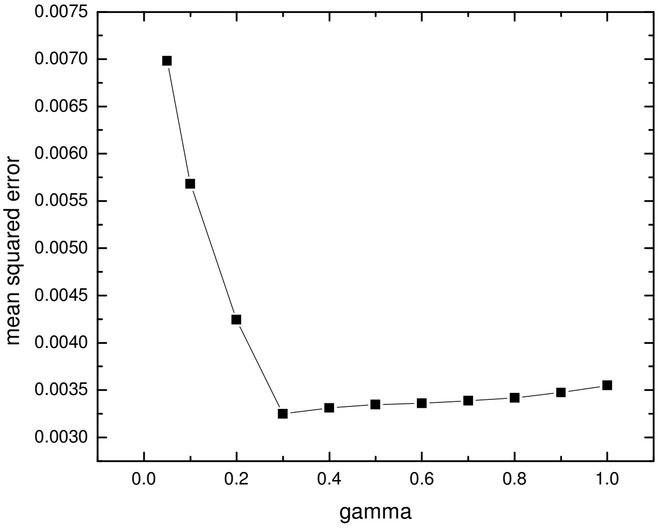

Fig. 2. The gamma *vs.* mean squared error in regression

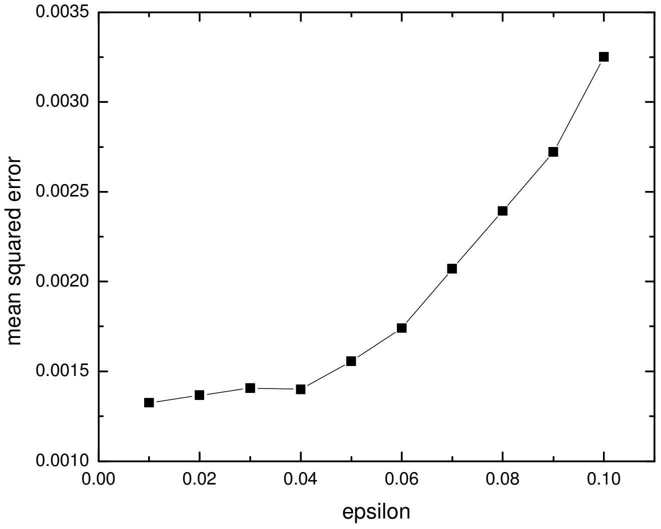

Fig. 3. The epsilon *vs.* mean squared error in regression

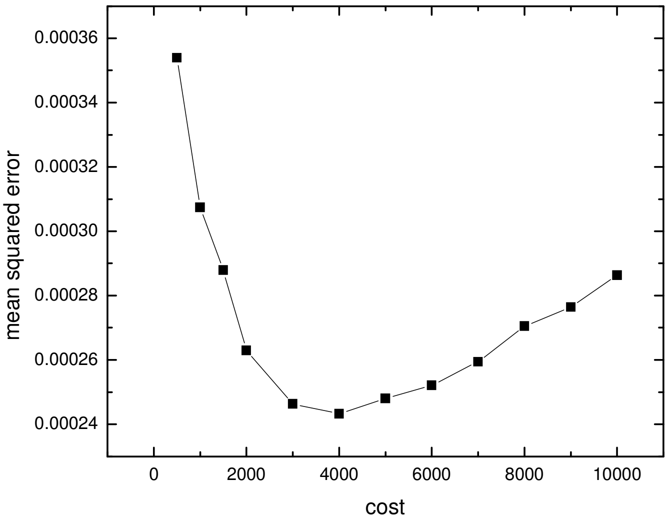

Fig. 4 The cost *vs.* mean squared error in regression

The results calculated with the nonlinear model built through SVR are shown in Table-3 and the scatter plot is shown in Fig. 1 to compare with the linear model. The AAE for our nonlinear model is 0.116 and the correlation coefficient is 0.996 indicating that the new nonlinear model has better accuracy than the existing models.

Conclusions

Linear and non-linear predictive QSPR models that based on high-order molecular connectivity indices are proposed in this work to correlate the aqueous solubility of 50 aliphatic alcohols. The optimal linear model obtained by stepwise regression has a correlation coefficient of 0.990 and an average absolute error of 0.149 ln units and is comparable with the existing models. The optimal nonlinear model obtained by support vector regression has a correlation coefficient of 0.996 and an average absolute error of 0.116 ln units and is better than the existing models. The new models are predictive and easy to apply for it requires only connectivity indices in the calculations and does not require any experimental physico-chemical properties in the calculation.

ACKNOWLEDGEMENT

The authors thank the support of the Natural Science Foundation of China (Contract: 20376078).

REFERENCES

1. J.S. Delaney, *Drug. Discov. Today*, **10**, 289 (2005).
2. S.R. Johnson and W. Zheng, *AAPS. J.*, **8**, E27 (2006).
3. W.L. Jorgensen and E.M. Duffy, *Adv. Drug. Deliver. Rev.*, **54**, 355 (2002).
4. S.H. Yalkowsky and R. Pinal, *Chemosphere*, **26**, 1239 (1993).
5. N. Jain and S.H. Yalkowsky, *J. Pharm. Sci.*, **90**, 234 (2001).
6. D.L. Peterson and S.H. Yalkowsky, *J. Chem. Inf. Model.*, **41**, 1531 (2001).
7. J. Tolls, J. Van Dijk, E.J.M. Verbruggen, J.L.M. Hermens, B. Loeprecht and G. Schüürmann, *J. Phys. Chem. A.*, **106**, 2760 (2002).
8. G. Yang, Y. Ran and S.H. Yalkowsky, *J. Pharm. Sci.*, **91**, 517 (2002).
9. G. Klopman, S. Wang and D.M. Balthasar, *J. Chem. Inf. Model.*, **32**, 474 (1992).
10. R. Kühne, R.-U. Ebert, F. Kleint, G. Schmidt and G. Schüürmann, *Chemosphere*, **30**, 2061 (1995).
11. G. Klopman and H. Zhu, *J. Chem. Inf. Model.*, **41**, 439 (2001).
12. T.J. Hou, T.J. Hou, K. Xia, W. Zhang and X.J. Xu, *J. Chem. Inf. Model.*, **44**, 266 (2004).
13. N.N. Nirmalakhandan and R.E. Speece, *Environ. Sci. Technol.*, **22**, 328 (1988).
14. T.M. Nelson and P.C. Jurs, *J. Chem. Inf. Model.*, **34**, 601 (1994).
15. P.D.T. Huibers and A.R. Katritzky, *J. Chem. Inf. Model.*, **38**, 283 (1998).
16. J. Huuskonen, M. Salo and J. Taskinen, *J. Chem. Inf. Model.*, **38**, 450 (1998).
17. M. Makino, *Environ. Int.*, **24**, 653 (1998).
18. B.E. Mitchell and P.C. Jurs, *J. Chem. Inf. Model.*, **38**, 489 (1998).
19. J. Huuskonen, *J. Chem. Inf. Model.*, **40**, 773 (2000).
20. R. Liu and S.S. So, *J. Chem. Inf. Model.*, **41**, 1633 (2001).
21. D. Yaffe, Y. Cohen, G. Espinosa, A. Arenas and F. Giralt, *J. Chem. Inf. Model.*, **41**, 1177 (2001).
22. D. Amic, S.C. Basak, B. Lucic, S. Nikolic and N. Trinajstic, *Sar. Qsar. Environ. Res.*, **13**, 281 (2002).
23. X.-Q. Chen, S.J. Cho, Y. Li and S. Venkatesh, *J. Pharm. Sci.*, **91**, 1838 (2002).
24. E.J. Delgado, *Fluid. Phase. Equilibr.*, **199**, 101 (2002).

25. O. Engkvist and P. Wrede, *J. Chem. Inf. Model.*, **42**, 1247 (2002).
26. G. Hua, S. Veerabahu and L. Pil, *Pharm. Res.*, **19**, 497 (2002).
27. A. Yan and J. Gasteiger, *QSAR Comb. Sci.*, **22**, 821 (2003).
28. M.N.D. Zakarya, *J. Mol. Model.*, **9**, 365 (2003).
29. C. Zhong and Q. Hu, *J. Pharm. Sci.*, **92**, 2284 (2003).
30. C.A.S. Bergstrom, C.M. Wassvik, U. Norinder, K. Luthman and P. Artursson, *J. Chem. Inf. Model.*, **44**, 1477 (2004).
31. J.S. Delaney, *J. Chem. Inf. Model.*, **44**, 1000 (2004).
32. J.R. Votano, M. Parham, L.H. Hall, L.B. Kier and L.M. Hall, *Chem. Biodivers.*, **1**, 1829 (2004).
33. A. Yan, J. Gasteiger, M. Krug and S. Anzali, *J. Comput. Aid. Mol. Des.*, **18**, 75 (2004).
34. C. Catana, H. Gao, C. Orrenius and P.F.W. Stouten, *J. Chem. Inf. Model.*, **45**, 170 (2005).
35. Y.-D. Hu and Y.-L. Wang, *Asian. J. Chem.*, **18**, 407 (2006).
36. D. Butina and J.M.R. Gola, *J. Chem. Inf. Model.*, **43**, 837 (2003).
37. Y.-L. Wang, Y.-D. Hu, L.-Y. Wu and W.-Z. An, *Int. J. Mol. Sci.*, **7**, 47 (2006).
38. D. Erös, G. Keri, I. Kovesdi, C. Szantai-Kis, G. Meszaros and L. Orfi, *Mini-Rev. Med. Chem.*, **4**, 167 (2004).
39. M. Randic, *J. Mol. Graph. Model.*, **20**, 19 (2001).
40. M. Randic, M. Pompe, D. Mills and S.C. Basak, *Molecules*, **9**, 1177 (2004).
41. K.V. Camarda and P. Sunderesan, *Indian Eng. Chem. Res.*, **44**, 4361 (2005).
42. S. Siddhaye, K. Camarda, M. Southard and E. Topp, *Comput. Chem. Eng.*, **28**, 425 (2004).
43. S. Chavali, B. Lin, D.C. Miller and K.V. Camarda, *Comput. Chem. Eng.*, **28**, 605 (2004).
44. L.B. Kier, L.H. Hall, W.J. Murray and M. Randic, *J. Pharm. Sci.*, **64**, 1971 (1975).
45. M. Randic, *J. Am. Chem. Soc.*, **97**, 6609 (1975).
46. L.B. Kier and L.H. Hall, Molecular Connectivity in Chemistry and Drug Research, New York, Academic Press (1976).
47. L.H. Hall and L.B. Kier, *J. Pharm. Sci.*, **66**, 642 (1977).
48. L.H. Hall and L.B. Kier, *J. Pharm. Sci.*, **67**, 1743 (1978).
49. L.B. Kier and L.H. Hall, Molecular Connectivity in Structure Activity Analysis, New York, Wiley (1986).
50. M. Randic, P.J. Hansen and P.C. Jurs, *J. Chem. Inf. Model.*, **28**, 60 (1988).
51. V. Vapnik, Statistical Learning Theory, New York, Wiley (1988).
52. V. Vapnik, The Nature of Statistical Learning Theory, New York: Springer-Verlag (1995).

(*Received*: 17 June 2006; *Accepted*: 21 April 2007) AJC-5588