MEDV−13 for QSRR of 62 Polychlorinated Naphthalenes

Shu Shen LIU*, Chun Sheng YIN, Lian Sheng WANG

State Key Laboratory of Pollution Control and Resources Reuse,
Department of Environmental Science & Engineering, Nanjing University, Nanjing 210093

Abstract: A molecular electronegativity distance vector based on 13 atomic types (MEDV−13) is used to describe the structures of 62 polychlorinated naphthalene (PCN) congeners and related to the gas chromatographic relative retention indices (*RI*s) of PCNs. Using multiple linear regression, a 4–variable quantitative structure–retention relationship (QSRR) with the correlation coefficient of estimations (*r*) being 0.9912 and the root mean square error of estimations (*RMSEE*) being 31.4 and the correlation coefficient of predictions (*q*) and the root mean square error of predictions (*RMSEP*) in the leave–one–out procedure are 0.9898 and 33.76, respectively.

Keywords: Molecular electronegativity distance vector (MEDV), polychlorinated naphthalene (PCN), gas chromatographic relative retention indice (*RI*), multiple linear regression (MLR).

How to describe the molecular structures is still one of the most important tasks in the quantitative structure–activity relationship (QSAR) techniques study. The current description methods include two–dimensional (2D) topological descriptors, energetic descriptors, quantum mechanical descriptors, and three–dimensional (3D) molecular field descriptors[1]. Although the QSAR methods based on the 3D structures such as CoMFA[2], GRID[3], COMPASS[4], and SOMFA[5], have been widely used in several scientific fields, implementing these methods is in general very difficult and time–consuming because of difficulty of generating optimal 3D conformation of the molecule under study. Many current excellent 2D QSAR methods such as the hologram QSAR (HQSAR) developed by Tong[6] and electrotopological state (E–state) index derived by Kier and Hall[7–8] have a comparable quality to the 3D methods. Recently, a novel molecular electronegativity distance vector based on 2D topological structure and 13 atomic types (MEDV−13) was reported[9–10] and used to derive several QSAR models between the MEDV−13 vector and the biological activities of some organic compounds including a set of 31 “benchmark” steroids binding to the corticosteroid–binding globulin (CBG), 58 dipeptides inhibiting angiotensin–converting enzyme (ACE), and 16 indomethacin amides and esters (ImAE) inhibiting cyclooxygenase–2 (COX–2) using the principal component regression (PCR) technique. In this paper, the MEDV−13 is employed to express the structures of 62 polychlorinated naphthalenes (PCNs), a series of environmental persistent compounds[11], and related to the gas chromatographic relative

[*] E-mail: ssliu@nju.edu.cn

retention indices (*RI*s) of the PCNs.

62 PCNs and their observed *RI*s taken from the literature[11] are listed in **Table 1**. The original MEDV–13 descriptors are calculated according to equation 4 shown in the literature[9]. In fact there are only 6 nonzero descriptors in the MEDV–13 of 62 PCNs due to absence of 10 atomic types of nos. 1, 4, 5, 6, 7, 8, 9, 10, 11, and 12. Here, the term "atomic type" is defined as the number of non–hydrogen atoms binding to that atom plus its identifying number (ID). The 6 nonzero descriptors are the 14th, 15th, 25th, 26th, 36th, and 91st element from the MEDV–13 descriptor (noted as $\boldsymbol{x}_{14}$, $\boldsymbol{x}_{15}$, $\boldsymbol{x}_{25}$, $\boldsymbol{x}_{26}$, $\boldsymbol{x}_{36}$, and $\boldsymbol{x}_{91}$). From the literaure[9], the 6 descriptors are closely related to three types of substructures such as –CH=, >C=, and –Cl and to 6 interactions between pairs of atomic types of nos. 2–2, 2–3, 3–3, 2–13, 3–13, and 13–13, which is just an description on the structures of PCNs.

Applying multiple linear regression (MLR), a quantitative structure–retention relationship (QSRR) model between the 6 MEDV–13 vector descriptors and the gas chromatography relative retention indices (*RI*s) of 62 PCNs is developed as follows:

$$\begin{aligned} RI = {} & (2860.8349 \pm 717.4991) + (3.3234 \pm 14.9545)\cdot\boldsymbol{x}_{14} \\ & - (77.0097 \pm 14.9741)\cdot\boldsymbol{x}_{15} + (91.7321 \pm 33.4925)\cdot\boldsymbol{x}_{25} \\ & - (231.2375 \pm 42.0688)\cdot\boldsymbol{x}_{26} + (175.7254 \pm 28.7429)\cdot\boldsymbol{x}_{36} \qquad (1) \\ & - (105.1958 \pm 142.1213)\cdot\boldsymbol{x}_{91} \end{aligned}$$

$$n = 62,\ m = 6,\ r{=}0.9922,\ RMSEE{=}\ 29.61,\ F{=}\ 578.16 \qquad \text{(Estimation)}$$

$$n = 62,\ m = 6,\ q{=}\ 0.9902,\ RMSEP{=}\ 33.04 \qquad \text{(LOO prediction)}$$

where *n* and *m* are respectively the number of samples and the nonzero MEDV–13 descriptors. The *r*, *RMSEE* and *F* are the correlation coefficient, the root mean square error, and *F* statistic of estimations, respectively. A good QSRR model should have not only an excellent estimation ability for the internal example but also a good prediction ability for the external example. So, a leave–one–out (LOO) cross validation procedure is used to test the prediction ability of the model built. The *q* and *RMSEP* refer to the correlation coefficient (*q*) and the root mean square error (*RMSEP*) of predictions obtained in the LOO procedure. From equation 1, the 6–variable QSRR model has high estimated and predicted ability. However, the 6–variable model is not robust because the absolute values of the regression coefficients (-3.3234 and -105.1958) of two descriptors, $\boldsymbol{x}_{14}$ and $\boldsymbol{x}_{91}$, are little than the corresponding standard deviations (14.9545 and 142.1213). It is essential to deleted them from the model in order to insure the stability of the model. Relationship analysis on the 6 nonzero MEDV descriptors also shows that the correlation coefficient between two descriptors, $\boldsymbol{x}_{36}$ and $\boldsymbol{x}_{91}$, is 0.9958, which also narrates one ($\boldsymbol{x}_{91}$) of two descriptors to be moved from the model. Thus a new 4–variable model are developed again using MLR as follows.

$$\begin{aligned} RI = {} & (2974.4148 \pm 334.5833) - (72.2978 \pm 15.0317)\cdot\boldsymbol{x}_{15} \\ & + (82.1856 \pm 13.6691)\cdot\boldsymbol{x}_{25} - (202.0907 \pm 42.1476)\cdot\boldsymbol{x}_{26} \end{aligned}$$

$$+ (135.9185 \pm 21.4291)\cdot\chi_{36} \qquad (2)$$

$n = 62$, $m = 4$, r=0.9912, $RMSEE$= 31.40, F= 797.66 (Estimation)

n =62, m = 4, q= 0.9898, $RMSEP$= 33.76 (LOO prediction)

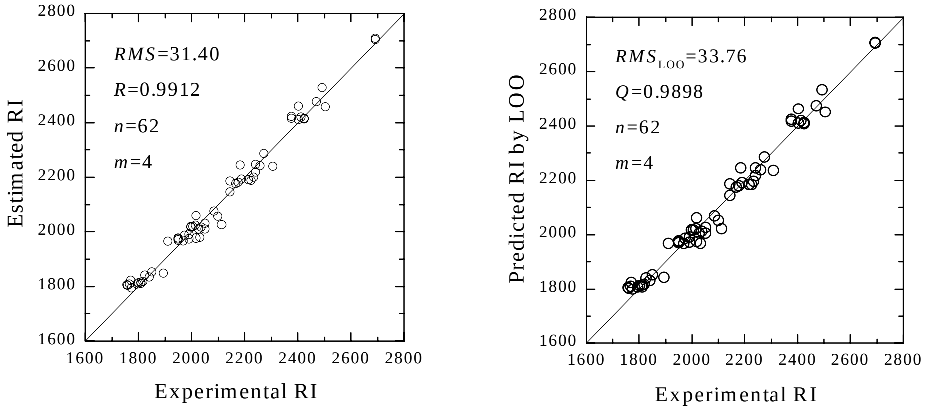
Equation 1 and 2 show that there is not significant difference between the whole qualities expressed by *r*, *RMSEE*, *q*, and *RMSEP* of two QSRR models. However, 4–variable model have higher robusticity than 6–variable equation due to its high *F* value of 797.66 and significant regression coefficients. The 4–variable model also shows that the structures of 62 PCNs under study can be completely described by only 4 MEDV−13 descriptors related to 4 interactions between 4 pairs of atomic types of nos. 2–3, 3–3, 2–13, and 3–13. The experimental *RI*s of 62 PCNs taken from the reference 11, the estimated *RI*s by 4–variable model, and the predicted *RI*s by LOO method are listed in **Table 1.** The values of 4 nonzero MEDV descriptors are seen from the note 12. The relationship and prediction profile are easily seen from **Figure 1a** and **1b** with the estimated and predicted *RI*s versus experimental *RI*s.

Table 1 The experimental, estimated and predicted relative retention indices of 62 PCNs

No	Compound*	RI_{EXP}	RI_{EST}	RI_{LOO}	No	Compound*	RI_{EXP}	RI_{EST}	RI_{LOO}
1	1,3,6-3Cl-NA	1759.0	1803.2	1806.1	32	1,2,5,8-4Cl-NA	2052.0	2028.5	2027.4
2	1,3,5-3Cl-NA	1761.0	1800.7	1802.9	33	1,2,6,8-4Cl-NA	2052.0	2006.8	2005.6
3	1,3,7-3Cl-NA	1769.0	1807.0	1809.7	34	1,4,5,8-4Cl-NA	2086.0	2073.0	2069.4
4	1,4,6-3Cl-NA	1772.0	1820.8	1823.9	35	1,2,3,8-4Cl-NA	2101.0	2054.9	2052.5
5	1,2,4-3Cl-NA	1776.0	1793.0	1796.7	36	1,2,7,8-4Cl-NA	2114.0	2024.9	2022.1
6	1,2,5-3Cl-NA	1796.0	1805.9	1806.9	37	1,2,3,5,7-5Cl-NA	2145.0	2184.8	2186.7
7	1,2,6-3Cl-NA	1802.0	1811.3	1811.8	38	1,2,4,6,7-5Cl-NA	2145.0	2144.7	2144.6
8	1,2,7-3Cl-NA	1812.0	1814.7	1814.8	39	1,2,4,5,7-5Cl-NA	2168.0	2173.9	2174.5
9	1,6,7-3Cl-NA	1812.0	1808.4	1808.2	40	1,2,4,6,8-5Cl-NA	2178.0	2180.1	2180.3
10	2,3,6-3Cl-NA	1819.0	1816.9	1816.7	41	1,2,3,4,6-5Cl-NA	2186.0	2243.3	2245.5
11	1,2,3-3Cl-NA	1827.0	1838.8	1840.8	42	1,2,3,5,6-5Cl-NA	2190.0	2190.4	2190.4
12	1,3,8-3Cl-NA	1842.0	1831.8	1831.2	43	1,2,3,6,7-5Cl-NA	2217.0	2187.7	2184.3
13	1,4,5-3Cl-NA	1852.0	1852.0	1852.0	44	1,2,4,5,6-5Cl-NA	2227.0	2186.8	2183.8
14	1,2,8-3Cl-NA	1896.0	1846.3	1841.8	45	1,2,4,7,8-5Cl-NA	2235.0	2197.4	2195.3
15	1,3,5,7-4Cl-NA	1911.0	1963.1	1967.1	46	1,2,3,5,8-5Cl-NA	2243.0	2245.1	2245.2
16	1,2,4,6-4Cl-NA	1950.0	1974.0	1975.0	47	1,2,3,6,8-5Cl-NA	2243.0	2218.1	2216.6
17	1,2,4,7-4Cl-NA	1950.0	1975.0	1976.0	48	1,2,4,5,8-5Cl-NA	2261.0	2240.1	2237.7
18	1,2,5,7-4Cl-NA	1950.0	1969.7	1970.6	49	1,2,3,4,5-5Cl-NA	2275.0	2285.0	2285.9
19	1,3,6,7-4Cl-NA	1970.0	1967.6	1967.4	50	1,2,3,7,8-5Cl-NA	2309.0	2238.0	2235.3
20	1,4,6,7-4Cl-NA	1974.0	1986.4	1986.9	51	1,2,3,4,6,7-6Cl-NA	2378.0	2420.0	2424.2
21	1,2,5,6-4Cl-NA	1993.0	1974.0	1972.5	52	1,2,3,5,6,7-6Cl-NA	2378.0	2412.5	2416.2
22	1,3,6,8-4Cl-NA	1993.0	1990.3	1990.1	53	1,2,3,4,5,7-6Cl-NA	2405.0	2457.3	2460.7
23	1,2,3,5-4Cl-NA	2000.0	2016.7	2017.3	54	1,2,3,5,6,8-6Cl-NA	2405.0	2409.2	2409.4
24	1,3,5,8-4Cl-NA	2000.0	2015.0	2016.4	55	1,2,3,5,7,8-6Cl-NA	2415.0	2418.7	2418.9
25	1,2,3,6-4Cl-NA	2006.0	2016.9	2017.5	56	1,2,4,5,6,8-6Cl-NA	2425.0	2410.6	2408.6
26	1,2,3,7-4Cl-NA	2017.0	2021.7	2022.0	57	1,2,4,5,7,8-6Cl-NA	2425.0	2413.9	2412.5
27	1,2,3,4-4Cl-NA	2018.0	2056.1	2062.6	58	1,2,3,4,5,6-6Cl-NA	2472.0	2473.7	2473.9
28	1,2,6,7-4Cl-NA	2018.0	1976.9	1974.7	59	1,2,3,4,5,8-6Cl-NA	2493.0	2526.4	2534.3
29	1,2,4,5-4Cl-NA	2029.0	2006.8	2005.5	60	1,2,3,6,7,8-6Cl-NA	2505.0	2455.3	2449.2
30	2,3,6,7-4Cl-NA	2034.0	1977.4	1967.8	61	1,2,3,4,5,6,7-7Cl-NA	2694.0	2702.7	2704.6
31	1,2,4,8-4Cl-NA	2038.0	2013.9	2012.6	62	1,2,3,4,5,6,8-7Cl-NA	2694.0	2706.0	2708.4

*NA refers to naphthalene.

Figure 1 Plot of the estimated (a) or predicted *RI*s (b) *versus* the experimental *RI*s.



Acknowledgment.

We are especially grateful to the China Postdoctoral Science Foundation and the National High Technology Project of China (No. 2001AA640601) for their financial supports.

References and Note

1. A. R. Katritzky, U. Maran, V. S. Lobanov, M. Karelson, *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1.
2. R. D. Cramer, D. E. Patterson, J. D. Bunce, *J. Am. Chem. Soc.* **1988**, *110*, 5959.
3. P. J. Goodford, *J. Med. Chem*. **1985**, *28*, 849.
4. A. N. Jian, K. Koile, D. Chapman, *J. Med. Chem*. **1994**, *37*, 2315.
5. D. D. Robinson, P. J. Winn, P. D. Lyne, W. G. Richards, *J. Med. Chem*. **1999**, *42*, 573.
6. W. Tong, D. R. Lowis, R. Perkins, Y. Chen, W. J. Welsh, D. W. Goddette, T. W. Heritage, D. M. Sheehan, *J. Chem. Inf. Comput. Sci*. **1998**, *38*, 669.
7. L. B. Kier, L. H. Hall, *Pharm. Res*. **1990**, *7*, 801.
8. L. B. Kier, L. H. Hall, *J. Math. Chem*. **1991**, *7*, 229.
9. S. S. Liu, C. S. Yin, Z. L. Li, S. X. Cai, *J. Chem. Inf. Comput. Sci*. **2001**, *41*, 321.
10. S. S. Liu, C. S. Yin, Y. Y. Shi, S. X. Cai, Z. L. Li, *Chin. J. Chem*. **2001**, *19*, 751.
11. J. Olivero, k. Kannan, *J. Chromatogr A.* **1999**, *849*, 621.
12. The values of MEDV−13 of 62 PCNs were deposited to editorial department of CCL.

Received 28 November, 2001