MEDV-13 for QSAR Studies on the COX-2 Inhibition by Indomethacin Amides and Esters

LIU, Shu-Shen*,[a,b,c](刘树深) YIN, Chun-Sheng[d](印春生) SHI, Yun-Yu[b](施蕴渝)
CAI, Shao-Xi[a](蔡绍皙) LI, Zhi-Liang[a](李志良)
[a] *College of Bioengineering, Chongqing University, Chongqing 400044, China*
[b] *Laboratory of Structural Biology, University of Science and Technology of China, Hefei, Anhui 230026, China*
[c] *Department of Applied Chemistry, Guilin Institute of Technology, Guilin, Guangxi 541004, China*
[d] *School of Environment, Nanjing University, Nanjing, Jiangsu 210093, China*

A molecular electronegativity distance vector based on 13 atomic types (MEDV-13), is a descriptor for predicting the biological activities of molecules based on the quantitative structure-activity relationship (QSAR). The MEDV-13 with 91 descriptors is employed to describe the structures of a series of selective cyclooxygenase-2 (COX-2) inhibitors including 16 indomethacin and its amide and ester derivatives (ImAE). A principal component regression (PCR) is used to derive a QSAR model relating the biological activities expressed by pIC_{50} values to the MEDV-13. With the number of principal components of 6, the correlation coefficient (R) and the root mean square error (RMS) are 0.9245 and 0.1682 in modeling stage, and 0.8417 and 0.2389 in leave-one-out prediction step, respectively.

Keywords Molecular electronegativity distance vector (MEDV-13), selective cyclooxygenase-2 (COX-2) inhibitor, indomethacin, quantitative structure-activity relationship (QSAR)

Introduction

The use of nonsteroidal antiinflammatory drugs (NSAIDs) for the treatment of inflammation and pain is often accompanied by gastrointestinal ulcerations and bleeding.[1] It is known that cyclooxygenase (COX) exists as two isoforms (COX-1 and -2). One of the isoforms, COX-2, could be induced by mitogens and inflammatory mediators. It is found that selective inhibition of COX-2 might avoid the side effects of currently available NSAIDs while retaining their therapeutic efficacy.[2-6] In the last several years, extensive libraries of selective COX-2 inhibitors have been developed by different laboratories.[7-12]

Quantitative structure-activity relationship (QSAR) techniques have become indispensable in all aspects of research into the molecular interpretation of biological properties.[13] QSAR studies should be used to identify the molecular parameters necessary for maximizing COX-2 inhibition while simultaneously minimizing the inhibition of constitutively expressed COX-1 and should facilitate the discovery and development of selective COX-2 inhibitors that should lead to safer nonsteroidal antiinflammatory drugs. The ability to produce quantitative correlation between structural information of the molecules and the biological activity of these compounds is of inestimable value in deciding upon the choice of future synthetic chemistry.[14] Six reports about QSAR work of COX-2 inhibitors have appeared in some literatures.[1,15-19] Until now, however, no one has demonstrated a direct and quantitative correlation between inhibiting COX-2 activity and structural characteristics of indomethacin and its amides and esters (ImAE) reported by Kalgutkar.[20]

In our previous paper,[21-22] a molecular electronegativity distance vector (MEDV) was reported and used to study the quantitative structure-property relationship of

* E-mail: ssliuhl@263.net
Received December 26, 2000; revised March 19, 2001; accepted April 27, 2001.
Project supported by the National High Technology Project of China (No. 863-103-13-03-01).

organic compounds such as the boiling points of alcohols. Taking invariability of relative electronegativity of element in the different molecular environment and complexity of calculating relative bond-length into account, the vector-type structural descriptor is modified by introducing 13 atomic types, the modified electrotopological state (E-state) index and 2D topological distance.[23] The modified MEDV is called MEDV-13. In this paper, the MEDV-13 is used for QSAR study of a set of amide and ester derivatives of indomethacin, a series of selective COX-2 inhibitors.

Methodology

Molecular electronegativity distance vector (MEDV-13)

With the relative electronegativity (q) expressed by modified E-state indices and the topological distances (d), the descriptor (h_{kl}) in the novel molecular electronegativity distance vector based on 13 atomic types (MEDV-13) can be calculated from Eq (1).[23]

$$h_v = h_{kl} = \sum_{i\in k, j\in l} \frac{q_i q_j}{d_{ij}^2} \qquad (1)$$

$$(k, l = 1,2,3,\cdots,13; l \geq k; v = 1,2,3,\cdots,91)$$

where k or l is the atomic type of the atom i or j in the molecule, i or j a coding number or series number and d_{ij} the shortest graph distance of various pathways passed from the ith to the jth atom. From the literature,[23] there are in general 13 atomic types for most of organic compounds, then the MEDV-13 has 91 elements. Table 3 in the literature shows the relation of the subscript v in h_v to k and l in h_{kl} for clarify. Here the term "atomic type" of an atom is defined as the number of non-hydrogen atoms binding to that atom plus its identifying number (ID). And the ID of carbon, nitrogen, phosphorus, oxygen, sulfur and halogen atoms is respectively 0, 4, 4, 8, 8, and 12. For example, the atomic types of No. 1—22 and 24 atoms in the skeleton structure (Fig. 1) of 16 ImAE compounds are 1, 10, 3, 2, 2, 3, 7, 3, 3, 2, 2, 3, 13, 2, 2, 9, 3, 1, 3, 3, 2, 2 and 9, respectively.

Principal component regression (PCR) *for* QSAR

Because the MEDV-13 has in general 91 descriptors, a principal component regression (PCR)[24,25] program developed in house is employed to derive a latent QSAR model. The PCR model is a projection method that relates a matrix X (containing the chemical structural descriptors) to a matrix Y (containing the biological activities). The method provides an approximation of an X matrix in terms of the product of two smaller matrices T (called score matrix) and V^T (called load matrix) as follows,

$$X = TV^T + E \qquad (2)$$

where E is a residual matrix. The projections can be calculated for any given number of variables, and there is a proper number of variables, called number of principal components (PC). After the projection, the matrix T is used (instead of the original matrix X) to explain or predict Y.

$$Y = TG + H \qquad (3)$$

where G is a coefficient matrix and H a residual matrix.

A convenient QSAR model can be obtained by following simple matrix transformation procedure.

$$Y = XB = TV^TB = TG \qquad (4)$$

where B is a coefficient matrix for original structural variables in the convenient QSAR model.

Results and discussion

Data set

16 COX-2 inhibitors including indomethacin and 15 amide and ester derivatives (ImAE) previously studied by Kalgutkar *et al*.[20] are selected. Their biological activities are expressed by IC_{50} values in μM representing time-dependant COX-2 inhibition and being average values from duplicate experiments. The structural descriptors of these compounds skeleton parent structure of which is shown in Fig. 1 are those in the MEDV-13 derived from Eq. 1. For most of organic compounds, the MEDV-13 has 91 descriptors in general.

Selection of variables

Because the number of variables having non-zero value in different molecules is different, it is essential to

select the variables having statistically significant in a QSAR model. A three-step procedure is performed to select the variables entering into the PCR analysis in our present paper. Firstly, the variables having zero value for all samples ($n = 16$) are left out from the data set consisting of n samples. Secondly, the variables having only a few such as 1 non-zero values in n samples are left out from the variable set. Finally, the correlation coefficients (R) between various pairs of variables are calculated and then one of two variables having $R > 0.98$ is deleted.

For the ImAE system consisting of 16 samples, only 42 variables, Nos. 1, 2, 3, 5, 6, 7, 9, 10, 13, 14, 15, 17, 18, 19, 21, 22, 25, 26, 28, 29, 30, 32, 33, 36, 49, 51, 52, 55, 56, 57, 59, 60, 63, 66, 67, 70, 77, 78, 81, 82, 85 and 91 descriptor, are not all zero values. Only one sample has non-zero value for Nos. 5, 17, 28, 49, 51, 52 and 55 descriptor. So, the 7 variables should be deleted from the data set. Various correlation coefficients (R) between the remaining 35 variables are less than 0.98 but 3 pairs and then 3 descriptor (Nos. 36, 81 and 91 descriptor) should be deleted. Then, only 32 variables enter into the PCR analysis.

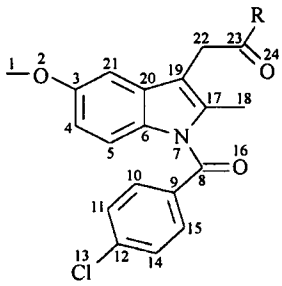

Fig. 1 Skeleton structure of ImAE.

Number of principal components (PC)

It is known that a high quality model should have not only a good ability of estimation for the internal samples but also an excellent ability of prediction for the external samples and the latter is often more important for a QSAR model. The number of principal components (PC) plays an important role in PCR analysis. More or less PC will reduce the quality of the QSAR model. It was found that the correlation coefficient (R) in modeling stage is always increasing with the number of PC while the root mean square error (RMS) reducing with PC, but the R_{LOO} and RMS_{LOO} in prediction stage for leave-one-out (LOO) procedure is irregularly changing and exists as one or more maximum points or minimum points in plot of R_{LOO}, RMS_{LOO} vs. PC. The results shown that RMS_{LOO} or R_{LOO} is a good criterion in determination of the number of principal components. And the lower the RMS_{LOO} or the higher the R_{LOO} is, the better the quality of the model is. It is evident that the number of PC of 6 is suitable for our ImAE system (Fig. 2).

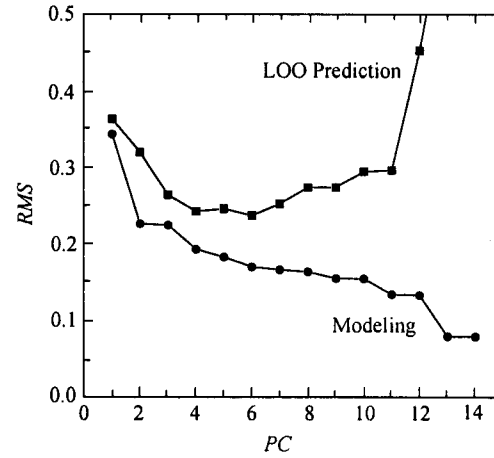

Fig. 2 Effects of PC on RMS for ImAE.

The principal component axe is a linear combination of some original descriptors such as the MEDV-13 variables in this paper. Analysis on the principal components shows that the first PC is mainly affected by the original variables of $v = 14$, 15, and 21; the second one by $v = 14$, 15, and 18; the third one by $v = 14$ and 18; the fourth one by $v = 10$, 15, 22, 25, and 32; the fifth one by $v = 22$, 25, 32 and 78; and last one by $v = 25$, 32, and 77. Then, all principal components are only related to ten original MEDV-13 descriptors ($v = 10$, 14, 15, 18, 21, 22, 25, 32, 77, and 78) whose values are listed in Table 1.

QSAR *studies*

The PCR procedure is performed on the data set containing ten MEDV-13 descriptors together with corresponding biological activities expressed by pIC_{50}, negative common logarithm of IC_{50}, to produce a latent QSAR model. The first QSAR model with $PC = 6$,

called M1, is derived from the data set consisting of all 16 ImAE compounds. The *R* and *RMS* between the estimated activities (pIC_{M1}) by the M1 and the observed activities (pIC_{OBS}) are 0.9245 and 0.1682, respectively. The pIC_{OBS} and pIC_{M1} values of 16 ImAEs are listed in Table 2. Plot of the pIC_{M1} estimated by M1 model versus pIC_{OBS} is shown in Fig. 3.

Table 1 Ten MEDV-13 descriptors entering into PCR analysis

No	$v = 10$	$v = 14$	$v = 15$	$v = 18$	$v = 21$	$v = 22$	$v = 25$	$v = 32$	$v = 77$	$v = 78$
1	4.1166	27.2199	28.7511	0	14.7927	3.0958	3.8357	-5.0467	8.9733	0.7131
2	4.2349	27.6464	30.1823	0.3310	14.0886	3.1548	3.8673	3.4161	1.4718	0.6060
3	7.7042	27.5477	29.7931	0	13.9557	3.7231	3.8602	1.9722	1.4640	3.8789
4	4.1987	28.2578	30.7318	1.4397	17.4820	3.2045	3.8970	4.7482	3.1829	0.6869
5	4.1730	46.5107	35.5921	3.4794	24.1532	3.3313	3.9779	-2.7957	2.7424	0.6932
6	8.5098	46.5234	35.6478	0	19.5192	10.3303	3.9861	1.7831	1.5754	5.0124
7	1.4592	48.4399	41.1427	0	19.8439	5.1380	4.0029	2.3241	1.5788	4.4271
8	1.2830	49.8131	39.8391	0	18.8146	4.7019	3.9940	3.6220	1.5960	4.4525
9	4.4011	44.3555	30.4854	0	19.1186	6.8060	11.7867	1.2890	1.5649	4.5196
10	4.3990	41.5792	31.4565	15.2247	18.3767	6.3584	3.9532	1.3563	1.5555	4.4668
11	1.0814	49.0027	42.9564	2.1390	20.0714	1.2391	4.0115	4.0349	1.5867	0.3792
12	4.1607	44.8690	32.0961	1.8707	19.3426	3.3280	11.8885	2.9856	1.5728	0.6238
13	4.1610	42.0046	32.8827	17.1458	18.5929	3.3042	3.9616	3.0425	1.5634	0.6221
14	4.1332	27.3497	29.2142	0	13.6571	3.1144	3.8451	-0.6524	1.4068	0.5900
15	4.2155	59.1980	38.1085	2.2895	18.6460	3.3753	4.0134	4.6597	1.5753	0.6293
16	4.5139	58.5035	37.3577	0	18.2822	6.6139	4.0030	3.1262	1.5673	5.3877

Table 2 pIC_{50} values observed and calculated using various models

No	R	pIC_{OBS}	pIC_{M1}	pIC_{M2}	pIC_{M3}	pIC_{LOO}
1	OH	0.125	0.125	0.133	0.122	0.290[a]
2	$NHCH_3$	0.155	0.424	0.433	0.417	0.500[a]
3	OCH_3	0.602	0.529	0.540	0.505[a]	0.493[a]
4	$NHCH_2CH_2OH$	0.602	0.457	0.448	0.430	0.408[a]
5	$NHC_6H_4(4\text{-}NHCOCH_3)$	0.921	0.839	0.876	0.866	0.855[a]
6	$OC_6H_4(4\text{-}OCH_3)$	1.398	1.179	1.234	1.205	0.974[a]
7	$OC_6H_4(4\text{-}SCH_3)$	0.523	0.856	0.925	0.930[a]	0.929[a]
8	$OC_6H_4(2\text{-}SCH_3)$	1.222	0.942	1.011	1.027	0.890[a]
9	$OC_6H_4(4\text{-}F)$	1.125	1.237	1.269	1.231	1.286[a]
10	$O(3\text{-}C_5H_4N)$	1.301	1.376	1.360	1.371	1.424[a]
11	$NHC_6H_4(4\text{-}SCH_3)$	0.921	0.743	0.784	0.812[a]	0.639[a]
12	$NHC_6H_4(4\text{-}F)$	1.222	1.136	1.142	1.125	1.031[a]
13	$NH(3\text{-}C_5H_4N)$	1.301	1.288	1.246[a]	1.278	1.268[a]
14	NH_2	0.155	0.349	0.372[a]	0.350	0.384[a]
15	$HCH_2CH_2C_6H_5$	1.222	1.309	1.363[a]	1.411[a]	1.353[a]
16	$OCH_2CH_2C_6H_5$	1.301	1.388	1.470[a]	1.495	1.475[a]

[a] Refer to the values predicted by models such as M2, M3, or LOO.

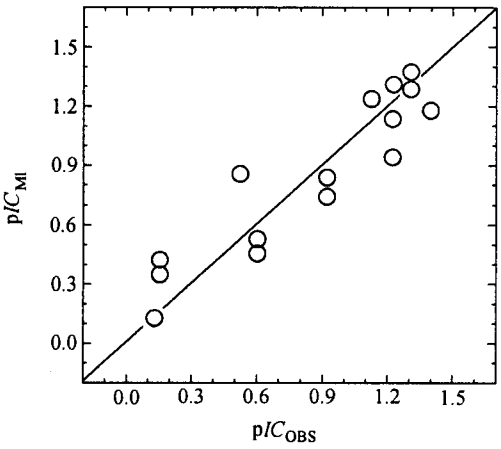

Fig. 3 Plot of pIC_{M1} vs. pIC_{OBS}.

A predictive ability of a QSAR model for the external samples is another important criterion in evaluating the quality of the model. The results estimated by the M1 model only explain the estimated ability for the internal samples, and the predictive ability of the M1 model need to still test. In this paper, a LOO statistical procedure, one of many cross-validation (CV) techniques, is employed to test the predictive ability. In such a cross-validation experiment involving n molecules, a model is built from all but the first molecule, and this model is used to predict the activity of the first molecule. Then all but the second molecule are used to create a model that predicts the second molecule, and so on. In this way, each molecule is predicted, as though the system had never seen it before, on the basis of all the other molecules. The activities predicted by 16 LOO models with $PC = 6$, pIC_{LOO}, are also listed in Table 2 (see pIC_{LOO} column). The R and RMS between the activities predicted by n LOO models and the activities observed experimentally are 0.8417 and 0.2389, respectively.

In order to further validate stability and predictive ability of the model (M1), a new model (M2) with $R = 0.9028$ and $RMS = 0.1801$ has been developed using the first 12 compounds from all 16 compounds as an internal training set and the M2 model is then employed to predict the biological activities of the last 4 ImAEs (constructing an external predictive set). The calculated activities (pIC_{M2}) are also listed in Table 2. And the R and RMS for 4 external predictive compounds are respectively 0.9808 and 0.1570. Furthermore, another new model (M3) with $R = 0.9488$ and $RMS = 0.1523$ has been created in such a way that three compounds are selected as an internal training set every four samples. Then the M3 is used to predict the activities of the remaining 4 samples and the predictive RMS is 0.2359. The calculated activities (pIC_{M3}) are also listed in Table 2.

Customary QSAR *model*

It has been well known that the PCR model is a latent one with regard to principal components and the principal component is a linear combination of some original descriptors. Such model is inconvenient for application. A customary linear model (Eq. (5)) about the original descriptors is derived from the QSAR model (M1) developed by the PCR technique (using Eq. (4)). Then, the MEDV-13 descriptors calculated from Eq. (1) can directly enter into Eq. (5) to estimate and predict the activities of the compounds under study.

$$\begin{aligned} pIC_{CAL} = & -1.7376\times10^{-4} + 0.016842\cdot h_{10} + \\ & 0.031453\cdot h_{14} - 0.024519\cdot h_{15} + \\ & 0.028700\cdot h_{18} + 1.9781\ 10^{-5}\cdot h_{21} + \\ & 0.024843\cdot h_{22} + 0.022432\cdot h_{25} + \\ & 0.019230\cdot h_{32} - 0.019533\cdot h_{77} + \\ & 0.019235\cdot h_{78} \end{aligned} \quad (5)$$

Conclusion

We have described a QSAR model between biological activities expressed by pIC_{50} values and the MEDV-13 of 16 indomethacin amides and esters using PCR technique, discussed the effects of various original MEDV-13 variables on the principal components and given the selection procedure of the variables entering into PCR analysis. It is believed that the MEDV descriptor can be widely applied in the QSAR study on many COX-2 inhibitor system and some researches are in progress.

References

1 Wilkerson, W. W.; Copeland, R. A.; Covington, M. B.; Trzaskos, J. M. *J. Med. Chem.* **1995**, *38*, 3895.
2 Fu, J. Y.; Masferrer, J. L.; Siebert, K.; Raz, A.; Needleman, P. *J. Biol. Chem.* **1990**, *265*, 16737.

3 Xie, W.; Chipman, J. G.; Robertson, D. L.; Erikson, R. L.; Simmons, D. L. *Proc. Natl. Acad. Sci. U. S. A.* **1991**, *88*, 1692.

4 Herchsman, H. R. *Biochim. Biophys. Acta* **1996**, *1299*, 125.

5 O'Banion, M. K.; Winn, V. D.; Young, D. A. *Proc. Natl. Acad. Sci. U.S.A.* **1992**, *89*, 4888.

6 Puig, C.; Crespo, M. I.; Godessart, N.; Feixas, J.; Ibarzo, J.; Jimenez, J.-M.; Soca, L.; Cardelus, I.; Hereedia, A.; Miralpeix, M.; Puig, J.; Beleta, J.; Huerta, J. M.; Lopez, M.; Segarra, V.; Ryder, H.; Palacios, J. M. *J. Med. Chem.* **2000**, *43*, 214.

7 Black, W. C.; Brideau, C.; Chan, C.-C.; Charleson, S.; Chauret, N.; Claveau, D.; Ethier, D.; Gordon, R.; Greig, G.; Guay, J.; Hughes, G.; Jolicoeur, P.; Leblanc, Y.; Nicoll-Griffith, D.; Ouimet, N.; Riendeau, D.; Visco, D.; Wang, Z.; Xu, L.; Prasit, P. *J. Med. Chem.* **1999**, *42*, 1274.

8 Futaki, N.; Takahashi, S.; Yoshikawa, K.; Arai, I.; Higuchi, S.; Otomo, S. *Prostaglandins* **1994**, *47*, 55.

9 Kalgutker, A. S.; Kozak, K. R.; Crews, B. C.; Hochgesang, Jr. G. P.; Marnett, L. J. *J. Med. Chem.* **1998**, *41*, 4800.

10 Li, J. J.; Norton, M. B.; Renhard, E. J.; Anderson, G. D.; Gregory, S. A.; Isakson, P. C.; Koboldt, C. M.; Masferrer, J. L.; Perkins, W. E.; Seibert, K.; Zhang, Y. Y.; Zweifel, B. S.; Reitz, D. B. *J. Med. Chem.* **1996**, *39*, 1846.

11 Marnett, L. J.; Kalgutkar, A. S. *Curr. Opin. Chem. Biol.* **1998**, *2*, 482.

12 Reitz, D. B.; Li, J. J.; Norton, M. B.; Reinhard, E. J.; Collins, J. T.; Anderson, G. D.; Gregory, S. A.; Koboldt, C. M.; Perkins, W. E.; Seibert, K.; Isakson, P. C. *J. Med. Chem.* **1994**, *37*, 3878.

13 Katritzky, A. R.; Maran, U.; Lobanov, V. S.; Karelson, M. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1.

14 Robinson, D. D.; Winn, P. J.; Lyne, P. D.; Riehards, W. G. *J. Med. Chem.* **1999**, *42*, 573.

15 Marot, C.; Chavatte, P.; Lesieur, D. *Quant. Struct. Act. Relat.* **2000**, *19*, 127.

16 Singh, P.; Kumar, R. *J. Enzyme Inhib.* **1999**, *14*, 277.

17 Singh, P.; Kumar, R. *J. Enzyme Inhib.* **1998**, *13*, 409.

18 Kumar, R.; Singh, P. *Indian J. Chem., Sect. B: Org. Chem. Incl. Med. Chem.* **1997**, *36*, 1164.

19 Wilkerson, W. W.; Copeland, R. A.; Covington, M. B.; Grubb, M. F.; Hewes, W. E.; Kerr, J. S.; Trzaskos, J. M. *Med. Chem. Res.* **1995**, *5*, 399.

20 Kalgutker, A. S.; Crews, B. C.; Rowlinson, S. W.; Marnett, A. B.; Kozak,, K. R.; Remmel, R. P. *Proc. Natl. Acad. Sci. U.S.A.* **2000**, *97*, 925.

21 Liu, S. S.; Liu, H. L.; Xia, Z. L.; Cao, C. Z.; Li, Z. L. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 951.

22 Liu, S. S.; Liu, Y.; Li, Z. L.; Cai, S. X. *Acta Chim. Sinica* **2000**, *48*, 1353 (in Chinease).

23 Liu, S. S.; Yin, C. S.; Cai, S. X.; Li, Z. L. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 321.

24 Geladi, P.; Kowalski, B. R. *Anal. Chim. Acta* **1986**, *185*, 1.

25 Liu, S. S.; Yi, Z. S. in *Basic Chemometrics*, Ed.: Li, Y. F., Science Press, Beijing, **1999**, pp. 118—121 (in Chinese).

(E200012290 LI, L.T.; DONG, L.J.)