Inter*net* Electronic Journal of Molecular Design

November 2002, Volume 1, Number 11, Pages 610–619

Editor: Ovidiu Ivanciuc

Special issue dedicated to Professor Haruo Hosoya on the occasion of the 65th birthday
Part 3

Guest Editor: Jun–ichi Aihara

A Novel Variable Selection and Modeling Method based on the Prediction for QSAR of Cyclooxygenase–2 Inhibition by Thiazolone and Oxazolone Series

Shu–Shen Liu,[1,2,3] Shi–Hai Cui,[1] Yun–Yu Shi,[2] and Lian–Sheng Wang[1]

[1] State Key Laboratory of Pollution Control and Resources Reuse, Department of Environmental Science & Engineering, Nanjing University, Nanjing 210093, P. R. China
[2] Laboratory of Structural Biology, University of Science and Technology of China, Hefei 230026, P. R. China
[3] Department of Applied Chemistry, Guilin Institute of Technology, Guilin 541004, P. R. China

Received: July 18, 2002; Revised: October 15, 2002; Accepted: October 23, 2002; Published: November 30, 2002

Citation of the article:
S.–S. Liu, S.–H. Cui, Y.–Y. Shi, and L.–S. Wang, A Novel Variable Selection and Modeling Method based on the Prediction for QSAR of Cyclooxygenase–2 Inhibition by Thiazolone and Oxazolone Series, *Internet Electron. J. Mol. Des.* **2002**, *1*, 610–619, http://www.biochempress.com.

Inter*net* Electronic Journal of Molecular Design

BioChem Press
http://www.biochempress.com

A Novel Variable Selection and Modeling Method based on the Prediction for QSAR of Cyclooxygenase–2 Inhibition by Thiazolone and Oxazolone Series[#]

Shu–Shen Liu,[1,2,3,*] Shi–Hai Cui,[1] Yun–Yu Shi,[2] and Lian–Sheng Wang[1]

[1] State Key Laboratory of Pollution Control and Resources Reuse, Department of Environmental Science & Engineering, Nanjing University, Nanjing 210093, P. R. China
[2] Laboratory of Structural Biology, University of Science and Technology of China, Hefei 230026, P. R. China
[3] Department of Applied Chemistry, Guilin Institute of Technology, Guilin 541004, P. R. China

Received: July 18, 2002; Revised: October 15, 2002; Accepted: October 23, 2002; Published: November 30, 2002

Internet Electron. J. Mol. Des. **2002**, *1* (11), 610–619

Abstract

A novel variable selection and modeling method based on prediction is developed to construct the quantitative structure–activity relationships (QSAR) between the molecular electronegativity distance vector based on 13 atomic types and the biological activities of a set of selective cyclooxygenase–2 inhibitory molecules, thiazolone and oxazolone series (TOS). Two parameters, the interrelation coefficient between the independent variables and the correlation coefficient in the leave–one–out cross–validation (q), are introduced into the classical all–subset regression to improve its searching course. Using multiple linear regression, a 4–variable linear model for the data set of 21 TOSs is developed with the correlation coefficient of 0.9248 and the root mean square error of 0.283 in modeling stage and the coefficient of 0.8849 and the error of 0.351 in the leave–one–out validation step, respectively. To further test the predictive ability of the model, 15 TOS compounds are picked up from the set of 21 TOSs to construct a training set which is used to build a QSAR model and then the model is employed to predict the biological activities of the remaining compounds.

Keywords. Variable selection and modeling method based on prediction; molecular electronegativity distance vector; cyclooxygenase–2 inhibitor; thiazolone; oxazolone; quantitative structure–activity relationships; QSAR.

1 INTRODUCTION

The discovery that there are two cyclooxygenase isoforms, cyclooxygenase–1 (COX–1) and cyclooxygenase–2 (COX–2) that are variably expressed in different tissues raised the possibility that the therapeutic effect of nonsteroidal antiinflammatory drugs (NSAIDs) could be separated from their toxic gastrointestinal effects. COX–1 is expressed constitutively in most tissues throughout the body, including the gastrointestinal mucosa. COX–2 is expressed at low levels in

[#] Dedicated to Professor Haruo Hosoya on the occasion of the 65th birthday.

* Correspondence author; phone: 86–025–3596509; E–mail: ssliuhl@263.net or ssliu@nju.edu.cn.

most cells, including the normal human stomach and intestine [1]. Unlike COX–1, however, the expression of COX–2 can be up–regulated at inflammatory sites by cytokines and bacterial products such as lipopolysaccharide. Thus, an NSAID that inhibits COX–2 while sparing COX–1 has the potential to be anti–inflammatory yet nontoxic to the gastrointestinal tract. Many selective COX–2 inhibitors have been developed by different laboratories. Quantitative structure–activity relationship (QSAR) studies related to the selective COX–2 inhibitors have also recently appeared [2–10]. It is well known that QSAR studies can identify the molecular parameters necessary for maximizing COX–2 inhibition while simultaneously minimizing the inhibition of constitutively expressed COX–1 and facilitate the discovery and development of selective COX–2 inhibitors that should lead to safer nonsteroidal antiinflammatory drugs.

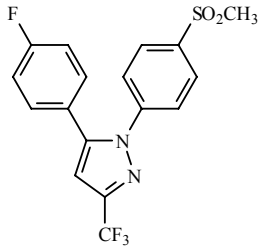
To develop a QSAR model, it is necessary to describe the chemical structure of the examined compound as one or a set of structural descriptors. Many topological indices, including Wiener index, Hosoya index, Randić connectivity indices, Balaban index, were developed and have been applied in QSAR studies widely [9–22]. In our previous paper [23,24], the electrotopological state (E–State) index which was introduced by Kier and Hall and widely used in QSAR models [25–33] was employed to generate a novel molecular electronegativity distance vector based on the 13 atomic types. The vector, called MEDV–13, contains 91 descriptors. The MEDV–13 was employed to derive the QSAR models of two panels of selective COX–2 inhibitors, a series of indomethacin and its amides and esters (ImAE) [34] using principal component regression (PCR) and a set of 2,3–diarylcyclopentenones (DAPs) [35] using genetic algorithm (GA), respectively. It has been shown that the MEDV–13 describes adequately the structures of COX–2 inhibitors of interest. Because it is difficult to determine the physical meaning of the principal components obtained by PCR and because usually GA does not find the best variable subset, a novel variable selection and modeling method based on prediction (VSMP) was recently developed in our laboratory [36]. In this paper, the VSMP technique will be employed to select the best subset from the MEDV–13 for another set of COX–2 inhibitors [37] consisting of 18 thiazolone and oxazolone series (TOS) together with SC–58125, NS–398 and indomethacin.

2 MATERIALS AND METHODS

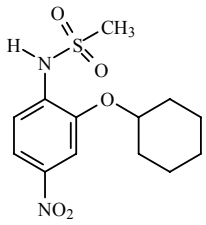
2.1 Data Set and the Molecular Descriptors MEDV–13

The data set used in this paper contains 21 COX–2 inhibitors, 18 TOSs together with SC–58125, NS–398 and indomethacin, whose skeleton structures are shown in Figure 1. All values are expressed in terms of pIC_{50} or $\log 1/IC_{50}$ where IC_{50} represents the drug concentration (μM) that inhibits 50% of activity. All IC_{50} values against purified enzymes in unit of μM were determined using recombinant human cyclooxygenase–2 (rh COX–2, purified from baculovirus–infected SF–9

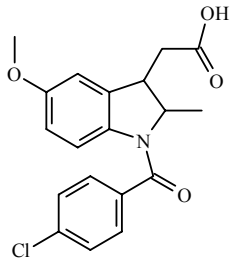
cells) [37]. The pIC_{50} values are widespread and homogeneous (Figure 2). Eight compounds display pIC_{50} values between 4.0 and 5.0 (low activity), 10 display pIC_{50} values between 5.0 and 6.0 (moderate activity), and 3 show pIC_{50} values between 6.0 and 7.0 (good activity). The structural descriptors of these compounds are the MEDV−13 descriptors shown in our previous paper [24].



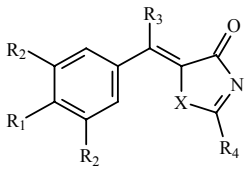
SC–58125 (No. **20**)



NS–398 (No. **21**)



indomethacin (No. **9**)



The parent structures of other 18 inhibitors

Figure 1. The structures of the 21 COX–2 inhibitors.

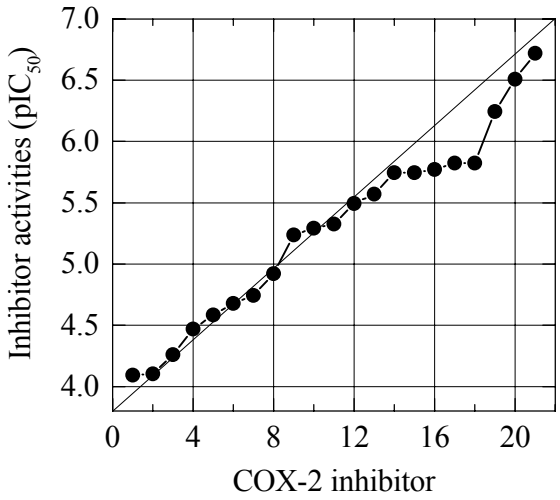


Figure 2. Distribution of pIC_{50} for the 21 COX–2 inhibitors.

From the literature [24], the original MEDV−13 descriptor, $\boldsymbol{x}_v$ (v = 1, 2, 3, …, 91), can be calculated. First, the relative electronegativity (e) of a non–hydrogen atom is calculated using the atomic type, atomic attributes, and intrinsic state (I) of the atom defined in Table 1:

$$e_i = I_i + \sum_{j \neq i}^{all\ j} (I_i - I_j) / d_{ij}^2 \tag{1}$$

where d_{ij} is the shortest graph distance between two atoms, atom i and j. Then, the MEDV−13 descriptor, $\boldsymbol{x}_v$, is calculated from the following formula:

$$x_v = x_{kl} = \sum_{i \in k, j \in l} \frac{e_i e_j}{d_{ij}^2} \qquad (k, l = 1,2,3,\cdots,13; l \geq k; v = 1,2,3,\cdots,91) \tag{2}$$

where k or l is the atomic type of the atom i or j in the molecule.

Table 1. The atomic types, atomic attributes and intrinsic state (*I*) for various non–hydrogen atoms.

atom	type	attribute	*I*	atom	type	attribute	*I*	atom	type	attribute	*I*
$-CH_3$	1	1	2.0000	~C≈	3	16	1.8333	≥N=	7	30	2.2361
$-CH_2-$	2	2	1.5000	–OH	9	17	2.4495	–SH	9	31	1.7691
–CH<	3	3	1.3333	–O–	10	18	1.8371	–S–	10	32	1.1567
>C<	4	4	1.2500	=O	9	19	3.6742	=S	9	33	2.3134
$=CH_2$	1	5	3.0000	~O	9	20	3.0619	>S=	11	34	1.1340
=CH–	2	6	2.0000	$-NH_2$	5	21	2.2361	≥S≤	12	35	1.1227
=C<	3	7	1.6667	–NH–	6	22	1.6771	–F	13	36	2.6458
=C=	2	8	2.5000	>N–	7	23	1.0882	–Cl	13	37	1.9108
≡CH	1	9	4.0000	=NH	5	24	3.3541	–Br	13	38	1.6536
≡C–	2	10	2.5000	=N–	6	25	2.2361	–I	13	39	1.5345
$\sim CH_2$	1	11	2.5000	≡N	5	26	4.4721	$-PH_2$	5	40	1.6149
~CH–	2	12	1.7500	~NH	5	27	2.7951	–PH–	6	41	1.0559
~C<	3	13	1.5000	~N–	6	28	1.9566	>P–	7	42	0.8696
~CH~	2	14	2.0000	~N~	6	29	2.2361	≥P<	8	43	0.9006
–C≈	3	15	1.6667								

[a] The symbols "~" and " ≈" represent one and two conjugated double bonds

Analyzing the MEDV–13 descriptors where each has 21 elements (samples), only 61 MEDV descriptors have one or more nonzero elements, while 7 descriptors ($\boldsymbol{x}_{45}$, $\boldsymbol{x}_{46}$, $\boldsymbol{x}_{49}$, $\boldsymbol{x}_{63}$, $\boldsymbol{x}_{85}$, $\boldsymbol{x}_{90}$, and $\boldsymbol{x}_{91}$) contain 1 nonzero element, 12 descriptors ($\boldsymbol{x}_{12}$, $\boldsymbol{x}_{13}$, $\boldsymbol{x}_{24}$, $\boldsymbol{x}_{25}$, $\boldsymbol{x}_{35}$, $\boldsymbol{x}_{36}$, $\boldsymbol{x}_{62}$, $\boldsymbol{x}_{67}$, $\boldsymbol{x}_{69}$, $\boldsymbol{x}_{70}$, $\boldsymbol{x}_{80}$, and $\boldsymbol{x}_{81}$) contain 2 nonzero elements, and the descriptor $\boldsymbol{x}_{40}$ has 3 nonzero elements. The 20 descriptors with too few nonzero elements should be first eliminated from the 61 descriptors with nonzero elements. So, there are in fact 41 nonzero MEDV descriptors to enter into successive VSMP analysis.

2.2 Variable Selection and Modeling based on the Prediction

To accelerate the speed of classical all–subsets regression (ASR) and to obtain the best variable subset based on the predictive quality, two statistic parameters, the interrelated coefficient (r_{int}) between the variables and the correlation coefficient in the leave–one–out (LOO) cross validation (q), are introduced into the ASR procedure to construct a novel computer program for the variable selection and modeling based on the prediction (VSMP). How to select and analyze the best subset from among a large independent variable matrix including n compounds which each has m descriptors, $\boldsymbol{x}(n,m)$? The optimal selection task is finished in two main phases in the VSMP program. In the first phase, an optimal subset is selected for a given number of variables (vn). This optimal subset is the best for a given vn but not always the best for the whole subset space including all subsets of different vn. The main steps are:

(1) Specify the values of several statistic parameters such as the number of independent variables (vn) in an optimal subset and the interrelated coefficient (r_{int}) between the independent variables. Then specify the initial values of two important iterative statistics such as r_{cri} and f_{max}. The former, r_{cri}, is a control parameter to decide whether the sequential LOO cross–validation step is run or not. The later, f_{max}, is defined as the maximum correlation coefficient obtained in the LOO cross–

validation. The selective rule of the initial values of the r_{cri} and f_{max} is not larger than the final optimal value of q^2. For example, if the optimal value of q^2 is 0.70 in the previous loop, the r_{cri} and f_{max} values of <0.70 are appropriate.

(2) Select systematically a subset, $\boldsymbol{x}(n,vn)$, from the whole independent variable set, $\boldsymbol{x}(n,m)$, and calculate various correlation coefficients (r_a) between all pair of variables.

(3) Compare the r_as with the r_{int} specified in step (1). If there is/are one/more r_as being larger than r_{int}, then return to the step (2) to continue selecting a subset.

(4) If all r_as are not larger than the r_{int}, then use multiple linear regression (MLR) to build a relationship model between the independent variable subset, $\boldsymbol{x}(n,vn)$, and the whole dependent variable set, $\boldsymbol{y}(n)$, and calculate the relevant statistics such as the correlation coefficient (r_m) in building model. If the r_m is little than r_{cri}, then return to the step (2) to select a subset again.

(5) If the r_m is lagrer than the r_{cri}, then call the LOO cross–validation algorithm to calculate the predictive correlation coefficient (q) and compare with the f_{max} determined in the former loop. If $q^2 \leq f_{max}$, then return to the step (2) to select a new subset again.

(6) If $q^2 > f_{max}$, then let both f_{max} and r_{cri} equal q^2. If there is still any other subset to be selected, then return to the step (2) to continue the selection of a new subset. Or, enter the second main phase of VSMP procedure.

In the second main phase, the best subset from among various optimal subsets of different *vn* (2, 3, 4, …) is decided. It has been known that a good QSAR model should possess not only high calibration statistics for the internal molecules but also a high predictive ability for the external molecules. It is found that the correlation coefficient in calibration step (r) monotonically increases for increasing *vn* and the LOO cross–validation correlation coefficient (q) gradually increases until a limited value and then decreases for increasing *vn*. For the root mean square errors (*RMS*), the similar results have been acquired. With the increase of *vn*, calibrated *RMS* (*RMSEE*) is monotonically decreasing and validated *RMS* (*RMSEP*) gradually decreases until a limit value and then increases. So, the determination of the best subset is mainly depended on the q or *RMSEP* in the LOO cross–validation procedure. The plot of *RMSEP* versus *vn* will be employed together with some statistic analysis to determine the best subset entering into the final QSAR model.

3 RESULTS AND DISCUSSION

3.1 Best Subset

With the initial values of r_{int} = 0.70, r_{cri} = 0.10, and f_{max} = 0, the VSMP program is run with the values of *vn* =2, 3, and 4, respectively. When *vn* equals 2, 3, and 4, respectively, the optimal variable subset together with several important statistic parameters such as the *RMSEE* and *RMSEP*

are listed in Table 2. Obviously, the best subset is the combination of 4 MEDV descriptors of nos. $\boldsymbol{x}_1$, $\boldsymbol{x}_7$, $\boldsymbol{x}_{29}$, and $\boldsymbol{x}_{52}$ due to the highest q^2 (0.7831) and lowest *RMSEP* (0.351).

Table 2. Some statistical indices for the optimal subsets of *vn* = 2, 3, and 4.

vn	q^2	*q*	*RMSEP*	r^2	*r*	*RMSEE*	descriptor			
2	0.5241	0.7239	0.517	0.6027	0.7763	0.469	1	32		
3	0.7243	0.8511	0.398	0.8210	0.9061	0.315	1	33	52	
4	0.7831	0.8849	0.351	0.8552	0.9248	0.283	1	7	29	52

The correlation between a MEDV descriptor and the atomic type *k* and *l* can be deduced from Eq. (2) (see Table 3). From Table 3, the 4 descriptors in the best subset are closely correlated with six atomic types (*k* or *l* =1, 3, 5, 6, 7, and 10) of various non–hydrogen atoms. Analyzing Table 1 and the molecular structures of compounds of interest, the six atomic types reflect the importance of some substructures, $-CH_3$ (no. 1), >C= (no. 3), $-NH_2$ or ≡N (no. 5), =N– (no. 6), >N– (no. 7), and –S– (no. 10) in the molecules, which are considered to be the main factors affecting the biological activity in this set of molecules.

Table 3. The correlation between $\boldsymbol{x}_v$ and atomic types *k* and *l*.

x_v	k=1	k=2	k=3	k=4	k=5	k=6	k=7	k=8	k=9	k=10	k=11	k=12	k=13
l=1	1	2	3	4	5	6	7	8	9	10	11	12	13
l=2		14	15	16	17	18	19	20	21	22	23	24	25
l=3			26	27	28	29	30	31	32	33	34	35	36
l=4				37	38	39	40	41	42	43	44	45	46
l=5					47	48	49	50	51	52	53	54	55
l=6						56	57	58	59	60	61	62	63
l=7							64	65	66	67	68	69	70
l=8								71	72	73	74	75	76
l=9									77	78	79	80	81
l=10										82	83	84	85
l=11											86	87	88
l=12												89	90
l=13													91

3.2 Best QSAR Model

Using the multiple linear regression (MLR) technique, the best QSAR model between the pIC_{50} values of all 21 COX–2 inhibitors and 4 optimal MEDV descriptors in the best subset is developed. The model (M_{21}) with $r = 0.9248$ and $RMSEE = 0.283$ is:

$$\begin{aligned} pIC_{50} = (5.9441 \pm 0.1753) - (0.08943 \pm 0.01263)\cdot\boldsymbol{x}_1 + (0.1261 \pm 0.0289)\cdot\boldsymbol{x}_7 + \\ + (0.1746 \pm 0.0281)\cdot\boldsymbol{x}_{29} + (0.7057 \pm 0.1781)\cdot\boldsymbol{x}_{52} \\ n = 21,\ m = 4,\ r^2 = 0.8552,\ r = 0.9248,\ RMSEE = 0.283,\ F = 23.620 \text{ (Calibration)} \end{aligned} \quad (3)$$

The values of the pIC_{50} calibrated (M_{21} column) by Eq. (3) are listed in Table 4 together with the pIC_{50} observed (pIC_{50} column) and the optimal descriptors ($\boldsymbol{x}_1$, $\boldsymbol{x}_7$, $\boldsymbol{x}_{29}$, $\boldsymbol{x}_{52}$) for the 21 inhibitors.

Table 4. Molecular descriptors and pIC_{50} (the unit of IC_{50}: μM) values observed and calculated for 21 inhibitors.

No	X	R_1	R_2	R_3	R_4	x_1	x_7	x_{29}	x_{52}	pIC_{50}	M_{21}	M_{15}
1	NMe	OH	t–Bu	H	NHC(=NH)NH2	15.7305	–8.8695	3.2238	0	4.09	3.98	4.10
2	O	OH	t–Bu	H	NHOEt	15.2047	0	1.1200	0	4.10	4.78	4.87*
3	S	OH	t–Bu	H	NHCN	14.7114	0	2.6450	–0.7586	4.26	4.55	4.61
4	O	OH	t–Bu	H	OH	14.7317	0	–1.5131	0	4.47	4.36	4.53
5	O	OH	t–Bu	H	NHO–allyl	15.3321	0	0.6886	0	4.59	4.69	4.79
6	S	OH	t–Bu	H	NHC(=NH)NH2	14.8691	0	3.5059	–0.7820	4.68	4.67	4.71*
7	S	OH	t–Bu	H	NMeOMe	16.8629	–8.2050	7.6716	0	4.74	4.74	4.73
8	S	OH	t–Bu	Me	NHC(=NH)NH2	16.2803	0	4.6106	–0.7924	4.92	4.73	4.72
9	Indomethacin					0.1325	–2.3090	0	0	5.24	5.64	6.01*
10	O	OH	t–Bu	H	NHC(=NH)NH2	14.7277	0	–1.2826	1.1783	5.29	5.23	5.39
11	S	OH	t–Bu	H	OH	14.8732	0	2.2143	0	5.33	5.00	5.07
12	S	OH	t–Bu	H	NHOEt	15.3518	0	6.4448	0	5.49	5.70	5.65
13	S	OH	t–Bu	H	NHO–allyl	15.4784	0	5.9000	0	5.57	5.59	5.55*
14	S	OH	i–Pr	H	NHC(=NH)NH2	4.8884	0	3.3863	–0.7620	5.74	5.56	5.75
15	S	OH	t–Bu	H	SMe	15.5736	0	7.8345	0	5.74	5.92	5.83
16	S	OH	t–Bu	H	NHOMe	15.3282	0	6.1220	0	5.77	5.64	5.60*
17	S	OH	t–Bu	H	SH	14.9590	0	5.3254	0	5.82	5.54	5.52
18	S	OH	t–Bu	H	NHOH	14.9243	0	5.0653	0	5.82	5.49	5.49
19	S	OH	i–Pr	H	NHOMe	5.1491	0	5.9931	0	6.24	6.53	6.65
20	SC–58125					0	–0.1808	4.2382	0	6.51	6.66	6.91*
21	NS–398					0	0.0076	0.9137	0	6.72	6.10	6.43

It has been known that a good QSAR model should possess not only a good calibrated statistics for the internal molecules but also a high predictive ability for the external molecules. A LOO cross–validation procedure is used to test the predictive ability of the model (Eq. 3). The results show a good predictive ability of the model with $q^2 = 0.7831$, $q = 0.8849$, and $RMSEP = 0.351$ between the pIC_{50} predicted by LOO procedure and the pIC_{50} observed.

In order to further validate the stability and predictive ability of the model, 15 compounds are picked out from the data set of 21 inhibitors to construct a training set and the remaining compounds form the testing set. The training set is used to build a QSAR model (M_{15}) that will be employed to predict the values of the pIC_{50} of 6 compounds in the testing set. The pIC_{50} values calibrated and predicted by the model M_{15} are also listed in Table 4. Comparing the values of the pIC_{50} predicted with the pIC_{50} observed, the correlation coefficient (R_P) and the root mean square errors (RMS_P) between the values of the pIC_{50} observed and the pIC_{50} predicted for the testing set are $R_P = 0.8770$ and $RMS_P = 0.479$, respectively. The model M_{15} is:

$$\begin{gathered} pIC_{50} = (6.2944 \pm 0.2139) - (0.1046 \pm 0.0153)\cdot\boldsymbol{x}_1 + (0.1155 \pm 0.0264)\cdot\boldsymbol{x}_7 + \\ + (0.1493 \pm 0.0275)\cdot\boldsymbol{x}_{29} + (0.7044 \pm 0.1693)\cdot\boldsymbol{x}_{52} \\ n = 15,\ m = 4,\ r^2 = 0.9041,\ r = 0.9508,\ RMSEE = 0.227,\ F = 23.569 \text{ (Estimation)} \\ n = 15,\ m = 4,\ q^2 = 0.8336,\ q = 0.9130,\ RMSEE = 0.305 \text{ (LOO cross–validation)} \end{gathered} \quad (4)$$

The above results show that the model developed in this paper has good calibrated statistics and high predictive ability. From Figures 3 and 4 obtained by plotting the values of the pIC_{50} calibrated or predicted by the model (M_{21} and M_{15}) versus the pIC_{50} observed experimentally, we obtain the same conclusion.

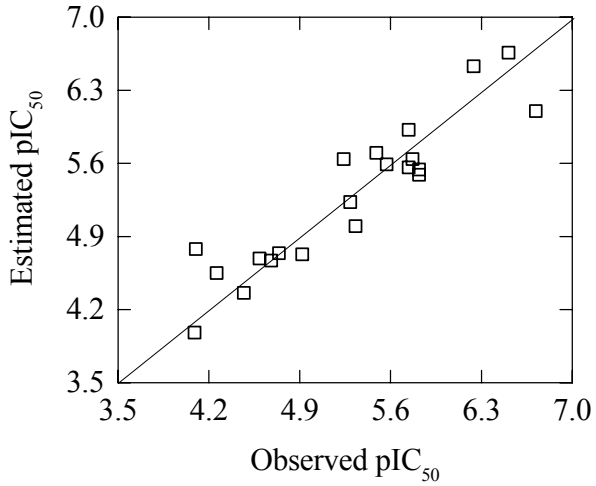


Figure 3. Plot of the pIC_{50} observed versus estimated by the model M_{21}.

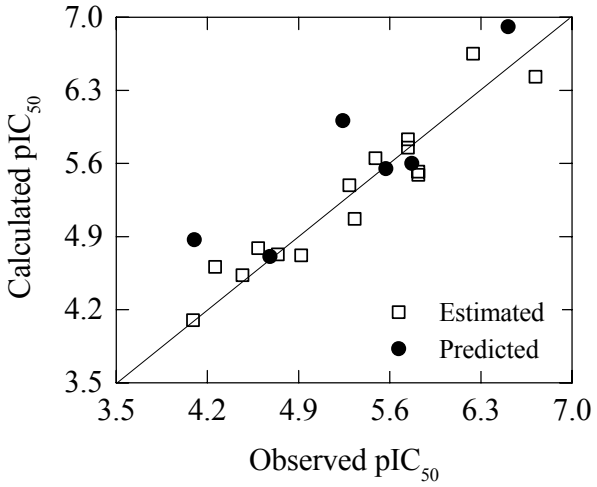


Figure 4. Plot of the pIC_{50} observed versus calibrated or predicted by the model M_{15}.

3.3 Variable Correlation

The correlation between the independent variables entering into the final QSAR model is an important nature of the model and must be validated. The absolute values of all *r* between various pairs of variables in the best set are less than $r = 0.25$. These inter–correlation coefficients are $r(\boldsymbol{x}_1,\boldsymbol{x}_7) = -0.1324$, $r(\boldsymbol{x}_1,\boldsymbol{x}_{29}) = 0.2243$, $r(\boldsymbol{x}_1,\boldsymbol{x}_{52}) = 0.0243$, $r(\boldsymbol{x}_7,\boldsymbol{x}_{29}) = -0.1625$, $r(\boldsymbol{x}_7,\boldsymbol{x}_{52}) = -0.0814$, and $r(\boldsymbol{x}_{29},\boldsymbol{x}_{52}) = -0.2428$, which shows that there are no significant correlations between the structural descriptors in the best subset.

4 CONCLUSIONS

We have described a novel four–variable QSAR model between biological activities expressed by pIC_{50} values and the MEDV–13 of 21 COX–2 inhibitors using a novel variable selection and modeling based on the predictions (VSMP). The results show that the model has not only high

calibrated quality with r = 0.9248 and *RMSEE* = 0.283 but also a good prediction ability with q = 0.8849 and *RMSEP* = 0.351 in LOO procedure.

Acknowledgment

We are especially grateful to the Guangxi Natural Science Fund (No. 0236063) and the National High Technology Research and Development Program of China (No. 863–103–13–03–01) for their financial supports.

5 REFERENCES

[1] B. Cryer and M. Feldman, Cyclooxygenase–1 and Cyclooxygenase–2 Selectivity of Widely Used Nonsteroidal Anti–Inflammatory Drugs, *Amer. J. Med.* **1998**, *104*, 416–421.
[2] W. W. Wilkerson, R. A. Copeland, M. B. Covington, and J. M. Trzaskos, Antiinflammatory 4,5–Diaryl– Pyrroles. 2. Activity as a Function of Cyclooxygenase–2 Inhibition, *J. Med. Chem.* **1995**, *38*, 3895–3901.
[3] R. Kumar, and P. Singh, Diarylspiro[2.4]Heptenes as Selective Cyclooxygenase–2 Inhibitors: A Quantitative Structure–Activity Relationship Analysis, *Indian J. Chem. B* **1997**, *36*, 1164–1168.
[4] P. Singh and R. Kumar, Novel Inhibitors of Cyclooxygenase–2: The Sulfones and Sulfonamides of 1,2–Diaryl–4,5–Difluorobenzene. Analysis of Quantitative Structure–Activity Relationship, *J. Enzyme Inhib.* **1998**, *13*, 409–417.
[5] P. Singh and R. Kumar, 1,2–Diarylimidazoles as Inhibitors of Cyclooxygenase–2: A Quantitative Structure–Activity Relationship Study, *J. Enzyme Inhib.* **1999**, *14*, 277–288.
[6] C. Marot, P. Chavatte, D. Lesieur, Comparative Molecular Field Analysis Of Selective Cyclooxygenase–2 (COX–2) Inhibitors, *Quant. Struct.–Act. Relat.* **2000**, *19*, 127–134.
[7] G. R. Desiraju, B. Gopalakrishnan, R. K. R. Jetti, D. Raveendra, J. A. R. P. Sarma, and H. S. Subramanya, Three–Dimensional Quantitative Structural Activity Relationship (3D–QSAR) Studies of Some 1,5–Diaryl–pyrazoles: Analogue Based Design of Selective Cyclooxygenase–2 Inhibitors, *Molecules* **2000**, *5*, 945–955.
[8] P. Chavatte, S. Yous, C. Marot, N. Baurin, and D. Lesieur, Three–Dimensional Quantitative Structure– Activity Relationships of Cyclo–Oxygenase–2 (COX–2) Inhibitors: A Comparative Molecular Field Analysis, *J. Med. Chem.* **2001**, *44*, 3223–3230.
[9] H. Wiener, Structure Determination of Paraffin Boiling Points, *J. Am. Chem. Soc.* **1947**, *69*, 17–20.
[10] H. Wiener, Correlation of heats of isomerization and differences in heats of vaporisation of isomers, among the paraffin hydrocarbons, *J. Phys. Chem.* **1948**, *52*, 2636–2638.
[11] H. Hosoya, Topological Index: A Newly Proposed Quantity Characterizing the Topological Nature of Structural Isomers of Saturated Hydrocarbons, *Bull. Chem. Soc. Japan* **1971**, *44*, 2332–2339.
[12] M. Randić, On Characterization of Molecular Branching, *J. Am. Chem. Soc.* **1975**, *97*, 6609–6615.
[13] M. Randić, Generalized Molecular Descriptors, *J. Math. Chem.* **1991**, *7*, 155–168.
[14] M. Randić, Novel molecular description for structure–property studies, *Chem. Phys. Lett.* **1993**, *211*, 478–483.
[15] A. T. Balaban, Application of graph theory in chemistry, *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 334–343.
[16] A. T. Balaban, High discrimination distance–based topological index, *Chem. Phys. Lett.* **1982**, *89*, 399–404.
[17] O. Ivanciuc, Building–Block Computation of the Ivanciuc–Balaban Indices for the Virtual Screening of Combinatorial Libraries, *Internet Electron. J. Mol. Des.* **2002**, *1*, 1–9, http://www.biochempress.com.
[18] D. J. G. Marino, P. J. Peruzzo, E. A. Castro, and A. A. Toropov, QSAR Carcinogenic Study of Methylated Polycyclic Aromatic Hydrocarbons Based on Topological Descriptors Derived from Distance Matrices and Correlation Weights of Local Graph Invariants, *Internet Electron. J. Mol. Des.* **2002**, *1*, 115–133, http://www.biochempress.com.
[19] O. Ivanciuc, T. Ivanciuc, D. Cabrol–Bass, and A. T. Balaban, Optimum Structural Descriptors Derived from the Ivanciuc–Balaban Operator, *Internet Electron. J. Mol. Des.* **2002**, *1*, 319–331, http://www.biochempress.com.
[20] R. García–Domenech, A. Catalá–Gregori, C. Calabuig, G. Antón–Fos, L. del Castillo, and J. Gálvez, Predicting Antifungal Activity: A Computational Screening Using Topological Descriptors, *Internet Electron. J. Mol. Des.* **2002**, *1*, 339–350, http://www.biochempress.com.
[21] E. Estrada, The Structural Interpretation of the Randić Index, *Internet Electron. J. Mol. Des.* **2002**, *1*, 360–366, http://www.biochempress.com.
[22] O. Ivanciuc, T. Ivanciuc, and A. T. Balaban, Quantitative Structure–Property Relationship Evaluation of Structural Descriptors Derived from the Distance and Reverse Wiener Matrices, *Internet Electron. J. Mol. Des.* **2002**, *1*, 467–487, http://www.biochempress.com.

[23] S. S. Liu, C. S. Yin, Y. Y. Shi, S. X. Cai, Z. L. Li, MEDV–13 For QSAR Studies on the COX–2 Inhibition By Indomethacin Amides and Esters, *Chin. J. Chem.* **2001**, *19*, 751−756.
[24] S.–S. Liu, H.–L. Liu, Y.–Y. Shi, and L.–S. Wang, QSAR of Cyclooxygenase–2 (COX–2) Inhibition by 2,3–Diarylcyclopentenones Based on MEDV–13, *Internet Electron. J. Mol. Des.* **2002**, *1*, 310–318, http://www.biochempress.com.
[25] L. B. Kier and L. H. Hall, An electrotopological state index for atoms in molecules, *Pharm. Res.* **1990**, *7*, 801−807.
[26] L. B. Kier and L. H. Hall, An index of electrotopological state for atoms in molecules, *J. Math. Chem.* **1991**, *7*, 229–241.
[27] C. de Gregorio, L. B. Kier, and L. H. Hall, QSAR modeling with the electrotopological state indices: Corticosteroids, *J. Comput.–Aid. Mol. Des.* **1998**, *12*, 557−561.
[28] L. H. Hall and T. A. Vaughn, QSAR of phenol toxicity using electrotopological state and kappa shape indices, *Med. Chem. Res.* **1997**, *7*, 407–416.
[29] J. K. Buolamwini, K. Raghavan, M. R. Fesen, et al. Application of the electrotopological state index to QSAR analysis of flavone derivatives as HIV–1 integrase inhibitors, *Pharmaceut. Res.* **1996**, *13*, 1892−1895.
[30] R. R. A. AbouShaaban, H. A. AlKhamees, H. S. AbouAuda, et al. Atom level electrotopological state indexes in QSAR: Designing and testing antithyroid agents, *Pharmaceut. Res.* **1996**, *13*, 129−136.
[31] L. H. Hall, B. Mohney, and L. B. Kier, The electrotopological state: An atom index for QSAR, *Quant. Struct.–Act. Relat.* **1991**, *10*, 43−51.
[32] L. H. Hall, B. Mohney, and L. B. Kier, Comparison of electrotopological state indexes with molecular orbital parameters: inhibition of MAO by hydrazides, *Quant. Struct.–Act. Relat.* **1993**, *12*, 44–48.
[33] L. H. Hall and L. B. Kier, Electrotopological state indexes for atom types: a novel combination of electronic, topological, and valence state information, *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 1039−1045.
[34] S. S. Liu, C. S. Yin, S. X. Cai, and Z. L. Li, QSAR Study of Steroid Benchmark and Dipeptides Based on MEDV−13, *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 321−329.
[35] S. S. Liu, C. S. Yin, and L. S. Wang, Combined MEDV–GA–MLR Method for QSAR of Three Panels of Steroids, Dipeptides, and COX−2 Inhibitors, *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 749−756.
[36] S. S. Liu, C. S. Yin, and L. S. Wang, VSMP: A Novel Variable Selection and Modeling Method based on the Prediction, *J. Chem. Inf. Comput. Sci.* (revised).
[37] Y. Song, D. T. Connor, R. Doubleday, R. J. Sorenson, A. D. Sercel, P. C. Unangst, B. D. Roth, R. B. Gilbertsen, K. Chan, D. J. Schrier, A. Guglietta, D. A. Bornemeier, and R. D. Dyer, Synthesis, Structure–Activity Relationships, and in Vivo Evaluations of Substituted Di–tert–butylphenols as a Novel Class of Potent, Selective, and Orally Active Cyclooxygenase–2 Inhibitors. 1. Thiazolone and Oxazolone Series, *J. Med. Chem.* **1999**, *42*, 1151−1160.