QSAR Studies on the COX-2 Inhibition by 3,4-Diarylcycloxazolones Based on MEDV Descriptor

LIU, Shu-Shen[*,a,b] **(刘树深)** **CUI, Shi-Hai**[a] **(崔世海)** **YIN, Da-Qiang**[a] **(尹大强)**
SHI, Yun-Yu[b] **(施蕴渝)** **WANG, Lian-Sheng**[a] **(王连生)**

[a] *State Key Laboratory of Pollution Control and Resource Reuse, School of the Environment, Nanjing University, Nanjing, Jiangsu 210093, China*

[b] *Laboratory of Structural Biology, University of Science and Technology of China, Hefei, Anhui 230026, China*

Selective inhibition of cyclooxygenase-2 (COX-2) might avoid the side effects of current available nonsteroidal antiinflammatory drugs while retaining their therapeutic efficacy. 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 (MEDV) based on 13 atomic types and the biological activities of a set of selective cyclooxygenase-2 inhibitory molecules, 3,4-diarylcycloxazolones (DAA) plus indomethacin, naproxen, and celecoxib. Using multiple linear regression, a 5-variable linear model is developed with the calibrated correlation coefficient of 0.9271 and root mean square error of 0.17 in modeling stage and the validated correlation coefficient of 0.9030 and root mean square error of 0.20 in leave-one-out validation step, respectively. To further test the predictive ability of the model, 20 DAA compounds are picked up 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 balance compounds. The predicted correlation coefficient and root mean square error are 0.9332 and 0.19, respectively.

Keywords variable selection, molecular electronegativity distance vector, 3,4-diarylcycloxazolones, COX-2 inhibitor, QSAR

Introduction

It is shown that there are two cyclooxygenase isoforms, cyclooxygenase-1 (COX-1) and cyclooxygenase-2 (COX-2). They 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 most cells, including the normal human stomach and intestine.[1] COX-2 is an immediate-early gene induced by mitogenic and proinflammatory stimuli,[2] and it is has been postulated to be the isoform involved in inflammatory processes.[3] Currently available NSAIDs inhibit both COX-1 and COX-2, most of them exhibiting a selectivity for COX-1.[4] The discovery and characterization of COX-2 suggested that selective inhibition of this enzyme might avoid the side effects of current available NSAIDs while retaining their therapeutic efficacy. So, extensive libraries of selective COX-2 inhibitors have been developed by different laboratories.

However, the quantitative structure-activity relationship (QSAR) study related to the selective COX-2 inhibitors only appeared in a few literatures.[5-13] Our group also built the QSAR models for several types of COX-2 inhibitors such as indomethacin and its amides and esters (ImAE) using principal component regression (PCR),[14] 2,3-diarylcyclopentenones (**DAP**) with a combinatorial method of genetic algorithm (GA) and multiple linear regression (MLR),[15] and thiazolone and oxazolone series.[16]

In this paper, a novel variable selection and modeling technique based on prediction (VSMP) developed in our laboratory[17] will be used to perform QSAR study of a series of selective COX-2 inhibitors including thirty-one 3,4-diarylcycloxazolone (**DAAs**) derivatives and other three compounds, indomethacin, naproxen and celecoxib. For convenience, in the present paper, all 34 COX-2 inhibitors (named **DAA**) have been adopted. At first, the molecular electronegativity distance vector (MEDV)[18,19] descriptors of **DAA** inhibitors were calculated. Then the VSMP was used to search the best subset of descriptors from 91 original MEDV ones. Finally, MLR was employed to build QSAR model of the inhibitors under study.

Methodology

MEDV *descriptor*

In the literatures,[18,19] the original MEDV descriptor based on 13 atomic types, x_v ($v = 1, 2, \cdots, 91$), can be

* E-mail: ssliuhl@263.net or ssliu@nju.edu.cn

Received December 12, 2002; revised and accepted July 2, 2003.

Project supported by the National High Technology Research and Development Program of China (No. 2001AA646010-4) and the National Natural Science Foundation of China (No. 20177008).

calculated. Firstly, the relative electronegativity (q) of a non-hydrogen atom was calculated using Eq. (1). The atomic type, atomic attributes, and intrinsic state (I) of the atom are defined as shown in Table 1.

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

where d_{ij} is the shortest graph distance between two atoms, atom i and j.

Then, the MEDV descriptor, x_v, was calculated from the following formula

$$x_v = x_{kl} = \sum_{i \in k,\, j \in l} \frac{q_i q_j}{d_{ij}^2} \quad (2)$$

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

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

VSMP *for optimal selection of* MEDV *descriptors*

To accelerate the running speed of classical all-subsets regression (ASR) and 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 (q) in the leave-one-out (LOO) cross validation, are introduced into 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 in which each has m descriptors, $x(n, m)$? 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 one for a given vn but not always best for the whole subset space including all subsets of different vn. The fundamental processes are as follows:

(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}, acted as a criterion of the correlation coefficient in modeling, 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 that of 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, $x(n, vn)$, from the whole independent variable set, $x(n, m)$, 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, $x(n, vn)$, and the whole dependent variable set, $y(n)$, and calculate the relevant statistics such as the correlation coefficient (r_m) in building model. If the r_m is less than r_{cri}, then return to the step (2) to select a subset again.

Table 1 Atomic types, atomic attributes and intrinsic state (I) for various non-hydrogen atoms located in various molecular environments

Atom[a]	Type	Attribute	I	Atom[a]	Type	Attribute	I	Atom[a]	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.

(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 $q^2 \leqslant f_{max}$, 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, 5, 6, and 7) 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 with increasing *vn* and the LOO cross-validation correlation coefficient (q) gradually increases until a limited value and then decreases with increasing *vn*. For the root mean square errors (*RMS*), the similar results have been acquired. With the increase of *vn*, calibrated *RMS* (*RMSC*) is monotonically decreasing and validated *RMS* (*RMSV*) gradually decreases until a limit value and then increases. Therefore, the determination of the best subset is mainly depended on the q or *RMSV* in the LOO cross-validation procedure. The plot of *RMSV* versus *vn* will be employed together with some statistic analysis to determine the best subset entering into the final QSAR model.

Results and discussion

Data set

The data set used in this paper contains thirty-four COX-2 inhibitors, thirty-one 3,4-diarylcycloxazolones together with indomethacin, naproxen and celecoxib, which are directly taken from Table 1 in the literature.[20] For convenience, any one of thirty-four inhibitors is called as **DAA** compound. Their hydrogen-suppressed molecular structures are shown in Fig. 1. All inhibiting activity data (Table 2) are indicated as IC_{50} (μM). All data are expressed in terms of pIC_{50} [$pIC_{50} = -\lg (IC_{50} \times 10^{-6})$]. The pIC_{50} values are widespread (Fig. 2). The values of pIC_{50} are from 4.72 (**DAA25** and **DAA33**) to 6.68 (**DAA20**) and the mean value is 5.89 and standard derivation 0.492. It should be stated that the set excludes two compounds of No. 21 and No. 22 in the original literature[20] due to their IC_{50} values incompatible with 34 compounds selected here.

The structural descriptors of these **DAA** compounds are the MEDV descriptors obtained from the methodology. Analyzing the MEDV descriptors having 34 elements (samples), only 51 MEDV descriptors have one or more nonzero elements where 2 descriptors ($\boldsymbol{x}_4$ and $\boldsymbol{x}_{38}$) only contain 2 nonzero elements and 8 ones ($\boldsymbol{x}_{16}$, $\boldsymbol{x}_{27}$, $\boldsymbol{x}_{40}$, $\boldsymbol{x}_{42}$, $\boldsymbol{x}_{43}$, $\boldsymbol{x}_{45}$, $\boldsymbol{x}_{46}$, and $\boldsymbol{x}_{82}$) include 3 nonzero elements. The 10 descriptors with too little nonzero elements should be first eliminated from the 51 descriptors with nonzero elements due to no enough significance statistically. So, there are in fact 41 nonzero MEDV descriptors to enter into successive VSMP analysis.

From the above analysis, the data set consists of 34 **DAA** inhibitors and each has 41 MEDV descriptors.

Best variable-subset based on VSMP *and* QSAR *studies*

With the initial values of $r_{int} = 0.85$, $r_{cri} = 0.30$, and $f_{max} = 0$, the VSMP program is performed for the data including 34 **DAA** inhibitors with 41 MEDV descriptors in a given value of *vn*. The results show when *vn* equals 2, 3, 4, 5, 6 and 7, respectively, the root mean square error varies with the number of the descriptors in the optimal variable-subset, as shown in Fig. 3, in which the best combination of the descriptors from 41 MEDV ones should be the 5-descriptor subset of Nos. $\boldsymbol{x}_9$, $\boldsymbol{x}_{14}$, $\boldsymbol{x}_{19}$, $\boldsymbol{x}_{32}$ and $\boldsymbol{x}_{91}$. Using MLR method, a linear relationship model between the pIC_{50} values of **DAAs** and the five optimal descriptors are developed as follows.

$$pIC_{50} = (4.7210 \pm 0.2955) + (0.03483 \pm 0.01214) \cdot \boldsymbol{x}_9 - (0.03795 \pm 0.00768) \cdot \boldsymbol{x}_{14} - (0.4076 \pm 0.0462) \cdot \boldsymbol{x}_{19} - (0.05622 \pm 0.01618) \cdot \pm \boldsymbol{x}_{32} - (0.04313 \pm 0.00994) \cdot \boldsymbol{x}_{91} \quad (3)$$

where the value after the symbol " ± " refers to the standard derivation of the regression coefficient. The calibrated root mean square error (*RMSC*) and correlation coefficient (r) between the pIC_{50} values calculated by Eq. (3) and ones observed experimentally are 0.23 and 0.8804 respectively, To examine the stability of the model [Eq. (3)], it is essential to perform a cross-validation procedure for the **DAA** data set. The leave-one-out (LOO) cross-validation was used. The validated root mean square error (*RMSV*) and correlation coefficient (q) between the pIC_{50} values predicted by LOO validation and ones observed experimentally are 0.26 and 0.8464, respectively. These results show that though the model has significant statistics, the root mean square errors, both *RMSC* and *RMSV*, still be a little high. Analyzing the pIC_{50} values obtained by LOO validation, it is found that three inhibitors have too high errors. They are **DAA02** (error of −0.49), **DAA20** (−0.65) and **DAA31** (−0.52).

Removing the 3 compounds from the data set of 34 inhibitors, selection of the optimal descriptors is repeated by using VSMP program. When *vn* equals 2, 3, 4, 5, 6, and 7, respectively, the root mean square error varies with the number of the descriptors in the optimal variable-subset, as shown in Fig. 3 from which the best variable-subset is still 5-descriptor combination. These descriptors are $\boldsymbol{x}_{14}$, $\boldsymbol{x}_{19}$, $\boldsymbol{x}_{32}$, $\boldsymbol{x}_{52}$ and $\boldsymbol{x}_{91}$ where $\boldsymbol{x}_{52}$ replaces $\boldsymbol{x}_9$ in 5-descriptor subset from the 34-DAA set. However, it is compatible with each other due to high correlation coefficient of 0.8811 between $\boldsymbol{x}_{52}$ and $\boldsymbol{x}_9$. Using MLR , a new

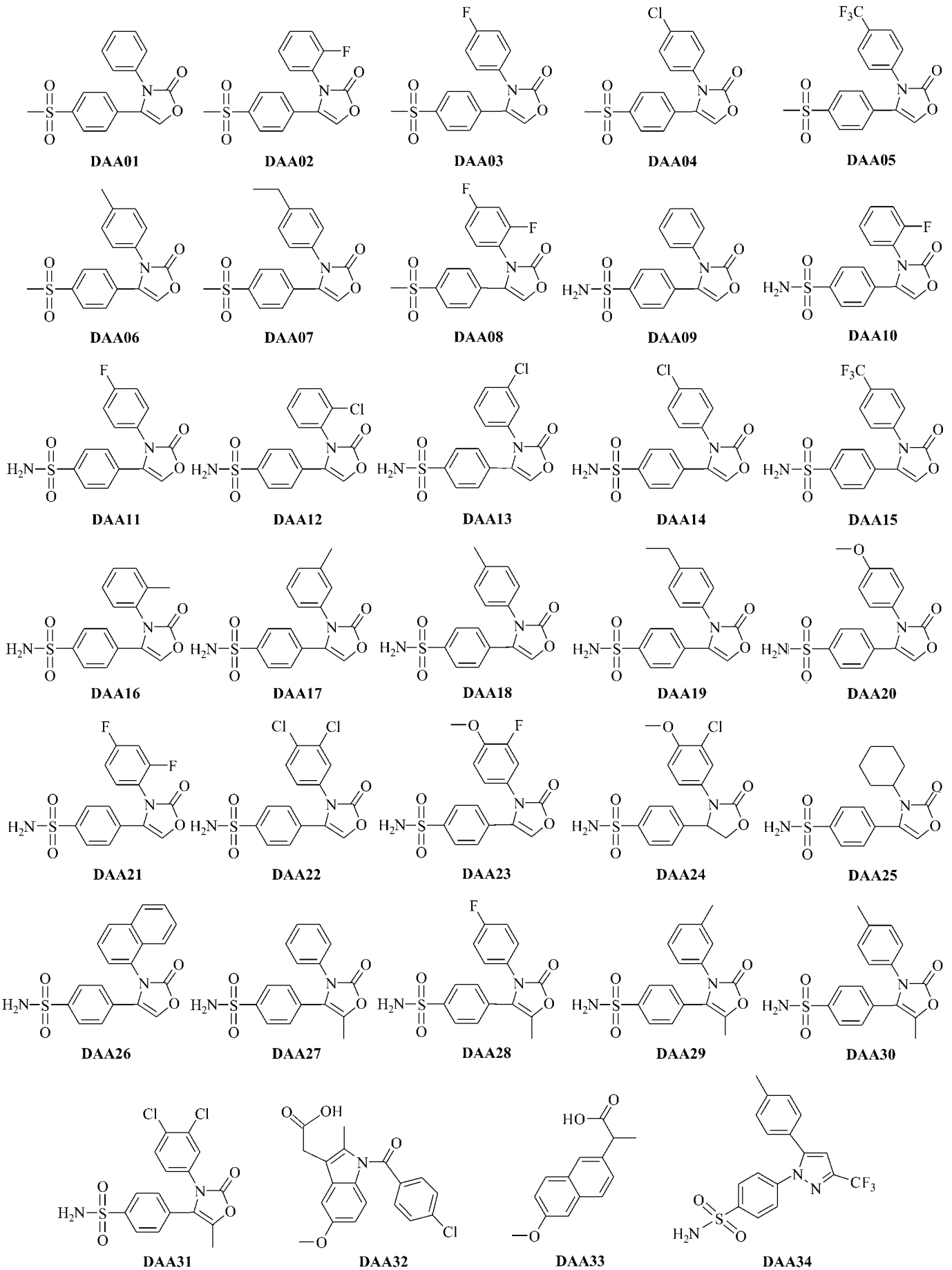

Fig. 1 Hydrogen-suppressed structures of 34 selective COX-2 inhibitors.

Table 2 Values of six descriptors and the observed IC_{50} and pIC_{50} as well as various calculated pIC_{50} for **DAA** inhibitor compounds

No.	Orig compd.[20]	IC_{50}[20]	pIC_{50}	Y_{MOD}	Y_{LOO}	Y_{TRN}	x_9	x_{14}	x_{19}	x_{32}	x_{52}	x_{91}
DAA01	1	1.6	5.80	5.92	5.96	5.81*	8.8749	47.4707	−7.6159	6.4769	0.0000	0.0000
DAA02	2	0.66	6.18	**5.69**		**5.62**	8.8819	37.6026	−6.1789	6.3616	0.0000	0.0000
DAA03	3	0.51	6.29	6.28	6.27	6.25*	8.8792	36.3984	−7.4975	6.2005	0.0000	0.0000
DAA04	4	0.32	6.49	6.23	6.18	6.18*	8.9063	38.5935	−7.6447	6.7831	0.0000	0.0000
DAA05	5	2.0	5.70	5.71	5.72	5.64*	8.9040	37.5993	−7.8159	6.1440	0.0000	16.6218
DAA06	6	0.48	6.32	6.25	6.24	6.21	9.5824	38.3224	−7.6272	6.7122	0.0000	0.0000
DAA07	7	0.69	6.16	6.24	6.26	6.18*	9.5303	40.6045	−7.8332	6.9841	0.0000	0.0000
DAA08	8	1.35	5.87	5.97	6.00	5.97	8.8862	26.9424	−6.0159	6.0551	0.0000	1.2997
DAA09	9	2.4	5.62	5.67	5.68	5.60	0.0000	47.1005	−7.6045	6.1316	0.0965	0.0000
DAA10	10	2.2	5.66	5.44	5.42	5.41*	0.0000	37.2615	−6.1655	6.0135	0.0960	0.0000
DAA11	11	1.5	5.82	6.02	6.04	6.04*	0.0000	36.0547	−7.4856	5.8546	0.0962	0.0000
DAA12	12	6.5	5.19	5.29	5.30	5.24	0.0000	39.1392	−6.1455	7.3192	0.0976	0.0000
DAA13	13	1.5	5.82	5.89	5.89	5.89	0.0000	37.1785	−7.3658	6.6156	0.0972	0.0000
DAA14	14	1.2	5.92	5.97	5.97	5.98	0.0000	38.2453	−7.6335	6.4356	0.0970	0.0000
DAA15	15	2.0	5.70	5.46	5.32	5.43	0.0000	37.2540	−7.8035	5.7966	0.0958	16.6119
DAA16	16	9.7	5.01	5.34	5.38	5.29	1.2626	38.9081	−6.1484	7.1610	0.0974	0.0000
DAA17	17	0.51	6.29	5.93	5.91	5.93	0.8916	36.9447	−7.3538	6.5157	0.0971	0.0000
DAA18	18	0.79	6.10	6.00	5.99	6.00*	0.6767	37.9748	−7.6159	6.3649	0.0969	0.0000
DAA19	19	1.56	5.81	5.98	6.00	5.97	0.5950	40.2494	−7.8220	6.6355	0.0974	0.0000
DAA20	20	0.21	6.68	**6.02**		**6.02**	0.5181	38.4585	−7.7175	6.3684	0.1321	0.0000
	21	>100		**6.45**		**6.44**	0.0000	40.8970	−7.6119	−3.4309	0.1344	0.5390
	22	59%		**6.51**		**6.51**	0.0000	41.6116	−7.7695	−3.8623	0.1347	0.5408
DAA21	23	2.3	5.64	5.71	5.72	5.76	0.0000	26.6260	−6.0021	5.7064	0.0956	1.2985
DAA22	24	0.4	6.40	6.05	6.00	6.12	0.0000	29.3827	−7.2319	7.1598	0.0977	0.5453
DAA23	25	1.13	5.95	6.17	6.21	6.25*	0.5192	28.5869	−7.2296	6.1849	0.1316	0.0000
DAA24	26	0.78	6.11	6.12	6.12	6.19	0.5306	29.4925	−7.3069	7.0847	0.1344	0.0000
DAA25	27	18.9	4.72	4.71	4.71	4.69	0.0000	29.9192	−4.1263	8.5996	0.1046	0.0000
DAA26	28	4.7	5.33	5.21	5.13	5.06*	0.0000	56.3691	−7.5120	7.8172	0.0968	0.0000
DAA27	29	1.8	5.74	5.71	5.71	5.63	2.2929	46.5209	−7.3548	5.2560	0.1025	0.0000
DAA28	30	0.51	6.29	6.07	6.06	6.08	2.2968	35.4606	−7.2323	4.9497	0.1021	0.0000
DAA29	31	0.93	6.03	5.97	5.97	5.96	3.2155	36.3277	−7.0922	5.6313	0.1031	0.0000
DAA30	32	1.7	5.77	6.04	6.06	6.04	2.9922	37.3533	−7.3627	5.4814	0.1029	0.0000
DAA31	33	0.23	6.64	**6.09**		**6.14**	2.3309	28.7444	−6.9630	6.2771	0.1037	0.5464
DAA32	indomethacin	0.22	6.66	6.62	6.59	6.68	3.5650	27.2191	−6.2956	−5.0467	0.0000	0.0000
DAA33	naproxen	18.9	4.72	4.74	4.84	4.67*	3.4816	23.0328	0.0000	−13.4793	0.0000	0.0000
DAA34	celecoxib	1.1	5.96	6.18	6.16	6.20	0.2670	38.3740	−9.8400	6.2813	0.1507	17.5772

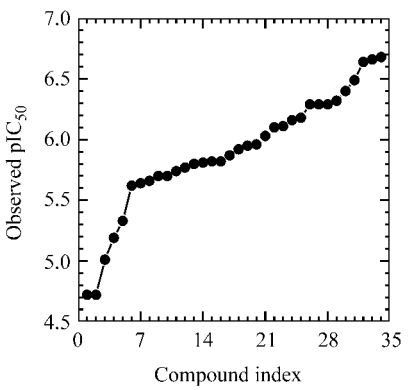

Fig. 2 Distribution of pIC_{50} of 34 COX-2 inhibitors.

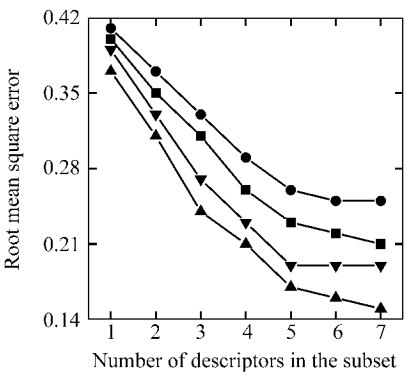

Fig. 3 Varies of the root mean square error with the number of descriptors in the optimal subset. –●– Validation for **34-DAA** set, –■– Calibration for **34-DAA** set, –▼– Validation for **31-DAA** set, –▲– Calibration for **31-DAA** set.

5-variable linear model between the pIC_{50} values and the five optimal descriptors is rebuilt as following Eq. (4a). If the descriptor x_9 is used instead of x_{52}, then another model [Eq. (4b)] is obtained.

$$pIC_{50} = (4.8247 \pm 0.2280) - (0.03617 \pm 0.00583) \cdot x_{14} - (0.4152 \pm 0.0345) \cdot x_{19} - (0.05249 \pm 0.01253) \cdot x_{32} - (2.7253 \pm 0.7397) \cdot x_{52} - (0.03888 \pm 0.00733) \cdot x_{91}$$
$$n = 31,\ m = 5,\ r^2 = 0.8703,\ r = 0.9329,$$
$$RMSC = 0.17,\ F = 33.55\ (\text{calibration})$$
$$n = 31,\ m = 5,\ q^2 = 0.8355,\ q = 0.9141,$$
$$RMSV = 0.19\ (\text{LOO validation}) \qquad (4a)$$

$$pIC_{50} = (4.6117 \pm 0.2302) + (0.03191 \pm 0.00979) \cdot x_9 - (0.03473 \pm 0.00605) \cdot x_{14} - (0.4030 \pm 0.0362) \cdot x_{19} - (0.06055 \pm 0.01256) \cdot x_{32} - (0.03947 \pm 0.00763) \cdot x_{91}$$
$$n = 31,\ m = 5,\ r^2 = 0.8595,\ r = 0.9271,$$
$$RMSC = 0.17,\ F = 30.595\ (\text{calibration})$$
$$n = 31,\ m = 5,\ q^2 = 0.8153,\ q = 0.9030,$$
$$RMSV = 0.20\ (\text{LOO validation}) \qquad (4b)$$

Comparing Eq. (4a) with Eq. (4b) no significant difference is found. The values of the descriptors (x_9, x_{14}, x_{19}, x_{32}, x_{52} and x_{91}) are listed in Table 2.

Comparing Eq. (4) with Eq. (3), the 5-descriptor model from the 31 **DAA** data set has not only better calibration power for the internal inhibitor samples but also higher validation ability for the external samples than the one from the 34 **DAA** data set. Table 2 lists the pIC_{50} values calibrated (Y_{MOD}) and validated (Y_{LOO}) by Eq. (4b) together with the experimental values (pIC_{50}). Table 2 also lists the coding number (Orig compd.) and IC_{50} values (IC_{50}) of all inhibitor molecules in the original literature.[20] Ten values expressed in boldface refer to ones predicted by the model. The plot of the Y_{MOD} values versus Y_{LOO} ones is shown in Fig. 4.

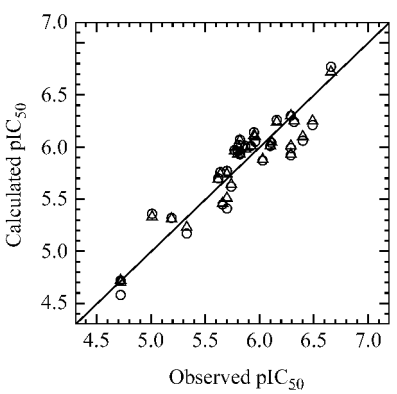

Fig. 4 Plot of pIC_{50} calculated by Eq. (4b) versus observed. –△– Calibration, –○– Vlidataion.

To further validate the stability and predictive ability of the above model, the pIC_{50} values of 31 **DAA** compounds are ranked in the order of small to large and 20 compounds are equidistantly picked out from 31 ranked inhibitors to construct a training set and the balance compounds comprise a testing set. The training set is used to build a new linear model between the pIC_{50} values and the optimal descriptors of x_9, x_{14}, x_{19}, x_{32} and x_{91}. The model with $RMSC = 0.18$ and $r = 0.9219$ is as following Eq. (5), which is compatible with Eq. (4). Then the model is employed to predict the pIC_{50} values of 11 compounds in the external testing set. The results show that the predicted root mean square error ($RMSP$) and correlation coefficient (r_P) between the pIC_{50} values predicted by the model [Eq. (5)] derived from the training set and the observed pIC_{50} values of 11 testing inhibitors are $RMSP = 0.19$ and $r_P = 0.9332$. The pIC_{50} values calibrated (Y_{TRN}) and predicted (Y_{TRN}^*) by the model [Eq. (5)] are also listed in Table 2 and the plot of the pIC_{50} calculated by Eq. (5) versus ones observed is shown in Fig. 5. Obviously, the training model has also high calibration and prediction power.

$$pIC_{50} = (4.7279 \pm 0.3948) + (0.02676 \pm 0.01829) \cdot x_9 - (0.04305 \pm 0.01028) \cdot x_{14} - (0.4316 \pm 0.0620) \cdot x_{19} - (0.06221 \pm 0.01924) \cdot x_{32} - (0.04192 \pm 0.01169) \cdot x_{91} \qquad (5)$$

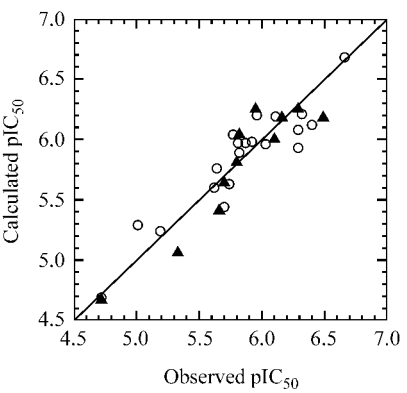

Fig. 5 Plot of pIC_{50} calculated by Eq. (5) versus observed. –○– Calibration, –▲– Prediction

Self-correlation and physical meaning of descriptors

The inter-correlation coefficient (r_{INT}) 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_{INT} between various pairs of variables in the best 5-variable subset are less than 0.70. These inter-correlation coefficients are listed in Table 3. The results show that there are no significant correlations between the structural descriptors both in Eqs. (3) and (4).

From Table 1 and Eq. (2), the descriptors entering into the final model are related to six atomic types in the **DAA** molecules. These are types 1, 2, 3, 7, 9 and 13, which characterize the atom segments such as $-CH_3$ (1), $-CH=$ or $-CH_2-$ (2), $>C=$ or $>CH-$ (3), $>N-$

Table 3 Correlation coefficient (r_{INT}) matrix between various pairs of the descriptors in the optimal combination

r_{INT}	x_{14}	x_{19}	x_{32}	x_{52}	x_{91}
x_9	0.0357	0.0110	0.0554	0.8811	0.0147
x_{14}		0.5070	0.4366	0.0614	0.0269
x_{19}			0.6966	0.2678	0.3115
x_{32}				0.3501	0.0505
x_{52}					0.0534

(7), =O or −OH (9), and −F or −Cl (13). The atomic type of an atom reflects its adjacent environments. Accordingly, specific contributions of branch-chain such as >C= and >CH−, functional groups such as =O and −OH, shape of molecular skeleton in **DAA** such as >N− can be learnt from the descriptors based on diverse atomic types. Five descriptors in Eq. (4b), x_9, x_{14}, x_{19}, x_{32} and x_{91}, show five interactions between five pairs of atom types, *i.e.* $-CH_3$ and =O or −OH ($x_9 = x_{1,9}$), −CH= or $-CH_2-$ and −CH= ($x_{14} = x_{2,2}$), −CH= or $-CH_2-$ and >N− ($x_{19} = x_{2,7}$), >C= or >CH− and =O or −OH ($x_{32} = x_{3,9}$), and −F or −Cl and −F ($x_{91} = x_{13,13}$). The other atom segments such as the common skeleton structure, ⩾S⩽, of **DAA** molecules has no effect on the pIC_{50} values due to the existence in all **DAA** molecules.

Conclusion

A novel 5-variable QSAR model between biological activities expressed has been described by pIC_{50} values and the MEDV descriptors for 34 **DAA** COX-2 inhibitors using a novel variable selection and modeling based on the predictions (VSMP), with the calibrated root mean square error of *RMSC* = 0.17 and correlation coefficient of r = 0.9271 and the validated root mean square error of *RMSV* = 0.20 and correlation coefficient of q = 0.9030. The results show that the model has not only high calibrated quality but also prediction ability.

References

1 Cryer, B.; Feldman, M. *Am. J. Med.* **1998**, *104*, 416.

2 Herchsman, H. R. *Biochem. Biophys. Acta* **1996**, *1299*, 125.

3 Dubois, R. N.; Abramson, S. B.; Crofford, L.; Gupta, R. A.; Simon, L. S.; van de Putte, L. B.; Lipsky, P. E. *FASEB J.* **1998**, *12*, 1063.

4 Mitchell, J. A.; Akarasereenont, P.; Thiemermann, C.; Flower, R. J.; Vane, J. R. *Proc. Natl. Acad. Sci. U.S.A.* **1994**, *90*, 11693.

5 Wilkerson, W. W.; Copeland, R. A.; Covington, M. B.; Trzaskos, J. M. *J. Med. Chem.* **1995**, *38*, 3895.

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

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

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

9 Kumar, R.; Singh, P. *Indian J. Chem., Sect. B* **1997**, *36*, 1164.

10 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.

11 Chavatte, P.; Yous, S.; Marot, C.; Baurin, N.; Lesieur, D. *J. Med. Chem.* **2001**, *44*, 3223.

12 Garg, R.; Kurup, A.; Mekapati, S. B.; Hansch, C. *Abstr. Pap. Am. Chem. Soc.* **2001**, *221*, 236.

13 Desiraju, G. R.; Gopalakrishnan, B.; Jetti, R. K. R.; Raveendra, D.; Sarma, J. A. R. P.; Subramanya, H. S. *Molecules* **2000**, *5*, 945.

14 Liu, S. S.; Yin, C. S.; Shi, Y. Y.; Cai, S. X.; Li, Z. L. *Chin. J. Chem.* **2001**, *19*, 751.

15 Liu, S. S.; Liu, H. L.; Shi, Y. Y.; Wang, L. S. *Internet Electron. J. Mol. Des.* **2002**, *1*, 310.

16 Liu, S. S.; Cui, S. H.; Shi, Y. Y.; Wang, L. S. *Internet Electron. J. Mol. Des.* **2002**, *1*, 610.

17 Liu, S. S.; Liu H. L.; Cui, S. H.; Wang, L. S. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 964.

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

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

20 Puig, C.; Crespo, M. I.; Godessart, N.; Feixas, J.; Ibarzo, J.; Jimenez, J.-M.; Soca, L.; Cardelus, I.; Heredia, 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.

(E0212121 SONG, J. P.; FAN, Y. Y.)