Combined MEDV-GA-MLR Method for QSAR of Three Panels of Steroids, Dipeptides, and COX-2 Inhibitors

Shu-Shen Liu,* Chun-Sheng Yin, and Lian-Sheng Wang

State Key Laboratory of Pollution Control and Resources Reuse, Department of Environmental Science & Engineering, Nanjing University, Nanjing 210093, People's Republic of China

Received November 19, 2001

The MEDV-13, molecular electronegativity distance vector based on 13 atomic types, has at best 91 descriptors. It is impossible to indirectly use multiple linear regression (MLR) to derive a quantitative structure−activity relationship (QSAR) model. Although principal component regression (PCR) or partial least-squares regression (PLSR) can be employed to develop a latent QSAR model, it is still difficult how to determine the principal components (PCs) and depict the physical meaning of the PCs. So, a genetic algorithm (GA) is first employed to select an optimal subset of the descriptors from original MEDV-13 descriptor set. Then MLR is utilized to build a QSAR model between the optimal subset and the biological activities of three sets of compounds. For 31 benchmark steroids, a 5-descriptor QSAR model (M1) between the corticosteroid-binding globulin (CBG) binding affinity of the steroids and 5-descriptor subset is developed. The root-mean-square error of estimations (*RMSEE*) and the correlation coefficient of estimations (r) between the CBG binding affinity (BA) observed and the BA estimated by M1 are 0.422 and 0.9182, respectively. The root-mean-square error of predictions (*RMSEP*) and the correlation coefficient of predictions (q) between the BA observed and the BA predicted by leave-one-out cross validations are 0.504 and 0.8818, respectively. For 58 dipeptides inhibiting angiotensin-converting enzyme (ACE), a 5-variable QSAR model (M2) between the pIC_{50} of peptides and 5-descriptor subset is derived. The M2 has a high quality with $RMSEE = 0.339$ and $r = 0.9398$ and $RMSEP = 0.370$ and $q = 0.9280$. For 16 indomethacin amides and esters (ImAE) inhibiting cyclooxygenase-2 (COX-2), a 6-variable QSAR model (M3) with $RMSEE = 0.079$ and $r = 0.9839$ and $RMSEP = 0.151$ and $q = 0.9413$ is built.

INTRODUCTION

How to describe the structure of a molecule is a very important task in the quantitative structure- activity relationship (QSAR) technique study. At present, the current methods about the description of the molecular structures include two-dimensional topological descriptors, energetic descriptors, quantum mechanical descriptors, and three-dimensional molecular field descriptors.[1] It has become evident that physical, chemical, or biological properties of a compound depend on the three-dimensional (3D) arrangements of the atoms in the molecule. In recent years, the QSAR methods based on the 3D structures of the molecules, such as CoMFA,[2] GRID,[3] COMPASS,[4] and SOMFA,[5] have been widely used in several scientific fields. However, implementing these methods based on 3D structure are in general difficult and time-consuming because of the difficulty in generating optimal 3D conformation of the molecule under study. And many current excellent QSAR methods based on two-dimensional properties of the molecule such as topological structural characteristics also have a comparable quality to the 3D methods. Clearly, it is still essential to improve the current 2D-QSAR techniques and develop new or/and better 2D-QSAR methods than the current ones. Tong[6] recently employed a relative simple approach, called a hologram QSAR (HQSAR), only encoding 2D structure information, to develop a potential QSAR model by using partial least squares regression (PLS). The QSAR models generated using HQSAR techniques have comparable quality to those of CoMFA. Another excellent 2D descriptor, an atom level electrotopological state (E-state) index, was introduced by Kier and Hall and used successfully for a variety of QSAR studies.[7−11]

In our previous paper,[12,13] a molecular electronegativity distance vector based on 13 atomic types (MEDV-13) was reported, and the principal component regression (PCR) technique was used to derive the QSAR models between the MEDV-13 vector and the biological activities of three panels of organic compounds including a set of 31 "benchmark" steroids binding to the corticosteroid-binding globulin (CBG) and a set of 58 dipeptides inhibiting angiotensin-converting enzyme (ACE) as well as a set of 16 indomethacin amides and esters (ImAE) inhibiting cyclooxygenase-2 (COX-2). However, how to explain and determine the principal components obtained by PCR method are still difficult. To determine which descriptors are main factors affecting the biological activities such as the pIC_{50} of the inhibitors, it is essential to select an optimal subset of the variables from original MEDV descriptor set.

In general, a variable set space has a complex landscape with many local solutions, and, therefore, many techniques[14−18] such as FRED (fast random elimination of descriptors) algorithm,[14] neural network algorithm,[15] and evolutionary algorithm[16] were used to search the solution as good as

* Corresponding author phone: (86)-025-3596509; e-mail: ssliu@nju.edu.cn or ssliuhl@263.net.

10.1021/ci010245a CCC: $22.00
Published on Web 04/24/2002

Table 1. 43 Atomic Attributes Used in MEDV-13

no.	attribute	δ^v	δ	I	no.	attribute	δ^v	δ	I	no.	attribute	δ^v	δ	I
1	$-CH_3$	1	1	2.0000	**16**	⊥C∀	4.5	1	1.8333	**30**	≥N=	5	3	2.2361
2	$-CH_2-$	2	1	1.5000	**17**	−OH	1	2	2.4495	**31**	−SH	1	3	1.7691
3	−CH<	3	2	1.3333	**18**	−O−	2	3	1.8371	**32**	−S−	2	1	1.1567
4	>C<	4	3	1.2500	**19**	=O	2	4	3.6742	**33**	=S	2	2	2.3134
5	$=CH_2$	2	1	3.0000	**20**	⊥O	1.5	1	3.0619	**34**	>S=	4	1	1.1340
6	=CH−	3	2	2.0000	**21**	$-NH_2$	1	2	2.2361	**35**	≥S≤	6	4	1.1227
7	=C<	4	1	1.6667	**22**	−NH−	2	3	1.6771	**36**	−F	1	1	2.6458
8	=C=	4	1	2.5000	**23**	>N−	3	2	1.0882	**37**	−Cl	1	1	1.9108
9	≡CH	3	2	4.0000	**24**	=NH	2	1	3.3541	**38**	−r	1	1	1.6536
10	≡C−	4	2	2.5000	**25**	=N−	3	2	2.2361	**39**	−I	1	1	1.5345
11	$\perp CH_2$	1.5	3	2.5000	**26**	≡N	3	1	4.4721	**40**	$-PH_2$	1	1	1.6149
12	⊥CH-	2.5	1	1.7500	**27**	⊥NH	1.5	2	2.7951	**41**	−PH−	2	2	1.0559
13	⊥C<	3.5	2	1.5000	**28**	⊥N−	2.5	3	1.9566	**42**	>P-	3	3	0.8696
14	⊥CH⊥	3	1	2.0000	**29**	⊥N⊥	3	2	2.2361	**43**	≥P<	5	4	0.9006
15	-C∀	4	3	1.6667										

Table 2. Series Number (v) of MEDV-13 Descriptors Related to the Atomic Type k and l

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

possible. Genetic algorithms (GAs) are very likely to be the most widely known type of evolutionary algorithms. Because of their simplicity, flexibility, easy operation, minimal requirements, and global perspective, GAs have been successfully used in many scientific fields.[16−21] In this paper, a method combined the optimal selection of the MEDV-13 descriptors using a genetic algorithm (GA) program developed in our laboratory with multiple linear regression (MLR), called MEDV-GA-MLR method, is developed. First, a GA is employed to select the optimal subset of the variables from original MEDV-13 descriptor set. Then MLR method is utilized to build a QSAR model between the optimal variables and the biological activities.

METHODOLOGY

MEDV-13 Descriptor. Using the *atomic types* and *atomic attribute*s, the molecular electronegativity distance Vector based on 13 atomic types (MEDV-13), can be obtained according to the method developed in the previous paper.[12] The term *atomic type* is used to represent indirectly the topological structural characteristics of the molecule of interest and defined as the number of non-hydrogen atoms binding to that atom plus its identifying number (ID). And the ID is used to specify the number of valence electrons (v) of the atom in the same local topological environment and ID = (v-4)*4. The *atomic attribute* is used to characterize the local chemical environment and element nature of a non-hydrogen atom. Because the nature of an alone C═O double bond is different from the one of a conjugated C═O bond, the original 42-atomic attribute project is modified as an 43-atomic attribute one by adding a conjugated C═O bond. For the convenience of application, the modified 43-atomic attribute project is listed in Table 1 where the symbols "⊥"and "∀" represent one and two conjugated double bonds. The meaning of δ^v, δ, and I and calculating method of MEDV descriptors ($\boldsymbol{x}_v$ or h_{kl}) are the same as the literature.[12] The correlations between subscript "v" and atomic type "k" and "l" are listed in Table 2.

GA for Descriptor Selection. From the literature,[12] there are at best 91 MEDV-13 descriptors for the molecules having all 13 atomic types. So, it is impossible to directly use multiple linear regression (MLR) to develop a QSAR model, and it is essential to employ the multivariable statistical methods based on factor analysis such as principal component regression (PCR)[22,23] or partial least-squares regression (PLSR)[24,25] technique. To acquire a transparent QSAR model, a genetic algorithm (GA) program designed in our laboratory is used to select an optimal subset from original MEDV-13 descriptor set. Theories and applications about various GAs have been reported by many authors,[26−29] and so here the basic procedures about GAs are not explained again.

However, there are several problems such as the coding project and the evaluation of the fitness for various individuals (chromosomes) in a population need to be still described in any a concrete GA program. In our GA computer program, a binary coding project is adopted. To search various optimal subsets in a different number of variables (Vn) here Vn = 2, 3, 4, 5, 6, 7, 8, 9, the GA program is designed to determine the optimal subset in a given Vn. The most important and first step in binary coding GA algorithm is how to generate a binary bit string corresponding to an individual. Taking the 31 steroid data set as a sample, the procedure of

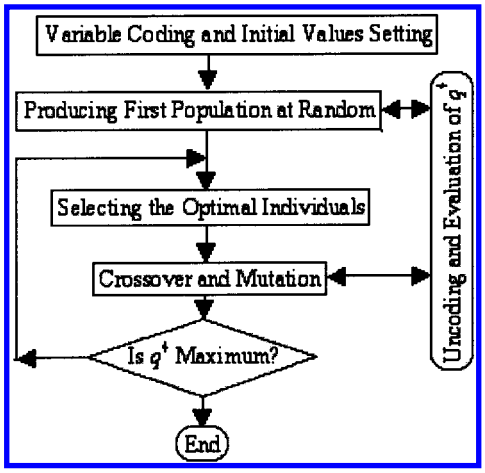

Figure 1. Flow chart for our GA procedure (GTVSGA).

generating a binary string is stated.

There are 15 original MEDV descriptors ($m = 15$) in the 31 steroid set. Now let the GA program search the optimal subset including five descriptors ($Vn = 5$) from an original descriptor set consisting of 15 variables. First, each of the 15 variables is respectively encoded as a series number. Then the length of the binary bit substring (*bm*) for such an anonymous variable is calculated as the following eq 1

$$bm = \text{int}\{[\log(m - 1) + bt - 1]/\log 2\} \qquad (1)$$

where "*bt*" refers to the effective bit number of the integer *m* and here $bt = 2$. The operator "*int*" makes the subsequent value in the bracket an integer. For the sample of 31 steroids, $bm = 13$. Finally, the length of total binary string for a chromosome including *Vn* equals $Vn \times bm$ and here $5 \times 13 = 65$.

Decoding a binary substring as a variable (**x**) is finished by eq 2.

$$\mathbf{x} = 1 + decimal(0101001001011)_{binary} \cdot \frac{m - 1}{2^{bm} - 1} \qquad (2)$$

On the other hand, determination of the fitness function for a GA process is also very important. It is well-known that the predictive ability for the external samples is more important than the estimation ability for the internal samples. So, quartic of the correlation coefficient of predictions (q) obtained in a leave-one-out (LOO) cross-validation (CV) step is taken as a fitness function in this paper. To eliminate the correlations between the independent variables, let the fitness value of an individual be zero if the absolute value of the correlation coefficient between arbitrary two variables in an individual is more than 0.95.

According to the above coding project and the fitness value criteria, a GA program (GTVSGA) is written with True BASIC language. The program not only includes various general steps in GA such as the selection of optimal individuals in an population, crossover, and mutation but also a combination with the LOO method in MLR calculation. The fundamental process of the GA program is shown in Figure 1.

RESULTS AND DISCUSSION

Data Set. To validate the feasibility of the MEDV-GA-MLR method, three data sets studied using PCR technique in our previous papers[12,13] are selected. The first set is taken from a steroid binding affinity prediction problem previously studied. The data set consists of 31 steroids assayed for binding affinity (BA) to one transport protein, corticosteroid-binding globulin (CBG).[5] The second one consists of 58 dipeptides inhibiting angiotensin-converting enzyme (ACE).[30] The final one is a series of 16 indomethacin amides and esters (ImAE) inhibiting cyclooxygenase-2 (COX-2) selected from Kalgutkar report.[31] Here, the symbols, BA, pIC_{50}, and pIC_{50}, are respectively employed to express the biological activities for 31 steroids, 58 dipeptides, and 16 ImAE compounds. The MEDV-13 descriptors above are used to characterize the structures of these compounds.

Preliminary Selection of MEDV-13 Descriptors. To ensure enough statistic significance of various MEDV-13 descriptors in a QSAR model, the descriptors having zero or a few (1 or 2) nonzero values in *n* samples are moved from original MEDV-13 descriptor set prior to a GA analysis. Furthermore, the descriptors having zero variance should be also eliminated from the set.

For the steroid system consisting of 31 molecules ($n = 31$), 25 MEDV-13 descriptors of nos. 1, 2, 3, 4, 9, 10, 13, 14, 15, 16, 21, 22, 25, 26, 27, 32, 33, 36, 37, 42, 43, 46, 77, 78, and 81 are not all zero values where five descriptors of nos. 10, 22, 33, 43, and 78 have only two nonzero values and the other five descriptors of nos. 13, 25, 36, 46, and 81 have only one nonzero values. So, the 10 descriptors should be moved from the 25 nonzero descriptor set. Then, there are in fact only 15 descriptors entering into the later GA analysis.

For the dipeptide system consisting of 58 samples ($n = 58$), 33 descriptors of nos. 1, 2, 3, 5, 6, 7, 9, 10, 14, 15, 17, 18, 19, 21, 22, 26, 28, 29, 30, 32, 33, 47, 48, 49, 51, 52, 56, 57, 59, 60, 66, 77, and 78 have nonzero values where six descriptors of nos. 10, 22, 33, 52, 60, and 78 have only two nonzero values and no. 57 descriptors has only one nonzero value. So, there are only 26 descriptors entering into the later GA analysis.

For the ImAE system consisting of 16 samples, 42 descriptors of 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 have nonzero values where seven descriptors of nos. 5, 17, 28, 49, 51, 52, and 55 have only one nonzero value. So, there are 35 descriptors entering into the GA analysis.

GA for Selection of Descriptors. A number of calculations show that a high number of individuals (N_I) in a population and little mutation probability (P_{MUT}) are propitious to search the optimal variables from *m* descriptor set. So, GA parameters such as N_I, crossover probability (P_{CRO}), and P_{MUT} are set to be 120, 0.75, and 0.01, respectively.

It is well-known that classical GA procedure is sometimes located into a localized optimum area so as to miss the best value. To search the best combination of the variables for a given *Vn*, a random number generation program is run prior to GA analysis to produce a different initial population, which will provide much chances to the best point.

The best combinations of variables and some statistic parameters obtained by above GA combined with MLR are respectively listed in Table 3A−C for three data set above.

Best Combination and QSAR. It has been known that a high quality model has not only a good ability of estimation

Table 3. Some Statistic Parameters in the Optimal Combinations for 31 Steroids, 58 Peptides, and 16 ImAEs

Vn	descriptor	r^2	RMSEE	*F*	q^2	RMSEP
	A. 31 Steroids					
2	2,21	0.5392	0.723	16.380	0.4730	0.775
3	2,16,21	0.6904	0.593	20.070	0.6146	0.663
4	2,16,21,32	0.7358	0.547	18.104	0.6480	0.635
5	**1,2,4,21,32**	**0.8431**	**0.422**	**26.873**	**0.7776**	**0.504**
6	1,2,4,21,32,37	0.8536	0.408	23.329	0.7651	0.522
7	1,2,4,21,26,32,37	0.8542	0.407	19.257	0.7484	0.542
8	1,2,4,15,21,26,32,37	0.8611	0.397	17.051	0.7327	0.561
9	1,2,4,14,16,21,26,27,32	0.8564	0.404	13.921	0.6686	0.641
	B. 58 Peptides					
2	21,26	0.6781	0.563	57.926	0.6379	0.598
3	21,26,32	0.8076	0.435	75.552	0.7787	0.467
4	5,21,26,32	0.8646	0.365	84.596	0.8416	0.395
5	**5,21,26,32,51**	**0.8832**	**0.339**	**78.671**	**0.8611**	**0.370**
6	5,21,26,32,51,59	0.8913	0.327	69.673	0.8667	0.363
7	3,5,6,21,26,32,51	0.8962	0.320	61.661	0.8674	0.362
8	3,5,6,19,21,26,32,51	0.9023	0.310	56.583	0.8700	0.358
9	1,3,5,6,21,26,32,48,59	0.9064	0.304	51.665	0.8708	0.358
	C. 16 ImAEs					
2	10,19	0.7697	0.212	21.727	0.6580	0.264
3	14,22,29	0.8298	0.182	19.504	0.7578	0.220
4	2,10,19,26	0.9090	0.133	27.477	0.7969	0.202
5	1,7,10,18,21	0.9389	0.109	30.722	0.8556	0.171
6	**2,10,19,26,29,57**	**0.9680**	**0.079**	**45.329**	**0.8861**	**0.151**
7	1,7,10,18,21,26,32	0.9676	0.079	34.181	0.9052	0.138
8	2,10,14,19,26,29,57,77	0.9766	0.068	36.518	0.9069	0.135
9	2,10,15,21,26,36,57,59,78	0.9828	0.058	38.010	0.8982	0.146

for the internal samples but also an excellent ability of prediction for the external samples, and the later is more important for a QSAR model. The determination of the best subset from among various optimal subsets have to examine not only the root-mean-square error and correlation coefficients of estimations (*RMSEE* or *r*) but also the root-mean-square error and correlation coefficients of predictions (*RMSEP* or *q*) in LOO prediction step. To examine intuitively the best subset, plots of the *RMSEE* and *RMSEP* values in Table 3A−C versus *Vn* are respectively shown in Figure 2A−C.

From Table 3A and Figure 2A, the best subset of the descriptors is a combination of five descriptors of nos. 1, 2, 4, 21, and 32 (υ). From Table 2, the five descriptors are related to five atomic types of nos. 1, 2, 3, 4, and 9 (*k* or *l*), which explains the effects of five substructures, CH_3-, =CH−, >CH−, >C<, and O= or O− on the biological activity. On the other hand, each MEDV-13 descriptor corresponds to a pair of atomic types, which emphasizes that two substructures related to atomic types have to coexist in the same molecule (seeing Table 2). For example, two substructures, =CH− and O=, corresponding to descriptor $\boldsymbol{x}_{21}$, both exist in the same molecule.

The MLR techniques are employed to derive a 5-variable QSAR model (M1) between the binding affinity (*BA*) to CBG data and five MEDV-13 descriptors ($\boldsymbol{x}_\upsilon$) of 31 steroids. The M1 model are shown in eq 3 where the values after symbol "±" refer to the standard deviation from regression coefficients.

$$\mathrm{BA} = -(2.9724 \pm 0.9279) + (0.8830 \pm 0.2298)\cdot\boldsymbol{x}_1 + (0.4145 \pm 0.0696)\cdot\boldsymbol{x}_2 + (0.4235 \pm 0.0685)\cdot\boldsymbol{x}_4 + (0.5345 \pm 0.0539)\cdot\boldsymbol{x}_{21} + (0.1094 \pm 0.0300)\cdot\boldsymbol{x}_{32} \quad (3)$$

$$n = 31,\ m = 5,\ r^2 = 0.8431,\ r = 0.9182,\ RMSEE = 0.422,\ F = 26.873 \text{ (estimation)}$$

$$n = 31,\ m = 5,\ q^2 = 0.7776,\ q = 0.8818,\ RMSEP = 0.504 \text{ (LOO prediction)}$$

Equation 3 shows that the correlation coefficient (*r*) and the root-mean-square error of estimations (*RMSEE*) between the activities (BA_{M1}) estimated by M1 and the observed activities (BA_{OBS}) are 0.9182 and 0.422, respectively. However, the results estimated by M1 only explain an estimated ability for the internal samples. The predictive ability of the M1 model still needs to be tested using a statistical procedure called cross-validation (CV). In this paper, a leave-one-out (LOO) method, one of many CV techniques, is employed to test the predictive ability. The correlation coefficient (*q*) and the root-mean-square error of predictions (*RMSEP*) between the activities (BA_{LOO}) predicted by *n* LOO models and the observed BA_{OBS} are 0.8818 and 0.504, respectively. The results above show that the M1 model has higher quality than the literature.[12]

For the 31 steroid set, the BA_{M1} estimated and BA_{LOO} predicted by M1 model are listed in Table 4A together with the BA_{OBS}. And Figure 3A shows the plot of BA_{M1} and BA_{LOO} versus BA_{OBS}.

From Table 3B and Figure 2B, the best subset of the descriptors is a combination of five descriptors of nos. 5, 21, 26, 32, and 51 (υ) for 58 peptides. From Table 2, the five descriptors are related to five atomic types of nos. 1, 2, 3, 5, and 9 (*k* or *l*), which explains the effects of five substructures, CH_3-, =CH-, >CH-, N-, and O= or O−, on the biological activity.

A 5-variable QSAR model (M2) with a high estimation ability of $r = 0.9398$ and $RMSEE = 0.339$ and an excellent

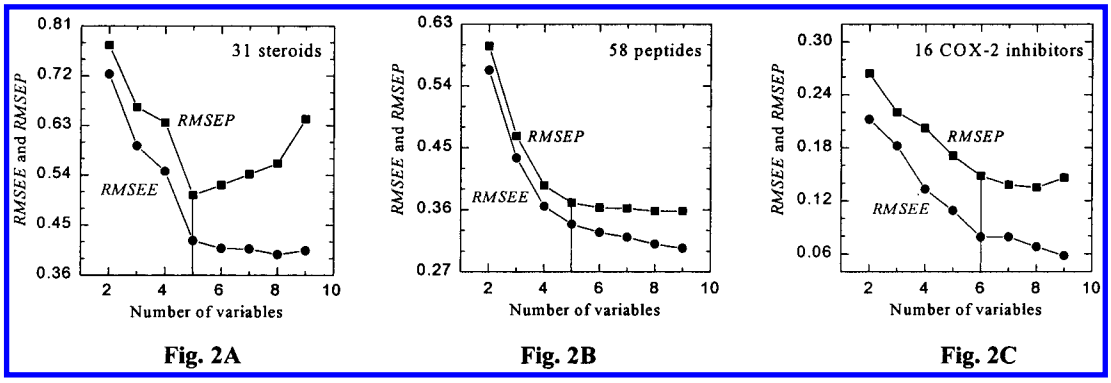

Figure 2. A. Varies of *RMSEP* and *RMSEE* with *Vn*. B. Varies of *RMSEP* and *RMSEE* with *Vn*. C. Varies of *RMSEP* and *RMSEE* with *Vn*.

Table 4. Optimal Descriptors and Various Activities of 31 Steroids, 58 Peptides, and 16 COX-2 Inhibitors

A. 31 Steroids

no.	compound	x_1	x_2	x_4	x_{21}	x_{32}	BA_{OBS}	BA_{EST}	BA_{LOO}
1	aldosterone	0	4.4910	−3.2899	16.5144	2.7832	6.279	6.628	6.686
2	androstanediol	0.3686	9.8221	−3.7198	10.9620	−0.7215	5.000	5.630	5.781
3	androstenediol	0.3518	9.9164	−5.5058	10.6579	−0.7575	5.000	4.732	4.662
4	androstenedione	0.3267	8.2825	−7.3663	15.6036	7.3250	5.763	6.772	6.953
5	androsterone	0.3428	8.6430	−6.0023	12.7573	4.2142	5.613	5.652	5.656
6	corticosterone	0.3485	7.6064	−6.6275	17.3463	3.4820	7.881	7.335	7.288
7	cortisol	0.3450	7.2822	−9.4078	20.1805	2.9187	7.881	7.473	7.419
8	cortisone	0.3231	6.0310	−10.7684	19.7789	6.3920	6.892	6.525	6.382
9	dehydroepiandrosterone	0.3269	8.7488	−7.7298	12.4690	4.0758	5.000	4.781	4.713
10	deoxycorticosterone	0.3504	9.0557	−5.9688	14.2826	5.1331	7.653	6.759	6.645
11	deoxycortisol	0.3469	8.6841	−8.7049	17.8790	5.1167	7.881	7.364	7.290
12	dihydrotestosterone	0.3595	8.6706	−4.2018	12.2316	2.4682	5.919	5.968	5.973
13	estradiol	0	4.2943	−2.1756	12.8824	1.8581	5.000	4.976	4.970
14	estriol	0	3.6353	−2.4907	14.6253	−0.8792	5.000	5.201	5.274
15	estrone	0	3.4620	−4.1470	14.6637	6.5670	5.000	5.263	5.364
16	etiocholanolone	0.3428	8.6430	−6.0023	12.7573	4.2142	5.225	5.652	5.697
17	pregnenolone	1.0256	10.9721	−6.4739	10.1981	−1.1058	5.225	5.070	5.035
18	17-hydroxypregnenolone	1.0093	10.5337	−10.3964	14.6547	−1.4247	5.000	5.560	5.668
19	progesterone	1.0250	10.4405	−6.0984	13.4649	2.1834	7.380	7.114	7.064
20	17-hydroxyprogesterone	1.0087	10.0004	−10.0190	17.7649	1.8204	7.740	7.516	7.493
21	testosterone	0.3516	9.4378	−5.1359	13.9623	2.5628	6.724	6.819	6.833
22	prednisolone	0.3373	6.6813	−9.9105	21.3338	5.3761	7.512	7.890	7.975
23	cortisol 21-acetate	0.5566	7.5711	−10.0895	21.8062	−5.1063	7.553	7.482	7.434
24	4-pregnene-3,11,20-trione	0.9634	7.3901	−8.2011	17.7418	4.0670	6.779	7.397	7.472
25	epicorticosterone	0.3485	7.6064	−6.6275	17.3463	3.4820	7.200	7.335	7.347
26	19-nortestosterone	0	4.3222	−2.0131	14.0158	3.0355	6.114	5.791	5.730
27	16a,17-dihydroxy-4-pregnene-13,20-dione	1.0031	8.8908	−10.6614	18.1206	−3.6344	6.247	6.372	6.404
28	17-methyl-4-pregnene-3,20-dione	3.4995	13.4937	−14.5044	13.4703	2.5057	7.120	7.043	6.722
29	19-norprogesterone	0.5394	5.4749	−2.7542	13.5168	2.6539	6.817	6.123	5.963
30	11b,17,21-trihydroxy-2a-methyl-4-pregnene-3,20-dione	1.1923	9.6885	−10.1378	18.9197	1.9913	7.688	8.135	8.214
31	11b,17,21-trihydroxy-2a-methyl-9a-fluoro-4-pregnene-3,20-dione	1.1690	9.1005	−13.4834	17.9188	−1.6154	5.797	5.524	5.422

B. 58 Peptides

no.	compd	x_5	x_{21}	x_{26}	x_{32}	x_{51}	pIC_{50}	$M2_{EST}$	$M2_{LOO}$
1	VW	2.2227	5.8186	5.9585	−19.1518	4.4844	5.80	5.66	5.61
2	IW	1.8040	6.4424	5.4654	−18.3091	4.6177	5.70	5.44	5.37
3	IY	1.7825	13.8150	1.9601	−21.1003	4.6686	5.43	4.51	4.35
4	AW	1.8425	5.4788	2.3472	−15.3665	23.5714	5.00	5.00	4.98
5	RW	0	8.1754	4.9664	−18.4966	6.7327	4.80	4.98	5.04
6	VY	2.1949	13.1357	2.5191	−21.9337	4.5338	4.66	4.76	4.78
7	GW	0	5.2231	4.5742	−16.2350	3.7329	4.52	4.46	4.45
8	VF	2.1940	7.0916	2.4575	−21.4447	4.3875	4.28	4.09	4.08
9	AY	2.1444	12.9315	1.7766	−22.2568	4.1377	4.06	4.35	4.41
10	IP	1.8698	6.6937	0.0908	−8.0810	4.7584	3.89	3.62	3.59
11	RP	0	8.5793	−0.1578	−8.3692	6.8737	3.74	3.27	3.20
12	AF	2.1435	6.9298	1.7293	−21.7812	4.0009	3.72	3.70	3.70
13	GY	0	12.4012	0.9756	−18.8683	3.7730	3.68	3.49	3.46
14	AP	2.3034	5.8776	−0.1607	−9.8053	4.2265	3.64	3.43	3.40
15	RF	0	9.4400	1.4626	−20.8333	6.5947	3.64	3.41	3.40
16	VP	2.3108	6.0003	0.4340	−9.0121	4.6232	3.38	3.77	3.83
17	GP	0	5.1804	0.7274	−12.3094	3.8026	3.35	2.94	2.92
18	GF	0	6.4362	0.9456	−18.4982	3.6455	3.20	2.84	2.83
19	IF	1.7818	7.7246	1.9040	−20.6389	4.5192	3.03	3.85	3.89
20	VG	2.1262	0.1219	1.2880	−17.8136	4.0404	2.96	2.97	2.97
21	IG	1.7294	0.6889	0.8060	−17.0572	4.1660	2.92	2.75	2.74
22	GI	0.4183	0.4265	1.4668	−17.8091	3.6242	2.92	2.58	2.56
23	GM	0.1472	3.0088	0.8468	−16.1593	3.6224	2.85	2.59	2.58
24	GA	0.3120	−0.3163	1.5096	−19.4148	3.4262	2.70	2.39	2.38
25	YG	0	11.1764	0.2504	−17.8335	4.4228	2.70	3.12	3.17
26	GL	0.3556	1.3524	1.1943	−17.7563	3.6059	2.60	2.54	2.54
27	AG	2.0626	0.1138	0.6769	−18.2614	3.6723	2.60	2.64	2.64
28	GH	0	3.2741	0.9934	−17.8054	3.6522	2.51	2.56	2.56
29	GR	0	2.7269	1.0663	−17.8530	6.1149	2.49	2.65	2.65
30	KG	0	2.9809	0.3337	−16.3438	5.0399	2.49	2.37	2.37
31	FG	0	5.0809	0.2333	−17.6012	4.1402	2.43	2.45	2.45
32	GS	0	0.1924	1.2152	−20.0050	3.8114	2.42	2.21	2.20
33	GV	0.4720	−0.3210	1.9982	−18.8924	3.5670	2.34	2.69	2.71
34	MG	0.4289	2.5477	0.3013	−15.9885	4.1424	2.32	2.41	2.41
35	GK	0	3.5757	0.9073	−16.6619	4.7884	2.27	2.67	2.68

Table 4. (Continued)

no.	compd	x_5	x_{21}	x_{26}	x_{32}	x_{51}	pIC_{50}	$M2_{EST}$	$M2_{LOO}$
36	GE	0	2.3413	1.4579	−32.2595	4.0718	2.27	1.92	1.88
37	GT	0.2254	−0.3875	2.8940	−27.4858	3.8930	2.24	2.58	2.60
38	WG	0	4.6690	1.3610	−16.1598	4.3858	2.23	3.00	3.03
39	HG	0	2.1287	0.1087	−16.8569	4.1530	2.20	2.11	2.11
40	GQ	0	0.8268	1.5039	−28.2946	10.0991	2.15	2.27	2.28
41	GG	0	−0.1261	0.0560	−15.4530	3.3355	2.14	1.88	1.86
42	QG	0	2.2573	0.6732	−25.0969	11.0031	2.13	2.26	2.28
43	SG	0	0.7700	0.5286	−19.0114	5.2604	2.07	2.09	2.09
44	LG	1.2345	1.3217	0.5574	−17.0800	4.1064	2.06	2.56	2.59
45	GD	0	−0.0837	2.0917	−36.4385	4.1663	2.04	1.72	1.67
46	TG	1.0050	0.0613	2.0817	−25.4053	5.5250	2.00	2.67	2.70
47	EG	0	2.5488	0.7303	−30.6833	5.4580	2.00	1.76	1.74
48	DG	0	0.4160	1.2153	−33.9767	6.0735	1.85	1.60	1.57
49	PG	0	3.4884	0.0706	−13.2920	0	1.77	2.24	2.29
50	LA	1.6113	1.1810	2.1536	−21.3640	4.2094	3.51	3.15	3.13
51	KA	0.4793	2.8293	1.8601	−20.5861	5.1639	3.42	2.96	2.95
52	RA	0.6076	2.2466	2.0626	−21.6131	6.2388	3.34	3.02	3.01
53	YA	0.3658	11.3048	1.9334	−22.4444	4.5279	3.34	3.75	3.79
54	AA	2.4200	0	2.3406	−22.4578	3.7686	3.21	3.25	3.26
55	FR	0	8.5003	1.3803	−20.5870	7.1046	3.04	3.31	3.33
56	HL	0.4169	4.4831	1.5692	−19.8334	4.4810	2.49	3.00	3.01
57	DA	0.3432	0.4939	3.1096	−39.0558	6.1964	2.42	2.29	2.27
58	EA	0.3555	2.5422	2.4902	−35.5239	5.5730	2.00	2.40	2.39

C. 16 COX-2 Inhibitors

no.	-R	x_2	x_{10}	x_{19}	x_{26}	x_{29}	x_{57}	pIC_{50}	$M3_{EST}$	$M3_{LOO}$
1	-OH	4.1826	4.1166	−6.2951	5.3851	0	0	0.125	0.038	−0.031
2	$-NHCH_3$	4.9925	4.2349	−6.3721	6.3054	−0.0123	−0.0985	0.155	0.216	0.250
3	$-OCH_3$	4.9057	7.7042	−6.3634	6.0652	0	0	0.602	0.667	0.731
4	$-NHCH_2CH_2OH$	4.4370	4.1987	−6.5066	6.4878	0.1064	−0.1542	0.602	0.554	0.523
5	$-NHC_6H_5(4\text{-}NHCOCH_3)$	6.2328	4.1730	−7.1102	5.7533	0.0999	−0.1336	0.921	0.888	0.851
6	$-OC_6H_5(4\text{-}OCH_3)$	7.1994	8.5098	−7.0778	5.8342	0	0	1.398	1.344	1.190
7	$-OC_6H_5(4\text{-}SCH_3)$	8.0573	1.4592	−7.1051	6.0357	0	0	0.523	0.637	0.733
8	$-OC_6H_5(2\text{-}SCH_3)$	7.4350	1.2830	−7.0257	7.2211	0	0	1.222	1.104	0.894
9	$-OC_6H_5(4\text{-F})$	4.9338	4.4011	−7.0309	5.6615	0	0	1.125	1.157	1.177
10	$-O(3\text{-}C_5H_4N)$	4.8424	4.3990	−6.9434	5.7079	0.4695	−0.1481	1.301	1.113	1.066
11	$-NHC_6H_5(4\text{-}SCH_3)$	8.1301	1.0814	−7.1198	6.3410	0.5551	−0.1146	0.921	0.924	0.926
12	$-NHC_6H_5(4\text{-F})$	4.9730	4.1607	−7.0458	5.9226	0.3414	−0.1087	1.222	1.296	1.313
13	$-NH(3\text{-}C_5H_4N)$	4.8808	4.1610	−6.9581	5.9729	0.9515	−0.2560	1.301	1.360	1.461
14	$-NH_2$	4.2211	4.1332	−6.3066	5.6929	0	0	0.155	0.181	0.194
15	$-NHCH_2CH_2C_6H_5$	5.0155	4.2155	−6.9987	6.5060	0.2278	−0.1574	1.222	1.287	1.317
16	$-OCH_2CH_2C_6H_5$	4.9626	4.5139	−6.9734	6.2527	0	0	1.301	1.329	1.363

Table 5. Correlation Coefficients between Pairs of Descriptors for 31 Steroids, 58 Dipeptides, and 16 COX-2 Inhibitors

A. 31 Steroids

r	$\mathbf{x}_1$	x_2	x_4	x_{21}	x_{32}
x_1	1.0000				
x_2	0.7097	1.0000			
x_4	−0.6842	−0.5702	1.0000		
x_{21}	0.0009	−0.2116	−0.5450	1.0000	
x_{32}	−0.2271	−0.2446	0.1457	0.0299	1.0000

B. 58 Dipeptides

r	$\mathbf{x}_5$	x_{21}	x_{26}	x_{32}	x_{51}
x_5	1.0000				
x_{21}	0.2080	1.0000			
x_{26}	0.2355	0.1103	1.0000		
x_{32}	0.1905	0.2465	−0.3227	1.0000	
x_{51}	0.0451	0.0574	0.1130	−0.0810	1.0000

C. 16 COX-2 Inhibitors

r	x_2	x_{10}	x_{19}	x_{26}	x_{29}	x_{57}
x_2	1.0000					
x_{10}	−0.3992	1.0000				
x_{19}	−0.6279	0.2778	1.0000			
x_{26}	0.3704	−0.3959	−0.1857	1.0000		
x_{29}	−0.0003	−0.1988	−0.3356	−0.0311	1.0000	
x_{57}	0.1822	0.1270	0.1944	−0.0883	−0.8061	1.0000

prediction ability of $q = 0.9280$ and $RMSEP = 0.370$, between the pIC_{50} inhibiting angiotensin converting enzyme (ACE) and the MEDV-13 descriptors (x_v) of 58 dipeptides, is developed using MLR technique. The M2 model is shown in eq 4.

$$pIC_{50} = (2.5322 \pm 0.1996) + (0.2843 \pm 0.0574)\cdot x_5 + (0.1083 \pm 0.0129)\cdot x_{21} + (0.4491 \pm 0.0414)\cdot x_{26} + (0.05318 \pm 0.00846)\cdot x_{32} + (0.04709 \pm 0.01634)\cdot x_{51} \quad (4)$$

$n = 58$, $m = 5$, $r^2 = 0.8832$, $r = 0.9398$, $RMSEE = 0.339$, $F = 78.671$ (estimation)

$n = 58$, $m = 5$, $q^2 = 0.8611$, $q = 0.9280$, $RMSEP = 0.370$ (LOO prediction)

For 58 dipeptides, the pIC_{50} estimated by the M2 model ($M2_{EST}$) and the pIC_{50} predicted by LOO models ($M2_{LOO}$) are listed in Table 4B together with the pIC_{50} observed. And Figure 3B shows the plot of the pIC_{50} estimated and predicted by LOO procedure versus the pIC_{50} observed.

From Table 3C and Figure 2C, the best subset is a combination of six variables of nos. 2, 10, 19, 26, 29, and 57 (v) for 16 COX-2 inhibitors (ImAE). The six variables are related to six atomic types of nos. 1, 2, 3, 6, 7, and 10

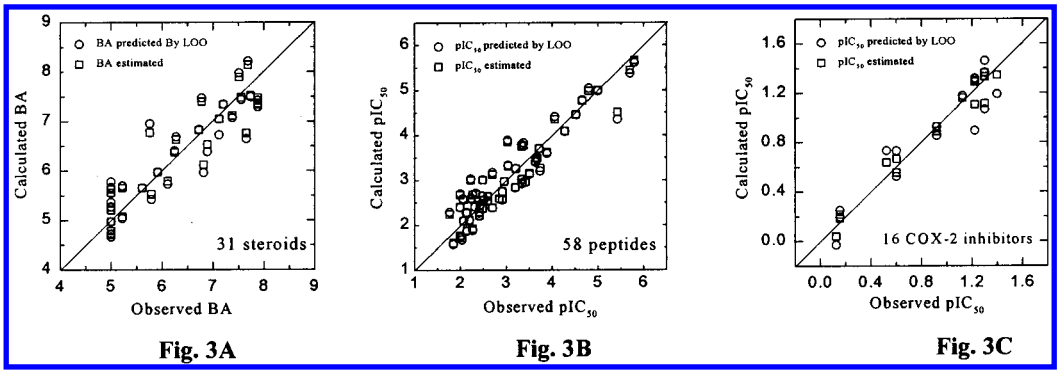

Figure 3. A. Plot of *BA* estimated by M1 and *BA* predicted by LOO versus *BA* observed. B. Plot of pIC_{50} estimated by M1 and *BA* predicted by LOO versus pIC_{50} observed. C. Plot of pIC_{50} estimated by M1 and *BA* predicted by LOO versus pIC_{50} observed.

Table 6. Comparison of MEDV-GA-MLR and PCR Methods

	31 steroids				58 peptides				16 COX-2 inhibitors			
method	*r*	RMSEE	*q*	RMSEP	*r*	RMSEE	*q*	RMSEP	*r*	RMSEE	*q*	RMSEP
MEDV-GA-MLR	0.9182	0.422	0.8818	0.504	0.9398	0.34	0.9280	0.37	0.9839	0.079	0.9413	0.151
PCR	0.9017	0.461	0.8737	0.519	0.9462	0.32	0.8847	0.47	0.9245	0.168	0.8417	0.239

(*k* or *l*) and characterize the effects of six substructures, CH_3-, =CH-, >CH=, −NH-, >N-, and −O-, on the biological activities inhibiting COX-2.

A 6-variable QSAR model (M3) with a high estimation ability of $r = 0.9839$ and $RMSEE = 0.079$ and an excellent prediction ability of $q = 0.9413$ and $RMSEP = 0.151$, between the pIC_{50} inhibiting COX-2 and the MEDV-13 descriptors ($\boldsymbol{x}_v$) of 16 COX-2 inhibitors, is developed using MLR technique. The M3 model is shown in eq 5.

$$pIC_{50} = -(11.3428 \pm 0.8575) - (0.1597 \pm 0.3180)\cdot\boldsymbol{x}_2 + (0.09870 \pm 0.01643)\cdot\boldsymbol{x}_{10} - (1.4866 \pm 0.1200)\cdot\boldsymbol{x}_{19} + (0.4242 \pm 0.0740)\cdot\boldsymbol{x}_{26} + (0.7508 \pm 0.1845)\cdot\boldsymbol{x}_{29} + (2.0331 \pm 0.6267)\cdot\boldsymbol{x}_{57} \quad (5)$$

$$n = 16,\ m = 6,\ r^2 = 0.9680,\ r = 0.9839,\ RMSEE = 0.079,\ F = 45.329 \text{ (estimation)}$$

$$n = 16,\ m = 6,\ q^2 = 0.8861,\ q = 0.9413,\ RMSEP = 0.151 \text{ (LOO prediction)}$$

For 16 COX-2 inhibitors, the pIC_{50} estimated by the M3 model ($M3_{EST}$) and the pIC_{50} predicted by LOO models ($M3_{LOO}$) are listed in Table 4C together with the pIC_{50} observed. And Figure 3C shows the plot of the pIC_{50} estimated and predicted by LOO procedure versus the pIC_{50} observed.

Correlation Analysis between Descriptors. In our GA process, if the maximum value among all correlation coefficients (*r*) between pairs of descriptors is higher than 0.95, then the fitness value of corresponding individual equals zero. So, in the best combination of descriptors, it is impossible to exist as significant correlations between descriptors. In fact, the maximum is $r(\boldsymbol{x}_1, \boldsymbol{x}_2) = 0.7097$, while the minimum is merely $r(\boldsymbol{x}_1, \boldsymbol{x}_4) = 0.0009$ for 31 steroids. For the 58 peptide set, all correlation coefficients (*r*) between pairs of descriptors are lower than 0.25. For the COX-2 inhibitors, all correlation coefficients are lower than 0.4 but $r(\boldsymbol{x}_2, \boldsymbol{x}_{19}) = 0.6279$ and $r(\boldsymbol{x}_{29}, \boldsymbol{x}_{57}) = 0.8061$. All relation results are in detail listed in Table 5A−C.

Comparison of MEDV-GA-MLR with PCR Method. The results obtained using the principal component regression (PCR) technique in the literature[12,13] are listed in Table 6 together with the results based on the MEDV-GA-MLR method developed in this paper.

Table 6 shows that the MEDV-GA-MLR method has higher predictive ability for the external samples than PCR method for all three data sets. On the other hand, the best number of variables (*Vn*) optimized by GA method equals the number of the principal components using PCR technique for two sets of 31 steroids and 16 COX-2 inhibitors. For 58 dipeptides, the best *Vn* is different from the number of the principal components (PCs) due to the existence of several minimum points in plot of *RMSEP* vs the number of PCs.

CONCLUSION

A MEDV-GA-MLR method combining the optimal selection of the MEDV-13 descriptors finished using the GA program with the modeling method using MLR based on LOO predictions is developed to estimate and predict the biological activities based on the QSAR model related the molecular structures to their known biological activities. Studies on three molecular systems of 31 steroids, 58 dipeptides, and 16 COX-2 inhibitors show that the MEDV-GA-MLR method has higher estimation and prediction abilities than PCR method. The meaning of the optimal variables is also more explicit than the principal components. It can be foreseen that the method will become one of the general QSAR research tools based MEDV descriptors. The related studies are in progress.

ACKNOWLEDGMENT

We are especially grateful to the China Postdoctoral Science Foundation and the National High Technology Project of China (No. 2001AA646010-4) for their financial supports.

REFERENCES AND NOTES

(1) Katritzky, A. R.; Maran, U.; Lobanov, V. S.; Karelson, M. Perspective: Structurally diverse quantitative structure−property relationship correlations of technologically relevant physical properties. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1−18.

(2) Cramer, R. D.; Patterson, D. E.; Bunce, J. D. Comparative molecular field analysis (CoMFA). 1. Effect of shape on binding of steroids to carrier proteins. *J. Am. Chem. Soc*. **1988**, *110*, 5959−5967.

(3) Goodford, P. J. A computational precedure for determining energetically favorable binding sites on biologically important molecules. *J. Med. Chem*. **1985**, *28*, 849−857.

(4) Jian, A. N.; Koile, K.; Chapman, D. Compass: predicting biological activities from molecular surface properties. Performance comparisons on a steroid benchmark. *J. Med. Chem*. **1994**, *37*, 2315−2327.

(5) Robinson, D. D.; Winn, P. J.; Lyne, P. D.; Richards, W. G. Self-Organizing Molecular Field Analysis: A Tool for Structure−Activity Studies. *J. Med. Chem*. **1999**, *42*, 573−583.

(6) Tong, W.; Lowis, D. R.; Perkins, R.; Chen, Y.; Welsh, W. J.; Goddette, D. W.; Heritage, T. W.; Sheehan, D. M. Evaluation of quantitative structure−activity relationship methods for large-scale prediction of chemicals binding to the estrogen receptor. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 669−677.

(7) Kier, L. B.; Hall, L. H. An electrotopological state index for atoms in molecules. *Pharm. Res*. **1990**, *7*, 801−807.

(8) Kier, L. B.; Hall, L. H. An index of electrotopological state for atoms in molecules. *J. Math. Chem*. **1991**, *7*, 229−241.

(9) de Gregorio, C.; Kier, L. B.; Hall, L. H. QSAR modeling with the electrotopological state indices: Corticosteroids. *J. Comput. Aid. Mol. Des.* **1998**, *12* (6), 557−561.

(10) Buolamwini, J. K.; Raghavan, K.; Fesen, M. R. et al. Application of the electrotopological state index to QSAR analysis of flavone derivatives as HIV-1 integrase inhibitors. *Pharm. Res.* **1996**, *13*(12), 1892−1895.

(11) Hall, L. H.; Kier, L. B. 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.

(12) Liu, S. S.; Yin, C. S.; Li, Z. L.; Cai, S. X. QSAR study of steoroid benchmark and dipeptides based on MEDV-13. *J. Chem. Inf. Comput. Sci*. **2001**, *41*, 321−329.

(13) Liu, S. S.; Yin, C. S.; Shi, Y. Y.; Cai, S. X.; Li, Z. L. MEDV-13 for QSAR studies on the COX-2 inhibition by indomethacin amides and esters. *Chin. J. Chem*. **2001**, *19*, 751−756.

(14) Waller, C. L.; Bradley, M. P. Development and validation of a novel variable selection technique with application to multidimensional quantitative structure−activity relationship studies. *J. Chem. Inf. Comput. Sci*. **1999**, *39*, 345−355.

(15) Wikel, J.; Dow, E. The use of neural networks for variable selection in QSAR. *Bioorg. Med. Chem. Lett*. **1993**, *3*, 645−651.

(16) Kubinyi, H. Variable selection in QSAR studies, 1. An evolutionary algorithm. *Quant. Struct. -Act. Relat*. **1994**, *13*, 285−294.

(17) Kubinyi, H. Variable selection in QSAR studies, 2. A highly efficient combination of systematic search and evolution. *Quant. Struct. -Act. Relat.* **1994**, *13*, 393−401.

(18) McFarland, J. W.; Gans, D. J. On identifying likely determinants of biological activity in high dimensional QSAR problems. *Quant. Struct. -Act. Relat*. **1994**, *13*, 11−17.

(19) Leardi, R.; Gonzalez, A. L. Genetic algorithms applied to feature selection in PLS regression: how and when to use them. *Chemom. Intell. Lab. Syst*. **1998**, *41*, 195−207.

(20) Hasegawa, K.; Funatsu, K. GA strategy for variable selection in QSAR studies: GAPLS and D-optimal designs for predictive QSAR model. *J. Mol. Struct. THEOCHEM* **1998**, *425*, 255−262.

(21) Barros, A. S.; Rutledge, D. N. Genetic algorithm applied to the selection of principal components. *Chemom. Intell. Lab. Syst*. **1998**, *40*, 65−81.

(22) Geladi, P.; Kowalski, B. R. Partial least-sqaures regression: a tutorial. *Anal. Chim. Acta* **1986**, *185*, 1−17.

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

(24) Baroni, M.; Clementi, S.; Cruciani, G.; Costantino, G.; Riganelli, P. Predictive ability of regression models part 2: selection of the best predictive PLS model. *J. Chemometr*. **1992**, *6*, 347−356.

(25) Cruciani, G.; Watson, K. A. Comparative molecular field analysis using GRID force field and GOLPE variable selection methods in a study of inhibitors of glycogen. *J. Med. Chem.* **1994**, *37*, 2589−2601.

(26) Holland, J. H. *Adaptation in Natural and Artificial Systems*; University of Michigan, Ann Arbor, MIT, 1975.

(27) Goldberg, D. E. *Genetic algorithms in search, optimization and machine learning*; Addison-Wesley: New York, 1989.

(28) Rogers, D.; Hopfinger, A. J. Application of genetic function approximation to quantitative structure- activity relationships and quantitative structure−property relationships. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 854−866.

(29) Leardi, R.; Boggia, R.; Terrile, M. Genetic algorithms as a strategy for feature selection. *J. Chemom*. **1992**, *6*, 267−281.

(30) Zaliani, A.; Gancia, E. MS-WHIM scores for amino acids: a new 3D-descriptor for peptide QSAR and QSPR studies. *J. Chem. Inf. Comput. Sci*. **1999**, *39*, 525−533.

(31) 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.

CI010245A