QSAR Study of Steroid Benchmark and Dipeptides Based on MEDV-13

Shu-Shen Liu,*[†,‡,§] Chun-Sheng Yin,[#] Zhi-Liang Li,[†] and Shao-Xi Cai[†]

College of Bioengineering, Chongqing University, Chongqing 400044, P. R. China, Laboratory of Structural
Biology and Department of Applied Chemistry, University of Science and Technology of China,
Hefei 230026, P. R. China, and Department of Applied Chemistry, Guilin Institute of Technology,
Guilin 541004, Guangxi Province, P. R. China

Received July 23, 2000

A molecular electronegativity distance vector based on 13 atomic types, called MEDV-13, is a descriptor for predicting the biological activities of molecules based on the quantitative structure−activity relations (QSAR). The MEDV-13 uses a modified electrotopological state (E-state) index to substitute for the relative eletronegativity (q) of non-hydrogen atoms in the molecule of interest in the MEDV and a topological distance for the relative distance (d) in the MEDV. For an organic molecule containing several chemical elements such as C, H, O, N, S, F, Cl, Br, I, and P, the MEDV-13 includes at best 91 descriptors. Then it is essential to employ a principal component regression (PCR) technique to derive a QSAR model relating the biological activities to the MEDV-13. The MEDV-13 is used to study the QSAR of the corticosteroid-binding globulin (CBG) binding affinity of the steroids and the activity inhibiting angiotensin-converting enzyme (ACE) of dipeptides, and resulting models have a comparable quality to the current three-dimensional (3D) methods such as CoMFA though the MEDV-13 is a descriptor based on two-dimensional topological information.

INTRODUCTION

Quantitative structure−activity relationship (QSAR) techniques have become indispensable in all aspects of research into the molecular interpretation of biological properties.[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. The ability to produce quantitative correlation between three-dimensional properties of the molecules and the biological activity of these compounds is of inestimable value in deciding upon the choice of future synthetic chemistry.[2]

Eminent among the techniques used is comparative molecular field analysis (CoMFA) introduced by Cramer[3] and used by several other authors.[4−6] As in the related GRID technique of Goodford,[7] molecules are sited in a 3D grid. At each grid point an interaction energy related to shape or electrostatic potential is calculated. Partial least-squares analysis (PLS)[8,9] is then used to extract the relationship between the interaction energies and the biological activity.

Molecular similarity was introduced as a concept by Carbo.[10] The use of similarity as a 3D-QSAR tool was introduced by Good et al.[11] and used by several other groups.[12,13] Similarity between pairs of molecules can be defined in terms of shape or of electrostatic potential, but instead of a large number of values at grid points of CoMFA it has a single numerical measure of overall similarity. A set of molecules may be compared to a single reference molecule to yield a predictive QSAR.[13,14] More information can be obtained from a matrix of the similarity indices between all pairs of molecules in a training set using PLS to derive a QSAR.[11,15,16]

COMPASS, introduced by Jian,[17] is another 3D method coding only the surface properties of the molecule. The COMPASS operate in three phases. The first phase constructs a set of initial guesses as to the bioactive conformation (called visually a pose) and alignment of each molecule; the second phase simultaneously chooses a bioactive pose for each molecule, starting from these guesses, and constructs a statistical model which explains quantitatively and predicts the relationship between the surface characteristics of the given molecules and their biological activity; the third phase predicts the activity and active pose of a new molecule and can also graphically display the basis of the prediction in a way that aids molecular design.

SOMFA, a self-organizing molecular field analysis introduced by Robinson,[2] is a novel technique for three-dimensional QSAR. It is simple and intuitive in concept and avoids the complex statistical tools and variable selection procedures favored by other methods. Like CoMFA, a grid-based approach is used. But no probe interaction energies need to be evaluated. Like the similarity methods it is the intrinsic molecular properties, such as shape, that are used to develop the QSAR models. Their results show the SOMFA to be as predictive as the best 3D-QSAR methods available.

However, implementing these methods based on 3D structure are in general difficult and time-consuming because

* Corresponding author phone: (86)-773-5896620; fax: (86)-773-5812796; e-mail: ssliu@glite.edu.cn.

[†] Chongqing University.

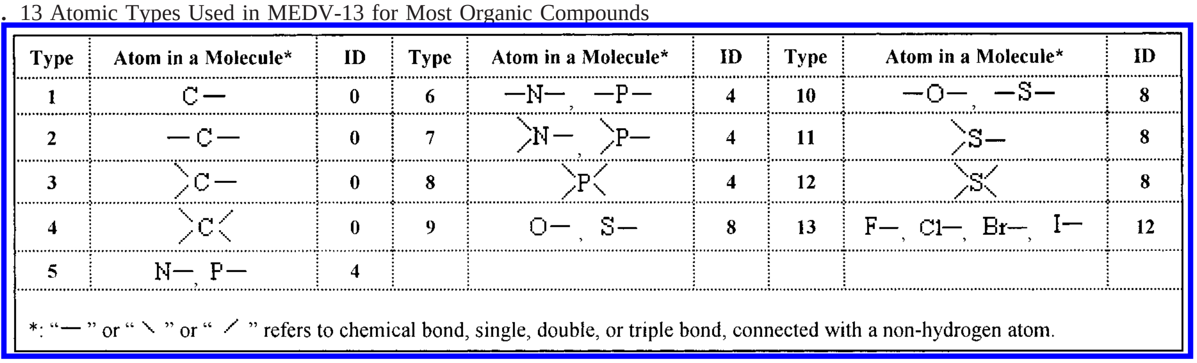
[‡] Laboratory of Structural Biology, University of Science and Technology of China.

[§] Guilin Institute of Technology.

[#] Department of Applied Chemistry, University of Science and Technology of China.

10.1021/ci0003350 CCC: $20.00
Published on Web 02/21/2001

Table 1. 13 Atomic Types Used in MEDV-13 for Most Organic Compounds



Type	Atom in a Molecule*	ID	Type	Atom in a Molecule*	ID	Type	Atom in a Molecule*	ID
1	C—	0	6	—N—, —P—	4	10	—O—, —S—	8
2	—C—	0	7	>N—, >P—	4	11	>S—	8
3	>C—	0	8	>P<	4	12	>S<	8
4	>C<	0	9	O—, S—	8	13	F—, Cl—, Br—, I—	12
5	N—, P—	4						

*: "—" or "\" or "/" refers to chemical bond, single, double, or triple bond, connected with a non-hydrogen atom.

of difficulty of 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 methods.

Tong[18] recently employed a relative simple approach, called hologram QSAR (HQSAR), only encoding 2D structure information, to develop potential QSAR model by using partial least-squares regression (PLS). In HQSAR, each molecule in the data set is divided into structural fragments that are then counted in the bins of a fixed length array to form a molecular hologram. The bin occupancies of the molecular hologram are structural descriptors encoding compositional and topological molecular information. And QSAR models generated based on the HQSAR technique have comparable quality to those of CoMFA, one of most excellent 3D QSAR techniques so far. However, it is still difficult to generate all linear, branched, and overlapping substructural fragments in the size range of fixed length, especially large length fragments. On the other hand, it appears to be not entirely reasonable to execute a hashing process.

Another excellent 2D descriptor, an atom level electrotopological state (E-state) index, was introduced by Kier and Hall[19,20] and used successfully for a variety of QSAR studies.[21−27] Afterward, an atom type E-state index which is useful for database characterization, molecular similarity analysis, and QSAR was proposed.[28] The new atom type indices can be used in a manner similar to group additive schemes.

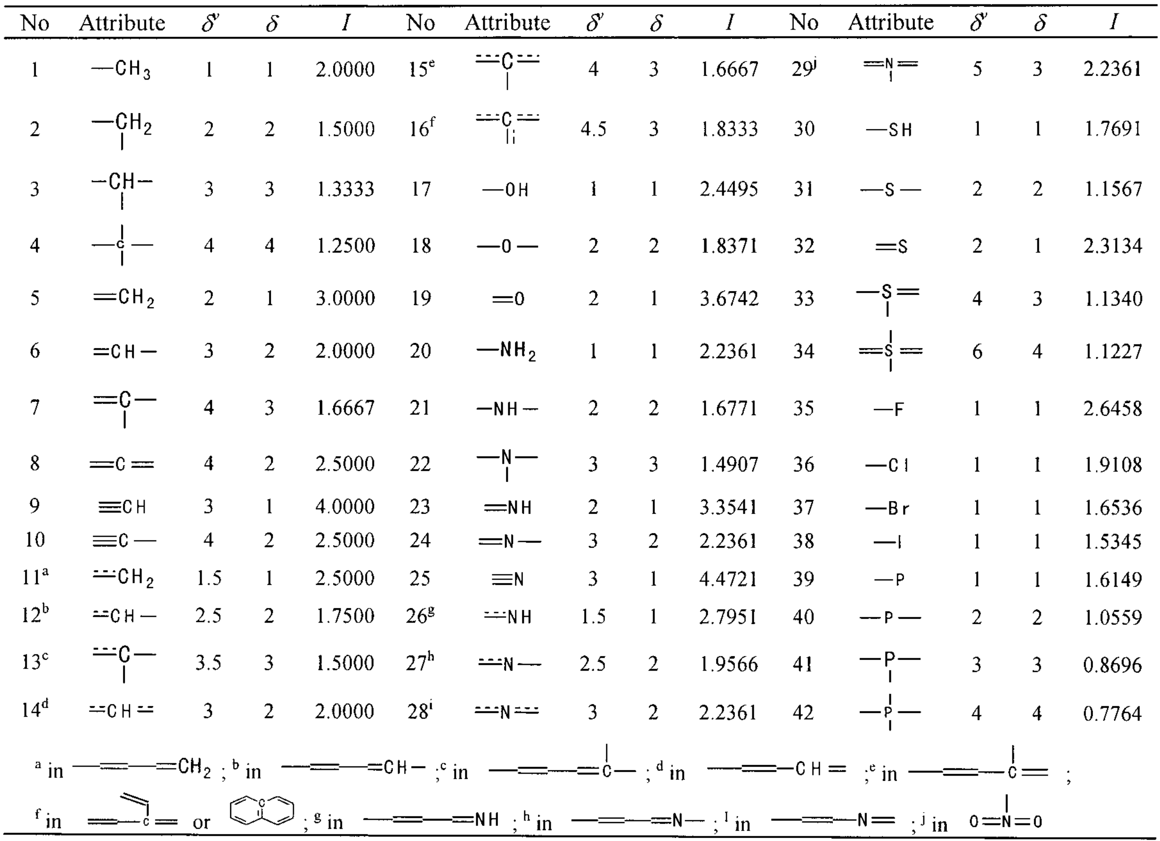
In our previous paper,[29,30] the molecular electronegativity distance vector (MEDV) was reported and used to study QSPR of organic compounds such as the boiling points of alcohols. Taking invariability of relative electronegativity of element in the different molecular enviorments and complexity of calculating relative bond length into account, the vector-type structural descriptor is modified by introducing 13 atomic types and the modified E-state index and 2D topological distance. The novel MEDV is called MEDV-13. In the MEDV-13, 13 atomic types are proposed to replace 4 atomic types in the MEDV[30] in order to enhance the discriminability of the MEDV-13 for various molecular structures. A modified E-state index is used to substitute for the relative electronegativity (q) in order to explain a case in which the q of an element atom in different chemical environments is variable. Because of the introduction of 13 atomic types and E-state index, direct topological distance can be used to replace for the relative bond length in order to simplify the MEDV-13 calculation. We now use the 13 atomic types and modified E-state index to construct a MEDV-13 vector containing 91 descriptors rather than 10 descriptors in the MEDV. The method is used to predict the corticosteroid-binding globulin (CBG) binding affinity of the "benchmark" steroids and the activity inhibiting angiotensin-converting enzyme (ACE) of dipeptides, and resulting QSAR models based on the MEDV-13 have a comparable quality to the current three-dimensional (3D) methods such as CoMFA.

METHOD

13 Atomic Types. In this paper, the term *atomic type* is used to represent indirectly the topological structural characteristics of the molecule of interest. For most organic molecules, non-hydrogen atoms are often carbon, nitrogen, phosphorus, oxygen, sulfur, halogen including fluorine, chlorine, bromine, and iodine. To distinguish the diverse effects of the non-hydrogen atoms in a molecule, the term "*atomic type*" of an atom is defined as the number of non-hydrogen atoms binding to that atom plus its identifying number (ID). That is, if an atom is linked to k ($k = 1,2,3,4$) non-hydrogen atom/atoms through chemical bond, then *atomic type* of the atom equals to k plus its ID. The number of non-hydrogen atoms bonded reflects the topological structural characteristics of a local environment of the atom. It has been known that the different atoms have a different number of principal quantum and electrons for the valence shell. And the ID is used to specify the number of valence electrons (υ) of the atom in the same local topological environment and ID = (υ-4)*4. That is, the element atoms in the same period but in a different group (υ) such as C, N, O, and F are identified by the ID. For example, the methene carbon located in butane molecule belongs to the second atomic type or type 2 because it is linked to two carbon atoms and ID is 0, while methyl carbon belongs to type 1. The atomic type of nitrogen atom in trimethylamine molecule equals 3 (linked three carbon atoms) plus 4 (ID of nitrogen), i.e., atomic type 7. Table 1 shows the IDs of carbon, nitrogen, phosphorus, oxygen, sulfur, and halogen atoms together with their atomic types. In general, any non-hydrogen atom in an organic molecule consisting of elements C, H, O, N, S, P and halogen can be falled into one of the 13 *atomic type*s.

42 Atom Attributes. As the same is Hall and Kier's scheme,[28] each atom in the molecule is identified by its valence state, including the number of attached hydrogen

Table 2. 42 Atomic Attributes and Intrinsic State Values for Organic Compounds



No	Attribute	δ^v	δ	I	No	Attribute	δ^v	δ	I	No	Attribute	δ^v	δ	I
1	—CH3	1	1	2.0000	15[e]	╌C╌ with — below	4	3	1.6667	29[j]	=N= with — below	5	3	2.2361
2	—CH2—	2	2	1.5000	16[f]	╌C╌ with ╌ below	4.5	3	1.8333	30	—SH	1	1	1.7691
3	—CH<	3	3	1.3333	17	—OH	1	1	2.4495	31	—S—	2	2	1.1567
4	>C<	4	4	1.2500	18	—O—	2	2	1.8371	32	=S	2	1	2.3134
5	=CH2	2	1	3.0000	19	=O	2	1	3.6742	33	—S(—)=	4	3	1.1340
6	=CH—	3	2	2.0000	20	—NH2	1	1	2.2361	34	=S(<)=	6	4	1.1227
7	=C<	4	3	1.6667	21	—NH—	2	2	1.6771	35	—F	1	1	2.6458
8	=C=	4	2	2.5000	22	—N<	3	3	1.4907	36	—Cl	1	1	1.9108
9	≡CH	3	1	4.0000	23	=NH	2	1	3.3541	37	—Br	1	1	1.6536
10	≡C—	4	2	2.5000	24	=N—	3	2	2.2361	38	—I	1	1	1.5345
11[a]	╌CH2	1.5	1	2.5000	25	≡N	3	1	4.4721	39	—P	1	1	1.6149
12[b]	╌CH—	2.5	2	1.7500	26[g]	╌NH	1.5	1	2.7951	40	—P—	2	2	1.0559
13[c]	╌C<	3.5	3	1.5000	27[h]	╌N—	2.5	2	1.9566	41	—P<	3	3	0.8696
14[d]	╌CH╌	3	2	2.0000	28[i]	╌N╌	3	2	2.2361	42	>P<	4	4	0.7764

[a] in —=—=CH2 ; [b] in —=—=CH— ; [c] in —=—=C(—)— ; [d] in —=—CH= ; [e] in —=—C(—)= ; [f] in =—C= or naphthalene ; [g] in —=—=NH ; [h] in —=—=N— ; [i] in —=—N= ; [j] in O=N(—)=O

atoms. In this paper, the atom type scheme introduced by Hall and Kier's[28] is used to describe the chemical environments of the atom in different molecules. The E-state index is also used in the MEDV-13. But there exist some different aspects from Hall and Kier. First of all, a concept *atomic attribute* is introduced to replace the atom type in Hall and Kier method because the concept *atomic type* has been used to express the number of non-hydrogens bonded to that atom plus the atom's identifying number (ID). Second, instead of the aromatic valence state, we use a conjugated system indicator (CSI) to distinguish various atoms in a conjugated system including an aromatic system because the atoms located in different positions of the conjugated system have different effects on the molecule. For example, nos. 11, 12, and 13 atomic attributes in Table 2 refer to the carbon atom located in both ends of a conjugated system, while attributes 14 and 15 refer to the carbon atom in the center of the conjugated system, respectively. Table 2 lists 42 *atomic attributes* of the organic compounds containing the elements such as carbon, nitrogen, oxygen, sulfur, and so on.

Electrotopological State (E-State) Index. The E-state index introduced by Hall and Kier,[28] which is modified a little in this paper, is used to substitute for the relative electronegativity in the MEDV for various atoms in the molecule under study. Imitating Hall and Kier's definition but introducing a coefficient, an intrinsic state of an atom in the molecule is defined as follows.

$$I = \sqrt{v/4}\cdot((2/n)^2\,\delta^v + 1)/\delta \qquad (1)$$

The symbol v is the number of valence electrons; n is the principal quantum number for the valence shell of that atom; and δ^v and δ are the molecular connectivity delta values which are given as follows

$$\delta = \sigma - h,\ \delta^v = \sigma + \pi - h \qquad (2)$$

where σ and π are respectively the number of electrons in σ and π orbitals and h is the number of hydrogen atoms bonded to the atom. Here no electrons in lone pairs enter into the δ^v because of no forming of a covalent bond, and the values of formula $((2/n)^2\ \delta^v + 1)/\delta$ will be equal for the element atoms in the same period. Then a coefficient, $\sqrt{v/4}$ where the 4 is the number of valence electrons of carbon atom, has to be introduced in the I to make the element atoms located in the same period, but a different main group such as C-, N-, O-, and F- having the same local chemical environment have different intrinsic state (I). And the ratio of I of an element atom to another one is close to the ratio of Pauling's electronegativity. Besides, the number of π-electrons of the atom in the conjugated π-electron system is multiplied by a factor of 0.5 because the π-electron not only belongs to an atom but also is shared by more atoms in these conjugated system. Thus the atoms having the same *atomic attribute*s but located in different molecules will have an invariable intrinsic state value (I_i). And the E-state index expressed by a graph distance rather than the graph distance plus 1 will act as a measure of relative electronegativity.

Then the E-state index for atom i, q_i, is defined as follows

$$q_i = I_i + \sum_{j\neq i}^{all\,j}(I_i - I_j)/d_{ij}^2 \qquad (3)$$

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

Table 3. Relation of the v to the k and l Values in the MEDV-13

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

Molecular Electronegativity Distance Vector (MEDV-13). With the relative electronegativity (q) expressed by modified E-state indices and the topological distances (d), molecular electronegativity distance vector based on 13 atomic types (MEDV-13) descriptors, h_{kl}, can be calculated from eq 4 being similar to the molecular electronegativity distance vector (MEDV) based on 4 atomic types in our previous report[30]

$$h_v = h_{kl} = \sum_{i\in k, j\in l} \frac{q_i q_j}{d_{ij}^2} (k,l = 1,2,3,\cdots,13; l \geq k; v = 1,2,3,\cdots,91) \quad (4)$$

where k or l is the *atomic type* of the atom i or j in the molecule and i or j is a coding number or series number and d_{ij} is the shortest graph distance of various pathways passed from the ith to the jth atom. Because there are in general 13 *atomic type*s in organic compounds, there are 91 elements in the MEDV-13 according to eq 4. Table 3 shows the relation of the subscript v in h_v to k and l in h_{kl} for clarify. The eq 4 is formally similar to (MEDV),[30] but here we use the E-state index substituting for the relative electronegativity and the graph distance for relative distance in the MEDV. On the other hand, the number of descriptors in the MEDV-13 is 91 rather than 10 in the MEDV. This is because the relative Pauling's electronegativity of an element atom is invariable in different molecules, and it then cannot describe the atoms located in different chemical environments in the same molecule or in different ones while the E-state index varies with the different chemical environments of the atoms. Besides, the graph distance replacing the relative distance makes the calculation more simple. The atomic types from 4 to 13 enable the description of the more complex molecules such as ones containing many heteroatoms.

Because the MEDV-13 have in general 91 descriptors, it is almost impossible to use multiple linear regression (MLR) to develop a QSAR model. Considering only one dependent variable, biological activity, in most QSAR models, a principal component regression (PCR) technique[31,32] developed in house is employed to derive a latent QSAR model. In the context of QSAR, the biological activity can be thought of as a function of the structural descriptors such as MEDV-13 of the compounds of interest.

RESULTS AND DISCUSSION

Data Set. Two data sets are selected in this paper. One is taken from a steroid binding affinity prediction problem previously studied by Cramer using CoMFA,[3] by Good employing a molecular similarity method,[11] by Jain using COMPASS,[17] by Bravi using MS−WHIM method,[33] and by Robinson using SOMFA.[2] The data set consists of 31 steroids (Figure 1) assayed for binding affinity to one transport protein, corticosteroid-binding globulin (CBG). Another is a series of 58 dipeptides inhibiting angiotensin-converting enzyme (ACE) selected from Zaliani's report.[35]

The structural descriptors of these steroid and dipeptide compounds are those MEDV-13 variables derived from eq 4. The MEDV-13 has in general 91 descriptors but in fact only 25 nonzero descriptors in the steroid benchmarks and 33 variables in the peptide molecules.

Selection of Variables. Because the number of variables having nonzero values in different molecules is different, it is essential to select the variables having a statistically significant difference in a QSAR model. A three-step procedure is performed to select the variables entering into the PCR analysis in our present paper. First, the variables having zero values for all samples (n) are left out from the data set consisting of n samples. Second, the variables having only a few nonzero values in n samples such as $n/20-n/15$ are also left out of the variable set. Finally, the correlation coefficients (R) between various pairs of variables are calculated, and one of two variables having $R > 0.95$ is deleted.

For the steroid system consisting of 31 samples, only 25 variables, nos. 3, 4, 9, 10, 13, 14, 15, 16, 21, 22, 25, 26, 27, 32, 33, 36, 37, 42, 43, 46, 77, 78, and 81 descriptors, are not all zero values. Only 2 samples have nonzero values for nos. 10, 22, 33, 43, and 78 descriptor variables, and 1 sample has nonzero value for nos. 13, 25, 36, 46, and 81 descriptors. So, the 10 variables should be also deleted from the data set. The various correlation coefficients (Rs) between the remaining 15 variables are less than 0.95, and the highest R of -0.9472 is the R between nos. 37 and 42 variables. Finally, only 15 variables enter into the PCR analysis.

For the peptide system consisting of 58 dipeptides, only 33 variables, 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 descriptors, are not all zero values. Only 5 samples have nonzero values for nos. 19, 30, 49, and 66 variables, 3 samples have nonzero values for the no. 7 descriptor, 2 samples have nonzero values for nos. 10, 22, 33, 52, 60, and 78 variables, and 1 sample has nonzero value for the no. 57 variable. Then the 12 variables are deleted from the data set. Various Rs between the remaining 21 variables are all less than 0.95 except for the R of -0.9521 between nos. 3 and 9 variables and the R of 0.9747 between nos. 47 and 48 variables. Leave out nos. 9 and 47 from the remaining data set. Then only 19 variables enter into the final PCR analysis.

Number of Principal Components (PCs). For a data set whose number of samples is smaller than the number of independent variables, it is impossible to model a QSAR relationship using the classical multiple linear regression method. A PCR program devised in house is employed to derive a QSAR model between the biological activity and

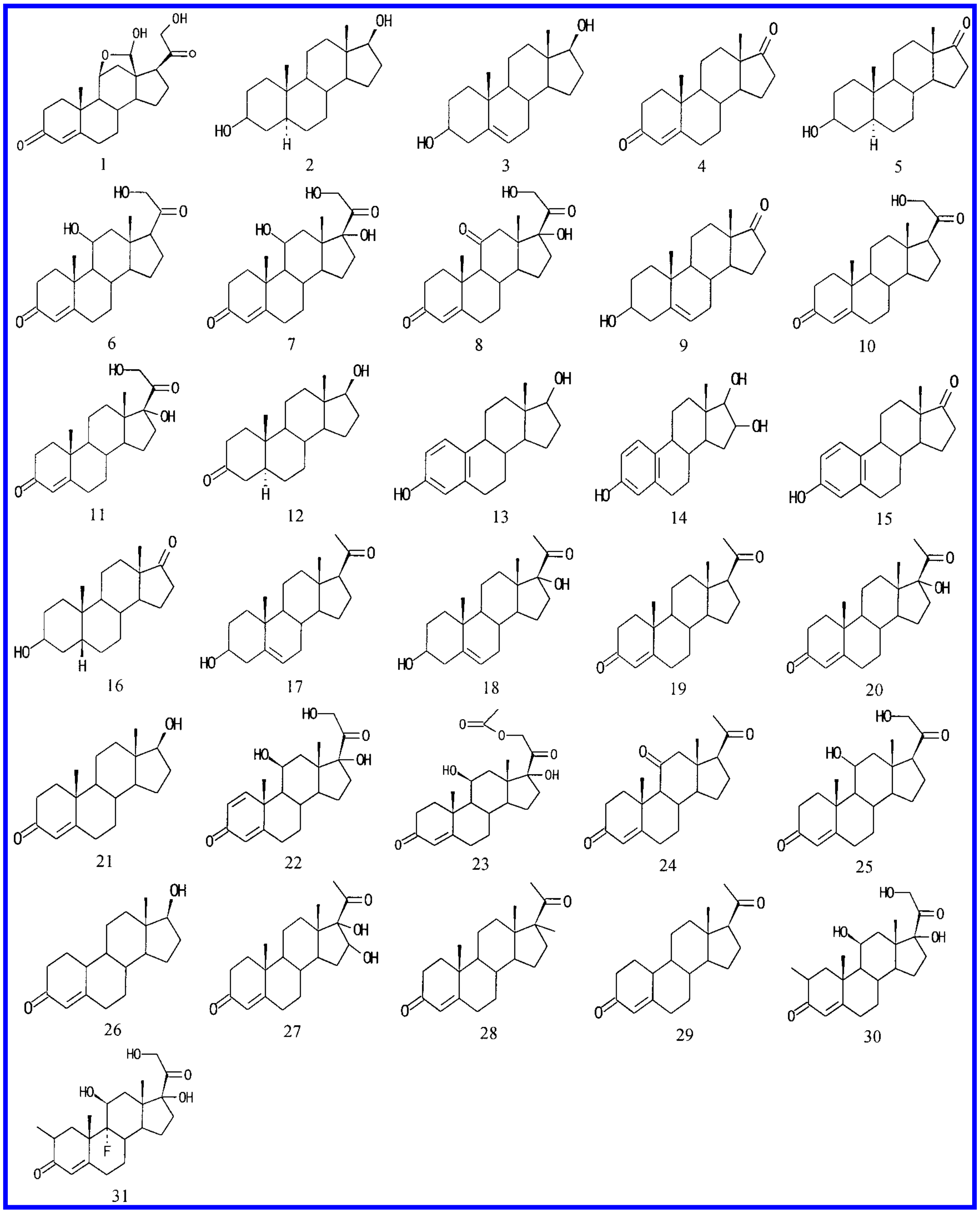


Figure 1. Structures of the 31 steroids in the benchmark data set.

the MEDV-13 descriptors. It has been known that a high quality model should have not only a good ability of estimation for the internal samples but also an excellent ability of prediction for the external samples, and the later is more important for a QSAR model.

It has been found that the number of principal components (PCs) plays an important role in PCR model analysis. More or less PCs reduce the quality of the QSAR model. Then it is essential to first determine the number of PCs in PCR analysis. In this work, it is found that the correlation coefficient (*R*) in the modeling stage is always increasing with the number of PCs, while the root-mean-square error (*RMS*) reduces with PCs (Figures 2 and 3), but the *R* and *RMS* in prediction step for cross-validation are irregularly changing and exist at one or more maximum points for the *R* and minimum points for the *RMS* in plots of *R* and *RMS* vs PCs. The above results show that RMS_{CV} or R_{CV} is a good criterion in determination of the number of principal components. And the lower the RMS_{CV} is or the higher the R_{CV} is, the better the quality of model is. From Figures 2 and 3, it is evident that the number of PCs of 5 is suitable for the steroid system. In the same way, the number of PCs in the peptide system is 19.

QSAR Studies on 31 Steroids. The PCR techniques are employed to model the QSAR relationship between the binding affinity (A) to CBG data and the MEDV-13 descriptors of 31 steroids. As stated above, only 15 variables enter into the PCR procedure. The first QSAR model (M1)

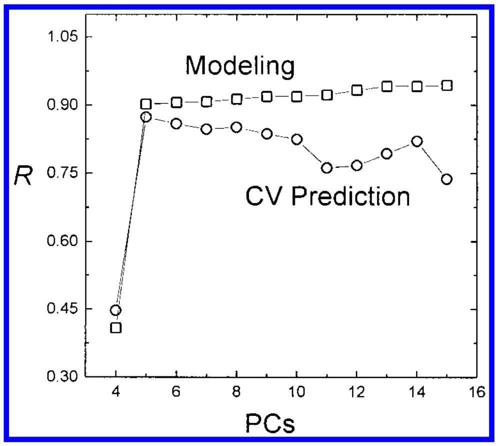


Figure 2. Plot of R_M or R_{CV} vs PCs.

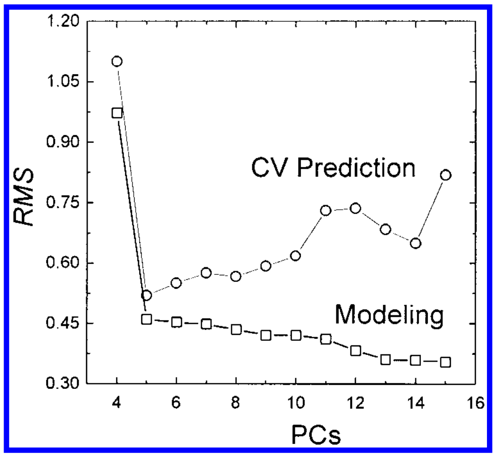


Figure 3. Plot of RMS_{CV} or RMS_M vs PCs.

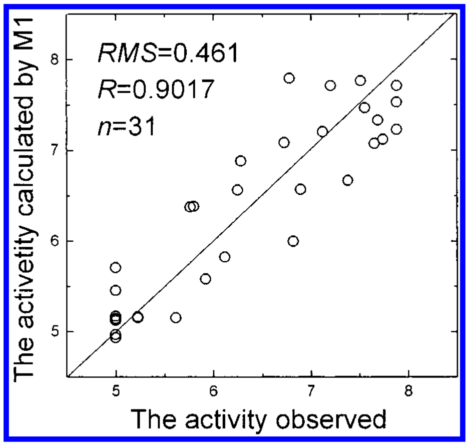


Figure 4. Plot of A_{M1} versus A_{OBS}.

is derived from a data set consisting of all 31 steroids. The *R* and *RMS* between the activities (A_{M1}) estimated by the M1 model with the number of PCs of 5 and the observed activities (A_{OBS}) are 0.9017 and 0.461, respectively. The A_{OBS} of 31 steroids are listed in Table 4 together with the A_{M1} estimated by the M1 model. Figure 4 shows the plot of A_{M1} versus A_{OBS}.

A predictive ability of a QSAR model for the external samples is another criterion in evaluating a quality of the model. The results estimated by the M1 model only explain an estimated ability for the internal samples, and the predictive ability of the M1 model is needed to still test 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. In such a cross-validation experiment involving *n* (here $n = 31$) molecules, a model is built from all but the first molecule, and this model is used to predict the activity of the first molecule. Then all but the second molecule are used to create a model that predicts the second molecule and so on. In this way, each molecule is predicted, as though the system had never seen it before, on the basis of all the other molecules. The activities (A_{CV}) predicted by 31 LOO models are also listed in Table 4. Here the number of PCs used in LOO procedures is still 5. The *R* and *RMS* between the A_{CV} predicted by *n* LOO models and the A_{OBS} observed experimentally are 0.8737 and 0.519, respectively.

In the previous studies,[3] systems were initially developed using the 21-molecule (no. 1−no. 21) set with a cross-validation experimental design. Only when development was complete were the 10 additional molecules (no. 22−no. 31) predicted on the basis of the model derived from the 21 compounds. We also use the same experimental design to develop the second QSAR model (M2) with $R = 0.9388$ and $RMS = 0.395$ ($R = 0.9471$ and $RMS = 0.391$ in the literature[3]). The *R* and *RMS* between the A_{OBS} and the LOO predictive activities are 0.8905 and 0.524, respectively ($R = 0.8136$ and $RMS = 0.707$ in the literature[3]). This shows that the M2 based on the MEDV-13 has a comparable quality with CoMFA model.

The M2 model above is used to predict the activities of 10 additional steroids which never appear in modeling the M2 model. The *R* and *RMS* in the predictive step are 0.6623 and 0.650, respectively. The results estimated and predicted by the M2 model are also listed in Table 4. To compare the results of the MEDV-13 with other methods, Table 5 gives the activities calculated by the M2 and CV models in this paper together with the activities obtained by the other methods such as CoMFA as well as CoMFA (FFD)[2,34] from the reports of Norinder,[34] COMPASS[2,17] from Jain's paper,[17] MS−WHIM[2,33] from Bravi,[33] and the similarity and SOMFA methods from Robinson.[2] Here all activities calculated are taken from the report of Robinson,[2] and all minuses are eliminated for the convenience. The results show that the *R* of the MEDV-13 is the highest among these methods, and the *RMS* of the MEDV-13 is lower than the ones of the four methods such as the two CoMFA, COMPASS, and MS−WHIM methods but higher than the ones of the similarity and SOMFA methods.

It should be indicated that selecting the former 21 molecules from a 31-molecule set to construct the M2 model could produce a chance correlation. So, another test procedure is also executed picking out 21 molecules in such a way as to select two out of every three from 31 steroids to construct a calibration set and using the remaining 10 molecules to form a predictive set. The 21-molecule calibration set is used to derive the third QSAR model (M3), and then the M3 model is employed to predict the activities of 10 samples in the predictive set. The activities estimated and predicted by the M3 model are also listed in Table 4. The *R* and *RMS* are respectively 0.9246 and 0.4097 for the modeling step and 0.8204 and 0.6063 in the predictive step, which shows that the predictive ability of the M3 is indeed higher than one of the M2. So, the comparison of results should be based on the CV procedure.

Table 4. Activities Observed and Estimated and Predicted by Several Models

steroid	A_{OBS}	A_{M1}	dA_{M1}	A_{CV}	dA_{CV}	A_{M2}	dA_{M2}	A_{M3}	dA_{M3}
1	6.279	6.880	0.601	6.942	0.663	6.894	0.615	6.915	0.636
2	5.000	5.169	0.169	5.171	0.171	5.390	0.390	5.339[a]	0.339[a]
3	5.000	4.966	−0.034	5.004	0.004	5.027	0.027	5.305	0.305
4	5.763	6.375	0.612	6.497	0.734	6.347	0.584	6.278	0.305
5	5.613	5.154	−0.459	5.071	−0.542	5.247	−0.366	5.085[a]	−0.528[a]
6	7.881	7.712	−0.169	7.654	−0.227	7.765	−0.116	7.802	−0.079
7	7.881	7.528	−0.353	7.460	−0.421	7.889	0.008	7.504	−0.377
8	6.892	6.565	−0.327	6.418	−0.474	6.787	−0.105	6.260[a]	−0.632[a]
9	5.000	4.932	−0.068	4.972	−0.028	4.865	−0.135	5.028	0.028
10	7.653	7.074	−0.579	6.985	−0.668	7.236	−0.417	7.258	−0.395
11	7.881	7.226	−0.655	7.135	−0.746	7.668	−0.213	7.302[a]	−0.579[a]
12	5.919	5.581	−0.338	5.536	−0.383	5.741	−0.178	5.760	−0.159
13	5.000	5.141	0.141	5.224	0.224	4.837	−0.163	5.254	0.254
14	5.000	5.453	0.453	5.609	0.609	5.075	0.075	5.522[a]	0.522[a]
15	5.000	5.121	0.121	5.217	0.217	4.715	−0.285	5.003	0.003
16	5.225	5.154	−0.071	5.142	−0.083	5.247	0.022	5.085	−0.140
17	5.225	5.162	−0.063	5.194	−0.031	5.122	−0.103	5.511	0.286
18	5.000	5.707	0.707	5.996	0.996	5.859	0.859	5.931[a]	0.931[a]
19	7.380	6.665	−0.715	6.582	−0.798	6.660	−0.720	6.814	−0.566
20	7.740	7.118	−0.622	7.071	−0.669	7.306	−0.434	7.142	−0.598
21	6.724	7.083	0.359	7.018	0.294	7.364	0.640	7.133[a]	0.409[a]
22	7.512	7.763	0.251	7.816	0.304	8.166[a]	0.654[a]	7.877	0.365
23	7.553	7.466	−0.087	7.460	−0.093	7.553[a]	0.000[a]	7.275	−0.278
24	6.779	7.791	1.012	7.688	0.909	7.652[a]	0.873[a]	7.788[a]	1.009[a]
25	7.200	7.712	0.512	7.767	0.567	7.765[a]	0.565[a]	7.802	0.602
26	6.114	5.825	−0.289	5.768	−0.346	5.659[a]	−0.455[a]	5.813	−0.301
27	6.247	6.559	0.312	6.651	0.404	6.666[a]	0.419[a]	6.444[a]	0.197[a]
28	7.120	7.203	0.083	7.275	0.155	7.340[a]	0.220[a]	7.470	0.350
29	6.817	5.999	−0.818	5.895	−0.922	5.740[a]	−1.077[a]	5.998	−0.819
30	7.688	7.330	−0.358	7.278	−0.410	7.642[a]	−0.046[a]	7.259[a]	−0.429[a]
31	5.797	6.380	0.583	6.272	0.475	6.845[a]	1.048[a]	6.160	0.363

[a] Refer to the activity predicted by the QSAR model such as M2 and M3.

Table 5. Comparison of Activities Calculated by Several Methods for 10 Steroids

steroid	measd activity	CoMFA	CoMFA (FFD)	similarity matrix analysis	COMPASS	MS−WHIM	SOMFA	M2 in MEDV-13
22	7.512	8.084	7.883	7.453	7.062	7.300	7.279	8.166
23	7.553	7.666	7.430	7.022	7.729	8.332	7.034	7.553
24	6.779	6.538	6.642	6.939	6.462	6.821	6.925	7.652
25	7.200	7.804	7.705	7.146	7.466	7.445	7.232	7.765
26	6.144	6.396	6.495	5.908	5.994	6.121	5.744	5.659
27	6.247	7.346	6.962	7.046	6.383	6.901	6.800	6.666
28	7.120	7.010	6.848	6.569	6.625	6.532	6.603	7.340
29	6.817	6.864	6.816	6.850	7.403	6.838	6.692	5.740
30	7.688	7.970	7.767	7.539	7.741	7.860	7.345	7.642
31	5.797	8.005	7.793	7.457	7.779	7.491	7.283	6.845
RMS		0.837	0.716	0.640	0.705	0.662	0.585	0.650
R		0.3659	0.3929	0.3429	0.3924	0.5265	0.4379	0.6623

QSAR Studies on 58 Dipeptides. The PCR techniques are also employed to model the QSAR relationship between the inhibiting angiotensin-converting enzyme (ACE) activity expressed by pIC_{50} and the MEDV-13 descriptors of 58 dipeptides. As stated above, only 19 variables enter into the PCR procedure. The first QSAR model (M4) is derived from a data set consisting of all 58 descriptors. The *R* and *RMS* between the activities (A_{M4}) estimated by the M4 model with the number of PCs of 19 and the observed activities (pIC_{50}) are 0.9462 and 0.32, respectively. The observed pIC_{50} of 58 dipeptides are listed in Table 6 together with the A_{M4} estimated by the M4 model. And Figure 5 shows the plot of A_{M4} versus pIC_{50}. A cross-validation technique is also used to test the stability and predictive ability of the M4 model, and the *R* and *RMS* obtained by LOO prediction are 0.8847 and 0.47, respectively. Table 6 also shows the activities (A_{CV4}) predicted by LOO procedure. Table 7 lists some statistics such as the *R* and *RMS* obtained using different QSAR methods. From Table 7, the MEDV-13 method is to

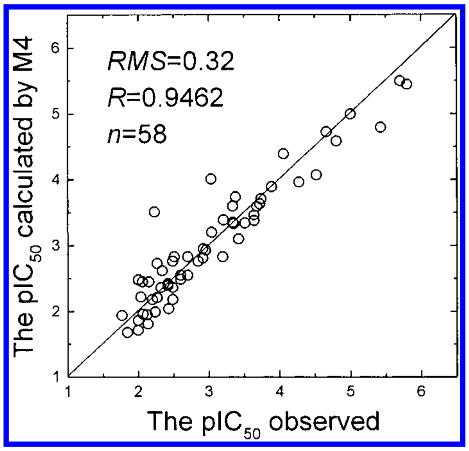


Figure 5. Plot of A_{M4} versus pIC_{50}.

Table 6. Activity Observed and Calculated by M4 and CV for 58 Dipeptides

no.	peptide	pIC_{50}	A_{M4}	A_{CV4}	no.	peptide	pIC_{50}	A_{M4}	A_{CV4}
1	VW	5.80	5.44	5.27	30	KG	2.49	2.18	1.99
2	IW	5.70	5.49	5.39	31	FG	2.43	2.04	1.86
3	IY	5.43	4.79	4.50	32	GS	2.42	2.39	2.37
4	AW	5.00	4.99	4.99	33	GV	2.34	2.62	2.71
5	RW	4.80	4.58	4.47	34	MG	2.32	2.36	2.37
6	VY	4.66	4.72	4.74	35	GK	2.27	2.73	2.95
7	GW	4.52	4.07	3.93	36	GE	2.27	2.21	2.19
8	VF	4.28	3.96	3.84	37	GT	2.24	1.99	1.76
9	AY	4.06	4.39	4.51	38	WG	2.23	3.51	4.10
10	IP	3.89	3.89	3.90	39	HG	2.20	2.18	1.92
11	RP	3.74	3.71	3.69	40	GQ	2.15	2.45	2.72
12	AF	3.72	3.63	3.60	41	GG	2.14	1.81	1.73
13	GY	3.68	3.59	3.55	42	QG	2.13	1.95	1.77
14	AP	3.64	3.38	3.25	43	SG	2.07	1.96	1.91
15	RF	3.64	3.46	3.39	44	LG	2.06	2.45	2.56
16	VP	3.38	3.74	3.95	45	GD	2.04	2.22	2.35
17	GP	3.35	3.33	3.17	46	TG	2.00	1.86	1.77
18	GF	3.20	2.83	2.71	47	EG	2.00	1.71	1.62
19	IF	3.03	4.01	4.43	48	DG	1.85	1.68	1.63
20	VG	2.96	2.93	2.92	49	PG	1.77	1.94	2.17
21	IG	2.92	2.95	2.96	50	LA	3.51	3.34	3.29
22	GI	2.92	2.81	2.78	51	KA	3.42	3.10	2.85
23	GM	2.85	2.76	2.75	52	RA	3.34	3.35	3.35
24	GA	2.70	2.55	2.51	53	YA	3.34	3.60	3.72
25	YG	2.70	2.83	2.88	54	AA	3.21	3.39	3.46
26	GL	2.60	2.49	2.46	55	FR	3.04	3.20	3.28
27	AG	2.60	2.55	2.53	56	HL	2.49	2.36	2.03
28	GH	2.51	2.83	2.88	57	DA	2.42	2.42	2.42
29	GR	2.49	2.76	2.89	58	EA	2.00	2.48	2.53

Table 7. Comparison of Different QSAR Models

statistic	M4 in this paper	t-scores method[35]	MS−WHIM scores method[35]
n	58	58	58
R	0.9462	0.8395	0.8643
RMS	0.32	0.54	0.50

date the best one for study on QSAR between inhibiting ACE activities and structural descriptors of 58 dipeptides.

CONCLUSION

We have described a new method based on the MEDV-13 method for predicting unknown biological activities based on relating molecular structures to their known biological activities. The study on steroids shows that the performance of the MEDV-13 method has a comparability with the previous methods containing 3D QSAR, and the peptide study gives the high quality of the QSAR model based on the MEDV-13 method. However, the MEDV-13 method only employs information about an element atom type, valance electron state, and chemical bond type from the 2D molecular topology and requires no information related to 3D structures or physicochemical properties or molecular alignment. So, the MEDV-13 descriptor is a fast, easy to use, reproducible, and predictable one for the QSAR studies.

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) 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.

(3) 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.

(4) DePriest, S A. 3D-QSAR of angiotensin-converting enzyme and thermolysin inhibitors: a comparison of CoMFA models based on deduced and experimentally determined active site geometries. *J. Am. Chem. Soc*. **1993**, *115*, 5372−5384.

(5) Horwitz, J. P. Comparative molecular field analysis of in vitro growth inhibition of L1210 and HCT-8 cells by some pyrazoloacridines. *J. Med. Chem*. **1993**, *36*, 3511−3516.

(6) Klebe, G.; Abraham, U. On the prediction of binding properties of drug molecules by comparative molecular field analysis. *J. Med. Chem*. **1993**, *36*, 70−80.

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

(8) 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.

(9) 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.

(10) Carbo, R.; Leyda, L.; Arnau, M. An electron density measure of the similarity between two compounds. *Int. J. Quantum Chem*. **1980**, *17*, 1185−1189.

(11) Good, A. C.; So, S.−S.; Richards, W. G. Structure−activity relationships from molecular similarity matrixes. *J. Med. Chem*. **1993**, *36*, 433−438.

(12) Good, A. C.; Peterson, S. J.; Richards, W. G. QSARs from similarity matrixes. Technique validation and application in the comparison of different similarity evaluation methods. *J. Med. Chem*. **1993**, *36*, 2929−2937.

(13) Burt, G.; Huxley, P.; Richards, W. G. The application of molecular similarity calculations. *J. Comput. Chem*. **1990**, *11*, 1139−1146.

(14) Seri-Levi, A.; Salter, R.; West, S.; Richards, W. G. Shape similarity as a single independent variable in QSAR. *Eur. J. Med. Chem*. **1994**, *29*, 687−694.

(15) Montanari, C. A.; Tute, M. S.; Beezer, A. E.; Mitchell, J. C. Determination of receptor bound drug conformations by QSAR using flexible fitting to derive a molecular similarity index. *J. Comput. Aid. Mol. Des*. **1996**, *10*, 67−73.

(16) Benigni, R.; Cotta Ramusino, M.; Giorgi, F.; Gallo, G. Molecular similarity matrixes and quantitative structure activity relationships: a case study with methodological implications. *J. Med. Chem*. **1995**, *38*, 629−635.

(17) 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.

(18) 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*, 8: 669∼677.

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

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

(21) Kier, L. B.; Hall, L. H. An atom-centered index for drug QSAR models. In *Advances in Drug Design*; Testa, B., Ed.; Academic Press: 1992; Vol. 22.

(22) 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.

(23) Hall, L. H.; Vaughn, T. A. QSAR of phenol toxicity using electrotopological state and kappa shape indices. *Med. Chem. Res*. **1997**, 7-(6−7), 407−416.

(24) 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. *Pharmaceut. Res*. **1996**, *13*(12), 1892−1895.

(25) AbouShaaban, R. R. A.; AlKhamees, H. A.; AbouAuda, H. S. et al. Atom level electrotopological state indexes in QSAR: Designing and testing antithyroid agents. *Pharmaceut. Res*. **1996**, *13*(1), 129−136.

(26) Hall, L. H.; Mohney, B.; Kier, L B. The electrotopological state: An atom index for QSAR. *Quant. Struct.−Act. Relat*. **1991**, *10*, 43−51.

(27) Hall, L. H.; Mohney, B.; Kier, L. B. Comparison of electrotopological state indexes with molecular orbital parameters: inhibition of MAO by hydrazides. *Quant. Struct.−Act. Relat*. **1993**, *12*, 44−48.

(28) 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.

(29) Liu, S. S.; Liu, H. L.; Xia, Z. L.; Cao, C. Z.; Li, Z. L. Molecular Distance-Edge Vector (μ): An Extension from Alkanes to Alcohols. *J. Chem. Inf. Comput. Sci.* **1999**, *39*(6), 951∼957.

(30) Liu, S. S.; Liu, Y.; Li, Z. L.; Cai, S. X. A Novel Molecular Electronegativity-Distance Vector, *Acta Chim. Sinica* **2000**, *48*(11), 1353−1357.

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

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

(33) Bravi, G.; Gancia, E.; Mascagni, P.; Pegna, M.; Todeschini, R.; Zaliani, A. MS−WHIM, new 3D theoretical descriptors derived from molecular surface properties: a comparative 3D QSAR study in a series of steroids. *J. Comput.-Aid. Mol. Des*. **1997**, *11*, 79∼92.

(34) Norinder, U. 3D-QSAR investigation of the Tripos benchmark steroids and some protein tyrosine-kinase inhibitors of styrene type using the TDQ approach. *J. Chemometr*. **1996**, *10*, 533−545.

(35) 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.

CI0003350