Peripheries of Molecular Projection in Three-dimensional Space and Studies of Quantitative Structure-activity Relationships

ZHANG, Qing-You(张庆友) LUO, Cao-Cao(罗操操) QI, Yu-Hua(齐玉华)
DONG, Lin(董林) WANG, Jun(王俊) XU, Lu*(许禄)

Changchun Institute of Applied Chemistry, Chinese Academy of Sciences, Changchun, Jilin 130022, China

A new index, *i.e.*, the periphery representation of the projection of a molecule from 3D space to a 2D plane is described. The results, correlation with toxicity of substituted nitrobenzenes, obtained by using periphery descriptors are much better than that obtained by using the areas (*i.e.*, shadows) of projections of the compounds. Even better results were achieved by using the combination of periphery descriptors and the projections areas as well as the indicated variable K reflecting the action of group NO_2 position on the benzene ring.

Keywords nitrobenzene, periphery, projection, QSAR, toxicity

Introduction

Quantitative structure-activity relationship studies have been applied widely for the prediction of organic compounds. Since the 1960’s, enormous efforts have been made by various investigators to develop quantitative parameters. During this period, Hansch and his co-workers[1,2] made important breakthrough for biological QSAR with electronic, stereo and hydrophobic parameters to be known as the extra-thermodynamic approach.

Since the 1980’s, several methods considering the reaction between three-dimensional (3D) molecular structure and receptor in QSAR have appeared. These methods are all called 3D-QSAR approaches. At present, the most widely used 3D-QSAR technique is CoMFA (comparative molecular field analysis).[3] However, recently, it was found that CoMFA alone can not obtain sufficiently strong equation to allow confident prediction for amino-benzenes. When some other parameters, such as heat of molecular formation of the compounds, were introduced into the CoMFA model, the results were improved greatly. It gives us a hint that a better description for molecular structures will give out a better prediction model, and this hint challenged us to look for other method of the projection areas of molecules in 3D space for 3D-QSAR. It is surprised that much better results than that obtained by using CoMFA were achieved.[4] We continue the research to describe the peripheries of the projections of molecules in 3D space for 3D-QSAR along this way. The satisfactory results can also be obtained. In the present study, two methods will be related, *i.e.* the method for the projection of a molecule, and the method for description of the periphery of a projection. Furthermore, the details about how to combine the above two methods for a real data set is be given in this paper.

Principles of the methods

The molecular shape profiles

Molecular shape is an important chemical concept. It has been used for QSAR. Some methods have been suggested for the representation of molecular shape.[5,6] We try to utilize the method suggested by Randić,[7] thus we here briefly introduce this algorithm as follows.

For catacondensed benzenoids all carbon atoms are on the molecular periphery. The simplest structure is benzene. In the case of benzene the (geometry) distance matrix is:

$$D=\begin{bmatrix} 0 & 1 & \sqrt{3} & 2 & \sqrt{3} & 1 \\ 1 & 0 & 1 & \sqrt{3} & 2 & \sqrt{3} \\ \sqrt{3} & 1 & 0 & 1 & \sqrt{3} & 2 \\ 2 & \sqrt{3} & 1 & 0 & 1 & \sqrt{3} \\ \sqrt{3} & 2 & \sqrt{3} & 1 & 0 & 1 \\ 1 & \sqrt{3} & 2 & \sqrt{3} & 1 & 0 \end{bmatrix} \tag{1}$$

From this matrix, the row (or column) sums: R_1, R_2, $\cdots$ R_6, can be obtained, and the average of row sums is $R=(R_1+R_2+\cdots+R_6)/6$, where 6 is the number of atoms in the structure. If all the elements in matrix (1) are squared, we can obtain matrix (2), ${}^2D$,

* E-mail: luxu@ciac.jl.cn; Tel.: 0086-0431-5265539; Fax: 0086-0431-5685653

Received July 15, 2003; revised Februry 9, 2004; accepted February 26, 2003.
Project supported by the National Natural Science Foundation of China (No. 20077026).

$$^{2}D=\begin{bmatrix} 0 & 1 & 3 & 4 & 3 & 1 \\ 1 & 0 & 1 & 3 & 4 & 3 \\ 3 & 1 & 0 & 1 & 3 & 4 \\ 4 & 3 & 1 & 0 & 1 & 3 \\ 3 & 4 & 3 & 1 & 0 & 1 \\ 1 & 3 & 4 & 3 & 1 & 0 \end{bmatrix} \qquad (2)$$

Again we consider the row sums and construct the average row sum: $^2R=(^2R_1+^2R_2+\cdots+^2R_6)/6$. Similarly, we can obtain ^3R, ^4R, $\cdots$. In order to reduce the role of ever interesting powers, the averaged row sum is normalized as below,

$$^0R,\ ^1R,\ ^2\mathrm{R}/2!,\ ^3R/3!,\ \cdots={}^0P,\ ^1P,\ ^2P,\ ^3P,\ \cdots$$

The sequence R for the first six powers of the matrix becomes

$$R=6,\ 7.46410162,\ 12,\ 20.3923049,\ 36,\ 65.1769145,\ 120$$

where the first item, 6 is the size of the system, and then R is normalized, $P=6$, 7.46410162, 6, 3.39871748, 1.5, 0.543140954, 0.166666667.

Orthogonal projection

In many cases, the biological activity or physicochemical property of interest is related to the 3D shape of the tested compounds. The shape parameters described here are molecular orthogonal projections. The projections were calculated by using a two-dimensional version of the point-encoded algorithm described by Stouch and Jurs.[8] In order to perform the calculation, the 3D atomic coordinates of the compounds must be available. These coordinates were computed by using SYBYL option BUILD and the conformations of the compounds were minimized by using the ENERGY MINIMIZE option (Tripos Associates, 1699 S. Hanley Road, Suite 303, St. Louis, MO63144). Once the molecular geometry has been defined, the structure is oriented in three-dimensional space according to some defined criterion. A molecule, represented by the 3D coordinates of its atoms and their van der Waals radii, is placed in a 3D grid of arbitrary density (in this article, the density is 4 bits per linear nanometer). Each point of intersection of the grid is checked to see if it lies within the molecule. If so, the point is put into a state, "1", if not so, the point is put into a state, "0". The molecule is then viewed from three orthogonal directions defined by the X, Y and Z coordinate axes. For each perspective, the coordinates are compressed into the plane defined by the remaining two axes. For the perspective along the Z-axis, the Z coordinates would be disregarded and the molecule projected onto the X-Y plane. A simple analogy, from which the name was derived, would be to obtain the shadow, which results from directing parallel rays of light along the axis of perspective. The area of this projection will be used as an index of molecular shape.

The first index calculated is the area of the shadow of the molecule projected on a plane defined by the X and Y axes (S_1). The second index is the area projected onto the Y-Z plane (S_2), and the third index is the area projected onto the X-Z plane (S_3). Each area is also normalized by dividing the index by the area of the rectangle defined by the maximum dimensions of the projection on the plane, *i.e.* S_4, S_5 and S_6, were obtained. Another index is the area of the rectangle of the maximum dimensions of the projected on the X-Y plane (S_7). The seven indexes derived from three orthogonal projections, which are approached by Jurs,[9] are adopted and called shadow indexes in this article. Furthermore, the volume parameter, V, is calculated based on the point encoded. In this work, the density of a grid is 4 divisions per linear angstrom, therefore, the volume represented by a "1" code is $(1/4)^3=1.5625\times10^{-5}$ nm^{-3}. The volume,[8] V of the entire molecule is computed by summing up all the "1" and multiplied by 1.5625×10^{-5}.

Periphery description of the projection of a molecule

The projection areas of a molecule have been used by us[4,10] for QSAR. The existing problem is that the different shapes of the projections may possess the same area. Consequently, we represent the peripheries of the projections using the Randić method described as above to replace the projection areas for QSAR, *i.e.* the combination of Randić and Jurs methods was used in this study. For example, the shape of nitrobenzene projected on Y-Z plane is shown in Figure 1. Since the interior points of the projection play no role for the shape profile, we considered only the points on the periphery of a projection. Because the shape of a molecule is the orthogonal projection, therefore all the points including the points of the periphery on the intersections of the squares, so the geometry distances can be obtained easily. As an example, the distance matrix in the rectangle in Figure 1 is

$$\begin{bmatrix} 0 & \sqrt{2} & \sqrt{5} & \sqrt{13} & 2\sqrt{5} & \sqrt{29} \\ \sqrt{2} & 0 & 1 & \sqrt{5} & \sqrt{10} & \sqrt{17} \\ \sqrt{5} & 1 & 0 & \sqrt{2} & \sqrt{5} & \sqrt{10} \\ \sqrt{13} & \sqrt{5} & \sqrt{2} & 0 & 1 & 2 \\ 2\sqrt{5} & \sqrt{10} & \sqrt{5} & 1 & 0 & 1 \\ \sqrt{29} & \sqrt{17} & \sqrt{10} & 2 & 1 & 0 \end{bmatrix}$$

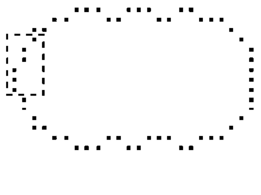


Figure 1 The shape of nitrobenzene projected on Y-Z plane.

Continually, based on Randić algorithm, six powers were calculated for the different projections on *X*-*Y*, *Y*-*Z* and *X*-*Z* planes, respectively. The total number of the parameters is 18. These parameters are:

$$^1P_{xy},\ ^2P_{xy},\ ^3P_{xy},\ ^4P_{xy},\ ^5P_{xy},\ ^6P_{xy},\ ^1P_{yz},\ ^2P_{yz},\ ^3P_{yz},$$

$$^4P_{yz},\ ^5P_{yz},\ ^6P_{yz},\ ^1P_{xz},\ ^2P_{xz},\ ^3P_{xz},\ ^4P_{xz},\ ^5P_{xz},\ ^6P_{xz}$$

Indicated descriptors

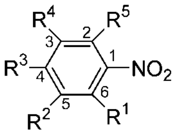
The number and the positions of group NO_2 on benzene ring play important roles for their activities. In order to reflect the influences, indicated descriptor, *K*, is introduced,

$$K=\begin{cases}0.5 & \text{(mononitrobenzene)}\\ 1.0 & \text{(}m\text{-dinitrobenzene)}\\ 3.0 & \text{(}p\text{- or }o\text{-dinitrobenzene)}\end{cases}$$

Experimental

Data

In this research, the toxicity LC_{50} of nitrobenzenes is the molar concentration of a nitrobenzene necessary to lead to death of the half tested fathead minnows[11] and−log LC_{50} is defined as the toxic descriptor. The nitrobenzene skeleton is



The different substitutions and their toxic activities are shown in Table 1. −log LC_{50} (obsd) is obtained from literature 11, and −log LC_{50} (cald) is calculated by artificial neural network -BFGS, and Diff. represent the difference between −log LC_{50} (obsd) and −log LC_{50} (cald).

Selection of the descriptors

The descriptors (*i.e.*, features) were derived from the structures of compounds, and used to create the predictive models. Abstraction of the features is the key step for a QSAR/QSPR study.

A good equation for structure and activity should possess high correlation coefficient *R*, low standard deviation *S*, and least variables. To this end, objective feature selection was done to weed out those descriptors that provide minimal or redundant information. The descriptors were analyzed using leaps-and-bounds regression.[12] Because in the subsequent statistical analysis we want to find the best subsets of the descriptors, *i.e.* when we want to take one, two, or three descriptors, we should find which one or one of the combinations is the best for the predictive model. Leaps-and-bounds regression provides us an effective approach, which can give out the answers quickly. This method is based on the fundamental inequality,

$$RSS\ (A) \leqslant RSS\ (B)$$

where *A* is any set of independent variables and *B* is a subset of *A*. The number of subsets evaluated in a search for the best subset regression can be restricted by the use of the inequality. For example, set A_1 contains 3 variables with *RSS*, 596, set A_2 contains 4 variables with *RSS*, 605. Thus, all the subsets of A_2 will be ignored, because of the regressions using these subsets with *RSS* greater than that for A_2, and also for A_1.

Results and discussion

Multiple regression

To compare projection areas (*i.e.*, shadows) and peripheries of the shadows for QSAR of nitrobenzenes, we will give out the results obtained by us in different cases.

Using variables *V*, S_1, S_2, S_3, S_4, S_5, S_6 and S_7

Variables *V*, S_1, S_2, S_3, S_4, S_5, S_6 and S_7 were assigned 1, 2,⋯, 8, respectively. From statistical viewpoint, the ratio of the number of sample (*N*) to the number of variables (*M*) should not be too low. Usually, it is recommended that $N/M \geqslant 5$. In the situation of this study, we have 35 samples, so *M* can not be great than 7. Table 2 shows the results of leaps-and-bounds regression for the best combinations of the 8 descriptors. For example, the best one-variable selection is 1 (*V*). The best two-variable selection is the combination of 1 and 7 (S_6), and the combination of 2 (S_1), 6 (S_5) and 8(S_7) is the best for three-variable selection, and so on. From this Table, we can see that 6-variable combination including variables 1, 2, 3, 4, 6 and 8 is the best one.

$$-\log\ LC_{50}=4.19-2.62\times V+3.41\times S_1+1.03\times S_2+0.58\times S_3-0.24\times S_5-1.58\times S_7 \qquad (1)$$

$$R=0.841, F=11.27, S=0.45, N=35$$

where *R* is correlation coefficient, *F* is significance test, *S* denotes standard deviation, and *N* stands for the number of samples.

Using variables *V*, S_1, S_2, S_3, S_4, S_5, S_6, S_7 and *K*

Because the above equation is not sufficiently strong to allow confident prediction, thus, the indicated variable *K* was added. The leaps-and-bounds regression results are shown in Table 3. This Table revealed that to put together the correlation coefficients, significance tests, and standard deviations, 7-variable combination including variables 1, 2, 3, 5, 6, 7 and 9 is the best one. The regression model is

Table 1 Substituted nitrobenzenes and their toxicity

No.	R^1	R^2	R^3	R^4	R^5	$-\log LC_{50}$ (obsd)	$-\log LC_{50}$ (calcd)	Diff.
1	CH_3	H	H	H	H	3.57	3.67	−0.10
2	H	CH_3	H	H	H	3.63	3.55	0.08
3	H	H	CH_3	H	H	3.76	3.89	−0.13
4	NO_2	H	H	H	H	5.45	5.40	0.05
5	H	NO_2	H	H	H	4.38	4.42	−0.04
6	H	H	NO_2	H	H	5.22	5.25	−0.03
7	CH_3	H	H	H	NO_2	5.01	5.10	−0.09
8	H	NO_2	CH_3	H	H	3.75	3.89	−0.14
9	CH_3	H	NO_2	H	H	5.15	5.20	−0.05
10	CH_3	NO_2	H	H	H	3.99	3.99	0.00
11	NO_2	H	CH_3	H	H	5.08	5.06	0.02
12	H	CH_3	H	NO_2	H	3.91	3.93	−0.02
13	H	NO_2	H	NO_2	H	5.29	5.27	0.02
14	H	H	H	H	H	3.02	3.21	−0.19
15	NH_2	H	H	H	H	3.70	3.51	0.19
16[a]	H	NO_2	NH_2	H	H	4.07	4.05	0.02
17	H	H	OH	H	H	3.36	3.46	−0.10
18	H	H	F	H	H	3.70	3.63	0.07
19	H	NO_2	CH_3	NO_2	H	4.88	4.89	−0.01
20	NO_2	CH_3	NO_2	H	H	6.37	6.37	−0.00
21	CH_3	NH_2	H	H	H	3.48	3.64	−0.16
22	H	CH_3	NH_2	H	H	3.24	3.32	−0.08
23	H	NH_2	CH_3	H	H	3.35	3.34	0.01
24[a]	NH_2	CH_3	H	H	H	3.80	3.57	0.23
25	NH_2	H	CH_3	H	H	3.80	3.50	0.30
26	NH_2	H	H	CH_3	H	3.79	3.73	0.06
27[a]	OH	H	H	NH_2	H	3.65	3.75	−0.10
28	CH_3	H	H	NH_2	H	3.77	3.76	0.01
29	H	NO_2	OH	H	H	4.04	3.91	0.13
30	H	NO_2	CH_3	NH_2	H	4.14	4.13	0.01
31	CH_3	NH_2	NO_2	H	H	5.34	5.21	0.13
32	CH_3	NO_2	NH_2	H	H	4.26	4.28	−0.02
33	NH_2	NO_2	CH_3	H	H	4.21	4.27	−0.06
34	NH_2	H	NO_2	CH_3	H	4.18	4.27	−0.09
35[a]	CH_3	NO_2	H	NH_2	H	4.46	4.37	0.09

[a] Samples of the test set.

$$-\log LC_{50}=4.19+0.58\times K-1.12\times V+2.66\times S_1-0.12\times S_3-0.65\times S_4-0.32\times S_5-1.77\times S_7 \quad (2)$$

$R=0.946$, $F=32.88$, $S=0.28$, $N=35$

Obviously, the results obtained by Eq. (2) are much better than that obtained by Eq. (1).

Using variable $^1P_{xy}$, $^2P_{xy}$, $^3P_{xy}$, $^4P_{xy}$, $^5P_{xy}$, $^6P_{xy}$, $^1P_{yz}$, $^2P_{yz}$, $^3P_{yz}$, $^4P_{yz}$, $^5P_{yz}$, $^6P_{yz}$, $^1P_{xz}$, $^2P_{xz}$, $^3P_{xz}$, $^4P_{xz}$, $^5P_{xz}$, $^6P_{xz}$

These variables were labeled from 1 to 18. For the selection of variables, leaps-and-bounds regression was also performed. The results are shown in Table 4. From this Table the results obtained by 7-variable combination are better than that obtained by 6-variable combination. Thus, we took 7-variable combination, and the

Table 2 Results of leaps-and-bounds regression for best combinations of the descriptors[a]

No.	Descriptor	R	S	F
1	1	0.690	0.56	30.05
2	1,7	0.782	0.49	25.13
3	2, 6, 8	0.826	0.45	22.21
4	1, 2, 6, 8	0.827	0.46	16.28
5	1, 2, 3, 4, 5	0.834	0.46	13.25
6	1, 2, 3, 4, 6, 8	0.841	0.45	11.27
7	1, 2, 3, 4, 5, 6, 8	0.842	0.46	9.39

[a] The variables are V, S_1, S_2, S_3, S_4, S_5, S_6 and S_7, which correspond to the numbers 1, 2, 3, …, 8, respectively.

Table 3 Results of leaps-and-bounds regression for best combinations of the descriptors[a]

No.	Descriptor	R	S	F
1	1	0.912	0.32	163.69
2	1, 2	0.920	0.31	88.35
3	1, 3, 8	0.929	0.29	65.57
4	1, 3, 5, 8	0.932	0.29	49.46
5	1, 2, 3, 7, 9	0.936	0.29	41.17
6	1, 2, 3, 6, 7, 9	0.941	0.28	36.03
7	1, 2, 3, 5, 6, 7, 9	0.946	0.28	32.88
8	1, 2, 3, 5, 6, 7, 8, 9	0.946	0.28	27.83

[a] The variables are K, V, S_1, S_2, S_3, S_4, S_5, S_6 and S_7, which correspond to the numbers 1, 2, 3, …, 9, separately.

Table 4 Results of leaps-and-bounds regression for best combinations of the descriptors[a]

No.	Descriptor	R	S	F
1	3	0.633	0.60	22.08
2	7, 13	0.671	0.58	13.09
3	6, 13, 14	0.767	0.51	14.74
4	6, 16, 17,18	0.873	0.40	24.05
5	6, 13, 15, 16, 17	0.901	0.36	24.87
6	5, 6, 13, 14, 15, 17	0.916	0.34	24.19
7	4, 5, 6, 13, 14, 15, 16	0.921	0.33	21.62
8	7, 8, 11, 12, 13, 14, 16, 18	0.928	0.32	20.15
9	6, 7, 8, 10, 12, 13, 14, 17, 18	0.941	0.30	21.62

[a] The variables are ${}^1P_{xy}$, ${}^2P_{xy}$, ${}^3P_{xy}$, ${}^4P_{xy}$, ${}^5P_{xy}$, ${}^6P_{xy}$, ${}^1P_{yz}$, ${}^2P_{yz}$, ${}^3P_{yz}$, ${}^4P_{yz}$, ${}^5P_{yz}$, ${}^6P_{yz}$, ${}^1P_{xz}$, ${}^2P_{xz}$, ${}^3P_{xz}$, ${}^4P_{xz}$, ${}^5P_{xz}$, ${}^6P_{xz}$, which correspond to the numbers 1, 2, 3, …, 18, separately.

regression model is

$$-\log\ \mathrm{LC}_{50}=4.19+28.36\times{}^4P_{xy}-63.67\times{}^5P_{xy}+36.00\times{}^6P_{xy}-63.46\times{}^1P_{xz}+224.29\times{}^2P_{xz}-256.60\times{}^3P_{xz}+95.58\times{}^4P_{xz} \qquad (3)$$

$$R=0.921,\ F=21.62,\ S=0.33,\ N=35$$

As case (3), if variable K is added, the correlation coefficient is $R=0.943$.

Using the combination of all the variables

We tried to further improve the predictive results, thus the combination of all the variables, *i.e.* the shadows, the peripheries of the shadows, V and K was observed. The different best combinations of the variables were calculated also by using leaps-and-bounds regression analysis, and the results are shown in Table 5. Though the better results can be obtained by using the 8-variable best combination, the 7-variable best combination was selected in this study, because the rule of thumb, *i.e.*, $N/M\geqslant 5$ was followed. The regression model is

$$-\log\ \mathrm{LC}_{50}=4.19-1.50\times S_2-1.78\times S_3-0.18\times S_6-1.99\times S_7+0.52\times K+3.32\times{}^1P_{yz}+2.67\times{}^1P_{xz} \qquad (4)$$

$$R=0.967,\ F=56.12,\ S=0.22,\ N=35$$

Fortunately, the results are greatly improved. It is interesting that all the shadows relate to the toxicity negatively, wherever all the peripheries relate to the toxicity positively. This means that compounds possessing lower S_2, S_3, S_6 and S_7 or higher ${}^1P_{yz}$ and ${}^1P_{xz}$ as well as K will be more toxic.

Table 5 Results of leaps-and-bounds regression for best combinations of the descriptors[a]

No.	Descriptor	R	S	F
1	9	0.912	0.32	163.69
2	1, 9	0.920	0.31	88.35
3	7, 9, 10	0.930	0.29	66.19
4	3, 7, 9, 16	0.937	0.28	53.59
5	4, 8, 9, 16, 22	0.959	0.23	65.73
7	3, 4, 7, 8, 9, 16, 22	0.967	0.22	56.12
8	3, 4, 6, 8, 9, 11, 16, 22	0.971	0.21	53.51
9	3, 4, 7, 8, 9, 17, 18, 19, 22	0.972	0.21	48.21

[a] The variables are V, S_1, S_2, S_3, S_4, S_5, S_6, S_7, K, ${}^1P_{xy}$, ${}^2P_{xy}$, ${}^3P_{xy}$, ${}^4P_{xy}$, ${}^5P_{xy}$, ${}^6P_{xy}$, ${}^1P_{yz}$, ${}^2P_{yz}$, ${}^3P_{yz}$, ${}^4P_{yz}$, ${}^5P_{yz}$, ${}^6P_{yz}$, ${}^1P_{xz}$, ${}^2P_{xz}$, ${}^3P_{xz}$, ${}^4P_{xz}$, ${}^5P_{xz}$, ${}^6P_{xz}$, which correspond to the numbers 1, 2, 3, …, 27, separately.

Neural networks

In recent years, artificial neural networks have been used widely. Among the neural network learning algorithms, the back-propagation (BP) method is one of the most commonly used methods. The drawback of BP is that the training processes slowly, because the gradient-descent algorithm is usually used for minimizing the sum-squared-error. In this research, the BFGS quasi-Newton method was used. The advantages of using the

BFGS method are that specifying rate or momentum is not necessary and training processes much more rapidly.[13]

The input nodes of the neural network are as the same as case (4) *i.e.*, S_2, S_3, S_6, S_7, K, ${}^1P_{yz}$, ${}^1P_{xz}$. The number of the output neuron is one. To avoid overtraining, the test set was used to monitor the training process for networks, that is, during the training of the network, the performance was monitored by predicting the values for the compounds in the test set. As long as test set results were improved, training was continued. However, when the test set results ceased to improve, the training was stopped. The results indicate that this method is an effective approach to avoid over-training.

Consequently, the entire data set was divided into two groups: 31 compounds as the training set and 4 compounds as the test set. As a usual rule of the thumb, the weights and bases should be less than the samples in number, thus the model achieved by the network is stationary. Therefore, the number of the hidden neurons should not be greater than 3. The experiments showed that in the situations of 1, 2 and 3 hidden neurons, better results could be obtained by using 3 hidden neurons. So the architecture of an over network was 7∶3∶1. The results obtained by neural network are R=0.990, F=1608.56 and S=0.12. Obviously, these are much better than those obtained by using multiple regression analysis.

The toxic activities of the 35 compounds in this paper were calculated using the model obtained in this part, as shown in Table 1. These indicate that the largest absolute error for compound **25** is 0.30, the corresponding relative error is 7.9% (Table 1). The toxic activities estimated by the QSAR model being plotted vs. the toxic activities observed are shown in Figure 2.

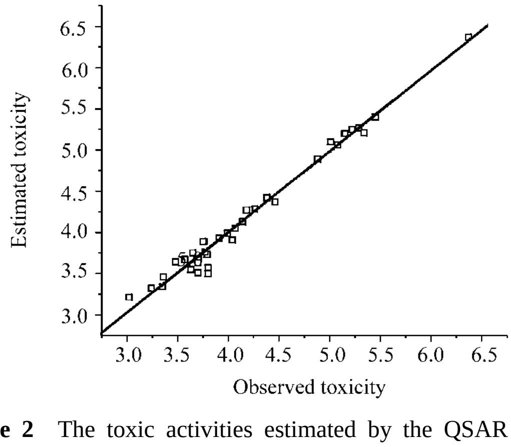


Figure 2 The toxic activities estimated by the QSAR model being plotted vs. the toxic activities observed.

Conclusion

Biological activity or physicochemical property of interest is often related to the 3D features of the test compounds. Because the different shapes may possess the same area, consequently, the periphery of a projection on a plane was described based on the algorithm suggested by Randić, and firstly applied to the real system, *i.e.*, to the studies on QSAR for nitrobenzenes. As we expected, the results obtained by using the representations of peripheries of the projections are much better than that obtained by using the projection areas. Especially by the combination of the peripheries with the shadows, and indicated variable K, quite excellent results were achieved. Periphery representation of a projection is an effective method for 3D QSAR.

References

1 Hansch, C.; Muir, R. M.; Fujita, T.; Maloncy, P. P.; Geiger, F. *J. Am. Chem. Soc.* **1963**, *85*, 2817.

2 Fujita, T.; Iwasa, J.; Hansch, C. *J. Am. Chem. Soc.* **1964**, *86*, 5175.

3 Cramer III, R. D.; Patterson, D. E.; Bunce, J. D. *J. Am. Chem. Soc.* **1988**, *110*, 5959.

4 Xu, L.; Yang, J. A.; Wu, Y. P. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 602.

5 Randić, M.; Razinger, M. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 429.

6 Randić, M.; Razinger, M. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 140.

7 Randić, M. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 373.

8 Stouch, T. R.; Jurs, P. C. *J. Chem. Inf. Comput. Sci.* **1986**, *26*, 4.

9 Rohrbaugh, R. H.; Jurs, P. C. *Anal. Chim. Acta* **1987**, *199*, 99.

10 Zhang, Q. Y.; Xu L. *Chem. J. Chin. Univ.* **2002**, *23*, 2125 (in Chinese).

11 Hall, L. H.; Maynard, E. L.; Kier, L. B. *Environ. Toxico. Chem.* **1989**, *8*, 431.

12 Furnival, G. M.; Wilson, R. W. *Technometrics* **1974**, *16*, 499.

13 Xu, L.; Ball, J. W.; Dixon, S. L.; Jurs, P. C. *Environ. Sci. Chem.* **1994**, *13*, 841.

(E0307158 LU, Y. J.)