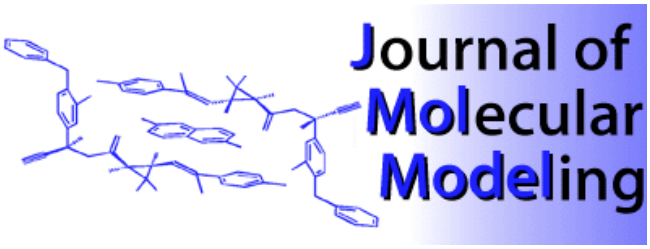
J. Mol. Model. **1999**, *5*, 252 – 262



FULL PAPER

Data Mining of Toxic Chemicals: Structure Patterns and QSAR

Jiansuo Wang, Luhua Lai, and Youqi Tang

Institute of Physical Chemistry, Peking University, Beijing 100871, P.R.China. Tel: +86-10-62751490; Fax: +86-10-62751725; E-mail: lai@ipc.pku.edu.cn

Received: 11 Januar 1999/ Accepted: 4 October 1999/ Published: 19 November 1999

Abstract We take a two-step strategy to explore noncongeneric toxic chemicals from the database RTECS: the screening of structure patterns and the generation of a detailed relationship between structure and activity. An efficient similarity comparison is proposed to screen chemical patterns for further QSAR analysis. Then CoMFA study is carried out on one structure pattern as an example of the implementation, and the result shows that QSAR studies of structure patterns can provide an estimate of the activity as well as a detailed relationship between activity and structure. From the performance of overall procedure, such a stepwise scheme is demonstrated to be feasible and effective to mine a database of toxic chemicals.

Keywords Toxic chemicals, Database mining, Structure patterns, QSAR

Introduction

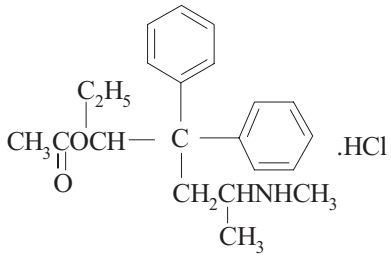
Due to the high cost and social objection to animal testing, and the large numbers of candidate chemicals for screening, it is important to predict the potential risk of chemicals with little or no empirical data. The ability to predict toxicity both quickly and accurately comes from an understanding of currently known knowledge. However, there exists a gap between the application to toxicity prediction and the mining of available databases, although there is a great wealth of information unexplored in the databases of toxic chemicals..

Current efforts for toxicity prediction mainly fall into the two categories [1-2]. One is knowledge-based systems that rely on a set of rules distilled from available knowledge or human experts, whose representative predictive systems are like DEREK [3-4], OncoLogic [5], etc. The other is correlative model approaches that rely on the use of statistics for exploring the relationship between structure and activity, which primarily include the technology of QSARs (Quantitative Structure-Activity Relationships) and are adopted by such programs as TOPKAT [6]. Knowledge-based systems are restricted to human experts and incapable of discovering new relationships automatically, while classic correlative QSARs are limited to within congeneric series of chemicals. Therefore, some new schemes should be proposed to explore the databases of noncongeneric chemicals effectively.

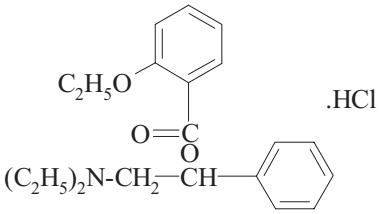
We attempt to adopt a two-step strategy to examine the chemicals in a database. The first step involves the screening of chemical categories, here referred to structural patterns shared by a series of molecules that possibly act in a toxicologically similar manner. This step is a coarse-grained and qualitative one, and an efficient method is introduced. The second step is the generation of a detailed relationship between structure and activity based on chemical clusters,

Correspondence to: L. Lai

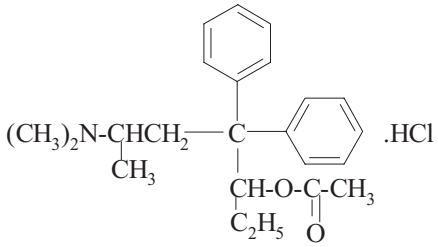
1.



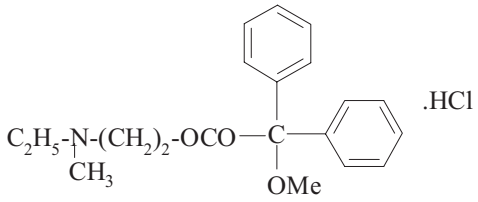
2.



3.



4.



Scheme 1 *Molecules that are used in an example of similarity computation*

which is a fine-grained and quantitative step. The scheme is a compromise between speed and accuracy.

In this paper, a molecular structure pattern is defined as a template comprising a given framework and some given groups; it represents a cluster of molecules sharing common structural features possibly required for a certain kind of property or activity. The notion of structure patterns of toxic chemicals arises from the specificity of action modes of chemicals in biological systems. In 1909, Paul Ehrlich demonstrated that drugs often induce physiological effects by binding to the highly specific target structures (receptors) at the cellular level. Now, the receptor theory and the concept of specificity have been universally accepted and profoundly enriched.

Table 1 *LD50 value range of toxic chemicals in the database RTECS [a]*

LD50 value-range	Number of chemicals
$\geq 10^4$ mg·kg^{-1}	254
10^3-10^4 mg·kg^{-1}	6,145
100-1,000 mg·kg^{-1}	23,188
10-100 mg·kg^{-1}	12,141
1-10 mg·kg^{-1}	4,453
100-1,000 ìg·kg^{-1}	558
10-100 ìg·kg^{-1}	137
<10 ìg·kg^{-1}	20
Total	46,896

[a] The statistic chemicals are the ones containing both WLN and LD50 in the database RTECS, totally 47,153 chemicals, but of them whose LD50 is labelled as mg·kg^{-1} or ìg·kg^{-1} are totally 46,896 as the above.

There is wide consensus that the shape and the chemical composition of a drug must complement those of the binding site on its receptor (here referring to critical macromolecules in the body such as proteins, DNA, etc.) [7-13]. For toxic chemicals, the biochemical basis is the same as that of drugs. Therefore, due to the specificity of action modes, toxic chemicals for certain kinds of activity will share common structural features.

QSAR refers to statistical analysis of potential relationships between chemical structure and biological activity, which provides a major form to summarise chemical and biological information so that we can generate and test hypotheses to facilitate an understanding of mechanism of molecular action [14-16]. Compared to classical two-dimensional (2D) QSAR, three-dimensional (3D) QSAR offers a more powerful tool to describe specific interaction between molecules. CoMFA [17-18] (Comparative Molecular Field Analysis) is one of the most widely used in 3D QSAR. Its basic idea is that the interactions between a series of chemicals and the target molecule in biological systems are usually noncovalent so that the differences of the steric and electrostatic field surrounding the series of chemicals might provide conditional requirements of molecular structure responsible for activity. Therefore, CoMFA can be introduced into the QSAR analysis of toxic chemicals after we identify structure patterns.

Data and methodology

The Registry of Toxic Effects of Chemical Substances (RTECS) [19] is a database of toxicological information compiled, maintained, and updated by the National Institute for Occupational Safety and Health (NIOSH). It contains information on over 130,000 chemicals. From RTECS we con-

Table 2 *The groups used in similarity computation*

No.	Groups	No.	Groups	No.	Groups
1.	H-(C=O)-NH-	34.	-SH	67.	-Br
2.	-NH-(C=O)-NH2	35.	-S-CH-	68.	-CI
3.	-NH-(C=O)-O-	36.	-SO2-	69.	-I
4.	-NH-(C=O)-CH2-	37.	-SO-CH3	70.	-P-
5.	-NH-(C=O)-H	38.	-S-	71.	-N-(phenyl)2
6.	-(C=O)-NH-OH	39.	-OPO-OC2H5	72.	-N=C-(phenyl)
7.	-(C=O)-NH-	40.	-OPO-OCH3	73.	-C=N-(phenyl)
8.	-O-NH2	41.	-PO-OCH3	74.	-NH-(C=O)-(phenyl)
9.	-O-(C=O)-NH2	42.	-PO-OC2H5	75.	-N=N-(phenyl)
10.	-(C=O)-NH2	43.	–P-(CH3)2	76.	-NH-SO2-(phenyl)
11.	-O-CN	44.	–P-(C2H5)2	77.	-NH-(phenyl)
12.	-CH=N-OH	45.	-PO2O-	78.	-P-(phenyl)2
13.	-ONO2	46.	-PO2OH	79.	-C-(phenyl)2
14.	-SO2-N-(CH3)2	47.	-C=C-(C=O)-H	80.	-CH-(phenyl)2
15.	-NO2	48.	-C=C-(C=O)-OH	81.	-C?C-(phenyl)
16.	-NO	49.	-CH-(OH)-CH-(OH)-	82.	-C=C-(phenyl)
17.	-N-(C=O)-(CH3)2	50.	-CH-OH	83.	-CH-(phenyl)
18.	-N-(C=O)	51.	-C-OH	84.	-(C=O)-(phenyl)
19.	-SCN	52.	-(C=O)-O-	85.	-(C=O)-O-(phenyl)
20.	-CN	53.	-(C=O)-H	86.	-SO2-(phenyl)
21.	-NNN	54.	-(C=O)-OH	87.	-CH2-O-(phenyl)
22.	-N=NN-(CH3)2	55.	-(C=O)-	88.	-CH(OH)-(phenyl)
23.	-NN	56.	-OH	89.	-S-(phenyl)
24.	$-NH3^+$	57.	-(O2)	90.	-CH2-(phenyl)
25.	-NH-(C=S)-NH2	58.	-O-	91.	= =
26.	-NH-(C=S)-	59.	-CF3	92.	=
27.	-NH-	60.	–CF	93.	-(phenyl)-(phenyl)
28.	-N-	61.	-F	94.	NH-cycle
29.	$-NH_2$	62.	$-CCl_3$	95.	N-cycle
30.	$-N^+-$	63.	-CCl	96.	O-cycle
31.	-SO2OH	64.	-Cl	97.	(C=O)-cycle
32.	-SO2O-	65.	-CBr3	98.	S-cycle
33.	-SO2-NH2	66.	-CBr		

struct a database in the following way: first select the chemicals containing both LD50 value and WLN structure representation [20] because of the need of structure information, totally 47,153 chemicals; then remove the chemicals containing metal atoms such as Se, Te, Bi, As, Sn, Zn, Pb, Au, Ag, Pa, W, Ni, and so on, because we only concern about organic chemicals; finally retain 46,544 chemicals to make up the database of toxic chemicals as the one that we will study. Here the toxicity of chemicals means that chemicals have LD50 values recorded. Table 1 displays LD50 value range of toxic chemicals in the database RTECS.

We attempt to analyse structure patterns of toxic chemicals in terms of structure similarity. We assume that every chemical in the database belongs to one of molecular structure patterns, then toxic chemicals for some kind of activity should be structurally similar in structure patterns so that we can screen such patterns.

We compute molecular structure similarity as follows: 1) two molecules cannot be compared until their frameworks are alike. The frameworks of two molecules are similar if they share the same framework indicators such as molecular monocycle count (classified into three-member cycle, five-member cycle, six-member cycle and others; saturation and non-saturation); fused-cycle count (classified into dicycle, tricycle, quad-cycle and others; saturation and non-saturation); branch-point atom count (classified into N-atom class, P-atom class and C-atom class) and so on. If the two molecules are not alike in the framework, the structure similarity value is regarded as zero. 2) structure similarity values, like the Tanimoto coefficient, are obtained by comparing group composition of the two molecules after comparing the frameworks. The group composition of the two molecules composes the group sets respectively (S_1 and S_2), and the structure similarity value equals to the sum of the cross set divided by the sum of the combine set, that is, Value = $(S_1 \cdot S_2)/(S_1+S_2)$. The

Table 3 *Similarity computation of the molecules in Scheme 1 (molecule **4** is reference molecule)*

No	Framework indicators	Group indicators [a]	Similarity values
1	phenyl 2,2	-NH-,0,1 -N-, 1,0 -COO-,1.1 -(C=O)-,1,1 -O-,2,1 -Cl, 1,1 -C-(phenyl)$_2$,1,1	5/8=0.62
2	phenyl 2,2	-N-,1,1 -COO-,1,1 -(C=O)-,1,1 -O-,2,2 -Cl, 1,1 -C-(phenyl)$_2$,1,0 -CH-phenyl, 0,1	6/8=0.75
3	phenyl 2,2	-N-,1,1 -COO-,1,1 -(C=O)-,1,1 -O-,2,1 -C-(phenyl)$_2$,1,1 -Cl, 1,1	6/7=0.86

[a] The first digit is the count index of reference molecule and the second digit is that of required molecules.

referred groups comprise common cyclic atoms (-N-, -NH-,-O-,-S-,-(C=O)-, etc.), common non-cyclic atoms (NH_2-, -NH-, -N-, -S-, -(C=O)-, -O-, F, Cl, Br, I ,etc.), and other groups with high occurrence frequency in the database RTECS. Table 2 lists the groups that are used. Table 3 and Scheme 1 give the examples of similarity computation.

We conduct CoMFA analysis for one of structure patterns screened out from the database RTECS. The representative molecule of the structure pattern is presented in Scheme 2. By computing molecular similarity, we get 189 chemicals from the database RTECS whose similarity values to the representative molecule are higher than 0.6. These chemicals have the same framework: a six-membered ring only containing one saturated hydrogen. Taking account of the experimental data of toxic effects, the chemicals mainly fall in the five major categories according to species observed and route of exposure. They are respectively: rat-intraperitoneal, 15 chemicals; mouse-intravenous, 19 chemicals; rabbit-intravenous, 37 chemicals; mouse-oral, 48 chemicals and mouse-intraperitoneal, 102 chemicals. We select the front three series to build CoMFA models between the structure and LD50 values about the chemicals.

By using the software SYBYL 6.4 [21], we carry out the following procedure to implement CoMFA analysis for each series of chemicals: 1) generating 3D structure of molecules. We build the structures and energy-minimise them using the modules in the SYBYL. The charges of molecules are computed by the Gasteiger-Marsili-Hückel method. 2) aligning the molecules. We use SYBYL's fitting capability on the atoms of the six-membered ring to align the molecules because they have this common framework. 3) investigating with CoMFA simultaneously considering flexible conformations of the molecules. For every molecule, the common framework is regarded as rigid and systematic searching of conformations is partially performed on the rotatable single bonds of side-chains. The conformations whose energies are not more than 10 kcal mol^{-1} above the lowest energy are accepted as energetically permitted. Then, the energetically permitted conformations for each molecule are classified and selected out about five representative conformations to be used in CoMFA study. The final conformation of each molecule is defined by iterative CoMFA analysis. 4) specifying the location of the region where CoMFA fields will be evaluated. Based on the automatically created region, the location of the region is optimised by translation in the all space. This translation optimisation of the region can be executed in the software AOP [22].

Results

Screening of structure patterns of toxic chemicals

We try to screen structure patterns of toxic chemicals from the database in the following way: Assign structure similarity limit value to 0.6; select the molecule if the molecule encounters more than 100 molecules with higher similarity values than 0.6 to it in the database; classify the molecules with similarity value larger than 0.6 to each other into one class; then take the representative molecules of every class as selected structure patterns with 0.6 as similarity limit and 100 as count limit. Totally 253 chemicals are found. Table 4 lists some structure patterns with higher chemical count than 150. To cover more of the original database, we also screen the structure patterns given 0.6 as similarity limit, 50 as count limit and 0.6 as similarity limit, 20 as count limit. Excluding the previous chemicals, 100 chemicals and 337 chemicals are screened out, respectively.

In order to analyse the structure patterns obtained, all of them are used to form a testing database to assess the original complete database. 17,181 out of 46,544 chemicals in the original database are predicted to be similar to the 690 structure patterns. This covers about 37% of the whole database. It indicates that these structure patterns have the characteristics to represent a cluster of toxic chemicals and provide some basis to make further QSAR analysis for toxic chemicals.

Table 4 *The representative molecules as structure patterns (with higher chemical count than 150 and given similarity limit as 0.6)*

Chemical -count	CAS -Number	Wisswesser Linear Notation
401	140-41-0	GR DMVN1&1 &GXGGVO
384	55-45-8	GR CMVO2UU2K &G &12/15
383	102585-42-2	GR BO1Y1&N2&V1N2&2 &GH
352	73972-98-2	L6TJ AX1&1&1 DOVR BG CG DG EG FVO &-NA-
323	2828-42-4	1Y1&UNOVMR
321	27585-47-3	4MVO2N2OVM4&2R BO4 CO1 &QV1VQ &QH
306	52174-08-0	1N1&V1SR BOVM1
274	94-02-0	2OV1VR
273	73623-41-3	ZR D1VO2Y1&1 &GH
252	66941-48-8	T6VMVNV FHJ FY1&S2U1 F2U1 &-NA- &29/6
246	102585-26-2	GR BG D1UN1YO2&O2
211	140-11-4	1VO1R
209	25561-56-2	T6N DNTJ ANU1VR DG& D1 &GH
203	63658-99-1	T56 BNYNJ B1 CUM D1O7 &GH
202	1508-65-2	L6TJ AXQR&VO2UU2N2&2 &GH
201	57166-13-9	L66J C1YM1&UN1 &GH
196	73664-67-2	L6TJ AN2U1&VR CO1 DO1 EO1
192	66227-09-6	3OV1OR CX1&1&1
191	2653-08-9	G1VMR DMV1G
189	115-44-6	T6VMVMV FHJ FY2&1 F2U1
188	25561-52-8	T6N DNTJ ANU1VR DOR&& D1 &GH
188	23564-06-9	2OVMYUS&MR BMYUS&MVO2
187	27591-74-8	2OPO&O2&OY1&U1VOY1&R
184	77791-43-6	T6NTJ AYVM1&2OR B1& B1 &GH
183	66827-36-9	T6N DOTJ A2OVXGR&R &GH
181	64058-98-6	GR DYO2&YUN4&M4
181	100758-54-1	L6TJ AMV3 D4
181	64058-98-6	GR DYO2&YUN4&M4
179	101651-68-7	QBQR BMV1 DBQQ
177	100310-78-9	T66 BO EOT&J C2N1&2G &GH
172	67205-34-9	NCR BG CG FG ECN DOV1
169	89-05-4	QVR BVQ DVQ EVQ
166	47003-79-2	FXFFR C1Y1&M1VM1
162	4746-61-6	Q1VMR
161	82394-11-4	T C676 IS&T&J B1N1&1 &GH
161	1159-83-7	T C676 BY IS JHJ BU3N1&1 NG &GH
160	5418-93-9	T56 BM DNJ CZ HG
159	74022-48-3	T56 BM DNJ CS3N1&1 GO1 HO1
159	21309-90-0	T56 BSNJ DO2M4 &GH
158	6550-57-8	T C676 BS INJ JN1&2N1&1 MG
155	615-16-7	T56 BNVNJ
154	61072-16-0	GR DO2NR B1&V1N2&2 &GH
152	73664-30-9	1U2MV1R CO1 DO1
151	63918-05-8	L6TJ AR DR D1O2N1&1 &GH
150	32210-23-4	L6TJ AX1&1&1 DOV1
150	991-30-0	T6N DNTJ AVR CO1 DO1 EO1& D1Y1&OVR CO1 DO1 EO1 &GH
150	32210-23-4	L6TJ AX1&1&1 DOV1

CoMFA analysis for one structure pattern

We carry out CoMFA studies using the QSAR module integrated in the software SYBYL 6.4. The biological data are the values of the acute medial lethal dose LD50 of chemicals. They are transformed to the logarithm form log(1/LD50) since the free energy is proportional to the logarithm of the equilibrium constant as the equation: ÄG = - RT lnK. The region for CoMFA is defined by performing AOP. The probe atom is chosen to be an sp^3-hybridized carbon atom with a charge of +1. The statistical analysis is performed by the means of the PLS (partial least-squares) technology. Leave-one-out cross-validation is used to check the analysis and the final PLS analysis serves as a CoMFA model, which can provide a numerical estimate of the activity as well as a qualitative graphical view of the most important three-dimensional aspects of the structure-activity relationship.

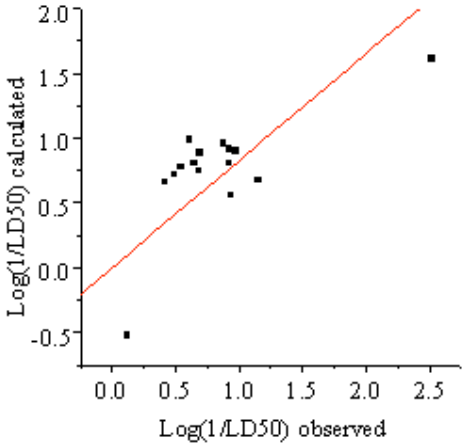


Figure 1a *Rat-intraperitoneal: cross-validated CoMFA analysis with three components; q^2 = 0.496*

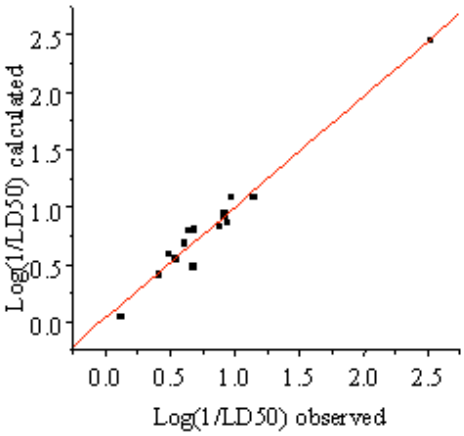


Figure 1b *Final CoMFA analysis with three components; r^2 = 0.967, F = 109 (Rat-intraperitoneal)*

Rat-intraperitoneal The 15 chemicals subjected to the leave-one-out cross-validated CoMFA yield a q^2 value of 0.496 with three components. And the final non-cross-validated CoMFA model has an r^2 value of 0.967 and an F value of 109. The biological activity data of chemicals, including observed and calculated, are listed in Table 5. And the results of the cross-validation and the fit are displayed by the graph of calculated activity versus observed activity in Figure 1a and b, respectively. Figure 1c is the contour map of the electrostatic and steric terms around the molecules with 80% of the signal by default, which provides where in space the QSAR terms have high or low values.

From Figure 1c, we can examine the 3D structure of the molecules associated to the activity. We can discern that the large red areas are the regions where positive potential is favourable for decreasing the toxicity; while the large yellow areas are the regions where bulky substituents are desirable to decrease the activity.

Mouse-intravenous The 18 chemicals yield a q^2 of 0.622with four components for cross validation, and the result of the final model has an r^2 of 0.978 and an F of 144 (Table 6, Figure 2).

For the molecules, the large yellow region can be introduced bulky substituents to decrease the toxicity, while the blue region is desirable to introduce negative potential.

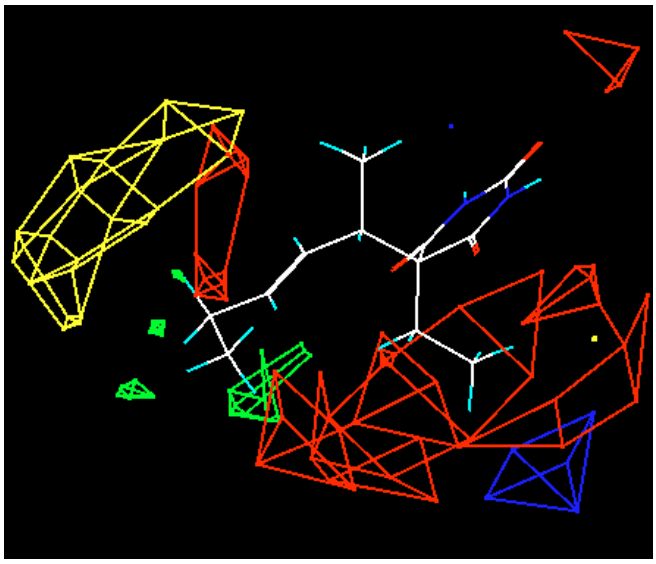


Figure 1c *Contour map of final CoMFA model; for steric effects, more bulk near green and less bulk near green is favorable to increase the active, while for electrostatic effects, more positive near blue and more negative near red is desirable for molecules to be more active.*

Table 5 *Rat -intraperitoneal: chemicals and biological activity data including observed and calculated*

No.	CAS number	LD50/10^{-3}kg kg^{-1} (observed)	log(1/LD50) (observed)	log(1/LD50) (cross-validated)	log(1/LD50) (fitted)
1	125_40_6	70	1.15	0.67	1.08
2	17013_35_3	3	2.52	1.61	2.45
3	52_43_7	121	0.92	0.80	0.94
4	57_43_2	115	0.94	0.56	0.86
5	57_44_3	246	0.61	0.99	0.69
6	60784_70_5	290	0.54	0.78	0.54
7	64038_27_3	205	0.69	0.89	0.80
8	66940_75_8	755	0.12	-0.52	0.04
9	66968_29_4	227	0.64	0.81	0.79
10	66968_31_8	376	0.42	0.66	0.41
11	66968_32_9	321	0.49	0.72	0.59
12	66968_81_8	210	0.68	0.75	0.48
13	76_74_4	108	0.97	0.90	1.08
14	76_75_5	120	0.92	0.92	0.92
15	780_59_6	132	0.88	0.96	0.82

Rabbit-intravenous The CoMFA of 37 chemicals is performed with five components: a q^2 of 0.608, an r^2 of 0.981 and an F of 323(Table 7, Figure 3). In Figure 3c, scattered yellow and green areas indicate that the changes of steric effects can help increase or decrease the activity of the molecules.

Discussion

It is essential to evaluate whether a QSAR model is valid. The "leave-one-out" cross-validation is used in the CoMFA and the cross-validation r^2 (that is q^2) from the performance is obtained to help validate the model. For the value of q^2, 1.0 corresponds to the perfect prediction while 0.0 implies that the average accuracy of the model is the same as no model at all. Thus, a CoMFA with a q^2 of about 0.5 has already been of signification to help in decision making. In the paper, according to the q^2 values of three CoMFA models, we can affirm the credibility of the results. In virtue of the models, we are able to acquire the understanding of the molecular interaction concealed in the series of molecules so as to be used to predict the potential activity of new molecules.

However, QSAR studies should be carried out on a series of chemicals that possibly act in a toxicologically similar manner. Compared with pharmacophores, it is somewhat different to identify toxicophores because of the difficulty to

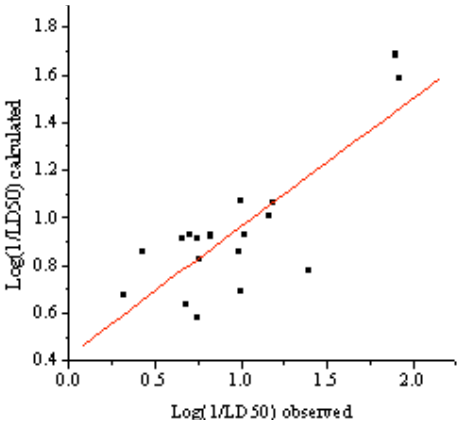


Figure 2a *Mouse-intravenous. cross-validated CoMFA analysis with four components; q^2 = 0.622*

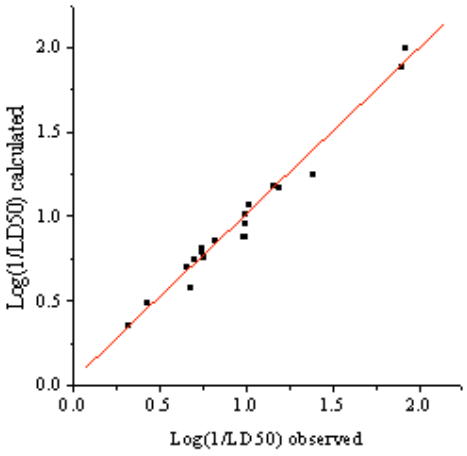


Figure 2b *final CoMFA analysis with four components; r^2 = 0.978, F = 144 (Mouse-intravenous)*

Table 6 *Mouse-intravenous: chemicals and biological activity data including observed and calculated*

No.	CAS number	LD50/10^{-3}kg kg^{-1} (observed)	log(1/LD50) (observed)	log(1/LD50) (cross-validated)	log(1/LD50) (fitted)
1	125_40_6	175	0.76	0.88	0.76
2	14077_86_2	100	1.00	0.69	1.01
3	2095_57_0	103	0.99	0.86	0.87
4	2095_58_1	150	0.82	0.92	0.86
5	21149_88_2	12.5	1.90	1.68	1.89
6	39847_06_8	210	0.68	0.63	0.57
7	480_68_2	180	0.75	0.58	0.79
8	52_43_7	218	0.66	0.92	0.70
9	66968_43_2	100	1.00	1.07	0.96
10	66968_97_6	95	1.02	0.93	1.07
11	67050_36_6	41	1.39	0.78	1.25
12	67050_51_5	375	0.43	0.85	0.49
13	67050_58_2	480	0.32	0.67	0.35
14	67050_60_6	200	0.70	0.93	0.74
15	67051_27_8	12	1.92	1.58	2.00
16	76_74_4	65	1.19	1.06	1.16
17	76_75_5	70	1.16	1.01	1.18
18	83_29_4	180	0.75	0.92	0.81

determine action targets for toxic chemicals in a database. How do we highlight different action patterns underlying these chemicals? We fish structure patterns in the pool of toxic chemicals by analysing molecular structure similarity. Each chemical is compared with the other chemicals in the database, then the chemicals having a large number of structurally neighbours pop out. It is simple and convenient to identify the features of large classes of molecules in the database.

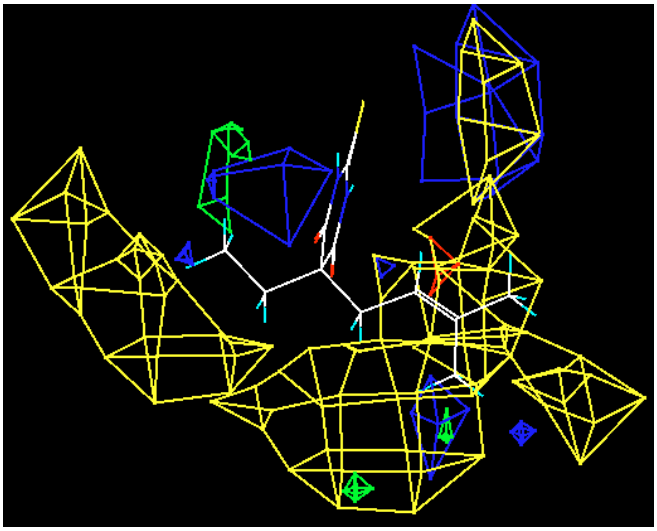
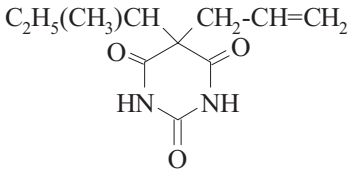


Figure 2c *Mouse-intravenous: contour map of final CoMFA model; the implication of color representation is described as in Figure 1c.*

The method by which we compute molecular structure similarity is just a two-dimensional (2D) similarity searching, compared with a variety of similarity approaches searching for a chemical database (see ref [23] for a review). But it is not only efficient but also effective to explore a large number of toxic chemicals. Firstly, similarity comparison will lead to pattern belonging of molecular structure because functional groups and the framework of a molecule determine its structure pattern. Secondly, the method compares functional group composition based on ring system framework of chemicals so that the integrity of molecules is taken account into without introducing computation complexity. Thirdly, the method intends to follow the notion that the shape and the chemical composition of chemicals have to complement those of the binding site on its receptor. The comparison of molecular frameworks intends to meet the shape requirements while the similarity of molecular groups provides some commonness of chemical composition. Thus, it may be advisable in such a flexible way to describe molecular structure patterns as rough skeletons of specific action.



Scheme 2 *The representative molecule of the structure pattern for CoMFA analysis. (WLN: T6VMVMV FHJ F2Y&1 F2U1; CAS-number: 115-44-6)*

Table 7 *Rabbit-intravenous: chemicals and biological activity data including observed and calculated*

No.	CAS number	LD50/10^{-3}kg kg^{-1} (observed)	log(1/LD50) (observed)	log(1/LD50) (cross-validated)	log(1/LD50) (fitted)
1	143_81_7	91	1.04	1.37	1.11
2	2537_29_3	1	3.00	2.65	3.04
3	39847_06_8	80	1.10	1.19	1.12
4	52_43_7	147	0.83	1.05	0.76
5	57_43_2	49	1.31	0.99	1.23
6	66843_08_1	103	0.99	1.09	0.99
7	66941_71_7	158	0.80	0.99	0.80
8	66941_87_5	73	1.14	1.11	1.11
9	66941_88_6	46	1.34	1.37	1.38
10	66941_89_7	73	1.14	1.40	1.22
11	66968_37_4	32	1.49	1.40	1.45
12	66968_38_5	65	1.19	1.62	1.31
13	66968_42_1	44	1.36	1.26	1.34
14	66968_51_2	104	0.98	1.21	0.92
15	66968_57_8	51	1.29	1.37	1.31
16	66968_58_9	61	1.21	1.06	1.20
17	66968_59_0	61	1.21	1.13	1.24
18	66968_66_9	10	2.00	1.32	1.94
19	67050_23_1	44	1.36	1.21	1.34
20	67050_69_5	58	1.24	1.09	1.25
21	67050_71_9	78	1.11	1.10	1.13
22	67050_73_1	162	0.79	1.03	0.95
23	67050_74_2	62	1.21	1.44	1.20
24	67050_75_3	53	1.28	1.10	1.27
25	67050_77_5	69	1.16	1.27	1.14
26	67050_78_6	65	1.19	1.05	1.13
27	67050_86_6	45	1.35	1.35	1.35
28	67050_87_7	53	1.28	1.22	1.18
29	67051_03_0	45	1.35	1.11	1.37
30	67051_04_1	58	1.24	1.23	1.26
31	67051_06_3	44	1.36	1.19	1.39
32	67051_23_4	110	0.96	1.29	0.89
33	67051_27_8	5	2.30	1.61	2.23
34	67114_20_9	58	1.24	1.26	1.31
35	67114_25_4	56	1.25	1.00	1.22
36	67124_90_7	124	0.91	1.03	0.93
37	76_74_4	33	1.48	1.20	1.47

As we know, the substances can interact, in a large variety of ways, with substances, tissues, and organs to cause toxic response, so it is difficult to define a set of chemical characteristics exactly that make a chemical toxic. However, in order to facilitate the understanding of toxic chemicals, helpful information should be obtained by mining the databases of toxicological data, which encode highly significant content. To make out the association among them and uncover the potential features offered by them will surely expand the effective informational content of currently available data. Based on this vision, we make an initial attempt to exploring the database RTECS. The scheme is a two-step strategy: discrimination of toxic chemicals and QSAR analysis of structure patterns. This is a compromise between efficient and effective: QSAR analysis to a database is effective but onerous; clustering of molecules is efficient but lack of accuracy. Our performance shows that the combination between the two is feasible to help us explore the databases of toxicological data.

In fact, there are many potential factors that will affect toxic effects of a chemical. Only considering action modes, toxic substances can be classified into several main categories: 1) substances that exhibit extremes of acidity, basicity, dehydrating ability, or oxidising power; 2) reactive substances that contain functional groups prone to react with biomolecules in a damaging way; 3) heavy metals; 4) lipid-

soluble compounds; 5) binding species in a reversible or irreversible way that bond to biomolecules and alter the normal function, and so on [24]. From their general modes of molecular interaction, they mainly fall into two types: specific and non-specific. Of the two types, specific interaction between molecules is considerably universal in biological systems and it is just our main concern in the present paper.

About the screening of structure patterns, there are several issues that should be noted. Firstly, the definition of structure patterns is flexible and subjective, which has advantages but also disadvantages. Secondly, molecular framework comparison is limited in common ring systems not more reasonably extended to similar ring systems. Thirdly, the exact locations of functional groups on the framework are not considered. These issues should be improved in further study. But the current method is convenient and relatively effective as an initial step to mine toxic chemicals.

Additionally, there still exist many other difficulties in the procedure of driving the mining of the databases to come into true. For example, the databases of toxicological data are always not suitable to make an analysis. This issue results from two aspects. On one side, the quality and standards for testing and representing toxic effects vary so greatly that the necessary reliability of data is difficult to reach. On the other side, the structure information of chemicals is usually not paid enough emphasis to in a database. It is obvious that the more significant the data from standardised toxicity tests are, the more rigorous and reliable the models developed are. However, in spite of these issues, it is no doubt that the advance in all related fields has already push us to start a mining of the database of toxicological data.

Acknowledgements This work is supported by the Department of Science and Technology of China, and the National Natural Science Foundation of China. We thank Liu Liang and Gao Ying for their help in using the CoMFA module and also Alan Gelberg for his helpful discussions about toxicity of chemicals.

References

1. Lewis, D.F.V. *Reviews in Computational Chemistry*, Volume III, VCH publishers: New York, 1992, 172-221.
2. Gombar, V. K.; Enslein, Kurt; Reid, D.A. *Network Science* **1996**. http://www.awod.com/netsci/Issues/Feb96/feature2.html
3. Sanderson, D.M.; Earnshaw, C.G. *Human & Exptl. Toxicol.* **1991**, *10*, 261-273.
4. Greene, N. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 148-150.
5. Woo, Y.; Lai, D.Y.; Argus, M.F.; Arcos, J.C. *Toxicol. Lett.* **1995,** *79*, 219-228.
6. TOPKAT 5.01, Health Designs: Oxford Molecular Ltd., UK. http://www.oxmol.com/prods/topkat/
7. DesJarlais, R.L.; Sheridan, R.P.; Seibel, G.L.; Dixon, J.S.; Kuntz, I.D.; Venkataraghavan, R. *J. Med. Chem.* **1988**, *31*, 722-729.

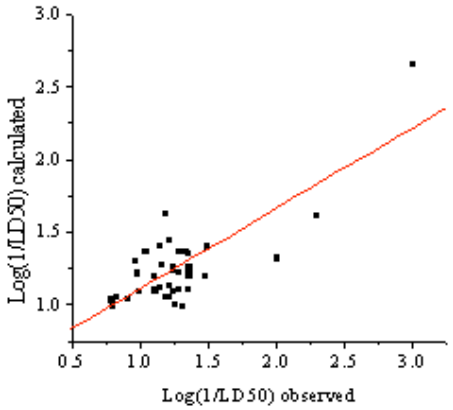


Figure 3a *Rabbit-intravenous. cross-validated CoMFA analysis with five components;* $q^2 = 0.608$

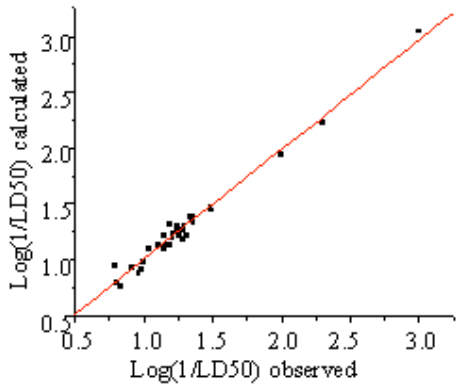


Figure 3b *final CoMFA analysis with five components;* $r^2 = 0.981$, $F = 323$ *(Rabbit-intravenous)*

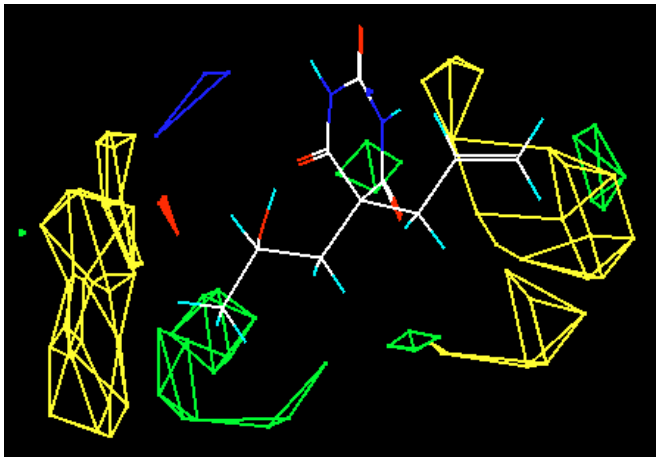


Figure 3c *Rabbit-intravenous: contour map of final CoMFA model; the implication of color representation is described as in Figure 1c*

8. Dixon, J.S. *TIBTECH* **1992**,*10*, 357-363.
9. Montgomery, J.A.; Niwas, S. *CHEMTECH* **1993**, *November*, 30-37.
10. Bugg, C.E.; Carson, W.M.; Montgomery, J.A. *Scientific American* **1993**, *December*, 60-66.
11. Kuntz, I.D.; Meng, E.C.; Shoichet, B.K. *Acc. Chem. Res.* **1994**, *27*, 117-123.
12. Verlinde, C.L.; Hol, W.G. *Structure* **1994,** *2*, 577-587.
13. Blundell, T.L. *Nature* **1996**, *384*, 23-26.
14. Bevan, D.R. *Network Science* **1996**. http://www.netsci.org/Science/Compchem/feature12.html
15. Coats, E.A. *Network Science* **1996.** http://www.netsci.org/Science/Compchem/feature13.html
16. Richon, A.B. *Network Science*, **1997**. http://www.netsci.org/Science/Compchem/feature19.html
17. Cramer, R.D.; Patterson, D.E.; Bunce, J.D. *J. Am. Chem. Soc.* **1988**, *110*, 5959-5967.
18. Cramer, R.D.; DePriest, S.A.; Patterson, D.E. In *3D QSAR in drug design: Theory, Methods, and Application*; Kubinyi, H., Ed.; ESCOM: Leiden, 1993; 443-485.
19. RTECS C2(96-4), National Institute for Occupational Safety and Health (NIOSH), U.S. Department of Health and Human Servies, 1996. http://www.ccohs.ca/
20. Smith, E.G.; Baker, P.A. *The Wisswesser Line-Formula Chemical Notation (WLN)*, third ed., Chemical Information Management: Cherry Hill, NJ, 1975.
21. SYBYL 6.4, Tripos Associates: St. Louis, MO, USA. http:// www.tripos.com/
22. Wang, R.X.; Gao, Y.; Liu, L.; Lai, L. *J. Mol. Model.* **1998**, *4*, 276-283.
23. Willett, P. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 983-996.
24. Manahan, S.E. *Toxicological chemistry*, second ed., Lewis Publishers: Michigan, 1992; chapter 9, p 217-247.