1173

Structural Features of Toxic Chemicals for Specific Toxicity

Jiansuo Wang, Luhua Lai,* and Youqi Tang

Institute of Physical Chemistry, Peking University, Beijing 100871, People's Republic of China

We have studied the structural features of toxic chemicals from the RTECS database associated with specific toxicity. The frameworks, functional groups, and structure patterns of molecules are taken into account. Potential active frameworks, groups, and structure patterns for specific toxicity are gained by computational chemistry approaches. These structural features of toxic chemicals will be helpful in understanding activities of toxic chemicals and useful in predicting the toxicity of chemicals, especially in the early stage of drug design.

I. INTRODUCTION

Judging potential toxicity of chemicals from their structure is what toxicologists desire. It is also of concern to the chemists who are working in the field of drug design. However, because the behavior of toxicants involves all aspects of their interaction with the biological system, from absorption, distribution, and metabolism to ultimate reaction and a cascade of biochemical and phenomenological changes, it is a challenging task to determine the structural features conferring toxicity to chemicals.

Efforts to predict the potential risk of chemicals based only on the structure information or easily available data have been made since the early 1980s.[1] The knowledge about structural features of toxic chemicals can be classified into three categories: (1) rules distilled from available knowledge and human experts, which are used in the knowledge-based expert systems such as DEREK,[2] OncoLogic,[3] etc.; (2) QSAR (quantitative structure−activity relationship) models, which are introduced to extensively model biological activity and utilized in the predictive systems such as TOPKAT,[4] etc.; (3) the findings from other technologies such as pattern cognition, neural network, and so on.[5,6] For example, Klopman defined chemical substructures capable of producing a toxic response as "toxicophores" and determined them by a comparative analysis between active chemicals and inactive chemicals.[7,8]

However, much effort is needed to understand the property of toxic chemicals. In the present paper, we try to make a systematic analysis of toxic chemicals from the RTECS database on a compact disk. RTECS[9] (The Registry of Toxic Effects of Chemical Substances) is a database of toxicological information compiled, maintained, and updated by the National Institute for Occupational Safety and Health (NIOSH). It now contains information on over 130 000 chemicals extracted from the open scientific literature, of which 58 682 are represented in the Wiswesser line notation (WLN).[10] Due to the need for structure information, we only use this subset of 58 682 chemicals in the present study. In terms of the type of toxicity data, we classify these chemicals into different data sets of specific toxicity, such as abnormality, tumor, mutation, and some organ or system toxicities
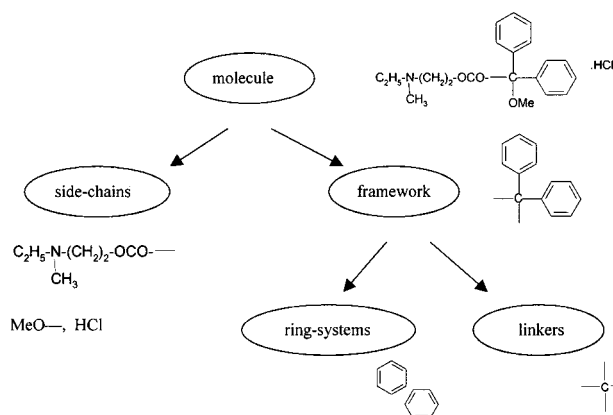
* Corresponding author. Fax: +86-10-62751725. E-mail: lai@ipc.pku.edu.cn.

Figure 1. Hierarchical description of molecules adopted by Bemis and Murcko[11] (one molecule is shown as a sample breakdown).

(Table 9 in the Appendix). Then we study the relationship between specific toxicity and frameworks, functional groups, and structure patterns of toxic chemicals. Thus, toxicity here means some form of specific biological activity that can be distinguished from the background of the full database. Compared with the cases that distinguish "toxic" from "nontoxic" chemicals, such specific toxicity is potentially more suitable for SAR investigation from a mechanistic standpoint.

II. FRAMEWORKS

Methods. We take an approach similar to that of Bemis and Murcko to dissect any molecule into four units: ring, linker, framework, and side chain (Figure 1).[11] Ring systems are defined as cycles or cycles sharing an edge within the graph representation of molecules; linkers are atom paths connecting two ring systems; ring systems and relevant linkers of a molecule constitute a framework. Frameworks can be viewed as basic skeletons of molecules; side chains include any nonframework atoms. Such disintegration of molecules is expected to help give prominence to basic molecular skeletons. From the WLN representation of molecules, we can obtain the molecular frameworks by removing nonlinker parts out of ring systems by complying with the rules and characteristics of the WLN representation.

We define the "distribution coefficient (DC)" of a framework as the percentage ratio of its occurrence frequency to

Table 1. Composition Frameworks in the Database RTECS and Their Functional Coefficients in Various Data sets[a]

frameworks	DC/%	abnormal	auto-nomic	blood	brain	cardiac	CNS	endo-crine	gas intestines	glands	kidney	liver	lung	mutation	skin	tumor	vascular
	154.22	0.78	0.84	1.07	0.92	3.43	0.85	1.02	1.15	1.58	1.01	1.00	0.90	0.99	1.15	0.82	0.97
	10.46	1.00		0.73	0.51			0.71	0.94	1.03	0.51	1.13	0.56	0.91	0.71	0.74	1.54
	10.09	0.62	1.00	0.94		1.88		0.84	0.86	0.78	1.44	1.13		1.23	0.84	0.83	
	7.77			0.98	0.63	2.04			0.75	1.26	1.52	1.02		0.51	1.77	1.39	0.73
	6.95	0.70	1.27		0.63	4.55			1.36	0.85	0.77	0.51		0.23	0.40	0.35	
	5.20	0.80							0.91		1.24	0.93		2.69	0.73	2.67	
	4.53	1.07	1.67	0.94		5.89			1.28	1.08	0.83	0.68		0.67		0.49	
	4.10	2.77		0.81		3.85			0.79	0.72	0.92	1.07	1.71	2.05	0.46	1.02	5.06
	3.39								2.03	1.16		1.04		1.31	1.09	1.30	
	3.39				2.08		2.56		0.75								
	2.83	1.23	2.18	0.84		3.59			0.77		1.14			0.49	0.45	0.39	
	2.73								1.60	2.96		1.45	4.72	4.88		2.83	
	2.51								1.01		1.05			0.51		0.79	1.02
	2.44			0.97			3.05		1.93		1.99	5.23	2.88	5.27	4.16	5.60	
	2.33								1.40			1.88		1.31	1.09	1.04	
	2.30			1.44					1.42	2.56				1.12	0.83	0.67	
	2.20								1.15		2.23			0.87			
	1.98			1.20			2.69		0.92	4.36		3.55		3.30	2.25	3.12	
	1.98			3.12					1.10		2.18	2.00			3.12	1.45	
	1.91													1.34			
	1.72													4.20		0.64	
	1.62	3.86												2.45		0.68	
	1.62			1.47					1.57			2.17		1.59	1.57	1.63	
	1.60			2.37					1.36		3.03	4.66	3.66	7.48	1.45	1.24	
	1.53	2.27		1.55					2.36		1.76	1.72	3.82	2.74	1.24	4.02	
	1.45									2.71	2.97					1.82	
	1.43													0.81			

Table 1. Composition Frameworks in the Database RTECS and Their Functional Coefficients in Various Data sets[a]

frameworks	DC/%	abnormal	auto-nomic	blood	brain	cardiac	CNS	endo-crine	gas intestines	glands	kidney	liver	lung	mutation	skin	tumor	vascular
[N — pyrrolidine ring]	1.31													1.42		1.01	
[caffeine-like ring, O, N]	1.26								1.44	2.33			4.64				
[O — tetrahydrofuran]	1.26												0.84				
[cyclohexene ring]	1.26	3.30		2.26					1.72			2.09	4.36			2.10	
[coumarin, O=C–O]	1.24														1.02		
[indole, N]	1.24											2.16					6.84
[phthalimide, O, N, O]	1.19	5.82													1.06		
[benzothiazole, S, N]	1.16			2.05										1.21	1.28		
[imidazole, H, N, N]	1.09			2.61					2.33					3.21	0.97	2.42	
[fused polycyclic (steroid)]	1.06													9.50		3.20	7.92
[dioxolane, O–O]	1.06														1.03		
[aziridine, N]	1.01													2.09		2.19	
[benzimidazole, N, N]	0.95				4.62												

[a] The listed frameworks exist in at least 50 toxic chemicals in the database RTECS; the blank entries indicate that the corresponding frameworks occur in less than five chemicals in specific-toxicity data sets. In the table, DC = (occurrence frequency/total chemical number) × 100%; the total number is 58 682.

the total number of chemicals in a database or data set and define "functional coefficient (FC)" of a framework as the ratio of its distribution coefficient in a specific toxicity subset of the database to its distribution coefficient in the whole database. The functional coefficient of a framework is always pertinent to some kind of activity. When the FC values of a framework in all the various data sets are always close to 1.00 (here, we set the range as "<10.00"), the framework is defined as a "composition framework". A composition framework means that the framework is uniformly scattered in the whole database so that it has no relation with any specific activity and just acts as one part of the molecular composition. On the other hand, the framework is regarded as an "activity framework" to this kind of specific toxicity when its FC value in a specific activity data set is much higher than 1.00 (here, we set the limit as >10.00). An "activity framework" indicates that the framework is concentrated in some activity subset of the database and may account for some kind of specific toxicity.

Results. We have studied the relationship between specific toxicity and molecular frameworks. The functional coefficients of frameworks in various data sets help to distinguish composition frameworks and activity frameworks (Tables 1 and 2).

Table 1 lists composition frameworks that exist in at least 50 chemicals in the RTECS database. From their functional

coefficients in various data sets, we can determine if these frameworks are uniformly distributed in data sets of different activity, implying that they are not related to any kind of activity. Such frameworks could be regarded as potentially "safe" in the absence of toxicity-conferring functional groups to be used as fragments in drug design. This is confirmed by the findings of Bemis and Murcko where the high-frequency fragments occurring in drugs are just the composition frameworks we have identified in the present study.[11]

Table 2 displays some potential active frameworks containing structural features for specific toxicity. The frameworks under consideration exist in at least five toxic chemicals and have a FC value of 10.00 or greater. For example, the frameworks for mutation and tumor activity are mainly aromatic fused cycles, N heterocycles, or those containing epoxy ethane, whereas the frameworks for autonomic nervous system toxicity display different structures, such as O heterocycles, N heterocycles, or P heterocycles. We postulate that every type of toxicity will have characteristic frameworks. CNS activity relates to relatively more N heterocycles, cardiac activity associates with some chemicals containing bridged rings, whereas some frameworks for liver activity contain aromatic dinitrogen, and some frameworks for lung activity have the fragment of epoxy ethane.

Table 2. Potential Active Frameworks for Specific Toxicity from the Database RTECS
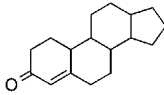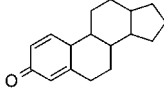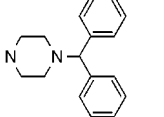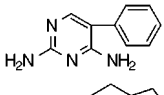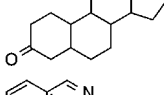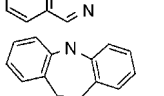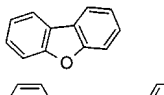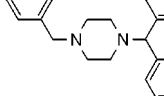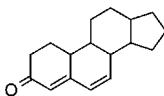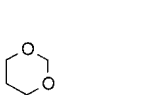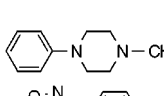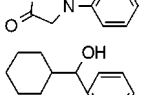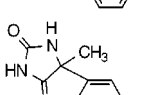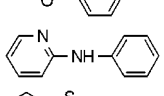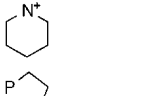
frameworks	frequency specific	DC specific/%	frequency total	DC total/%	functional coeff
		Abnormality			
(structure)	27	18.76	64	1.09	17.20
(structure)	18	12.51	29	0.49	25.31
(structure)	6	4.17	12	0.20	20.39
(structure)	6	4.17	10	0.17	24.47
(structure)	5	3.47	14	0.24	14.56
(structure)	5	3.47	13	0.22	15.68
(structure)	5	3.47	19	0.32	10.73
(structure)	5	3.47	19	0.32	10.73
(structure)	5	3.47	5	0.09	40.78
(structure)	5	3.47	10	0.17	20.39
		Autonomic Nervous System			
(structure)	45	56.89	60	1.02	55.64
(structure)	24	30.34	25	0.43	71.22
(structure)	11	13.90	13	0.22	62.77
(structure)	8	10.11	10	0.17	59.35
(structure)	8	10.11	9	0.15	65.94
(structure)	8	10.11	19	0.32	31.24
(structure)	8	10.11	14	0.24	42.39
(structure)	7	8.85	11	0.18	47.21
(structure)	6	7.59	39	0.66	11.41
(structure)	5	6.32	5	0.09	74.19
(structure)	5	6.32	5	0.09	74.19

STRUCTURAL FEATURES OF TOXIC CHEMICALS

J. Chem. Inf. Comput. Sci., Vol. 39, No. 6, 1999 **1177**

Table 2. (Continued)
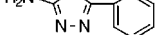
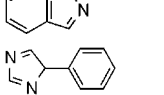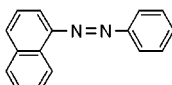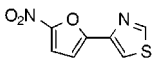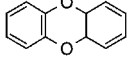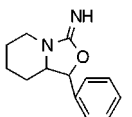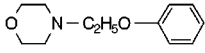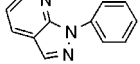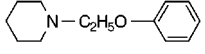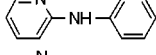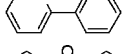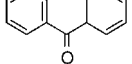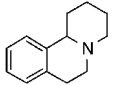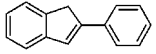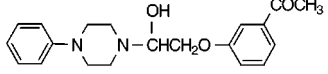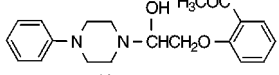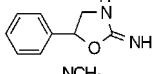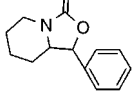| frameworks	frequency specific	DC specific/%	frequency total	DC total/%	functional coeff
		Blood			
	8	3.80	20	0.34	11.14
	7	3.32	10	0.17	19.50
	7	3.32	19	0.32	10.26
		Brain and Covering/CNS			
	18/18	15.87/22.30	18	0.31	51.75/72.72
	13/13	11.46/16.11	37	0.63	18.18/25.55
	11/11	9.70/13.63	11	0.19	51.75/72.72
	11/11	9.70/13.63	27	0.46	21.08/29.63
	8/8	7.05/9.91	19	0.32	21.79/30.62
	8/8	7.05/9.91	14	0.24	29.57/41.55
	8/8	7.05/9.91	19	0.32	21.79/30.62
	7/7	6.17/8.67	18	0.31	20.12/28.28
	7/7	6.17/8.67	7	0.12	51.75/72.72
	7/7	6.17/8.67	18	0.31	20.12/28.28
	6/6	5.29/7.43	7	0.12	44.36/62.33
	5/5	4.41/6.20	7	0.12	36.96/51.94
	5/5	4.41/6.20	5	0.09	51.75/72.72
	5/5	4.41/6.20	11	0.19	23.52/33.05
		Cardiac			
	16	50.63	94	1.62	31.28
	8	25.32	21	0.36	70.74
	6	18.99	7	0.12	159.17
	6	18.99	84	1.43	13.26
	6	18.99	7	0.12	159.17

Table 2. (Continued)
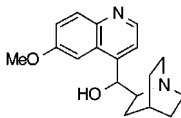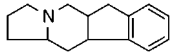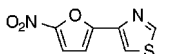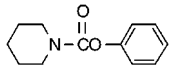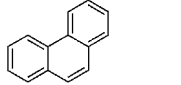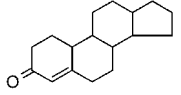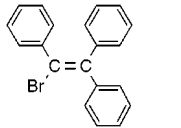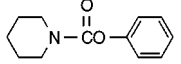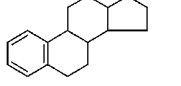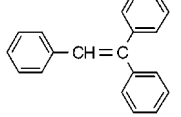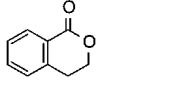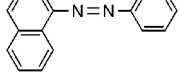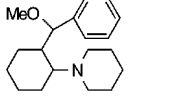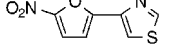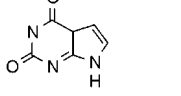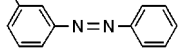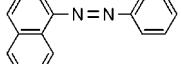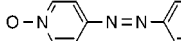
frameworks	frequency specific	DC specific/%	frequency total	DC total/%	functional coeff
		Cardiac (Continued)			
	5	15.82	21	0.36	44.21
		Gastrointestinal			
	10	3.62	10	0.17	21.27
	8	2.90	10	0.17	17.02
	8	2.90	8	0.14	21.27
	5	6.20	35	0.60	10.39
		Glands			
	19	18.65	64	1.09	17.10
	9	8.83	9	0.15	57.59
	8	7.85	8	0.14	57.59
	8	7.85	21	0.36	21.94
	5	4.91	5	0.09	57.59
		Kidney, Ureter, Bladder			
	10	5.38	12	0.20	26.33
	8	4.31	20	0.34	12.64
	7	3.77	7	0.12	31.60
	6	3.23	10	0.17	18.96
	5	2.69	5	0.09	31.60
		Liver			
	46	20.20	101	1.72	11.74
	10	4.39	16	0.27	16.11
	8	3.51	20	0.34	10.31
	6	2.64	6	0.10	25.77

Table 2. (Continued)
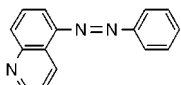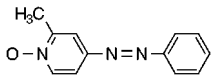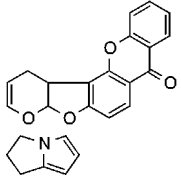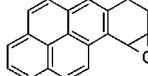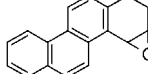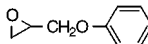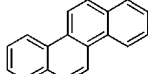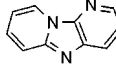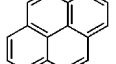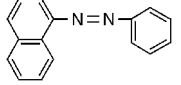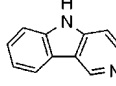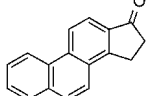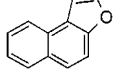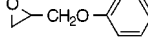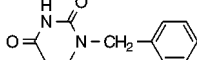
frameworks	frequency specific	DC specific/%	frequency total	DC total/%	functional coeff
		Liver			
	6	2.64	7	0.12	22.09
	5	2.20	5	0.09	25.77
	5	2.20	5	0.09	25.77
	5	2.20	5	0.09	25.77
		Lung			
	20	23.42	80	1.36	17.18
	10	11.71	62	1.06	11.08
	8	9.37	8	0.14	68.71
	6	7.03	6	0.10	68.71
	5	5.85	15	0.26	22.90
	5	5.85	13	0.22	26.43
		Mutation			
	18	4.20	22	0.37	11.20
	16	3.73	16	0.27	13.69
	16	3.73	19	0.32	11.53
	15	3.50	20	0.34	10.27
	14	3.27	15	0.26	12.78
	14	3.27	15	0.26	12.78
	14	3.27	15	0.26	12.78
	11	2.57	11	0.19	13.69
	11	2.57	11	0.19	13.69
	10	2.33	13	0.22	10.53
	10	2.33	10	0.17	13.69

Table 2. (Continued)
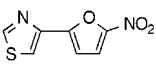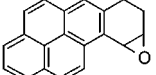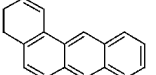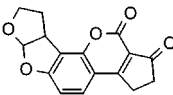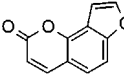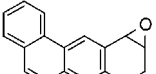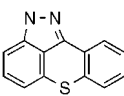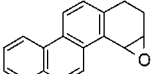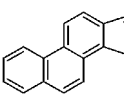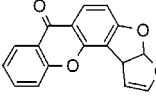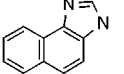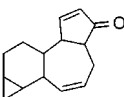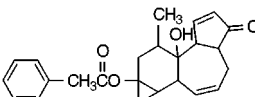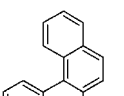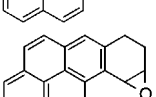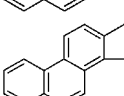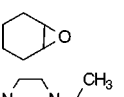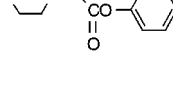
frameworks	frequency specific	DC specific/%	frequency total	DC total/%	functional coeff
		Mutation (Continued)			
	9	2.10	10	0.17	12.32
	8	1.87	8	0.14	13.69
	7	1.63	7	0.12	13.69
	7	1.63	7	0.12	13.69
	7	1.63	7	0.12	13.69
	7	1.63	9	0.15	10.65
	6	1.40	6	0.10	13.69
	6	1.40	6	0.10	13.69
	6	1.40	6	0.10	13.69
	5	1.17	5	0.09	13.69
	5	1.17	5	0.09	13.69
	5	1.17	5	0.09	13.69
		Skin			
	16	3.38	16	0.27	12.41
	13	2.75	13	0.22	12.41
	11	2.33	11	0.19	12.41
	7	1.48	8	0.14	10.86
	6	1.27	7	0.12	10.63
	6	1.27	7	0.12	10.63
	6	1.27	6	0.10	12.41
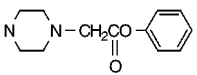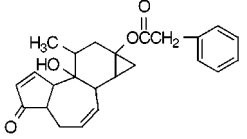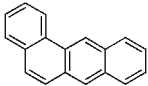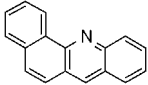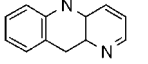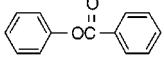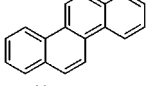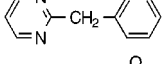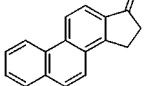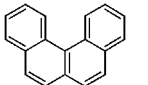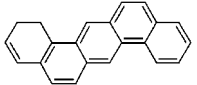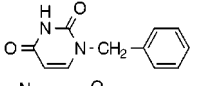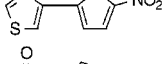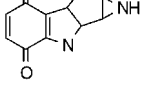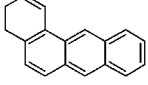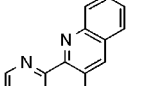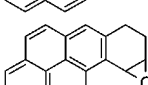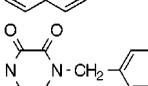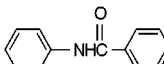
STRUCTURAL FEATURES OF TOXIC CHEMICALS

J. Chem. Inf. Comput. Sci., Vol. 39, No. 6, 1999 **1181**

Table 2. (Continued)

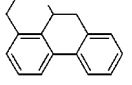| frameworks	frequency specific	DC specific/%	frequency total	DC total/%	functional coeff
		Skin (Continued)			
	6	1.27	6	0.10	12.41
	5	1.06	5	0.09	12.41
		Tumor			
	118	25.99	128	2.18	11.92
	23	5.07	29	0.49	10.25
	22	4.85	27	0.46	10.53
	19	4.19	23	0.39	10.68
	18	3.96	22	0.37	10.58
	15	3.30	19	0.32	10.20
	12	2.64	15	0.26	10.34
	11	2.42	11	0.19	12.93
	11	2.42	13	0.22	10.94
	11	2.42	12	0.20	11.85
	9	1.98	10	0.17	11.63
	9	1.98	11	0.19	10.58
	8	1.76	10	0.17	10.34
	7	1.54	7	0.12	12.93
	7	1.54	8	0.14	11.31
	6	1.32	6	0.10	12.93
	6	1.32	6	0.10	12.93

Table 2. (Continued)
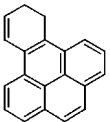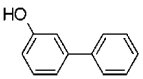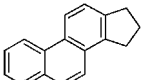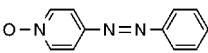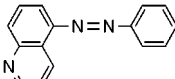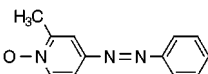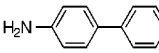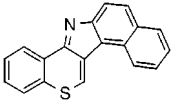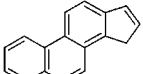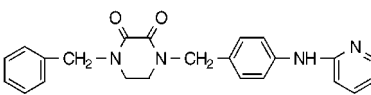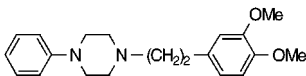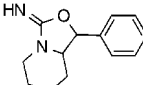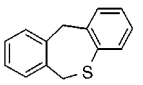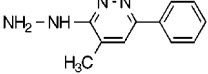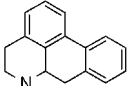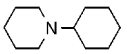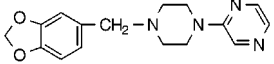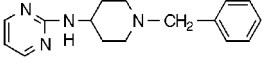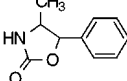
frameworks	frequency specific	DC specific/%	frequency total	DC total/%	functional coeff
		Tumor (Continued)			
	6	1.32	7	0.12	11.08
	6	1.32	7	0.12	11.08
	6	1.32	7	0.12	11.80
	6	1.32	6	0.10	12.93
	6	1.32	7	0.12	11.08
	5	1.10	5	0.09	12.93
	5	1.10	6	0.10	10.77
	5	1.10	5	0.09	12.93
	5	1.10	5	0.09	12.93
	5	1.10	5	0.09	12.93
		Vascular			
	24	22.68	25	0.43	53.25
	18	17.01	18	0.31	55.47
	12	11.34	36	0.61	18.49
	9	8.51	15	0.26	33.28
	9	8.51	9	0.15	55.47
	8	7.56	21	0.36	21.13
	7	6.62	23	0.39	16.88
	7	6.62	13	0.22	29.87
	7	6.62	10	0.17	38.83
	5	4.73	5	0.09	55.47
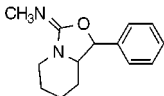
STRUCTURAL FEATURES OF TOXIC CHEMICALS

J. Chem. Inf. Comput. Sci., Vol. 39, No. 6, 1999 **1183**

Table 2. (Continued)

frameworks	frequency specific	DC specific/%	frequency total	DC total/%	functional coeff
		Vascular (Continued)			
	5	4.73	11	0.19	25.21

Table 3. Comparison of Occurrence of Unit Groups in Mutation Set and the Whole Database

group	distribution coeff/% total	distribution coeff/% mutation	group	distribution coeff/% total	distribution coeff/% mutation
Br	7.74	6.49	$-$OH	50.65	58.45
F	12.51	4.95	phenyl	66.44	41.55
Cl	44.77	35.65	S	26.95	20.77
H	33.36	14.47	$=$	29.68	24.13
I	7.08	1.47	C$=$O	76.93	57.40
N$^+$	11.73	6.74	OO	11.88	22.35
$-$NH$-$	47.62	38.94	X	14.02	4.36
$-$N$-$	105.37	109.71	Y	38.03	20.53
O	87.76	82.99	$-$NH$_2$	16.87	23.54
P	6.54	7.00			

In summary, Tables 1 and 2 listing composition frameworks and activity frameworks show that the frameworks with high frequency in the general database always are part of molecule composition, whereas the frameworks with relatively low frequency in the general database but high concentration in specific data sets are more likely to be responsible for specific activity.
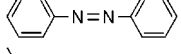
III. FUNCTIONAL GROUPS

Methods. We have examined functional groups in the toxic chemicals by a group-growing method. Firstly, we define 19 unit groups: X (C attached to four non-hydrogen atoms), Y (C attached to three non-hydrogen atoms), phenyl, $-$H, $-$N$-$ (N attached to three non-hydrogen atoms), $-$NH$-$, $-$NH$_2$, N$^+$ (N charged positive), $-$O$-$, $-$C$=$O, OO (representing "O$_2$" in molecules such as $-$SO$_2$, $-$NO$_2$ etc.), $-$OH, S, P, F, Cl, Br, I, and multiple bonds ("$=$" is double bond, "$\equiv$" is triple bond) using WLN notation. Because C, H, O, N, S, P, and halogens are major composition elements

of drugs and toxicants, we take their usually existing forms as unit groups. Here, we ignore such atoms as C and H of hydrocarbons. Then we examine group combination during the growing procedure where unit groups are connected together. DCs and FCs of groups are defined similarly to those of frameworks to help distinguish activity groups from composition groups.

Results. First we examine the association of groups with specific activity during the group-growing course. The state of mutation activity is displayed as an example and enlightens us to the possible magnitude of the hit groups. Table 3 and Table 4 indicate that the distribution of unit groups or two-group combinations is decentralized and nearly uniform. Hence, the groups tend to be part of molecular composition. Additionally, the occurrence frequency of the groups combined by more than six unit groups is low. Therefore, we take medium-size groups (three$-$five group combination) into consideration. Interestingly, this is somewhat reminiscent of the CASE approach[7] in which fragments ranging from 2 to 10 heavy atoms are candidates for "biophores".

We listed the potential active groups by four-unit-group combination (Table 5). The groups are with a higher frequency than 10 (i.e., contained in at least 10 chemicals) and a higher functional coefficient than 5.00. The results show that this medium-size combination of unit groups displays some association with specific activity. We find that N atoms are rich in the listed groups, especially N connected to aromatic cycles or unsaturated bonds. For example, dinitrogen groups as the feature to tumor or mutation toxicity have been confirmed by the information of known cancinogens or mutagens.[12] It can be anticipated that more information can be obtained from relative arrangement of several groups in three-dimensional space.

Table 4. Comparison of Combination Occurrence of Two Unit Groups in the Mutation Set and the Whole Database

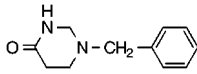| FC[a]	Br	F	Cl	H	I	N$^+$	$-$NH$-$	N$-$	O	P	$-$OH	phenyl	S	$=$	C$=$O	OO	X	Y	$-$NH$_2$
Br	1.90	0.00	0.00	0.15	0.00	0.00	0.00	0.00	2.35	0.00	0.00	0.48	0.00	1.03	0.00	0.00	0.00	2.97	0.00
F	1.49	0.21	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.42	0.00	0.00	0.00	0.00	0.15	0.00	0.00
Cl	1.37	0.00	1.71	0.24	0.00	0.00	0.00	0.00	1.30	0.00	0.00	0.53	0.00	3.28	1.40	0.00	1.03	2.12	0.00
H	0.00	0.00	0.67	1.96	0.00	0.00	0.00	2.74	0.00	0.00	0.00	0.91	0.00	0.00	0.00	0.00	0.00	1.89	0.00
I	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.07	0.00	0.00	0.00	0.00	0.00	0.00	0.00
N$^+$	0.00	0.00	2.11	0.00	0.00	0.00	1.91	0.70	1.30	0.00	4.29	0.89	4.17	1.31	1.97	0.00	0.00	0.00	3.19
$-$NH$-$	0.00	0.00	1.51	0.00	0.00	0.00	0.72	1.11	1.95	0.65	3.50	0.68	2.06	0.40	0.89	0.00	0.25	0.55	2.01
$-$N$-$	0.00	0.00	0.00	0.00	2.28	0.00	1.34	1.76	3.93	1.74	1.77	2.44	0.47	2.19	0.49	3.03	0.00	0.57	2.11
O	0.00	9.96	0.00	6.32	0.00	0.00	1.68	4.08	1.39	0.84	3.23	0.32	1.84	0.75	0.62	0.00	0.56	0.53	0.00
P	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.55	0.91	0.00	1.58	0.00	0.65	0.00	0.00	0.00	0.00	0.00	0.00
$-$OH	0.00	0.00	0.00	0.42	0.00	0.00	3.66	1.15	1.28	0.56	1.47	0.60	0.00	2.21	0.52	0.00	0.00	0.84	0.00
phenyl	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.63	2.33	0.00	3.10	0.00	0.00	3.60	1.38	0.00	0.00	0.00	0.00
S	0.00	0.00	0.00	0.61	0.00	0.00	0.00	0.82	0.75	0.64	0.67	0.28	0.34	1.18	0.63	1.32	0.89	0.47	0.00
$=$	0.00	0.00	0.00	0.00	0.00	5.56	0.59	1.75	0.00	0.00	0.00	0.00	0.68	0.31	0.00	0.00	0.00	0.56	0.00
C$=$O	0.00	0.00	0.00	1.78	0.00	0.00	0.59	0.69	0.64	0.00	0.55	0.34	0.42	0.00	0.37	0.00	0.25	0.45	0.97
OO	0.00	0.00	0.00	0.00	0.00	0.00	0.59	2.10	2.51	0.00	1.34	0.44	0.56	0.00	0.00	0.00	0.00	0.00	0.00
X	0.99	0.22	1.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.08	0.21	0.00	0.00	0.00	0.00	0.00	0.00	1.49
Y	2.48	0.62	2.20	0.00	0.00	0.00	1.00	0.22	0.62	0.00	0.58	0.21	0.57	0.64	0.88	0.00	0.00	0.00	0.68
$-$NH$_2$	0.00	0.00	0.00	1.65	0.00	0.00	0.89	2.55	1.64	0.00	0.00	2.89	0.32	0.65	1.29	0.00	0.00	0.35	0.00

[a] FC, functional coefficient of a group, is the ratio of the distrihution coefficient of the group in the mutation set to that in the database.

Table 5. Potential Active Groups for Specific Toxicity

groups and unit-group sequence of group combination									
group	no. 1	no. 2	no. 3	no. 4	freq specific	DC specific/%	freq total	DC total/%	FC
Autonomic									
phenyl−COO−NH−	−NH−	C=O	O	phenyl	13	16.44	109	1.86	8.85
−N−N−O−(C=O)−	−N−	−N−	O	C=O	12	15.17	31	0.53	28.72
HOOC-CH(O)H)-	−OH	C=O	Y	−OH	11	13.91	140	2.39	5.83
−(C=O)−C(OH)−CH−	C=O	X	−OH	Y	14	17.70	28	0.48	37.09
−C(OH)−CH(OH)−	X	−OH	Y	−OH	12	15.17	12	0.20	74.19
Endocrine									
−NH−SO$_2$.phenyl	−NH−	S	OO	phenyl	19	20.17	184	3.14	6.43
−NH−(C=C))−(C=O)−NH−	−NH−	C=O	C=O	−NH−	14	14.86	66	1.13	13.21
Gastrointestinal									
−NH−(C=O)−O−phenyl	−NH−	C=O	O	phenyl	42	15.23	109	1.86	8.20
−O−(C=O)−CH−phenyl	O	C=O	X	phenyl	31	11.24	75	1.28	8.79
Glands									
−NH−(C=O)−O−phenyl	−NH−	C=O	O	phenyl	42	41.22	109	1.86	22.19
−C(Br)=CH−	Y	Br	=	Y	19	18.65	20	0.34	54.71
Kidney									
H−SO$_2$−O−	H	S	OO	O	11	5.92	29	0.49	11.99
Liver									
−N=N−phenyl	−N−	=	−N−	phenyl	185	81.25	451	7.69	10.57
Mutation									
−N=N−N−	−N−	=	−N−	−N−	25	5.83	67	1.14	5.11
−N−CH−NH−(C=O)−	−N−	Y	−NH−	C=O	11	2.57	17	0.29	8.86
−O−SO$_2$−O	O	S	OO	O	13	3.03	35	0.60	5.09
−(C=O)−NH−O−(C=O)−	C=O	−NH−	O	C=O	10	2.33	11	0.19	12.45
H$_2$N−(C=O)−N−N−	−NH$_2$	C=O	−N−	−N−	17	3.97	26	0.42	8.95
Skin									
H$_2$N−(C=O)−N−N−	−NH$_2$	C=O	−N−	−N−	11	2.33	26	0.42	5.25
Tumor									
−NH−CH−(C=O)−O−	−NH−	Y	C=O	O	17	3.74	37	0.63	5.94
−N=N−phenyl	−N−	=	−N−	phenyl	212	46.70	451	7.69	6.08
−N−(C=O)−(C=O)−N−	−N−	C=O	C=O	−N−	19	4.19	22	0.38	11.16
−(C=O)−NH−P−NH$_2$	C=O	−NH−	P	−NH$_2$	11	2.42	14	0.24	10.16
H$_2$N−(C=O)−N−N−	−NH$_2$	C=O	−N−	−N−	13	2.86	26	0.44	6.46

Table 6. Potential Structure Patterns for Specific Toxicity (Given 0.6 as Similarity Limit, 10 as Count Limit)

freq	CAS no.	WLN	freq	CAS no.	WLN
Abnormality					
16	521-18-6	LE5 B666 OVTJ A1 E1 FQ	11	117-81-7	4Y2&1OVR BVO1Y4&2
12	382-67-2	L E5 B666 OV AHTTT&J A1 BF CQ E1 FV1Q G1	10	94-75-7	QV1OR BG DG
11	2698-38-6	GR C1 DO1VO2	10	155-91-9	T6N CNJ BMSWR DZ&DO1 FO1
11	25389-94-0	T6OTJ B1Z CQ DQ EQ FO- BL6TJ AZ CQ EZ DO- BT6OTJ CQ DZ EQ F1Q&...H2...S-O4			
Autonomic					
34	22644-66-2	T6O COTJ BY1&1 EY1&O1 E2U1	11	51071-52-4	T6NJ CVMR DO1
21	1049-81-6	T6N DNTJ A2R CO1 DO1&DR BQ	10	82058-41-1	T66 AAV DM HNTJ CR BG&ER BG&GR BG&HVM2 IR BG
20	2032-54-4	T6N DNTJ A2OR DG&D- BT6NJ	10	695-63-6	T5P BUTJ AO2 AO
19	64059-39-8	4K2&2&2O2K2&1&1&Q 2&E 2	10	7614-53-1	T5NNOVJ AR DVQ
19	63982-53-6	4K4&1&R COVM&I			
Blood					
13	924-16-3	ONN4&4	11	614-45-9	1X1&1&OOVR
12	97-00-7	WNR BG ENW	10	3374-04-7	QVY1&N2G2G&GH -R
12	55112-89-5	G2N4&2G&GH	10	99-09-2	ZR CNW
11	3653-48-3	OV1OR DG B1&-NA-	10	58-71-9	T46 ANV ES GUTJ G1OV1 HVO CMVl- BT5SJ&-NA-
11	76-43-7	L E5 B666 OV MUTJ A1 BF CQ E1 FQ F1	10	98-07-7	GXGGR
11	108-42-9	ZR CG	10	54220-95-0	ZR B1 F1 DO2
Brain					
31	63990-66-9	T6N DNTJ AlYQ1OR BV1&DR BG	11	34580-64-8	T56 BNN INJ BR CXTFFI&DO1VQ
22	7716-73-6	2N2&2OR CQ DV1&GH	11	75343-60-1	T56 ANYOTJ BUN4 DR -DL -C&GH
22	20800-09-3	T6N DOTJ A2OR CQ DV1&GH	11	75343-67-8	T56 ANYOTJ BUM DR CG -T&EH
20	52582-80-6	L66 B6/B-H/DI A B- C 1B ITJ B3MX1&1&1&GH	10	52673-65-1	L66 B6/B-H/DI A B- C lB ITJ B2N1&1 D2N1&l&1&GH 2
17	23771-09-7	T6N DNTJ A3VR CO1 DO1 EO1&DR CF&GH	10	31897-81-1	L66 B6/B-H/DI A B- C 1B ITJ B1Yl&N1&1 CG&GH
13	99901-13-0	T6VMVMV FHJ F2UY1&1 F2U1			
Cardiac					
20	75883-40-8	T56 BOJ FOl GMVl IO1 HO2M- AL6TJ&GH&QH	10	476-70-0	T C6666 1A Q KN&TT&J EO1 FQ K1 PQ QO1
18	102585-30-8	L66 A BTJ AMV1N2&2&GH	10	13062-95-8	Z1Y1&R COl&GH
14	75883-59-9	T56 BOJ FO1 GMVM1 IO1 HO2- AT4NTJ&QVVQ&QH 4/3	10	58-55-9	T56 BM DN FNVNVJ Fl H1
11	67427-59-2	MUYZM1YQ1MR 2&...H2...S-O4			

STRUCTURAL FEATURES OF TOXIC CHEMICALS

J. Chem. Inf. Comput. Sci., Vol. 39, No. 6, 1999 **1185**

Table 6. (Continued)

freq	CAS no.	WLN	freq	CAS no.	WLN
		CNS			
31	63990-66-9	T6N DNTJ A1YQ1OR BV1&DR BG	11	99901-13-0	T6VMVMV FHJ F2UY1&1 F2U1
24	7716-73-6	2N2&2OR CQ DV1&GH	11	75343-60-1	T56 ANYOTJ BUN4 DR -DL -C&GH
22	20800-09-3	T6N DOTJ A2OR CQ DV1&GH	11	75343-67-8	T56 ANYOTJ BUM DR CG -T&EH
20	52582-80-6	L66 B6/B-H/DI AB- C 1B ITJ B3MX1&1&1&GH	10	52673-65-1	L66 B6/B-H/DI AB- C 1B ITJ B2N1&1 D2N1&1&GH 2
17	23771-09-7	T6N DNTJ A3VR CO1 DO1 EO1&DR CF&GH	10	31897-81-1	L66 B6/B-H/DI A B- C 1B ITJ B1Y1&N1&1 CG&GH
11	34580-64-8	T56 BNN INJ BR CXFFF&DO1VQ			
		Endocrine			
14	76-43-7	L E5 B666 OV MUTJ A1 BF CQ E1 FQ F1	11	33755-46-3	L E5 B666 OV AHTT&J A1 BF CQ E1 FV1Q FOV4 G1
11	108-42-9	ZR CG	10	63884-46-8	T6KJ A1 C1 D- CT5NOJ E1&G
		Gastrointestinal			
82	63884-46-8	1MVOR B1K1&1&1&Q&I	10	3043-03-6	T5NTJ A2 COVXR&R&O1&GH
52	7417-67-6	ONN1&V1	10	28743-45-5	T C666 BNJ GNW IN1&1&GH
19	2912-86-9	4N2&2OVXR&R&O1&GH	10	138-89-6	ONR DN1&1
15	97-00-7	WNR BG ENW	10	58-08-2	T56 BN DN FNVNVJ B1 F1 H1
12	32976-88-8	ZVMVN2&NO	10	3123-89-5	L B677 MV&T&J CO1 DO1 EO1 JMV1 NZ
12	67590-46-9	T66 A B AO EOPOTJ CX1&1&1	10	108-39-4	QR C1
11	73239-98-2	T6N DOTJ ANO CO1 C1 E1	10	117-81-7	4Y2&IOVR BVO1Y4&2
11	868-85-9	1OPHO&O1	10	63906-08-1	T6KJ A1VR DR DV1- AT6KJ&E 2
11	20674-99-1	T B656 DN HNT&&J D4 F1 H2- ET6NJ B1	10	26049-68-3	T5OJ BNW E- ET5N CSJ BMZ
11	27807-62-1	G2N2G2O1	10	50285-72-8	ONN1&VN2&2
		Glands			
81	63884-46-8	1MVOR B1K1&1&1&Q&I	11	1229-66-9	FR B1YR&R
15	521-18-6	L E5 B666 OVTJ A1 E1 FQ	10	796-13-4	GR DYEUYR&R
12	67590-46-9	T66 A B AO EOPOTJ CX1&1&1	10	63906-08-1	T6KJ A1VR DR DV1- AT6KJ&E 2
		Kidney			
15	4549-43-3	ONN1&2U1	10	54952-08-8	T666 BVOT&J D1 GG IVMY1&VQ JQ
12	108-42-9	ZR CG	10	555-29-3	QVXZ1&1R CQ DQ
12	25389-94-0	T6OTJ B1Z CQ DQ EQ FO- BL6TJ AZ CQ EZ DO- BT6OTJ CQ DZ EQ F1Q&...H2...S-O4	10	4920-79-0	WNR CG DO1
11	93-58-3	1OVR	10	4150-95-2	1N1&2XQR&R B1
11	4499-40-5	T56 BM DN FNVNVJ F H&622			
		Liver			
41	56986-36-8	ONN4&1OV1	12	117-81-7	4Y2&1OVR BVO1Y4&2
41	17576-88-4	ER CNUNR DN1&1	11	108-39-4	QR C1
31	3684-97-7	ONN1&1CN	10	63019-67-0	L B656 HHJ EMV1Z
26	607-59-0	L66J BNUNR DN1&1	10	94-75-7	QV1OR BG DG
25	33804-48-7	1VN1&R DNUNR DN1&1	10	94-80-4	GR CG DO1VO4
19	7347-46-8	T6NJ AO B1 DNUNR DN1&1	10	101-14-4	ZR CG D1R DZ CG
17	106-47-8	ZR DG	10	608-93-5	GR BG CG DG EG
15	2058-67-5	2MR DNUNR	10	32598-14-4	GR BG DR BG CG DG
14	97-02-9	WNR BZ ENW	10	83768-87-0	QVYZ2S1U1
13	99-09-2	ZR CNW	10	2402-77-9	T6NJ BG CG
13	1456-28-6	T6N DOTJ ANO C1 E1			
		Lung			
23	52731-39-2	ONN4&4OV1	17	100700-14-9	T E3 D6 B666 FOTT&&J HQ IQ
		Mutation			
60	64057-51-8	ONNV1&2G	12	1792-40-1	T56 FM DNJ C1 GNW
41	97-02-9	WNR BZ ENW	12	33350-73-1	WNR DR B1
36	69242-95-1	T C666 BNJ DVZ IMR BO1 DMSW2&GH	12	67730-10-3	T B656 AN CN HNJ DZ
27	121-89-1	WNR CV1	12	2238-07-5	T3OTJ B1O1- BT3OTJ
27	55-51-6	G2N2G1R	12	54749-90-5	T6OTJ BQ CMVNNO&2G DQ EQ F1Q
25	121-87-9	ZR BG DNW	12	73239-98-2	T6N DOTJ ANO CO1 C1 E1
23	108-42-9	ZR CG	11	53222-15-4	T C666 BNJ EMV1 IMR DMSW1&WSQ1
23	100700-14-9	T E3 D6 B666 FOTT&&J HQ IQ	11	52416-18-9	1VR DNUNN1&1
22	480-44-4	T66 BO EVJ CR DO1&GQ IQ	11	621-90-9	1MR DNUNR
21	581-29-3	T C666 BNJ EZ	11	63438-27-7	T G3 D6 B666 HOTT&&J EQ FQ
21	17405-06-0	ONN1&R CE	11	608-32-2	ZR BZ CZ
19	69292-84-8	T C666 BNJ IMR DM4&GH	11	533-73-3	QR BQ DQ
18	16230-87-8	T5N CNJ A1VQ B1 DNW	11	100333-37-7	L D6 C6566 1A T EUT&&&J GQ HQ
16	53-96-3	L B656 HHJ EMV1	11	523-50-2	T B566 EO LVOJ
16	4010-74-6	T6NVMVTJ A1R DO4	11	83-67-0	T56 BN DN FNVMVJ B1 F1
15	38915-14-9	T C666 BNJ DO1 IM3N2&2G&GH 2	10	62-44-2	2OR DMV1
15	962-32-3	T B3 G6 E666 COT&&T&J	10	10517-47-2	QMR DV1
15	38915-33-2	T C666 BN GNJ FO1 IM2O2N2&2G MG&GH 2	10	581-28-2	T C666 BNJ FZ
15	95-86-3	ZR CZ DQ	10	1874-12-0	T5OJ BNW E1U1VO2
15	50-89-5	T6NVMVJ E1 A- ET5OTJ B1Q CQ	10	19900-65-3	ZR B2 D1R DZ C2
14	21679-14-1	T56 BN DN FN HNJ GF IZ D- BT5OTJ CQ DQ E1Q	10	104-94-9	ZR DO1
14	607-59-0	L66J BNUNR DN1&1	10	1825-21-4	GR BG CG DG EG FO1
14	55112-89-5	G2N4&2G&GH	10	478-08-0	L C666 BV IVJ DQ E1Q FQ
14	4382-33-6	T66 BO EVT&J CR CQ DQ EQ&DQ IQ	10	225-51-4	T D6 B666 CNJ
13	2578-75-8	T5OJ BNW E- ET5NN DSJ CMV1	10	96-09-3	T3OTJ BR
13	59665-12-2	1VM1NNO&1R	10	531-85-1	ZR DR DZ&GH 2
13	33804-48-7	1VN1&R DNUNR DN1&1	10	934-32-7	T56 BM DNJ CZ
13	538-41-0	ZR DNUNR DZ	10	61785-57-7	T56 BNONJ BO FNW I- AT3NTJ
13	40855-12-7	T5OJ BNW E1UN2Q	10	13860-69-0	ZVMVN1&NO
13	39884-53-2	T5N COTJ ANO B1	10	563-52-0	GY1&1U1
12	602-87-9	L566 1A LT&&J HNW	10	63-25-2	L66J BOVM1
12	56986-37-9	ONNY2&1&1OV1	10	106-92-3	T3OTJ B1O2U1
12	1499-54-3	QMV1MVR	10	88518-80-3	T66 BNJ EVMQ
12	81-60-7	L C666 BV IVJ DQ GQ KQ NQ	10	932-52-5	T6MVMVJ EZ

Table 6. (Continued)

freq	CAS no.	WLN	freq	CAS no.	WLN
			Skin		
46	103-09-3	4Y2&1OV1	14	10032-00-5	1Y1&U3Y1&U2OV1V1
30	1319-77-3	QR X1	14	7790-07-0	2Y2&1O2OV4VO2O1Y2&2
24	3766-81-2	4R BOVM1	14	4468-93-3	Q2O1Y2&2
23	4812-23-1	WNY1&U2	13	871-22-7	4OYO4
23	2014-26-8	T6N DOTJ A1VOR B1 F1&GH	13	102-79-4	Q2N4&2Q
21	66227-09-6	3OV1OR CX1&1&1	13	22224-92-6	1Y1&MPO&O2&OR C1 DS1
21	63991-22-0	1VOY1M1&R COV1	13	52736-58-0	T6NTJ AV- DL6UTJ
19	21725-46-2	T6N CN ENJ BMX1&1&CN DM2 FG	12	2120-70-9	VH1OR
18	32210-23-4	L6TJ AX1&1&1 DOV1	12	97-95-0	Q1Y2&2
18	56530-50-8	L E3 B675 MV IU NUTJ BQ C1 EOV11 F1 F1	12	195-84-6	T C66 K66 S66 1A B&AN IN QN B&NJ
		J1OV1 LQ N1			
17	1929-73-3	GR CG DO1VO2O4	11	65405-73-4	VH1O2UY1&3UY1&1
17	137-05-3	1UYCN&VO1	11	98-86-2	1VR
17	107-88-0	QY1&2Q	11	8001-54-5	AK1&1&1R&G
16	4395-92-0	VH1R DY1&1	11	108-42-9	ZR CG
16	23184-66-9	G1V1O4&R B2 F2	11	98-80-6	QBQR
16	577-59-3	WNR BV1	11	1464-53-5	T3OTJ B- BT3OTJ
16	122-57-6	1V1U1R	10	5096-17-3	L B656 HHJ EMV1 KG
16	1892-43-9	Q2OR DG	10	28314-03-6	L B656 HHJ FMV1
15	2813-95-8	2Y1&R CNW ENW BOV1	10	79-43-6	QVYGG
15	97-96-1	VHY2&2	10	37169-10-1	WNR CG EG BOV1
15	60391-92-6	ZVNNO&1VQ	10	63977-47-9	1OR DO1 BY1&1K1&1&1&G
14	15823-55-9	3N3&V1O1R C1	10	51-75-2	G2N1&2G
14	105-34-0	NC1VO1			
			Tumor		
82	7417-67-6	ONN1&V1	12	50-07-7	T D3 B556 BN EM JV MVTTT&J GO1 H1OVZ KZ L1
42	75775-68-7	T C666 BNJ IMR CO1&WSQ1	12	64598-81-8	T F3 D6 C666 GO EU HH&T&&J HQ IQ
40	17576-88-4	ER CNUNR DN1&1	12	63019-29-4	L D6 B666J J1O2
37	64253-15-2	G2N2GR DNUNR	12	834-24-2	ZR D1U1R
33	66232-25-5	G2N2GR D1VO4	12	195-84-6	T C66 K66 S66 1A B& AN IN QN B&NJ
31	607-59-0	L66J BNUNR DN1&1	11	38479-08-2	L B677 MV&T&J CO1 DO1 EO1 JMV1G NS1
31	39020-49-0	T C666 BN GNJ FO4 IM2O2N2&2G MG&GH 2&QH	11	58-18-4	L E5 B666 OV MUTJ A E FQ F -B&AEF
29	24812-96-2	QVR DO1 CN2G2G	11	138-89-6	ONR DN1&1
24	33804-48-7	1VN1&R DNUNR DN1&1	11	21600-51-1	2OVR DNUNN1&1
23	4831-62-3	T66 BNJ BO C1 ENW	11	1464-53-5	T3OTJ B- BT3OTJ
22	13045-94-8	QVYZ1R DN2G2G -D	11	534-52-1	WNR BQ C1 ENW
22	13991-74-7	L6TJ AVO2 AMVNNO&2G	11	1233-89-2	G2N2GR DOVR
22	54749-90-5	T6OTJ BQ CMVNNO&2G DQ EQ F1Q	11	120-83-2	QR BG DG
20	106-47-8	ZR DG	11	83-67-0	T56 BN DN FNVMVJ B1 F1
20	100482-48-2	G2N2G2V1	10	67625-01-8	ZR CG EG DMV1N2&2
19	7347-46-8	T6NJ AO B1 DNUNR DN1&1	10	2113-47-5	1VMR BR
17	108-39-4	QR C1	10	16339-18-1	NC1NNO&1CN
16	63019-67-0	L B656 HHJ EMV1Z	10	38914-98-6	T C666 BNJ IM2S2G&GH
16	553-27-5	G2NR&2G	10	100-01-6	ZR DNW
15	65277-78-3	R2VMYVO&1O1R&-NA-	10	225-51-4	T D6 B666 CNJ
15	15044-98-1	T3NTJ APO&MVR CO2U1&- AT3NTJ	10	4210-69-4	T D6 C666 BNJ MM2N2&2G&GH 2
15	73239-98-2	T6N DOTJ ANO CO1 C1 E1	10	24813-07-8	ZVR CN2G2G
14	99-80-9	ONN1&R DNO	10	69884-94-2	G2N2GR CVMR
14	32976-88-8	ZVMVN2&NO	10	20268-52-4	L D6 B666J JG
14	21070-33-7	T66 BNJ BO EMQ H4	10	18429-71-5	L D6 B666J F1 M1 O1
13	28314-03-6	L B656 HHJ FMV1	10	101-14-4	ZR CG D1R DZ CG
13	4128-71-6	1VMR DNUNR	10	93-51-6	QR D1 BO1
13	2037-00-5	2N2&1OV1	10	38925-89-2	T66 BNJ EM2M2G HO1&GH 2
13	97805-02-2	NC1N2G2G&GH	10	72238-02-9	T D6 B656 FN LMJ C1 EM3N2&2 J1 PO1
13	100482-70-0	G4KO&2G2G	10	64678-03-1	T6N CNJ B1R DO4&DQ E1VO2 F1
13	22954-10-5	E2N2ER DOVR	10	76765-32-7	1Y1&MVR D1 CNUNN1&1
12	531-82-8	T5OJ BNW E- ET5N CSJ BMV1	10	72586-67-5	ONN1&VMR DV1
12	58658-13-2	T C666 BNJ IMR CMV1			
			Vascular		
21	1049816	T6N DNTJ A2R CO1 DO1&DR BQ	11	75343-60-1	T56 ANYOTJ BUN4 DR -DL -C&GH
20	2032-54-4	T6N DNTJ A2OR DG&D- BT6NJ	11	75343-67-8	T56 ANYOTJ BUM DR CG -T&EH
19	35288-47-2	4OR BZ DVO2N2&2&GH	11	60855-66-5	T6NNJ CNZ1 D1 FR&GH
15	106-47-8	ZR DG	10	476-70-0	T C6666 1AQKN&TT&J EO1 FQ K1 PQ QO1
14	76167-77-6	T56 BO DO CHJ G1- AT6NTJ DM- BT6N CNJ	10	102585-98-8	3OR CZ FVM2N2&2&GH
11	25433-48-1	T5OJ B1Y1&1K2&1&1&1	10	68263-35-4	T C676 BY IS JHJ BU3N1&1 E4&GH
11	3198-07-0	ZY2&YQR CQ DQ&GH			

IV. STRUCTURE PATTERNS

Methods. Independent analyses of frameworks and functional groups have the potential to reveal structural features conferring toxicity. However, synergetic effects of the groups and the framework of a molecule are not negligible, and we should examine structural biophores in the whole molecule context as well.

How can we take the whole molecule integrity into account? Here, a molecular structure pattern, which concerns molecule integrity, is defined as a template composed of a given framework and some given groups. It represents a cluster of molecules sharing common structural features. For a given database of toxic chemicals, many molecular structure patterns may exist and each of them will account for some specific toxicity.

In the present study, we have analyzed the structure patterns of toxic chemicals in terms of structure similarity. If we assume that every chemical belongs to one of the
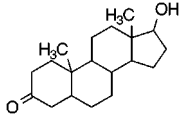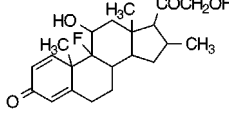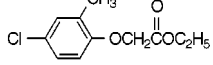
STRUCTURAL FEATURES OF TOXIC CHEMICALS

J. Chem. Inf. Comput. Sci., Vol. 39, No. 6, 1999 **1187**

Table 7. Structural Patterns for Abnormality
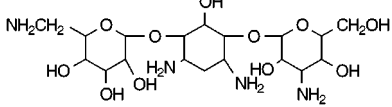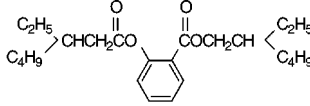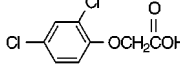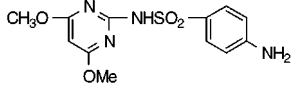
no.	structure patterns	CAS no.
1		521-18-6 2
2		386-67-2
3		2698-38-6
4		25389-94-0
5		117-81-7
6		94-75-7
7		155-91-9

Table 8. Prediction Results of Specific Structural Features as the Measure for Specific Toxicity

toxicity	right prediction of the chemicals with this specific toxicity[a]		wrong prediction of the chemicals without this specific toxicity[b]	
	no.	rate/%	no.	rate/%
abnormality	174	12.09	388	0.86
autonomic	244	30.85	848	1.85
blood	197	9.35	676	1.52
brain	269	23.72	777	1.71
cardiac	134	15.02	535	1.17
CNS	272	33.71	783	1.71
endocrine	117	12.42	435	0.95
gastrointestinal	410	14.86	801	1.83
glands	256	25.12	317	1.26
kidney	176	9.48	509	1.14
liver	411	18.05	385	1.87
lung	104	12.18	60	0.13
mutation	1154	26.92	471	1.11
skin	743	15.71	1027	2.46
tumor	1396	30.75	576	1.37
vascular	244	23.06	719.	1.58

[a] The data of specific toxicity can be seen in Table 9. [b] The chemicals not included in corresponding specific data sets but in the database are selected for this prediction.

molecular structure patterns, then the chemicals for certain kinds 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 a three-member cycle, a five-member cycle, a six-member cycle, etc.; saturation and nonsaturation), fused-cycle count (classified into dicycle, tricycle, and quadracycle, etc.; saturation and nonsaturation), branch-point

Table 9. Recorded Health Effects and Sizes of Data Sets for Some Specific Toxicity

no.	toxicity	health effects	size of the data set/ chemicals
1	abnormality	reproductive, specific developmental abnormality to various systems and organs	1439
2	autonomic	recordings from autonomic nervous system	791
3	blood	blood system changes such as serum composition, leukocyte count, platelet count, and so on	2107
4	brain	recordings from specific areas of central nervous system and other changes	1134
5	cardiac	cardiac system changes such as heart weight, pulse rate, cardiac output, and so on	892
6	CNS	recordings from specific areas of central nervous system	807
7	endocrine	changes in endocrine organs and hormones	942
8	gastro-intestinal	changes in gastrointestinal system	2759
9	glands	changes in glands and hormones	1019
10	kidney	concern about kidney, ureter, bladder	1857
11	liver	recordings from liver	2277
12	lung	recordings from lung	854
13	mutation	mutation testing in cells, microorganism, and so on	4286
14	skin	recordings from skin and appendages	4730
15	tumor	recording from tumorigenic testing	4540
16	vascular	changes in vascular system	1058

atom count (classified into N atom class, P atom class, and C atom class), and so on. If the two molecules do not resemble each other in their frameworks, the structure similarity value is regarded as zero. (2) Structure similarity values are obtained by comparing the 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 is equal to the sum of the cross-set divided by the sum of the combine set; that is, Value $= S_1 \cap S_2/S_1 \cup S_2$. The referred groups comprise common cyclic atoms ($-N-$, $-NH-$, $-O-$, $-S-$, $-(C=O)-$, etc.), common noncyclic 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 10 of the Appendix lists the groups that are used; Table 11 of the Appendix gives some examples of similarity computation.)

Results. We screen out the potential structure patterns for specific toxicity in the following way: assign the structure similarity limit to 0.6; select the molecule out if the molecule encounters more than 10 molecules with a higher similarity value than 0.6 to it in the data set; classify the molecules with a similarity value larger than 0.6 to each other into one class; then take a representative molecule out of every class to serve as a selected structure pattern, given 0.6 as the similarity limit and 10 as the count limit. These representative structure pattern molecules are listed in Table 6 by Chemical Abstracts Service number and WLN code.

We display the seven molecular structures of patterns for abnormality activity in Table 7. These representative structures provide us with the basis for further study. For example, each pattern can be a potential starting point for QSAR research. On the basis of these structure patterns, a series of similar molecules to any selected pattern can be extracted from the database, facilitating the study of structure—toxicity relationships and toxicity mechanisms. Meanwhile, these patterns can be used as indicators of potential toxicity in

Table 10. 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	−NH²	62	−CCl₃	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		

Table 11. Examples of Similarity Computation with the Following Reference Molecule

C₂H₅−N−(CH₂)₂−OCO− −C(phenyl)(phenyl) •HCl
with CH₃ and OMe substituents
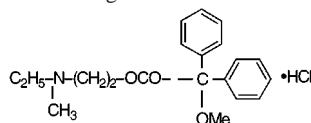
molecule	structure	framework indicators	group indicators	similarity values
1	(structure: CH₃COCH−C(C₂H₅)(phenyl)(phenyl CH₂CHNHCH₃CH₃) •HCl)	phenyl 2,2ᵃ	−NH−, 0,1* −N−,1,0 −COO−,1,1 −(C=O)−,1,1 −O−,2,1 −Cl,1,1 −C−(phenyl)₂,1,1	5/8 = 0.62
2	(structure: C₂H₅O / O=C−O / (C₂H₅)₂N−CH₂−CH(phenyl)(phenyl) •HCl)	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	(structure: (CH₃)₂N−CHCH₂−C(CH₃)(phenyl)(phenyl)−CH−O−C−CH₃ / C₂H₅ O •HCl)	phenyl 2,2	−N−,1,1 −COO−,1,1 −(C=O)−,1,1 −O−,2,1 −C−(phenyl)₂,1,1 −CH−phenyl,1,1	6/7 = 0.86

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

terms of similarity comparison between molecules; that is, the molecules with high similarity to structure patterns would be regarded as potentially "toxic" for screening purposes.

V. CONCLUSION

We have made an extensive survey of the structural properties of toxic chemicals in the RTECS database

STRUCTURAL FEATURES OF TOXIC CHEMICALS

J. Chem. Inf. Comput. Sci., Vol. 39, No. 6, 1999 **1189**

associated with specific toxicity. Toxic chemicals were analyzed by taking account of frameworks, functional groups, and structure patterns of the molecules for identifying structural features and similarity measures, respectively, associated with specific toxicity.

When examining frameworks and functional groups for their biofunctional significance, we differentiate the causes of high occurrence frequency as composition factors and activity factors. Composition factors are associated with uniformity of distribution, whereas activity factors tend to concentrate in specific toxicity subsets. By use of statistical frequencies, we identify some potential active frameworks and groups for specific toxicity. Additionally, we use similarity analysis to examine structure patterns and screen some potential structure patterns for specific toxicity. These structural features associated with frameworks, functional groups, and structure patterns offer us useful information for better understanding the structural basis for chemical toxicity.

There is a great deal of information in the literature on "structural fragments" or "toxicophores" associated with some types of toxicity; of them mutagenicity is one of extensively explored areas. The structural features to mutagenicity we found, such as aromatic fused cycles, N heterocycles, or fragments containing epoxy ethane, have accordance with the characteristics of known cancinogens.[12] However, a general comparative analysis of existing knowledge still needs more effort.

Our findings concerning structural features of toxic chemicals for specific toxicity display considerable specificity when used as a standard for evaluating specific toxicity (for frameworks and groups, their presence indicates toxicity, and for structure patterns, 0.6 is similarity limit (Table 8)). They incorrectly identify only 1% of the nonspecific chemicals (more than 40 000 chemicals from RTECS) and cover nearly 10% of specific-toxicity chemicals. Thus, these structural features are characteristic of specific toxicity and potentially useful for prediciton, especially for the chemicals in the early stage of drug design.

Although much work has been done in the area of developing structure−toxicity relationships,[1,13,14] much work remains to be done. Traditional QSAR technologies, including Free−Wilson and Hansch approaches, have proven powerful and effective.[15] The findings derived from them provide us with a great deal of information to reveal and confirm the understanding of the structural basis for chemical toxicity, but when confronted with a database consisting of noncongeneric toxic chemicals, these technologies reveal their limitations. Fortunately, it has been becoming feasible to further explore large databases of toxic chemicals when we introduce advances in chemometrics, such as similarity analysis, three-dimensional (3D) QSAR, and new statistic approaches.

ACKNOWLEDGMENT

APPENDIX

Tables 9−11 contain health effects and data set sizes, similarity computation groups, and similarity computation examples, respectively.

REFERENCES AND NOTES

(1) Goldberg, L. *Structure-activity correlation as a predliclive tool in toxicology*; Hemisphere: Washington, D.C., 1983.
(2) Sanderson, D. M.; Earnshaw, C. G. Computer prediction of possible toxic action from chemical structure: The system DEREK. *Hum. Exp. Toxicol.* **1991**, *10*, 261−273.
(3) Woo, Y.; Lai, D. Y.; Argus, M. F.; Arcos, J. C. Development of structure-activity relationship rules for predicting carcinogenic potential of chemicals. *Toxicol. Lett.* **1995**, *79*, 219−228.
(4) TOPKAT 5.0, Oxford molecular group. *Nature* **1998**, *391*, 719 (http://www.oxmol.com/prods/topkat/).
(5) Lewis, D. F. V. Computer-assisted methods in the evaluation of chemical toxicity. In *Reviews in Computational Chemistry*; VCH New York, 1992; Vol. III, pp 172−221.
(6) Gombar, V. K.; Enslein, K.; Reid, D. A. Computer-assisted toxicity assessment: Criteria for acceptance. *Network Sci.* **1996** (hlttp://www.awod. com/netsci/Issues/Feb96/feature2.html).
(7) Klopman, G. Artificial intelligence approach to structure−activity studies: Computer automated structure evaluation of biological activity of organic molecules. *J. Am. Chem. Soc.* **1984**, *106*, 7315−7320.
(8) Klopman, G. MultiCASE: 1. A hierarchical computer automated structure evaluation program. *Quant. Struct.-Act. Relat.* **1992**, *11*, 176−184.
(9) *RTECS C2(96-4)*; National Institute for Occupational Safety and Health (NIOSH), U.S. Department of Health and Human Services: Washington, D.C., 1996 (http://www.ccohs.ca/).
(10) Smith, E. G.; Baker, P. A. *The Wisswesser Line-Formula Chemical Notation (WLN)*, 3rd ed.; Chemical Information Management Inc.: Cherry Hill, NJ, 1975.
(11) Bemis, G. W.; Murcko, M. A. The properties of known drugs: l.Molecular frameworks. *J. Med. Chem.* **1996**, *39*, 2887−2893.
(12) Medicine Encyclopaedia of China: 8, Toxicology (in Chinese). Shanghai Publishing House of Technology and Science: Shanghai, 1982; pp 38−39.
(13) Howe, W. J.; Milne, M. M.; Pennell, A. F. *Retrieval of Medicinal Chemical Information*; ACS Symposium Series 84; American Chemical Society: Washington, D.C., 1978.
(14) Horvath, A. L. Relations between structure and properties. In *Molecular Design: Chemical structure generation from the properties of pure organic compounds*; Elsevier Science: New York, 1992; Chapter 3, p 575.
(15) Craig, P. N. Structure/property correlations. In *Chemical Information System*; Ash, J. E., Hyde, E., Ellis Horwood: Chichester, U.K., 1975; Chapter 16, p 259.

CI990039R