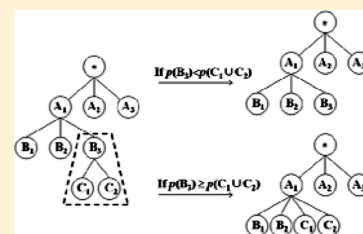


## Estimation of Carcinogenicity Using Molecular Fragments Tree

Yong Wang,<sup>†,||</sup> Jing Lu,<sup>†,||</sup> Fei Wang,<sup>†</sup> Qiancheng Shen,<sup>†</sup> Mingyue Zheng,<sup>\*,†</sup> Xiaomin Luo,<sup>\*,†</sup> Weiliang Zhu,<sup>†</sup> Hualiang Jiang,<sup>†,‡</sup> and Kaixian Chen<sup>†</sup><sup>†</sup>Drug Discovery and Design Center, State Key Laboratory of Drug Research, Shanghai Institute of Materia Medica, Chinese Academy of Sciences, 555 Zuchongzhi Road, Shanghai 201203, China<sup>‡</sup>School of Pharmacy, East China University of Science and Technology, Shanghai 200237, China

## S Supporting Information

**ABSTRACT:** Carcinogenicity is an important toxicological endpoint that poses high concern to drug discovery. In this study, we developed a method to extract structural alerts (SAs) and modulating factors of carcinogens on the basis of statistical analyses. First, the Gaston algorithm, a frequent subgraph mining method, was used to detect substructures that occurred at least six times. Then, a molecular fragments tree was built and pruned to select high-quality SAs. The *p*-value of the parent node in the tree and that of its children nodes were compared, and the nodes that had a higher statistical significance in binomial tests were retained. Finally, modulating factors that suppressed the toxic effects of SAs were extracted by three self-defining rules. The accuracy of the 77 SAs plus four SA/modulating factor pairs model for the training set, and the test set was 0.70 and 0.65, respectively. Our model has higher predictive ability than Benigni's model, especially in the test set. The results highlight that this method is preferable in terms of prediction accuracy, and the selected SAs are useful for prediction as well as interpretation. Moreover, our method is convenient to users in that it can extract SAs from a database using an automated and unbiased manner that does not rely on a priori knowledge of mechanism of action.



## ■ INTRODUCTION

The definition of structural alerts proposed by Ashby<sup>1</sup> is a new starting point for the research of toxicity. Structural alerts (SAs) are defined as molecular functional groups that can cause toxicity, and their presence alerts investigators to the potential toxicities of test chemicals.<sup>2</sup> Compared to SAR models based on physicochemical descriptors, SAs provide intuitive structural information for toxicologists and chemists and are readily applied to mechanism investigation. Furthermore, some properties of compounds, such as mutagenicity, can be better interpreted with SAs than with other descriptors.<sup>3,4</sup>

Some groups of SAs have been reported in the literature for carcinogenicity prediction. Following the electrophilic theory of Miller,<sup>5</sup> 19 SAs were proposed by Ashby<sup>6</sup> and served as a reference for subsequent investigators. Then 33 SAs were compiled by Bailey et al. on the basis of Ashby's SAs.<sup>7</sup> The conventional SAs are often proposed on the basis of expert opinion of toxicologists without the use of any statistics.<sup>8</sup> If SAs can be generated and extracted from a database by computer automatically, this will be conveniently employed by users. To this aim, 29 SAs produced on the basis of a data mining method by Kazius et al. were judged and accepted by expert knowledge from Ashby.<sup>9</sup> The classification error of the Kazius' model was 18%, and the predictive ability was equivalent to experiment. One disadvantage of this model is that the data set used for SAs identification was a mutagenicity database, so the derived SAs were restricted to genotoxic carcinogens. Helma developed a procedure: lazy

structure–activity relationships (lazar) for the prediction of rodent carcinogenicity.<sup>10</sup> This procedure uses the MolFea algorithm<sup>11</sup> and automatically searches for SAs with discriminating capacity. However, MolFea has practical limitations that can only investigate paths but not trees and graphs. Kazius et al. developed a new searching algorithm, Gaston (GrAph/Sequence/Tree extractiON),<sup>12</sup> to solve the problem. Meanwhile, to limit the output numbers of fragments with potential interest, constraints such as minimum frequency can be applied. However, if a fragment satisfies this threshold, its substructures must also meet this threshold. Thus, Gaston would retain a large number of redundant fragments, and this would cause some difficulties to analyze and extract SAs of high reliability. In this study, we develop a method to build and prune a molecular fragments tree for efficiently selecting SAs from redundant fragment candidates. This method can automatically and unbiasedly select molecular fragments with statistical significance in binomial test. In addition, modulating factors that can suppress the toxic effects of SAs<sup>8</sup> is also an interesting point. Here, we defined three rules to extract SAs and their corresponding modulating factors. The results showed that our group of SAs and modulating factors had better performance than the Benigni's model, especially in the test set. Details are described in the Materials and Methods section.

Received: June 11, 2012

Published: July 26, 2012

Table 1. Atomic Type of Compounds

atomic type	definition
C.3	carbon sp <sup>3</sup>
C.2	carbon sp <sup>2</sup>
C.1	carbon sp
C.ar	carbon aromatic
N.3	nitrogen sp <sup>3</sup>
N.2	nitrogen sp <sup>2</sup>
N.1	nitrogen sp
N.ar	nitrogen aromatic
N.pl3	nitrogen trigonal planar
N.4	nitrogen sp <sup>3</sup> positively charged
O.3	oxygen sp <sup>3</sup>
O.2	oxygen sp <sup>2</sup>
O.ar	oxygen aromatic
S.3	sulfur sp <sup>3</sup>
S.2	sulfur sp <sup>2</sup>
S.ar	Sulfur aromatic
S.o	sulfoxide sulfur
S.o2	sulfone sulfur
P.3	phosphorus sp <sup>3</sup>
F	fluorine
Cl	chlorine
Br	bromine
I	iodine

Table 2. Bond Type of Compounds

bond type	definition
1	single bond
2	double bond
3	triple bond
ar	aromatic bond

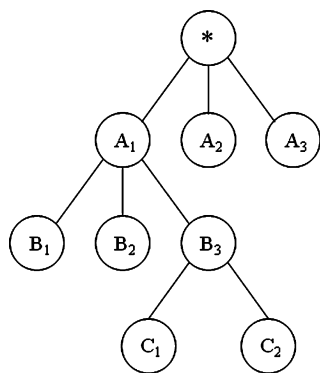
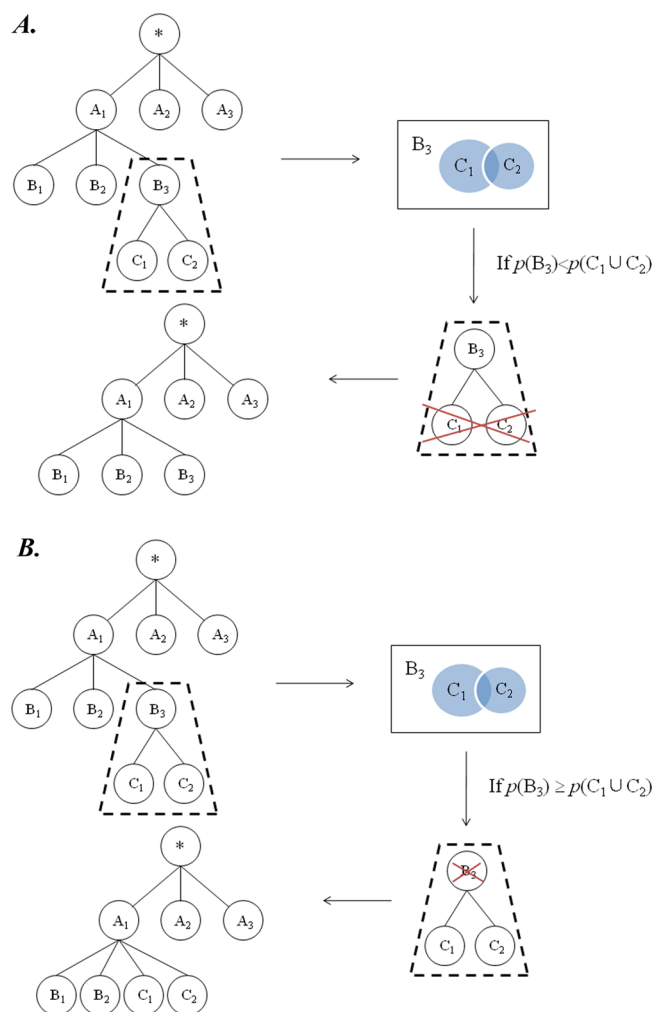


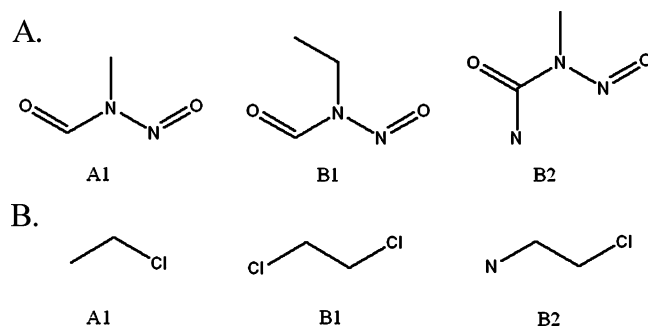
Figure 1. An illustrated molecular fragments tree.

## MATERIALS AND METHODS

**Data Set.** The ISSCAN database<sup>13,14</sup> was selected as the training set because the quality of its data met the requirement of structure–activity relationship studies.<sup>8</sup> The preliminary treatment is described below: (a) Remove compounds that have inconsistent results in different experimental groups.<sup>15</sup> (b) Remove mixtures, inorganic and organometallic compounds. (c) The salts are transformed to neutral compounds. (d) Remove compounds whose molecular weights are more than 500. (e) Only one stereoisomer is retained because the Gaston algorithm can only generate two-dimensional fragments, and fragments from a pair of stereoisomers are identical. The training set from ISSCAN consisted of 975 compounds (651 positives and



**Figure 2.** Two possibilities of pruning a molecular fragments tree. (A) If  $p(B_3) < p(C_1 \cup C_2)$ , the  $C_1$  and  $C_2$  node are disregarded because the statistical significance of  $B_3$  is higher than its child nodes. (B) If  $p(B_3) \geq p(C_1 \cup C_2)$ , the  $B_3$  node is deleted, and the  $C_1$  and  $C_2$  fragments become the child nodes of  $A_1$  instead of  $B_3$ .



**Figure 3.** Structures of Parent  $A_1$  and Children  $B_1$  and  $B_2$

324 negatives). To test the practicability of our method, 178 compounds (73 positives and 105 negatives) from Predictive Toxicology Challenge (PTC) 2000–2001,<sup>16</sup> which were unduplicated with ISSCAN, were chosen as the test set. The SMILES files of the training set and test set are provided in Supporting Information.

**Frequent Subgraph Mining.** To use Gaston, each chemical should be represented as a graph. The vertices of a graph correspond to the atoms of a compound and the edges to the bonds. Each vertex and edge has a label, which can be regarded

Table 3. Performance of Set I, Set II, and Set III

		TP	FN	TN	FP	sensitivity	specificity	accuracy
training set	Set I	493	158	187	137	0.76	0.58	0.70
	Set II	351	300	302	22	0.54	0.93	0.67
	Set III	439	212	239	85	0.67	0.74	0.70
test set	Set I	50	23	58	47	0.68	0.55	0.61
	Set II	28	45	86	19	0.38	0.82	0.64
	Set III	34	39	81	24	0.47	0.77	0.65

Table 4. Performance of Set III versus Benigni's Model

		TP	FN	TN	FP	sensitivity	specificity	accuracy
training set	Benigni	<sup>a</sup>	<sup>a</sup>	<sup>a</sup>	<sup>a</sup>	0.74	0.64	0.70
	Set III	439	212	239	85	0.67	0.74	0.70
test set	Benigni	26	47	74	31	0.36	0.71	0.56
	Set III	34	39	81	24	0.47	0.77	0.65

<sup>a</sup>The numbers of TP, TN, FP, and FN are not reported in ref 8.

as atomic type and bond type, respectively. In this study, the types of atoms and bonds that are similar to the Tripos atom typing scheme<sup>17</sup> are used in representing molecular graphs (Tables 1 and 2).

At every step, Gaston iteratively performs both substructure generation and the corresponding substructure search.<sup>12</sup> Fragments extracted from Gaston include paths, trees, and graphs. In the process of "quickstart" searching, first paths are found, then trees are identified, and finally graphs are considered.<sup>18</sup> First, the threshold of minimum frequency was set to 6 to ensure the statistical significance of every fragment is more than the threshold value ( $\alpha = 0.1$ ). A total of 36,859 fragments were generated in all. Then, the ratio of carcinogens to compounds is  $651/975 = 0.667$  in the training set. To improve the reliability of fragments, only fragments with accuracy  $Q$  significantly higher than 0.667 ( $p < 0.1$ ) are retained. It should be noted that this ratio does not correspond to the background level of observing carcinogens in nature or a typical pharmaceutical company's compound library, which are much more lower.<sup>19</sup> Generally, for a data-driven approach, a relatively small proportion of positives would lead to the calculations of accuracy heavily biased by the correct prediction of negatives, which always belies the existence of very poor sensitivities. Here, the high threshold value (of judging whether a fragment is a potential SA) is expected to enhance the sensitivity of resulted toxicity model.

$Q$  is defined as<sup>12</sup>

$$Q = \frac{TP}{TP + FP} \quad (1)$$

where TP is the number of carcinogens containing the fragment, and FP is the number of noncarcinogens containing the fragment.

**Building of a Molecular Fragments Tree.** Altogether 4659 fragments passed the above filtering procedure and were used to build a molecular fragments tree. Such a tree has the following properties:

- The root node (marked as \*) is an abstract node that stands for an arbitrary atom, so the root can match all of the compounds.
- Every nonroot node stands for a concrete molecular fragment.

- For a parent node (denoted as "A"), its child node (denoted as "B") must meet the following two conditions: (i) A is a substructure of B ( $A \subset B$ ). (ii) There is not such a fragment C that contains A and at the same time belongs to B, i.e.,  $(A \subset C \text{ and } C \subset B) = \text{False}$ . Consequently, the set of compounds that a node can match is a subset of the compounds that its parent node can match.

Such a molecular fragments tree can be built from top to bottom recursively. First, the root node is generated, and all of the smallest fragments become the children nodes of the root node. Then, for each existing node, all of the fragments that satisfy the Property c are constructed in the tree as its children. The building process is carried out recursively until no children can be found. Suppose that a set of fragments contains  $A_1, A_2, A_3, B_1, B_2, B_3, C_1$ , and  $C_2$ . The relationship is  $A_1 \subset B_1, A_1 \subset B_2, A_1 \subset B_3, B_3 \subset C_1$ , and  $B_3 \subset C_2$ . An example of the molecular fragments tree is shown in Figure 1.

**Pruning of the Molecular Fragments Tree.** Pruning of the molecular fragments tree is necessary to select the most significant fragments. We began to prune the complicated tree from the bottom. For a parent node  $B_3$  and its child nodes  $C_1$  and  $C_2$ , the  $p$ -value of  $B_3$  and  $C_1 \cup C_2$  were computed, respectively. Here,  $C_1 \cup C_2$  means that the set of compounds include at least one of  $C_1$  and  $C_2$ . If  $p(B_3) < p(C_1 \cup C_2)$ , the two fragments on the  $C_1$  and  $C_2$  node were discarded because the statistical significance of  $B_3$  is higher than its child nodes (Figure 2A). Alternatively, if  $p(B_3) \geq p(C_1 \cup C_2)$ , the  $B_3$  node was deleted, and the  $C_1$  and  $C_2$  fragments became the child nodes of  $A_1$  instead of  $B_3$  (Figure 2B). This pruning process continued from bottom to top until all nodes became the child nodes of the root node. In the end, 165 fragments were retained after being pruned.

In some cases, compounds matched by one fragment were a subset of another, but these two fragments had no structure-containing relationship. To ensure the reliability of the model, the fragment with a smaller  $p$ -value was retained. Finally, 77 individual fragments were kept as SAs to predict the carcinogenicity of compounds.

**Extraction of Structural Alerts and Their Corresponding Modulating Factors by Three Self-Defined Rules.** As mentioned before, the existence of modulating factors could suppress the toxicity effect of an SA. Thus, some SAs may not

be found by our molecular fragments tree due to their co-occurrence with modulating factors. From the data set, we can

also see that the discriminating capacities of some SAs are not obvious. Here, to find such hidden SAs, we investigated

**Table 5. Details of 77 SAs and 4 SA/Modulating Factor Pairs in Our Model**


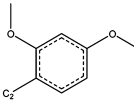
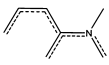
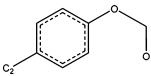
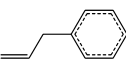
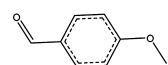
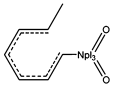

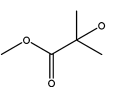

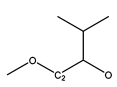
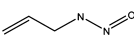
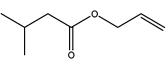

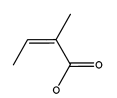
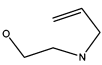
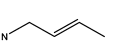
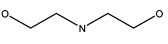
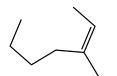
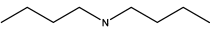
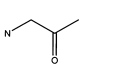
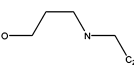
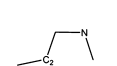

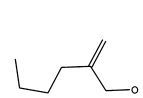
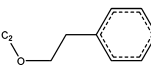
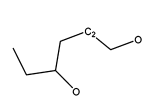

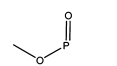

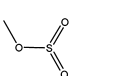
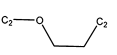
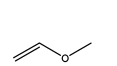
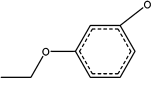
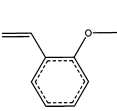

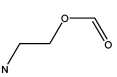

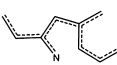
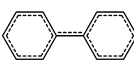


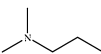







no <sup>a</sup>	Structural alerts <sup>b</sup> and modulating factors	Positive	negative	total	Accuracy <sup>c</sup>	p value <sup>c</sup>	no <sup>a</sup>	Structural alerts <sup>b</sup> and modulating factors	Positive	negative	total	Accuracy <sup>c</sup>	p value <sup>c</sup>
SA_1		7	0	7	1.000	0.059	SA_19		8	0	8	1.000	0.039
SA_2		6	0	6	1.000	0.088	SA_20		6	0	6	1.000	0.088
SA_3		10	0	10	1.000	0.017	SA_21		7	0	7	1.000	0.059
SA_4		6	0	6	1.000	0.088	SA_22		6	0	6	1.000	0.088
SA_5		8	0	8	1.000	0.039	SA_23		25	3	28	0.893	0.006
SA_6		7	0	7	1.000	0.059	SA_24		10	0	10	1.000	0.017
SA_7		7	0	7	1.000	0.059	SA_25		6	0	6	1.000	0.088
SA_8		7	0	7	1.000	0.059	SA_26		7	0	7	1.000	0.059
SA_9		13	1	14	0.929	0.028	SA_27		10	1	11	0.909	0.075
SA_10		7	0	7	1.000	0.059	SA_28		6	0	6	1.000	0.088
SA_11		7	0	7	1.000	0.059	SA_29		9	0	9	1.000	0.026
SA_12		9	0	9	1.000	0.026	SA_30		17	3	20	0.850	0.061
SA_13		6	0	6	1.000	0.088	SA_31		6	0	6	1.000	0.088
SA_14		7	0	7	1.000	0.059	SA_32		6	0	6	1.000	0.088
SA_15		11	1	12	0.917	0.054	SA_33		7	0	7	1.000	0.059
SA_16		7	0	7	1.000	0.059	SA_34		12	0	12	1.000	0.008
SA_17		7	0	7	1.000	0.059	SA_35		9	0	9	1.000	0.026
SA_18		6	0	6	1.000	0.088	SA_36		11	0	11	1.000	0.012
							SA_37		7	0	7	1.000	0.059

Table 5. continued

no <sup>a</sup>	Structural alerts <sup>b</sup> and modulating factors	Positive	negative	total	Accuracy <sup>c</sup>	<i>p</i> value <sup>c</sup>
SA_38		9	0	9	1.000	0.026
SA_39		78	6	84	0.929	1.24e-08
SA_40		10	1	11	0.909	0.075
SA_41		10	1	11	0.909	0.075
SA_42		10	1	11	0.909	0.075
SA_43		6	0	6	1.000	0.088
SA_44		6	0	6	1.000	0.088
SA_45		6	0	6	1.000	0.088
SA_46		11	1	12	0.917	0.054
SA_47		11	1	12	0.917	0.054
SA_48		6	0	6	1.000	0.088
SA_49		10	1	11	0.909	0.075
SA_50		10	1	11	0.909	0.075
SA_51		30	1	31	0.968	5.82e-05
SA_52		13	2	15	0.867	0.080
SA_53		29	1	30	0.967	8.46e-05
SA_54		6	0	6	1.000	0.088
SA_55		13	0	13	1.000	0.005
SA_56		7	0	7	1.000	0.059
SA_57		8	0	8	1.000	0.039
SA_58		16	0	16	1.000	0.002
SA_59		7	0	7	1.000	0.059
SA_60		17	0	17	1.000	0.001
SA_61		10	0	10	1.000	0.017
SA_62		10	1	11	0.909	0.075
SA_63		6	0	6	1.000	0.088
SA_64		12	1	13	0.923	0.039
SA_65		12	1	13	0.923	0.039
SA_66		6	0	6	1.000	0.088
SA_67		8	0	8	1.000	0.039
SA_68		14	0	14	1.000	0.003
SA_69		7	0	7	1.000	0.059
SA_70		7	0	7	1.000	0.059
SA_71		7	0	7	1.000	0.059
SA_72		7	0	7	1.000	0.059

Table 5. continued

no <sup>a</sup>	Structural alerts <sup>b</sup> and modulating factors	Positive	negative	total	Accuracy <sup>c</sup>	<i>p</i> value <sup>c</sup>
SA_73		12	1	13	0.923	0.039
SA_74		8	0	8	1.000	0.039
SA_75		10	1	11	0.909	0.075
SA_76		7	0	7	1.000	0.059
SA_77		6	0	6	1.000	0.088
S1-M1	 S1  M1	9	0	9	1.000	0.026
S2-M2	 S2  M2	27	7	34	0.794	0.078
S3-M3	 S3  M3	22	3	25	0.880	0.015
S4-M4	 S4  M4	164	61	225	0.729	0.027

<sup>a</sup>77 SAs are marked as SA, respectively. Four pairs of SAs without the occurrence of their corresponding modulating factors are marked as S–M, respectively. <sup>b</sup>C<sub>2</sub> means sp<sup>2</sup> carbon, and N<sub>pl3</sub> means trigonal planar nitrogen. Other atomic types are assigned according to their element types. <sup>c</sup>The accuracy is TP/(TP + FP). The threshold is  $Q > 0.667$  ( $p < 0.1$ ).

modulating factors (denoted as “M”) that SAs (denoted as “S”) may involve through the following three requirements:

- The ratio of positives in the data set including S is not significantly higher than 0.667. This rule filters out the SAs that had already been found.
- The ratio of positives in the data set including S but not M is significantly higher than 0.667. This rule demonstrates that S has a toxic effect without the co-occurrence of M, thus S is a structural alert.
- The ratio of negatives in the compounds including S and M simultaneously is significantly higher than 0.710. This rule indicates that M can suppress the toxicity elicited by S. Following these three conditions, M is defined as a modulating factor of S. The ratio of observing negatives in the overall training set is 0.333. This threshold led to the calculations of accuracy heavily biased by the correct prediction of positives and belied the existence of very poor specificity. Thus, we searched a threshold to get the best balanced accuracy (= (sensitivity + specificity)/2) in the training set, and the threshold was defined as 0.710.

Moreover, the molecular fragments tree made sensitivity lower substantially. So the smallest SA was retained if a series of SAs corresponded with the same modulating factor, and the biggest modulating factor was retained if a series of modulating factors corresponded with the same SAs. Thus, four pairs of SAs and their corresponding modulating factors were extracted from the training set and used to construct the classification model.

## RESULTS AND DISCUSSION

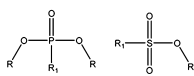
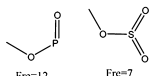

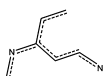
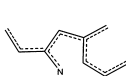
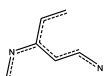
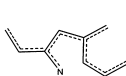



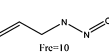


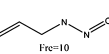



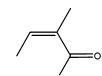
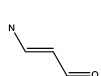







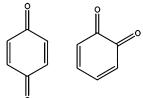
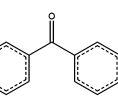
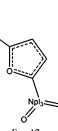
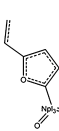
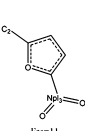
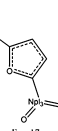
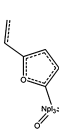
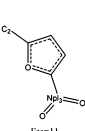


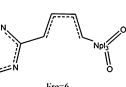

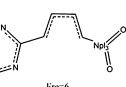

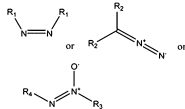





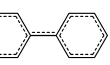
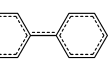








**Performance of Models.** Altogether 77 SAs were selected from the original 4659 fragments via the molecular fragments tree. In the pruning process, the fragments with a higher statistical significance were retained. For example, in Figure 3A, the parent (denoted as “A<sub>1</sub>”) matched 31 compounds in which 30 were true positives ( $p = 0.00006$ ), and the two children (denoted as B<sub>1</sub> and B<sub>2</sub>) matched 29 compounds in which 28 were

true positives ( $p = 0.00013$ ). Thus, the children B<sub>1</sub> and B<sub>2</sub> were discarded. In another example (Figure 3B), the parent A<sub>1</sub> matched 58 compounds in which 45 were true positives ( $p = 0.05040$ ), and the two children (B<sub>1</sub> and B<sub>2</sub>) matched 31 compounds in which 28 were true positives ( $p = 0.00250$ ), so the parent A<sub>1</sub> was discarded. The original fragment set (Set I), the selected set (Set II), and the 77 SAs plus four SA/modulating factor pairs (Set III) were tested for their predictivity. Here, a compound containing at least one of these 77 SAs is considered to be positive (i.e., potentially toxic). A compound containing at least one of the four pairs of SAs without the co-occurrence of their corresponding modulating factors is considered to be positive. Others are negative. For Set I, 4659 fragments without pruning matched 630 compounds in which only 493 were true positives, and the accuracy  $Q = 0.783$ . For Set II, 77 SAs matched 373 compounds in which 351 compounds were true positives, and  $Q$  was 0.941. The results demonstrate that our molecular fragments tree can efficiently extract SAs with higher accuracy from redundant fragment candidates. However, the number of true positives matched by Set II is smaller than Set I, which indicates higher risk in predicting positive samples as negatives (more unfavorable consequences as for toxicity evaluation). This problem is significantly alleviated by introducing four modulating factors in Set III, of which the TP number has been enhanced to 439. All statistics of Set III are more balanced than Set I, representing a better set of SAs for carcinogenicity prediction. The performance statistics of these three sets are shown in Table 3.

The performance of the model is measured by true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). Accuracy (= (TP + TN)/(TP + TN + FP + FN)) is an usual index for the overall classification performance. Sensitivity (= TP/(TP + FN)) and specificity (= TN/(FP + TN)) can assess the ability of the model to correctly identify positives and negatives. Table 4 shows the details of Set III and Benigni's model.<sup>8</sup> Of note is that Benigni's model was only applied to the ISSCAN database (of which the statistics were extracted from the original literature). Here, we predicted the



Table 6. Some Representative Examples of Our SAs Are Illustrated To Compare with References

Structural alerts in references	Structural alerts in our model <sup>a</sup>	Structural alerts in references	Structural alerts in our model <sup>a</sup>
 alkyl (C < 5) or benzyl ester of sulphonic or phosphonic acid <sup>8</sup>	 Fre=12  Fre=7	heterocyclic polycyclic aromatic hydrocarbons <sup>8</sup>  Fre=8  Fre=8	heterocyclic polycyclic aromatic hydrocarbons <sup>8</sup>  Fre=8  Fre=8
 aliphatic halogens <sup>8</sup>	 Fre=20  Fre=11	 Fre=10  Fre=84  Fre=31	 Fre=10  Fre=84  Fre=31
 alpha,beta-unsaturated carbonyls <sup>7,8</sup>	 Fre=7  Fre=15	 Fre=7	 Fre=7
 simple aldehyde <sup>8</sup>	 Fre=7  Fre=7	 Fre=7	 Fre=7
 Quinones <sup>8</sup>	 Fre=14	 Fre=17  Fre=10  Fre=11	 Fre=17  Fre=10  Fre=11
 Hydrazine <sup>8</sup>	 Fre=28	 Fre=6  Fre=6	 Fre=6  Fre=6
 aliphatic azo and azoxy <sup>8</sup>	 Fre=7	 Fre=6  Fre=8	 Fre=6  Fre=8
 polycyclic aromatic hydrocarbons <sup>8</sup>	 Fre=11	 Fre=13  Fre=16  Fre=7	 Fre=13  Fre=16  Fre=7
		 Fre=6	 Fre=6

<sup>a</sup>Here, the frequency of occurrence of every SA is marked as "Fre".

PTC set as an external validation using the 31 SAs in Benigni's model. The analysis demonstrates that the accuracy of Set III is equivalent to Benigni's model for the training set, but the sensitivity and specificity of our model are more balanced. More importantly, for the test set we may observe an overall performance improvement of Set III over Benigni's model.

At the same time, it should be also noticed that the SAs in Benigni's model were derived from experience of toxicologists, which need abundant professional knowledge. In contrast, our approach is completely automated, and potential SAs are extracted in unbiased manner, which does not need any manual intervention or a priori knowledge of action. This feature provides users a convenient way to do self-tailoring or expanding of an existing SA list, which may become more and more necessary with the fast growth of available chemical structures without their associated carcinogenicity data.

### Analysis of Structural Alerts Plus Modulating factors.

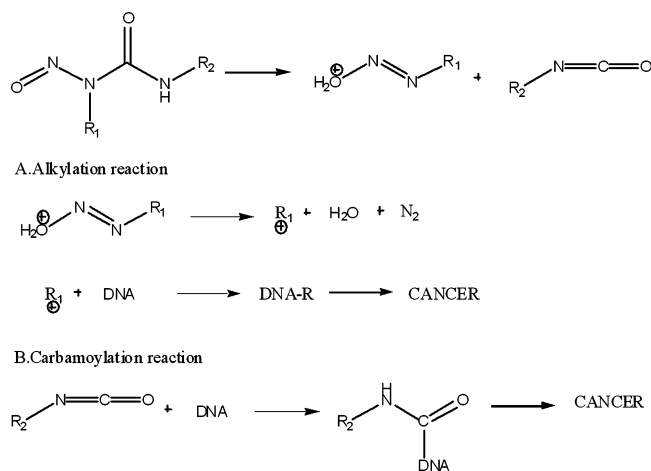
Table 5 displays the statistical details of 77 SAs and four SAs without the co-occurrence of their corresponding modulating factors for the training set. Some SAs of Benigni's model cannot match more than six compounds or satisfy  $Q > 0.667$  ( $p < 0.1$ ), so they cannot be extracted by our method. Table 6 shows some representative examples of our SAs that are equivalent to those defined in refs 7–9, and 20. SAs extracted using our method contain the majority of acknowledged carcinogenic fragments, such as sulfonic or phosphonic acid, aliphatic halogens,  $\alpha,\beta$ -unsaturated carbonyls, nitro-aromatic, aromatic amines, and so on. The results manifest the usefulness of our method with high accuracy and unambiguous mechanism interpretability.

However, instead of functional groups in the traditional sense, our SAs are randomly split structural fragments, which does not always accord with chemical intuition. Compared to the expert

rules defined by toxicologists, this is a common defect of SAs derived from the database. Whereas, our SAs have an advantage that can fully explore the molecular fragments generated from the data set and finds new information. For example, the SA “ester of phosphonic acid” (Table 6), which derived from the reference, limited the matched condition that the phosphate atom is connected to at least two oxygen atoms. However, from the data set, we found that just the structure of methyl phosphenite (SA\_15) was necessary, and another connection could link an oxygen or nitrogen atom. Moreover, the review<sup>21</sup> shows that the electrophilic sites of this SA are P or O of the structural element P–O–C. Thus, SA\_15 can correctly predict more positive compounds. Another example is hydrazine (SA\_23). Hydrazine derivatives are closely related to human population and have applications as precursors to some pharmaceuticals. However, their use are limited by toxicity side effects, such as hepatotoxicity and carcinogenicity. In the metabolism of alkyl-substituted hydrazine derivatives, the alkyl is oxidized to carbocations or alkyl radicals by different catalytic pathways,<sup>21</sup> and these intermediates can induce DNA lesions.<sup>22,23</sup> This indicates that the alkyl carbon atom attached to the hydrazine nitrogen atom is important for its toxic effects, and the structure of SA\_23 has also verified this. A similar situation exists in the SA of “ester of sulfonic acid” (SA\_16). The methylene carbon atom to the ester oxygen atom is attacked by nucleophiles, and the sulfonic acid anion is displaced.<sup>21</sup> Our SA\_16 also extracted this feature through the statistical method.

Some SAs without sufficient bond information were extracted and analyzed. The SA\_53 contains a sp<sup>2</sup> nitrogen atom (N<sub>2</sub>) that may be connected through a double bond to an oxygen or carbon atom according to the data set. If a N<sub>2</sub> atom connects to an oxygen atom, this SA is a typical SA of *N*-nitrosoureas. The reactions of *N*-nitrosoureas contain the intra- and intercellular alkylation and carbamylation with various biomolecules, especially DNA (Scheme 1),<sup>24</sup> which can seriously affect the

**Scheme 1. Carcinogenic Mechanism of *N*-Nitrosoureas**



health of normal cells and induce cancers.<sup>25</sup> A similar situation exists in SA\_38 and SA\_41 as well.

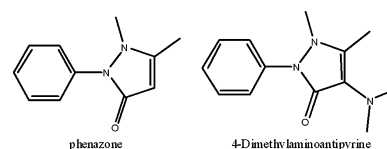
Detailed inspection of the resulted SAs reveals some yet unnoticed correlations between chemical motifs and biological activities. For example, a series of methylformylhydrazone analogues that matched by SA\_50 are experimentally verified as carcinogens. However, because these chemicals do not possess “classic” structural alerts, they cannot be correctly identified in

conventional rule-based expert systems. Another interesting point to note is that the structure of SA\_50 itself represents a carcinogen, which also implicates the validity of this SA. Similar situations can be observed in the cases of SA\_7 and SA\_27. The experiments data show that diethanolamine (SA\_27) is a carcinogen in multispecies. It can induce liver neoplasms through induction of the choline deficiency.<sup>26</sup> These three SAs and compounds that contain these SAs are illustrated in Table 7.

**Table 7. SA\_50, SA\_7, SA\_27 and Compounds That These SAs Can Match in the Training Set**

Structural alerts	Compounds that SAs can match
	Acetaldehyde methylformylhydrazone, Hexanal methylformylhydrazone, 3-Methylbutanal methylformylhydrazone, Pentanal methylformylhydrazone
	Allyl isovalerianate
	Diethanolamine, Triethanolamine

A shortcoming of our method, as well as Benigni’s model, is that the metabolites of compounds are ignored. Thus, the correct identification of compounds, which have just minor differences in structures but different classifications, are usually difficult. For example, SA\_23 can match phenazone and 4-dimethylaminoantipyrine (Figure 4). Phenazone is positive because the



**Figure 4. Structure of phenazone and 4-dimethylaminoantipyrine.**

metabolites of phenazone but not phenazone may result in the carcinogenic action.<sup>27</sup> The tertiary amino group may hinder this metabolic process, so 4-dimethylaminoantipyrine is negative. However, our method that extracts information from current data set cannot deal with this situation. Thus, in future work, considering the metabolites when building QSAR models can not only improve the results but also interpret reaction mechanisms.

Some SAs detected more than six compounds in the training set, but the ratio of positives was not significantly higher than the background positive rate. Table S1 of the Supporting Information lists the performance of S1, S2, S3, and S4. Table S2 of the Supporting Information shows the performance statistics of four pairs of SAs and their corresponding modulating factors. In the reference, aromatic *N*-acyl amine is a SA, and the carcinogenic potency disappears with ortho-disubstitution or a carboxylic acid substituent in the ortho position or a sulfonic acid on the ring with aromatic amino groups.<sup>8</sup> Through three self-defined rules, we found that M4 connected S4 (which met the matched conditions of aromatic *N*-acyl amine) could match a series analogs of diazepam, which were negatives. The methyl of M4 makes diazepam react with demethylation, and the major metabolite, nordazepam, is noncarcinogen in CCRIS.<sup>28</sup>

Considering the size of the data set and the chemical diversity it covers, the number of resulted modulating factors is small.



However, it does not imply an insignificant role of modulating factors. It can be found that introducing these three extra rules indeed increased the sensitivity of our model. Methodologically, our approach has implicitly taken into account diverse mechanisms, thus it is not unexpected that the potential mechanism is not straightforward as to how a modulating factor suppresses the toxic effect caused by a corresponding SA. Further work is required to improve the knowledge about the inhibition mechanism of modulating factors.

## CONCLUSION

In this study, a method by building and pruning a molecular fragments tree was introduced to estimate the carcinogenic potential of compounds. Fragments were extracted using the Gaston algorithm, which detected efficiently substructures of any size, shape, atomic, and bond information. Then the method of building and pruning a molecular fragments tree was applied to select SAs with high quality. It differs from a conventional rule-based expert system and uses a completely automated and unbiased manner to select SAs with statistical significance. Modulating factors help us identify some SAs that cannot efficiently predict carcinogenicity. Moreover, our model has higher predictive ability than the reference model, especially in the test set. So we expect the developed method will be a useful tool to identify structural features of carcinogens and provide some inspiration for drug development and cancer research.

## ASSOCIATED CONTENT

### Supporting Information

Supporting Information Table.doc: Table S1 for the performance of S1, S2, S3, and S4, and Table S2 for the performance of 4 SA/modulating factor pairs. Supporting Information Trainingset\_smiles.txt: The SMILES of the training set. Supporting Information Testset\_smiles.txt: The SMILES of the test set. This material is available free of charge via the Internet at <http://pubs.acs.org>.

## AUTHOR INFORMATION

### Corresponding Author

\*E-mail: myzheng@mail.shcnc.ac.cn (M.Z.); xmluo@mail.shcnc.ac.cn (X.L.). Phone: +86-21-50806705 (Z.L.). Fax: +86-21-50807188 (Z.L.).

### Author Contributions

<sup>||</sup>These authors contributed equally to this work.

### Notes

The authors declare no competing financial interest.

## ACKNOWLEDGMENTS

This work was supported by the Hi-TECH Research and Development Program of China (Grant 2012AA020308), National S&T Major Project (Grant 2012ZX09301-001-002), National Natural Science Foundation of China (21021063), and State Key Program of Basic Research of China (Grant 2009CB918502).

## ABBREVIATIONS

SA, structural alert; Gaston, GrAph/Sequence/Tree extraction; lazar, lazy structure–activity relationships; PTC, Predictive Toxicology Challenge; TP, true positives; TN, true negatives; FP, false positives; FN, false negatives

## REFERENCES

- (1) Ashby, J. Fundamental structural alerts to potential carcinogenicity or noncarcinogenicity. *Environ. Mutagen.* **1985**, *7* (6), 919–921.
- (2) Kruhlik, N. L.; Contrera, J. F.; Benz, R. D.; Matthews, E. J. Progress in QSAR toxicity screening of pharmaceutical impurities and other FDA regulated products. *Adv. Drug Delivery Rev.* **2007**, *59* (1), 43–55.
- (3) Llorens, O.; Perez, J. J.; Villar, H. O. Toward the design of chemical libraries for mass screening biased against mutagenic compounds. *J. Med. Chem.* **2001**, *44* (17), 2793–2804.
- (4) Helma, C.; Cramer, T.; Kramer, S.; De Raedt, L. Data mining and machine learning techniques for the identification of mutagenicity inducing substructures and structure activity relationships of non-congeneric compounds. *J. Chem. Inf. Comput. Sci.* **2004**, *44* (4), 1402–1411.
- (5) Miller, E. C.; Miller, J. A. Searches for ultimate chemical carcinogens and their reactions with cellular macromolecules. *Cancer* **1981**, *47* (10), 2327–2345.
- (6) Ashby, J.; Tennant, R. W. Chemical structure, Salmonella mutagenicity and extent of carcinogenicity as indicators of genotoxic carcinogenesis among 222 chemicals tested in rodents by the U.S. NCI/NTP. *Mutat. Res.* **1988**, *204* (1), 17–115.
- (7) Bailey, A. B.; Chanderbhan, R.; Collazo-Braier, N.; Cheeseman, M. A.; Twaroski, M. L. The use of structure–activity relationship analysis in the food contact notification program. *Regul. Toxicol. Pharmacol.* **2005**, *42* (2), 225–235.
- (8) Benigni, R.; Bossa, C. Structure alerts for carcinogenicity, and the Salmonella assay system: A novel insight through the chemical relational databases technology. *Mutat. Res.* **2008**, *659* (3), 248–261.
- (9) Kazius, J.; McGuire, R.; Bursi, R. Derivation and validation of toxicophores for mutagenicity prediction. *J. Med. Chem.* **2005**, *48* (1), 312–320.
- (10) Helma, C. Lazy structure–activity relationships (lazar) for the prediction of rodent carcinogenicity and Salmonella mutagenicity. *Mol. Divers.* **2006**, *10* (2), 147–158.
- (11) Stefan Kramer, L. D. R.; Helma, C. Molecular Feature Mining in HIV Data. In Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-01), 2001; pp 136–143.
- (12) Kazius, J.; Nijssen, S.; Kok, J.; Back, T.; Ijzerman, A. P. Substructure mining using elaborate chemical representation. *J. Chem. Inf. Model.* **2006**, *46* (2), 597–605.
- (13) SDF Download Page, U.S. EPA. [http://www.epa.gov/ncct/dsstox/sdf\\_isscan\\_external.html](http://www.epa.gov/ncct/dsstox/sdf_isscan_external.html) (accessed July 8, 2012).
- (14) Istituto Superiore di Sanità Website. <http://www.iss.it/ampp/dati/cont.php?id=233&lang=1&tipo=7> (accessed July 8, 2012).
- (15) Benigni, R.; Bossa, C.; Vari, M. R. Chemical Carcinogens: Structures and Experimental Data. <http://www.iss.it/binary/ampp/cont/ISSCANv2aEn.1134647480.pdf> (accessed July 8, 2012).
- (16) The Predictive Toxicology Challenge (PTC) for 2000–2001. <http://www.predictive-toxicology.org/ptc/> (accessed July 8, 2012).
- (17) Sybyl, version 6.8; Tripos, Inc.: St. Louis, MO.
- (18) Nijssen, S.; Kok, J. N. A Quickstart in Frequent Structure Mining Can Make a Difference. In 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), 2004, pp 647–652.
- (19) Snyder, R. D.; Green, J. W. A review of the genotoxicity of marketed pharmaceuticals. *Mutat. Res.* **2001**, *488* (2), 151–169.
- (20) Gonzalez, J. A.; Holder, L. B.; Cook, D. J. Predictive Toxicology Challenge: Model of Toxicology Prediction for Male Rats. [http://www.predictive-toxicology.org/ptc/comprehensible\\_models/uta.pdf](http://www.predictive-toxicology.org/ptc/comprehensible_models/uta.pdf) (accessed July 8, 2012).
- (21) Benigni, R.; Bossa, C. Mechanisms of chemical carcinogenicity and mutagenicity: A review with implications for predictive toxicology. *Chem. Rev.* **2011**, *111* (4), 2507–2536.
- (22) Gamberini, M.; Cidade, M. R.; Valotta, L. A.; Armelin, M. C.; Leite, L. C. Contribution of hydrazines-derived alkyl radicals to cytotoxicity and transformation induced in normal c-myc-overexpressing mouse fibroblasts. *Carcinogenesis* **1998**, *19* (1), 147–155.

- (23) Kovacic, P.; Jacintho, J. D. Mechanisms of carcinogenesis: Focus on oxidative stress and electron transfer. *Curr. Med. Chem.* **2001**, *8* (7), 773–96.
- (24) Helguera, A. M.; Gonzalez, M. P.; Cordeiro, M. N.; Perez, M. A. Quantitative structure-carcinogenicity relationship for detecting structural alerts in nitroso compounds: Species, rat; sex, female; route of administration, gavage. *Chem. Res. Toxicol.* **2008**, *21* (3), 633–642.
- (25) Gnewuch, C. T.; Sosnovsky, G. A critical appraisal of the evolution of N-nitrosoureas as anticancer drugs. *Chem. Rev.* **1997**, *97* (3), 829–1014.
- (26) Lehman-McKeeman, L. D.; Gamsky, E. A.; Hicks, S. M.; Vassallo, J. D.; Mar, M. H.; Zeisel, S. H. Diethanolamine induces hepatic choline deficiency in mice. *Toxicol. Sci.* **2002**, *67* (1), 38–45.
- (27) Johansson, S. L. Carcinogenicity of analgesics: Long-term treatment of Sprague-Dawley rats with phenacetin, phenazone, caffeine and paracetamol (acetamidophen). *Int. J. Cancer.* **1981**, *27* (4), 521–529.
- (28) TOXNET, Toxicology Data Network, U.S. National Library of Medicine. <http://toxnet.nlm.nih.gov/cgi-bin/sis/htmlgen?CCRIS> (accessed July 8, 2012).