JOURNAL OF CHEMICAL INFORMATION AND MODELING

ARTICLE

pubs.acs.org/jcim

Identification of Toxifying and Detoxifying Moieties for Mutagenicity Prediction by Priority Assessment

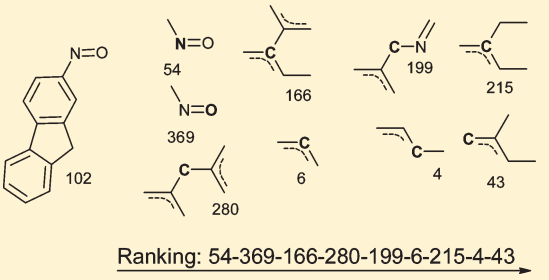
Mose′ Casalegno,[†] Emilio Benfenati,[‡] and Guido Sello*[,§]

[†]Department of Chemistry, Materials, and Chemical Engineering "Giulio Natta", Via Mancinelli 7, I-20131 Milano, Italy

[‡]Istituto di Ricerche Farmacologiche "Mario Negri", Via La Masa 19, I-20156 Milano, Italy

[§]Dipartimento di Chimica Organica e Industriale, Universita' degli Studi di Milano, via Venezian 21, I-20133 Milano, Italy

ABSTRACT: The search for structural subunits that affect compound toxicity cannot be manually performed on large databases. In addition, the a priori definition of important groups is impossible. Structural diversity requires the analysis of the complete data space and the selection of the details there present. A single substructure cannot be considered sufficient when assigning compound toxicity. In contrast, if we consider all the substructures in the database as the elements of a complete collection and if we can build a working hierarchy, the identification of the best feasible result using the available data is possible. If the database includes several significant examples, the results will be valuable. The use of a fragment-based description of a mutagenicity database together with the realization of a general hierarchy allows for the identification of the moieties that control the toxifying/detoxifying action of each compound.



INTRODUCTION

Toxicity of chemical compounds is a side effect that should be minimized to limit negative consequences to humans and the environment. Lawmakers are considering this issue everyday with increasing interest,[1] requiring strict control before a compound can be commercialized. Part of the control can be performed using well-developed models, at least in a preliminary search for toxic compounds. It is thus clear that the study and evaluation of new models and/or new applications of models are important to refining our knowledge concerning the possibility of predicting toxicity.

Mutagenicity is a form of toxicity that deals with the damages that compounds cause to the DNA structure, both affecting its spatial super structure and changing the chemical identity of its components. DNA mutations can have important consequences on cell functioning, decreasing the cell's ability to produce the enzymes needed for its correct functioning. Experimentally, a fast, inexpensive, and efficient test to determine the mutagenicity of chemicals is the Ames test.[2,3] It is well-known and has been applied to many compounds; consequently, much data are available in the literature. However, the available data present a marked level of uncertainty due to the heterogeneity of the experimental procedures and to lab-to-lab differences. Nevertheless, the Ames test is still the best test available for rapid screening. Experimental data are represented in terms of presence/absence (binary numbers) and can only separate definitely toxic from safe compounds. This aspect has the consequence of including a gray area with a greater chance of erroneous assignment when working with slightly toxic compounds, which is always present when using binary selection. Currently, an overall reliability of 85% is accepted by NTP.[4]

Modeling toxicity is a fascinating research activity. It is different from other modeling activities mostly because toxicity is caused by a combination of actions at the molecular level that are often scarcely known and difficult to represent. Yet, many models of mutagenic effects have been developed,[5−35] both knowledge-based and statistics-driven, whose application often has been successful. All the models share the common principle that compound mutagenicity is determined by its structure. This principle has the consequence of assigning all the causes to the structure-neglecting mechanisms of action. Given this common premise, the distinctive characteristic is the technique that is chosen to connect the structure to the mutagenicity; this is the main difference between knowledge-based and statistics-driven models. Dealing with a high number of events (as is the case with large databases), decreases the chances of successful application of knowledge as the core of the model, or better yet, knowledge should be validated by statistics. In 2005, Kazius et al. introduced a fragment-based model combining statistics with mechanistic and chemical knowledge.[11,36] The approval of a candidate substructure as toxicophore (e.g., a substructure indicating an increased potential for mutagenicity) was performed taking into account its chemical similarity with knowledge-based reported toxicophores. The model gave very good results once applied to a large data set made up by 5000 compounds. However, the preliminary selection of possible toxic candidates is a limit that cannot be overcome by this approach. In addition, the selection of a single substructure cannot describe the interaction between toxifying and detoxifying moieties in complex molecules. According to work of

Received: February 15, 2011
Published: June 28, 2011

Klopman et al.,[29] with the term "detoxifying", we indicate a structural feature that prevents the mutagenic activity from being expressed. Consequently, the main objective of our study is the development of a method that can insert all substructures into a common space, compare each substructure to all the others, and select the most appropriate one inside the subgroup of substructures that belong to the single compound. At the root of the model is the principle that we should determine the position of each fragment in the database substructure space considering all possible situations there present. This approach permits the evaluation of all interactions between group pairs and the exclusive use of those interactions that can be applied to the specific case. Moreover, if the pairs are representative (i.e., they have a sufficient number of occurrences in the database), the method implicitly considers all relevant interactions. The more the data are complete, the more the results are reliable. The next step, i.e., the inference of the mechanism of action, is still implicit, and currently, there is no automated method to extract this information. In 2007, Serafimova et al.[37] reported the development of a knowledge-based model where the use of metabolic activation and modes of action were explicitly used. The model is fascinating because it tries to find a mechanistic explanation to most of the facts that can intervene during toxicity occurrence. However, the problem of overcoming the role of a priori decisions remains unsolved. In 2008, Langham and Jain[38] used the same data set to develop a rule-based system to assign mutagenicity. The rules use information coming from a very large set of fragments of different lengths in terms of atoms and bonds. The system does not use expert knowledge to build the rules; in consequence, this approach is more similar to ours. The obtained result is highly interesting because the number of remaining variables is very low in comparison to the starting variable set. In addition, the rules used to assign toxicity are directly interpretable in terms of structural features.

The challenging idea to assign a tag to each compound using all its substructures is intriguing: it can be seen similar to the registration of an IR spectrum, where each signal can be found in many compounds but the whole spectrum is unique. The tag is the condensed description of the whole and is the first approximation; the second is the conjecture that all compounds that share the same tag have a description enough similar to show the same conduct. There is no need to define toxicophores or to model metabolic paths, nor to explicitly calculate similarity.

On the basis of this principle, in a previous work,[39] we introduced a methodology called the top priority fragment (TPF) approach capable of identifying among a pool of substructures those likely to be responsible for the observed toxicity. The method, based on the use of atomic centered substructures,[39−42] was successfully tested in the context of pesticide toxicity. Here, we further develop this strategy to a more general extent, thereby extending its applicability to binary end points such as mutagenicity. To test the reliability of our method, we examine the data set assembled by Kazius et al. The most representative classes of toxicophores found are reviewed in the light of the available literature. Special emphasis is placed in the analysis of detoxifying moieties for which little or no information can be found in the literature.

EXPERIMENTAL PROCEDURE

Method. In this section, we describe a fragment-based approach where fragment contributions to toxicity are evaluated by means of priority relationships. The idea to prioritize fragment contributions to toxicity was originally introduced by our group in the context of pesticide toxicity by means of a method called the TPF approach.[39] Hereafter, we further develop this strategy to a more general extent, extending its application toward binary end points.

Although the method we present in this section closely resembles its predecessor, some important novelties have been introduced that deserve a detailed description, highlighting the main differences between the two methods. The "new" TPF method comprises six sequential steps.

1. Preprocessing of Molecular Structures. In this preliminary stage, all the molecular structures—the whole data set, including both training and test set molecules—are preprocessed so as to optimally restructure the chemical information. This procedure, already described,[42] includes the identification of all heterocyclic and/or condensed aromatic systems and the correction of the bond orders according to their aromatic character. To all bonds belonging to a five- or six-member aromatic ring, a bond order of 4 is assigned and increased by 1 every time the bond is common to two rings.

2. Fragment Generation. Once preprocessed, the molecular structures are broken down into atomic centered units (ACUs) following the described[39] procedure. These fragments are made up by a central (parent) atom and its nearest neighbors. All atoms are specified by atom type and bond order to neighbors; thus, the description is highly defined, generating several different fragments. The total number of ACUs resulting from the breakdown of the entire data set is equal to the total number of atoms in the data set. Having eliminated redundant ACUs (i.e., keeping only one copy of each diverse fragment), the remains are used to build a fragment-based representation for each compound, which consists of a binary fingerprint where fragment occurrence is expressed in terms of integer numbers (1 or 0, indicating presence or absence). Mathematically, fragment occurrence can conveniently be expressed by introducing the occurrence matrix, O, whose elements $O(i,j)$ are defined as follows

$$O(i,j) = \begin{cases} 1; & \textit{if the } j^{th} \textit{ fragment is present in the } i^{th} \textit{ compound} \\ 0; & \textit{otherwise} \end{cases} \quad (1)$$

To simplify the model implementation, the possibility of multiple fragment occurrence has not been considered in eq 1. We shall nonetheless consider its inclusion in the future. Hereafter, we will indicate N_{acu} as the total number of fragments needed to describe the whole data set.

3. Local Priority Ranking. The whole data set is divided into training and test sets. Only training set molecules are used to build the prioritization scheme, which will be further applied to the test set compounds in the prediction stage. The aim of this stage is to build for each compound a hierarchy of ACUs, based on their ability to approximate the observed toxicity value. To this end, we assign to each ACU a toxicity coefficient equal to its arithmetic mean toxicity in the training set. For the generic j^{th} fragment, the average toxicity is defined as

$$T_{ave}(j) = \frac{\sum_{i=1,N_{tr}} T_{obs}(i) \times O(i,j)}{\sum_{i=1,N_{tr}} O(i,j)} \quad (2)$$

where $T_{obs}(i)$ is the observed toxicity value for the i^{th} compound, and N_{tr} is the total number of training compounds. In general,

T_{obs} can be a continuous value, although in this application it only assumes the 0−1 values. To evaluate the position of the j^{th} ACU in the i^{th} molecular hierarchy (i.e., locally), we first compute the absolute difference between its toxicity coefficient and the observed molecular toxicity

$$D(i,j) = |T_{ave}(j) - T_{obs}(i)|;\ for\ every\ j\ such\ that\ O(i,j) = 1 \quad (3)$$

Then, the values of $D(i,j)$ are sorted in increasing order, and the positions in the hierarchy are accordingly assigned. A rank matrix, R, with elements $R(i,j)$, is used to store the positions of the fragments as integer numbers, starting from the number "1", the highest rank. The same position is assigned to all fragments with equal values of $D(i,j)$.

The criterion used to rank the fragments at the local, i.e., molecular level, is analogous to that described in our previous work to identify top priority fragments, TPFs.[39] Here, however, no selection is performed. All fragments, excluding singly occurring ones (i.e., those present only once in the whole training set), are inserted into the molecular hierarchy and may potentially play a role in compound toxicity prediction.

4. Global Priority Ranking. The procedure described above provides N_{tr} molecular hierarchies, one for each training set molecule. These hierarchies define the priority relationships among fragments describing specific chemical environments and cannot be directly used to predict the toxicity of unknown compounds. We may, however, process the chemical information provided by the local hierarchies in order to develop a general, or global, scheme that can rank the fragments, including those of the test compounds, identifying those likely to be responsible for the observed toxicity, i.e., the TPFs. The arithmetic mean of the average TPF(s) toxicities finally provides the predicted toxicity value for the test compound under investigation.

In order to develop the above-mentioned scheme, we introduce the priority matrix P, consisting of $N_{acu} \times N_{acu}$ elements, whose generic elements, $P(j,jj)$, retrieve the priority relationship between the j^{th} and jj^{th} fragments. Different than our previous work, the priority relationships are here expressed as real numbers in the interval $[0, 1]$. To fill the priority matrix, the roles of fragments j and jj are investigated in the training molecules where they are *concurrently* contained. This is addressed by comparing their ranks within these molecules. For the generic i^{th} molecule, a winning score $W(i,j,jj)$ is defined as follows

$$W(i,ii,jj) = \begin{cases} 1, & if\ R(i,j) < R(i,jj) \\ 0, & if\ R(i,j) > R(i,jj) \end{cases} \quad (4)$$

According to the above equation, the j^{th} fragment "wins" when its rank is lower than the jj^{th} one. To provide $P(j,jj)$ with a numerical value, all winning scores are summed up and normalized

$$P(i,jj) = \frac{\sum_{i=1,N_{tr}} W(i,j,jj) \times O(i,j) \times O(i,jj)}{\sum_{i=1,N_{tr}} O(i,j) \times O(i,jj)} \quad (5)$$

The product $O(i,j) \times O(i,jj)$ ensures that the sum is performed only over the molecules where both fragments are concurrently present. It is easy to note that the elements of the priority matrix satisfy the following relationship:

$$P(j,jj) + P(jj,j) = 1 \quad (6)$$

In case the fragments j and jj never occur within the same molecule, the values of $P(j,jj)$ and $P(jj,j)$ are both set to zero for mathematical convenience. Use of eq 5 provides us a simple way to assess whether, on average, $T_{ave}(j)$ would lead to more accurate predictions than $T_{ave}(jj)$, when used to approximate the observed toxicity of a subset of training compounds.

5. Prediction. During this stage, the information retrieved in the priority matrix is exploited for predicting the molecular toxicity. The goal is to identify, for a given compound, the fragment(s) having the highest priority and use its/their toxicity coefficient(s) to predict compound toxicity. To this end, all the fragments occurring in the molecule of interest are considered. To each of them, a score, $S(i,j)$, is assigned and computed as

$$S(i,j) = \sum_{jj=1,N_{acu}} P(j,jj),\ if\ O(i,jj) = 1 \quad (7)$$

The highest score, S_{top}, identifies the fragment with the highest priority, e.g., the TPF (hereafter indicated as J_{top}). In order to check whether other fragments have the same score and should also be regarded as TPF(s), the absolute difference between their score and S_{top} is checked against a small numerical threshold, here equal to 10^{-5}. If only one TPF has been identified, the predicted toxicity of the i^{th} compound, $T_{pre}(i)$, is given by $T_{ave}(J_{top})$; otherwise, it is given by the average of the toxicity coefficients of all TPFs found. The method so far described can generally be applied to any data set to provide continuous toxicity predictions. At this stage, its application to binary mutagenicity data would give real-valued estimates in the range $[0,1]$. An additional step is therefore required to classify a compound either as mutagen or nonmutagen.

6. ROC Analysis and Final Prediction. The ROC (receiver operating characteristics) analysis is a useful tool for evaluating and optimizing the performance of classification models.[43] In the present work, the ROC analysis is performed with the aim at finding a numerical threshold that can be used to discriminate between mutagenic and nonmutagenic character. To begin the analysis, the modeling route described above is applied to the training set compounds. The predicted toxicities are then discretized according to

$$T_{int} = INT[T_{pre}(i) - T_{cut} + 1] \quad (8)$$

where T_{cut} is a numerical cutoff allowed to change in the interval $[0,1]$. Different models can be obtained by simply varying the value of T_{cut}. We are interested in finding the value of T_{cut} leading to optimal model performances in terms of specificity and sensitivity, defined as follows

$$sens(T_{cut}) = \frac{TP}{(TP + FN)} \quad (9)$$

$$spec(T_{cut}) = \frac{TN}{(TN + FP)} \quad (10)$$

where TP is the number of true positives; TN, that of true negatives; FP, that of false positives; and FN, that of false negatives. Taking into account that perfect classification is obtained by setting sens = 1 and spec = 1, the optimal value of T_{cut} can be found by minimizing the distance

$$D = \sqrt{(spec - 1)^2 + (sens - 1)^2} \quad (11)$$

This is done numerically by increasing T_{cut} until the minimum distance is found. At this point, the value of T_{cut} is inserted into

eq 8 and the final toxicities recalculated as binary values. The same cutoff is also used to predict the toxicity of test compounds.

RESULTS

In this work, two databases were investigated. The first, the same studied by Kazius et al.[11] comprising 4337 molecular structures, was used as the training set; the second, derived from the Young database[44] made up by 715 compounds, provided the test set. The total number of compounds investigated was therefore 5052. The two data sets contain mutagenicity data reported in the NTP database; here, the data do not refer to a unique assay, as evidenced by the different standardized experimental methods, bacterial strains, and metabolic activation mixtures that are used. Nevertheless, the average interlaboratory reproducibility was determined to be 85%. In addition, the two data sets were controlled in order to eliminate any overlap. Ames tests either with or without a metabolic activation mixture were considered in assembling these data sets.

Once preprocessed, all molecular structures were broken down into ACUs. A total of 83652 fragments were obtained that after removal of redundant and singly occurring ones reduced to N_{acu} = 2154. The priority relationships among these fragments were evaluated by submitting training set compounds to the local (stage 3) and global (stage 4) priority ranking. The procedure outlined in stage 5 was then used to identify the TPFs.

The ROC analysis was performed with the aim of finding a threshold to distinguish between mutagenic and nonmutagenic character. The ROC curve reported in Figure 1 was obtained increasing the value of T_{cut} from 0 to 1 by a small amount (0.01). Using eq 11, we found optimal model performances to occur at a threshold of 0.69. Binary mutagenicity values for all compounds (training and test ones) were finally obtained by means of eq 8.

Table 1 summarizes the results obtained with our model for the training and test sets. As for the accuracy (Acc), we considered the fraction of correctly predicted compounds. Specificity and sensitivity were estimated by means of eqs 9 and 10. The error percentage was simply taken as $(1 - \mathrm{Acc}) \times 100$. The number of true and false positives and negatives are also reported.

The mutagenicity classification of the training compounds resulted in a total classification error of 9.8%. A total of 2213 compounds, over 2401, were correctly identified as mutagens. A similar result was also obtained for nonmutagens, where the application of our method led to the successful recognition of 1699 true negatives over 1936. Only four compounds, entirely made up by singly occurring fragments, were not predicted.

Model validation was carried out considering a second, independent, data set consisting of 715 compounds (404 mutagens and 311 nonmutagens). The test molecules were directly submitted to the prediction stage (stage 5) and classified by means of eq 8. A classification error of 22% was observed in this case. The fraction of misclassified compounds was higher for mutagens (24.6%) than for nonmutagens (17.9%), in contrast to the training set, where mutagens (7.8%) show a better result than nonmutagens (12.1%).

To further assess the predictive capability of our model, we reported in Table 2 our results and those obtained by Kazius et al.[11] for the training set. The accuracy achieved by both models, very close to the experimental limit of Ames test data, testifies the efficacy of fragment-based methods in determining

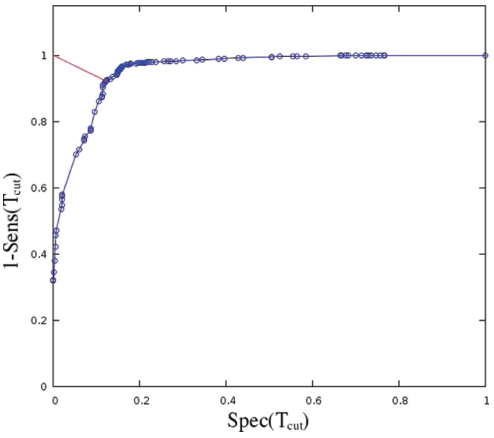


Figure 1. Receiver operating characteristics (ROC) curve shows the variation of the mutagenicity assignment in relation to optimal performance.

Table 1. Results for Training (TR) and Test (TS) Sets

parameter	TR set	TS set
number of compounds	4337	715
cutoff	0.69	0.69
accuracy	0.9020	0.7804
sensitivity	0.9225	0.7537
specificity	0.8785	0.8208
error (%)	9.8	22.0
true positives (TP)	2213	303
true negatives (TN)	1699	255
false positives (FP)	235	55
false negatives (FN)	186	100
unpredicted	4	2

compounds mutagenic character. Our model, nonetheless, gave a smaller classification error, its application resulting in approximately half the number of misclassified compounds.

Given this comparison, we may now focus in greater detail on the found toxicophores. Because of their huge number (759), our discussion will be limited to the most representative TPFs. These are collected in Figure 2.

EXAMPLES

Fragment 54: Nitroso Group. The nitroso group is present in several compounds, and it is considered a typical toxicophore. It is part of several different fragments in our analysis (11 partially overlapping fragments, Figure 3). In the training set, it is present in 122 compounds; 117 of them are classified as toxic compounds, while 5 are nonmutagenic. Our procedure identifies 117 toxic compounds and 5 nontoxic compounds; 101 compounds are found toxic because they contain the nitroso group, 16 are found toxic because they contain other toxicophores. Nonmutagenic compounds are correctly classified because they contain detoxifying fragments that prevent the action of the nitroso group. Comparing this result to the Kazius et al. result, it is possible to note the different weight that the two approaches give to a specific group. When the group is considered a toxicophore without considering the other part of a compound, the choice is

dx.doi.org/10.1021/ci200075g |J. Chem. Inf. Model. 2011, 51, 1564–1574

Table 2. Comparison between Our and Kazius's Results for the Training Set

	error (%)	true positives	true negatives	false positives	false negatives
TR set (this work)	9.8	2213	1699	235	186
TR set (Kazius)	18.0	2019	1539	397	382

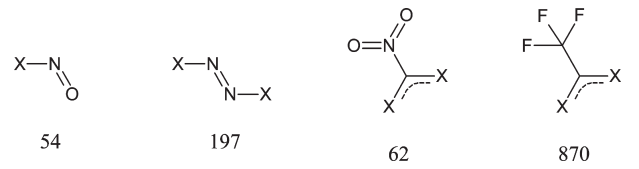


Figure 2. Selection of TPFs discussed in this work. Numbers refer to internal numbering scheme.

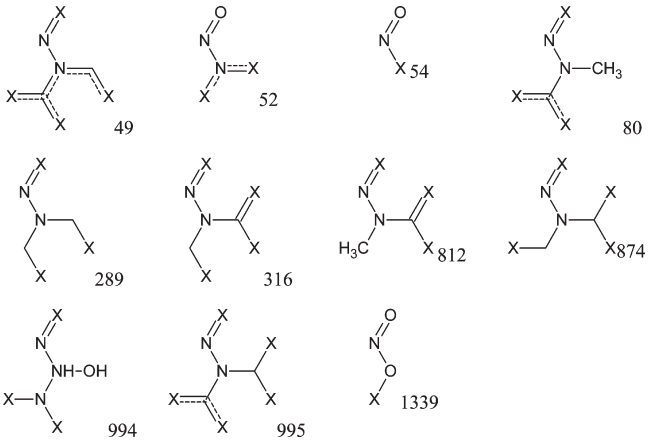


Figure 3. All fragments identified in the database containing the nitroso group. Fragment 54 is the parent ACU.

restricted to a black/white alternative. This can lead to a wrong prediction; in addition, the attribution of the toxicity to a toxicophore excluding other toxicophores can be an oversimplification. In the nitroso case, the overall result is equally good (4% erroneous prediction is a negligible error); nevertheless, it is intriguing that even in this case the straightforward analysis of all the available data could have had a beneficial effect. Another interesting point is that looking at the more specific groups used by Kazius et al. it is possible to note that the number of compounds containing the −N═O group decreases to 118 (aromatic nitrosos, nitrosoamines, alkyl nitrites); where have the other four compounds gone? They have not become nonmutagenic (5 compounds are still in error); consequently, either they are mispredicted (but this is not the case), or they have been inserted in a different toxic group, without an explicit reason.

Concerning the result of this group when present in the test set, 67 compounds are correctly found mutagenic and 1 compound is wrongly predicted mutagenic (448). This compound (Figure 4) does not show any clear reason to be nonmutagenic; it contains a nitrosoamine group and no other group that can be assumed as detoxifying. As a matter of fact, the same compound has been found positive in other similar tests (see, for example, the CCRIS database).[45]

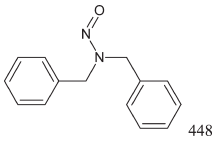


Figure 4. Compound 448 is wrongly predicted; however, its structure contains the N-nitroso alert group.

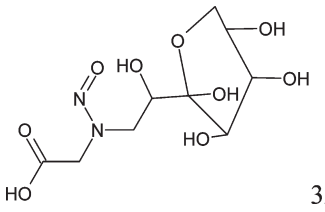


Figure 5. Compound 3246 is correctly predicted; it contains the N-nitroso alert group but it also contains detoxifying moieties.

In our approach, there are 11 training set fragments containing the −N═O group. The parent fragment (−N═O) is present in 122 compounds (in the training set), but it is chosen as the principal fragment only in 38 compounds. In all other cases, a more specific fragment prevails. When these alternative fragments are present, their effect seems to be more important than the general fragment effect. In fact, they are the fragment of choice with high frequency. Also excluding those fragments that are rare, the two fragments that have a good frequency (26 and 24 compounds) often prevail (26 and 23 times, respectively). These fragments (316, always toxic, and 289, 23/24 toxic) are representative of two different classes of nitroso amines, that are present in 49 toxic compounds, i.e., they are very strong toxicophores. Compound 3246 that contains fragment 289 is correctly predicted nonmutagenic (Figure 5). In this molecule, the presence of many hydrophilic groups prevents the toxic action of the N−N═O substructure, increasing the compound water solubility. This result is clearly a special case that does not limit the overall toxicophore character of the group.

Also in the test set, the two fragments that have a good frequency (13 and 35 compounds) often prevail (11 and 17 times, respectively). Then, these fragments (316, always toxic, and 289, 34/35 toxic) are confirmed as very strong toxicophores. When fragments 289 and 316 are not selected in predicted toxic compounds, the parent fragment 54 is chosen; compound 448 is wrongly predicted mutagenic (see above).

Fragment 197: Azo Group. The azo group is present in 158 compounds of the training set, of which 120 are mutagenic and 38 are nonmutagenic compounds. The accuracy level is thus 76%. In Kazius' study, specific fragments 88 azo aromatic compounds and 9 azoxy aromatic compounds are selected, showing 77% (67 mutagenic) and 33% (3 mutagenic) accuracy, respectively. When applying our approach, the 9 azoxy compounds are automatically separated, decreasing the number of azo compounds to 149. However, in contrast to Kazius' result, the azo group is used only 40 times to assign mutagenicity. In all other occurrences, a different group is selected to predict mutagenicity. This allows for the correct nonmutagenic assignment to 26 compounds (only three wrong predictions). In addition, the separation of the azoxy group permits the correct assignment of nonmutagenicity to five compounds that are classified using it (that, consequently, is not a toxicophore;

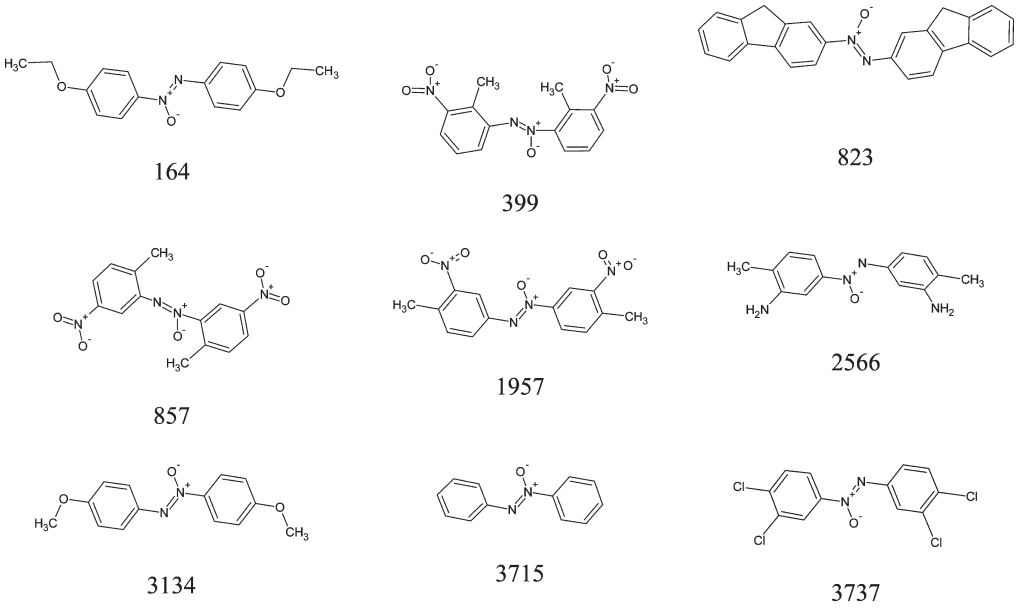


Figure 6. Compounds containing the group $N=N^{+}-O^{-}$. Six of them are nonmutagenic (164, 399, 823, 857, 1857, 3737); three are mutagenic (2566, 3134, 3715).

nonmutagenic: 164, 399, 823, 857, 1957), of mutagenicity (using a different group) to three azoxy-containing compounds (mutagenic: 2566, 3134, 3715), and of nonmutagenicity to the remaining compound (nonmutagenic: 3737) (all correct predictions). (Figure 6)

In the test set, the azo group is not frequent (21 cases); its use (in four cases, only) gives four correct predictions. All other cases are predicted using different groups. In particular, four of the five cases where the compounds are nonmutagenic are correctly predicted; only four cases of the mutagenic compounds are wrongly predicted. The direct use of the azo compounds would have had 76% accuracy, similar to the training set. Our result is exactly the same (76% accuracy); however, the errors have different origin and are partially connected to very complicated structural situations (e.g., compounds 454 and 458, mutagenic compounds containing a sulfonate group; 455, a nonmutagenic compound containing both an azo and an aromatic nitro group). (Figure 7)

The azo group is a typical example of where the use of an a priori fixed mechanism of action can create some problems that cannot be directly uncovered using the statistical approach. A 76% accuracy seems trustworthy.

Fragment 62: Aromatic Nitro Group; fragment 870: Aromatic Trifluoro Group. The aromatic nitro group is very frequent in both training (643 compounds) and test (38 compounds) sets. It is considered a strong toxicophore; in fact, 559 compounds in the training set and 33 compounds in the test set are mutagenic (87% accuracy in both). Aromatic trifluoromethyl group is less common (21 compounds in the training set, 0 compounds in the test set). It is present in 20 nonmutagenic compounds, and it is used in 17 cases to assign toxicity (always correct assignment). What is important here is that in four cases (839, 1857, 1874, 2525) the group is present together with the nitro group, but it always prevails giving the expected mutagenicity value: nonmutagenic.

This situation shows how the hierarchy works: even in the presence of a strong toxicophore ($Ar-NO_2$), if the data are

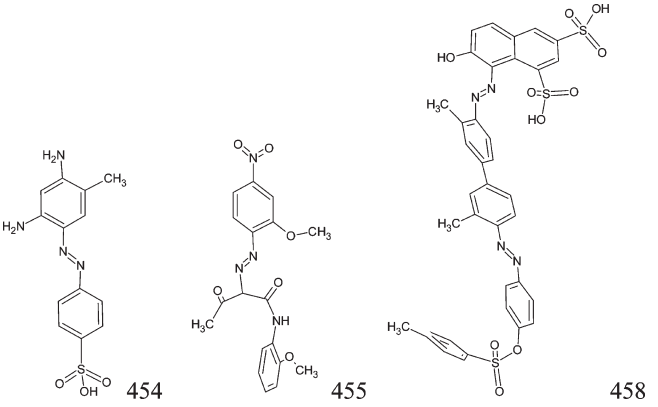


Figure 7. Compounds containing the azo group that are wrongly predicted.

complete enough, the result can be corrected. Is the trifluoromethyl group detoxifying? Looking at the critical four compounds, it seems that there is no other explicit reason that can explain the experimental result. The only toxic compound (3228) that contains $ArCF_3$ is a policyclic aromatic compound, and it is correctly predicted mutagenic. (Figure 8)

Looking at the complete result, we can draw some more conclusions. The total number of TPFs is 759, a number that seems too large. However going into detail, we can see that the first 100 TPFs explain 60% of the predictions, and the first 50 TPFs explain 47% of them. Moreover, 486 TPFs are used 3 times or less. This means that the number of significant TPFs is limited. There are six TPFs that are used most frequently: 62 (toxic) used 311 times, 126 (toxic) used 124 times, 329 (toxic) used 88 times, 271 (nontoxic) used 75 times, 403 (toxic) used 48 times, and 99 (toxic) used 42 times. (Figure 9)

TPF 99 is present 42 times in the training set and 3 times in the test set. It is always selected, and all the predictions are correct. This represents the best overall performance.

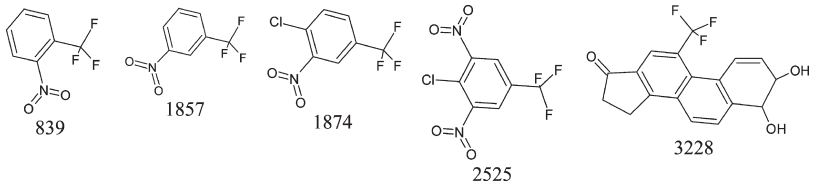


Figure 8. Compounds containing the CF_3 group. They are all correctly predicted. The first four are nonmutagenic, even in presence of an aromatic nitro group; the last is mutagenic, even in presence of the CF_3 group.

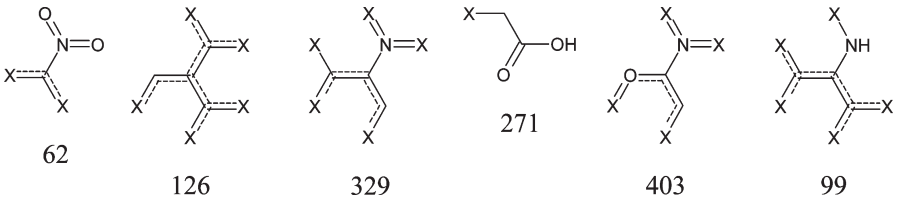


Figure 9. Most frequently used TPFs.

DISCUSSION

Two of the three most recent approaches (Kazius[11,36] and Serafimova[37]) have similar cores: both use a toxicophore-based model. Their target is the selection of the minimal set of atomic groups that can explain the compound toxicities. Both studies apply the concept of expert identification of fundamental groups together with statistical validation of the selected groups. Serafimova et al. add to this condition some descriptors of different origin and, more importantly, a simulation of metabolism modifications of parent compounds. Without going into details, the final result is very similar for both approaches, essentially identifying the same toxicophores and showing equal statistical return.

The absence of a toxicophore automatically means nontoxicity. Some partial corrections to this principle have been introduced without changing the core procedure. These models are attractive because they explicitly connect the compound toxicity to an assumed mode of action and thus are immediately understood by everyone. In contrast, this advantage implicitly contains a weakness: the absence of a recognized mechanism assigns a label of nontoxicity to the compounds, ignoring all new unprecedented situations. In addition, the interactions between groups are only partially considered.

When considering all fragments that are obtained from the dissection of the database compounds, there is a different risk: the number of independent variables is too high to have a statistical significance, thus limiting the possibility of reliable solutions. This problem is generally solved by the adoption of methods such as the principal component analysis or the neural networks, with the negative consequence of losing a direct connection between data and facts.

This problem is solved by Langham and Jain using an automatic selection of rules.[38] The result is very good. The original feature concerns the presence of the rules. In our approach, no explicit rule is created; this allows for a greater system flexibility. In fact, there is no need to create a new rule for each new molecular occurrence, and moreover, there is no need to check the complete rule set when adding a new instance. In contrast, rules are direct explanations of molecularly important features, while TPFs should be investigated in order to get an explicit rationale.

As already mentioned in the Introduction, our approach aims at assigning a tag to each compound using all the available data. In a large database, the combinations of groups in molecules are abundant, and we can suppose that, theoretically, all pair interactions should be present. If this was true, we could order the members of each pair, and if each possible pairing was present, we could order all the groups using the transitive principle, i.e., if A > B and B > C, then A > C. There are two fundamental problems: The first is the absence of completeness of the data (e.g., all groups that are present in a single molecule cannot be ordered). The second is the value that the symbol ">" has. While we cannot solve the first problem, we must focus on the second; the function that we are going to substitute for this symbol will control the model performance.

In a previous paper, we used a similar approach to predict the aquatic toxicity of a set of pesticides.[39] The database was small, and the diversity of molecular structures was high. Nevertheless, we obtained a good overall result. There, the toxicity was determined by lethal doses; consequently, the fragment toxicities were real numbers. In the present case, mutagenicity is a binary number; the assignment of a value to a fragment is therefore more complex. The rule we chose is based on this line of reasoning: if a group is exclusively present in toxic compounds, it is either a toxicophore or neutral (it does not affect toxicity); in contrast, a group that is exclusively present in nontoxic compounds is either detoxifying or neutral (it does not affect toxicity). If a compound only contains toxicophores, it should be toxic. If it does not contain toxicophores, it should be nontoxic. If it contains both toxicophore and detoxifying groups, its toxicity should be calculated. Here, a problem can be raised; it concerns the physical meaning of the term "detoxifying". As mentioned in the Introduction, it is not possible to define a group causing nontoxicity; we can only consider groups that prevent or limit the toxic action of other groups. Their action is mainly correlated to an improvement in the elimination of a compound from the organism (pK_a or water solubility; but also consider the CF_3 group above). However, our approach does not explicitly refer to a mechanism; the model only uses a group as a tag to assign the compound toxicity. The interpretation of the results is only possible a posteriori.

The next step concerns the method for ranking groups with hypothetical toxicity (either toxic or detoxifying). How can we decide if the simultaneous presence of contrasting groups permits the prediction of the compound toxicity? Ideally, if all

possible situations were sufficiently represented in the database, the answer would be the compilation of a complete ranking, and the result should be correct. In our real world (that formed by the current available data), we can nevertheless compile an order, being aware that the result will be an approximation and will be sometimes in error. The ranking is a two step procedure: first, in all group pairs the winning group is selected, and then all pair results are weighted to give a final global order. The global order concerns the single compound. The global winner assigns the position to the compound: either toxic or nontoxic. Later on, we will discuss the results

From the data elaboration of our model, we get some more details. First of all, we can check how many times a group is present in toxic compounds. This gives an indication concerning the possibility that the group is a strong toxicophore that is difficult to detoxify. It is clear that the composition of the database strongly influences the answer. This measure can be used in a successive moment to assign a reliability index to chosen groups.

A second marker is the number of times a group is present in the compounds weighted by the number of times it is chosen as the principal group. This is again an index of the importance of each group. It is clear that a group can be chosen only a few times even if it is present only in toxic compounds; i.e., this marker is different from the previous one. However, in a complete hierarchy, this group will have a bad position, and therefore, could be considered less influential than others.

Competition versus Cooperation. The biological activity of a chemical is usually a complex mechanism that includes several elements: (1) The compound should reach the site where its action takes place. (2) Its structure can experience some transformations before its action can occur. (3) It should position inside the action site in order to have the correct orientation (4) The result of its action should last as much as necessary in order to be detected. As a consequence, a complete model of biological activity requires knowing many features of the path taken by the compound to its final destination. This is not always the case; frequently, the known data are only the compound structure and the final biological outcome. However, it is common to assume that the compound structure (made by atoms and bonds) causes all of the many processes that will take place; i.e., the final evolution of the compound is contained in its structure that will respond to the environment perturbations in a deterministic way. The difficult step is the selection of the appropriate model that can be used to describe the molecular structure. This notwithstanding, even the best description should adapt to the current environment, i.e., it is very uncommon that a model can fit to different situations.

In the case of mutagenicity, it is possible to imagine that different groups can compete for interaction with the biological target, either directly transforming it or facilitating its transformation. In general, a group can compete or cooperate with the interaction progress. Moreover, when competing, a group can elicit a similar response (e.g., mutagenic effect, positive competition) or can contrast the action of another group (e.g., detoxification, negative competition). In the first case, the final result is not affected; in the second case, the final result can be completely different. Negative competition can be revealed by the experimental data, while positive competition can remain unmasked. Group cooperation is even more difficult to determine, in particular, when the biological response is represented by a binary number because cooperation supports the action and only changes its extent.

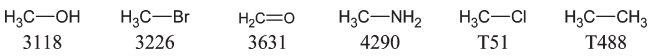


Figure 10. Compounds in the training set that cannot be predicted by our approach.

In this scenario, the difficulty of determining the role of each group present in a molecule is evident. In our study, we tried to solve, as a first step, the problem of negative competition. Using the available mutagenicity data, we detected all the cases where two groups with contrasting effects (toxifying or detoxifying) were present, and we verified if one group prevailed over the other. All the obtained data were then used in an intense comparison, and a complete hierarchy was realized. Finally, scouting about the list, we applied the hierarchy to the current subset of groups present in each single compound.

The application of the model to the training set (4337 compounds) gives a high level of accuracy. It should be pointed that the prediction is made using the general scheme; thus, it correctly considers the hierarchy (i.e., the training set is used to prepare the hierarchy, and then the hierarchy is applied to predict the toxicity). The first result that should be commented on concerns the compounds that cannot be predicted. In fact, after the elimination of the fragments that are present only once, four compounds remain without eligible groups and cannot be predicted. (3118, 3226, 3631, 4290; Figure 10). The accuracy, sensitivity, and specificity of the remaining predictions are 0.9028, 0.9225, and 0.8785, respectively. These values show that the use of all the data content can result in a prediction even better than expected. However, it should be noted that part of this prediction power could be more difficult to rationalize because the selection of some groups is not immediately amenable to biological mechanisms.

When applied to the test set (715 compounds, selected from the compilation of Young et al.[44]), the result can be similarly interpreted. First, two compounds are not predicted (T51, T488; Figure 10). Second, the accuracy, sensitivity, and specificity of the remaining predictions are 0.7826, 0.7537, and 0.8208, respectively. This result cannot be directly comparedto those of either to Kazius'[11] or to Serafimova'[37] or to Langham and Jain's[38] because the test sets are different. What can be noted is that in the Kazius case the test set is derived from the same source (Young's database), but it only contains 535 compounds. However, their accuracy is better (0.85).

A different comparison is nevertheless possible. We randomly selected 72 (10%) of the data set structures and compared our prediction (7 errors) to that obtained using the Serafimova's criteria (10 errors): all compounds that do not contain alert groups are considered nonmutagenic. The result agrees for 58 compounds (81%). The 14 different predictions are examined in the following. (Figure 11)

Compounds 21, 51, and 61 are halogenated compounds. Using Serafimova's criteria, they should be toxic (this can be questioned because in Serafimova's criteria only methylene chloride derivatives are considered toxic); this is true only for methane chloride 51. This compound is not predicted by our model because it cannot be compared to any other fragment, being a single fragment compound. In contrast, the special nonmutagenicity of 21 and 61 cannot be explained by toxicophore approaches.

Compounds 101, 381, and 671 are naphthalene derivatives. Two are nonmutagenic; one is mutagenic. No clear explanation is

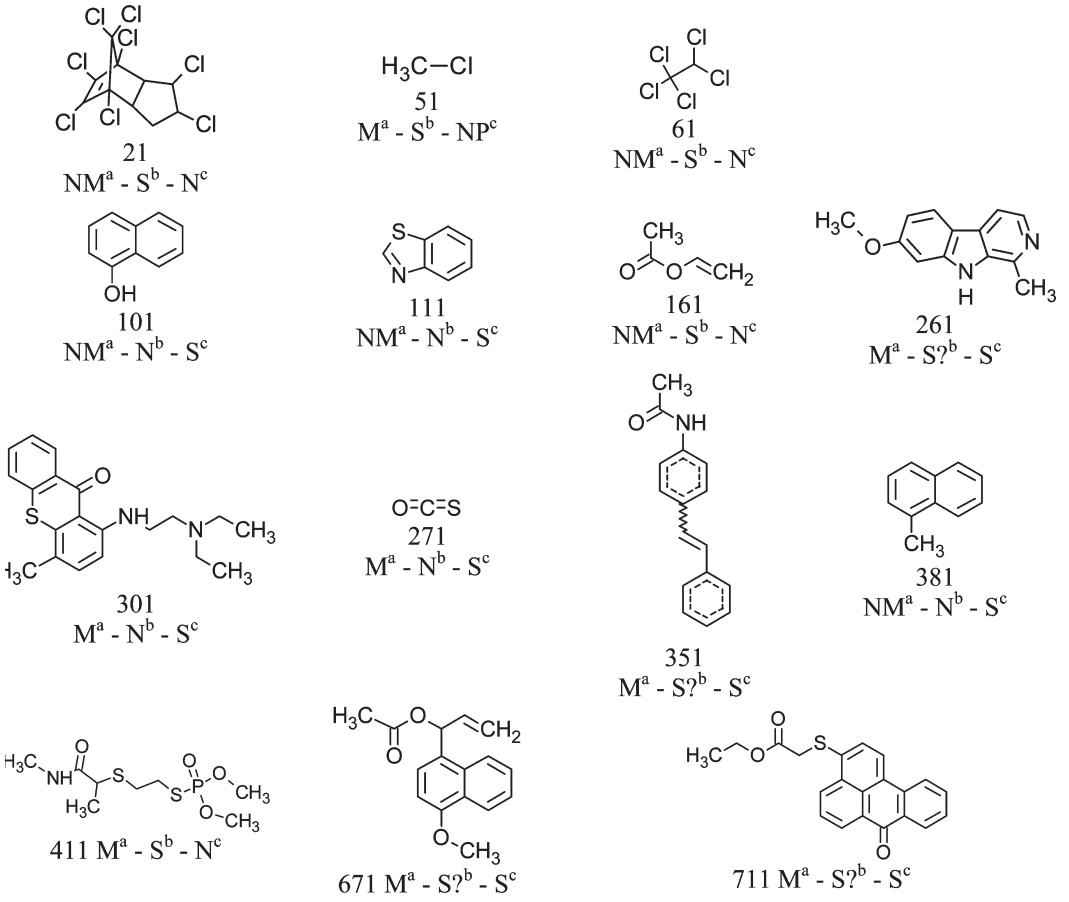


Figure 11. Compounds in the test set used to compare our approach to Serafimova's. (a) Experimental mutagenicity (M = mutagenic, NM = nonmutagenic). (b) Predicted by Serafimova's rules (S = mutagenic, S? = probably mutagenic, N = nonmutagenic). (c) Predicted by this approach (S = mutagenic, N = nonmutagenic, NP = nonpredicted).

available. Because they do not contain any toxicophore, they are nonmutagenic for Serafimova et al.; in contrast, our model predicts them as mutagenic compounds. The presence of two fused aromatic rings is sufficient to predict toxicity. A similar discussion can be also proposed for compounds 111 and 261 (this last is a polycyclic aromatic and may be toxic by Serafimova's criteria).

Compound 161 is nontoxic. Is the acetoxy group an alert group? In Serafimova, it seems to have a role in toxicity. This is not really a clear statement. As a result of our model, we do not have proof that this group is definitely toxic.

Compounds 271, 301, and 351 do not contain toxicophores, but they are toxic. This result is clear in our model that can find sufficient examples of toxic fragments, allowing for their correct classification. Compound 351 contains an acetoxy group and may be considered toxic by Serafimova's criteria.

Compound 411 contains the methyl phosphate toxicophore. However, no sufficient cases are present in the training set to attach the toxic label to this fragment; consequently, it is wrongly predicted.

Compound 711 is correctly predicted by our model. What can be the prediction of these expert approaches? It depends on interpretation; this is a clear example that demonstrates that the expert approach can raise many questions when applied to general structures.

This comparison (that does not aim at rewarding any of the approaches) is useful to better understand the differences between the the two approaches. The expert approach is outstanding when the molecular structure is within its boundaries, but it cannot be safely used outside them. In contrast, our model can give an acceptable answer even in new cases, if there are sufficient data inside the structure set; but, the result can be hard to explain.

CONCLUSION

In this paper, we described an approach to predict mutagenicity on the basis of a hierarchy of molecular fragments. The method does not use a priori knowledge and can be considered data-driven. Results are statistically of good quality both for the training and test sets. The discussion shows that the use of an unbiased analysis can give suggestions even for unprecedented cases. It is nevertheless clear that there is still room for many improvements in toxicity prediction; part of these should concern a greater attention to modes of action, even considering the relatively scarce knowledge of the complex mechanisms involved in toxicity.

AUTHOR INFORMATION

Corresponding Author

*E-mail: guido.sello@unimi.it.

ACKNOWLEDGMENT

The authors thank Prof. S. S. Young for sharing his database and a reviewer for raising the problem of detoxifying groups.

REFERENCES

(1) National Research Council (NRC). *Toxicity Testing in the 21st century: A Vision and a Strategy*; National Academies Press, Washington, DC, 2007.

(2) Ames, B. N.; McCann, H.; Yamasaki, E. Methods for detecting carcinogens and mutagens with the Salmonella/mammalian microsome mutagenicity test. *Mutat. Res.* **1975**, *31*, 347–364.

(3) Josephy, P. D.; Cruz, P.; Nohmi, T. Recent advances in the construction of bacterial genotoxicity assays. *Mutat. Res.* **1997**, *386*, 1–23.

(4) Piegorsch, W. W.; Zeiger, E. Measuring intra-assay agreement for the Ames *Salmonella* assay. In *Lecture Notes in Medical Informatics*; Springer-Verlag: Heidelberg, 1991; pp 35−41.

(5) Ariens, E. J. Domestication of chemistry by design of safer chemicals: Structure−activity relationships. *Drug Metab. Rev.* **1984**, *15* (3), 425–504.

(6) Benigni, R.; Bossa, C. Structural alerts of mutagens and carcinogens. *Curr. Comput.-Aided Drug Des.* **2006**, *2*, 1–19.

(7) Ashby, J. Fundamental structural alerts to potential carcinogenicity and non-carcinogenicity. *Environ. Mutagen.* **1985**, *7*, 919–921.

(8) Ashby, J.; Tennant, R. 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.

(9) Tennant, R.; Ashby, J. Classification according to chemical structure, mutagenicity to *Salmonella* and level of carcinogenicity of a further 39 chemicals tested for carcinogenicity by the U.S. national toxicology program. *Mutat. Res.* **1991**, *257* (3), 209–227.

(10) Ashby, J.; Tennant, R. Definitive relationships among chemical structure, carcinogenicity and mutagenicity for 301 chemicals tested by the U.S. NTP. *Mutat. Res.* **1991**, *257* (3), 229–306.

(11) Kazius, J.; McGuire, R.; Bursi, R. Derivation and validation of toxicophores for mutagenicity prediction. *J. Med. Chem.* **2005**, *48* (1), 312–320.

(12) Sanderson, D.; Earnshaw, C. Computer prediction of possible toxic action from chemical structure: The DEREK system. *Hum. Exp. Toxicol.* **1991**, *10* (4), 261–273.

(13) Ridings, J. E.; Barratt, M. D.; Cary, R.; Earnshaw, C. G.; Eggington, C. E.; Ellis, M. K.; Judson, P. N.; Langowski, J. J.; Marchant, C. A.; Payne, M. P.; Watson, W. P.; Yih, T. D. Computer prediction of possible toxic action from chemical structure: An update on the DEREK system. *Toxicology* **1996**, *106* (1−3), 267–279.

(14) Woo, Y.-T.; 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* (1−3), 219–228.

(15) Singer, B.; Kusmierek, J. Chemical mutagenesis. *Annu. Rev. Biochem.* **1982**, *51*, 655–693.

(16) Klopman, G.; Frierson, M.; Rosenkranz, H. The structural basis of the mutagenicity of chemicals in *Salmonella typhimurium*: The GeneTox Data Base. *Mutat. Res.* **1990**, *228*, 1–50.

(17) Rozenkranz, H.; Klopman, G. The structural basis of the mutagenicity of chemicals in *Salmonella typhimirium*: The National Toxicological Program Data Base. *Mutat. Res.* **1990**, *228*, 51–80.

(18) Dearden, J.; Barrat, M.; Benigni, R.; Bristol, D.; Combes, R.; et al. The development and validation of expert systems for predicting toxicity. *ATLA* **1997**, *25*, 223−252; http://www.frame.org.uk/page.php?pg_id=18 (accessed October, 13, 2009).

(19) Greene, N. Computer systems for the prediction of toxicity: An update. *Adv. Drug Delivery Rev.* **2002**, *54*, 417–431.

(20) Pearl, G.; Livingston-Carr, S.; Durham, S. Integration of computational analysis as a sentinel tool in toxicological assessments. *Curr. Top. Med. Chem.* **2001**, *1*, 247–255.

(21) Johnson, D.; Wolfgang, G. Predicting human safety: Screening and computational approaches. *Drug Discovery Today* **2000**, *5*, 445–454.

(22) Benfenati, E.; Gini, G. Computational predictive programs (expert systems) in toxicology. *Toxicology* **1997**, *119*, 213–225.

(23) Cronin, M.; Jaworska, J.; Walker, J.; Comber, M.; Watts, C.; et al. Use of QSARs in international decision-making frameworks to predict health effects of chemicals substances. *Environ. Health Perspect.* **2003**, *111*, 1391–1401.

(24) Benigni, R.; Richard, A. Quantitative structural-based modeling applied to characterization and prediction of chemical toxicity. *Methods* **1998**, *14*, 264–276.

(25) Benigni, R.; Guliani, A. QSAR studies in genetic toxicology: Congeneric and non-congeneric chemicals. *Arch. Toxicol. Suppl.* **1992**, *15*, 228–237.

(26) Richard, A. Application of SAR methods to non-congeneric data bases associated with carcinogenicity and mutagenicity: Issues and approaches. *Mutat. Res.* **1994**, *305*, 73–97.

(27) Enslein, K.; Borgstedt, H.; Tomb, M.; Blake, B.; Hart, H. A structure-activity prediction model of carcinogenicity based on NCI/NTP assays and food additives. *Toxicol. Ind. Health* **1987**, *3*, 267–287.

(28) Enslein, K. An overview of structure-activity relationships as an alternative to testing in animals for carcinogenicity, mutagenicity, dermal and eye irritation, and acute oral toxicity. *Toxicol. Ind. Health* **1988**, *4*, 479–498.

(29) Klopman, G. MULTICASE. 1. A hierarchical computer automated structure evaluation program. *Quant. Struct.−Act. Relat.* **1992**, *11* (2), 176–184.

(30) Schultz, T. W.; Cronin, M. T. D.; Netzeva, T. I. The present status of QSAR in toxicology. *J. Mol. Struct. (THEOCHEM)* **2003**, *622*, 23–38.

(31) Purdy, R. A mechanism-mediated model for carcinogenicity: Model content and prediction of the outcome of rodent carcinogenicity bioassays currently being conducted on 25 organic chemicals. *Environ. Health Perspect.* **1996**, *104*, 1085–1094.

(32) Lewis, D. F. V.; Ioannides, C.; Parke, D. V. Validation of a novel molecular orbital approach (COMPACT) for the prospective safety evaluation of chemicals, by comparison with rodent carcinogenicity and Salmonella mutagenicity data evaluated by the US NCI/NTP. *Mutat. Res.* **1993**, *291*, 61–77.

(33) Mekenyan, O. G.; Ivanov, J. M.; Karabunarliev, S. H.; Bradbary, S. P.; Ankley, G. T.; Karcher, W. A computationally-based hazard identification algorithm that incorporates ligand flexibility. 1. Identification of potential androgen receptor ligands. *Environ. Sci. Technol.* **1997**, *31*, 3702–3711.

(34) Mekenyan, O.; Nikolova, N.; Schmieder, P.; Veith, G. COREPA-M: A multi-dimensional formulation of COREPA. *QSAR Comb. Sci.* **2004**, *23* (1), 5–18.

(35) Mekenyan, O.; Dimitrov, S.; Pavlov, T.; Veith, G. A systematic approach to stimulating metabolism in computational toxicology. I. The TIMES heuristic modelling framework. *Curr.Pharm. Des.* **2004**, *10* (11), 1273–1293.

(36) Kazius, J.; Nijssen, S.; Kok, J.; Baeck, T.; Ijzerman, A. P. Substructure mining using elaborate chemical representation. *J. Chem. Inf. Model.* **2006**, *46*, 597–605.

(37) Serafimova, R.; Todorov, M.; Pavlov, T.; Kotov, S.; Jacob, E.; Aptula, A.; Mekenyan, O. Identification of the structural requirements for mutagencitiy, by incorporating molecular flexibility and metabolic activation of chemicals. II. General Ames mutagenicity model. *Chem. Res. Toxicol.* **2007**, *20*, 662–676.

(38) Langham, J. J.; Jain, A. N. Accurate and interpretable computational modeling of chemical mutagenicity. *J. Chem. Inf. Model.* **2008**, *48*, 1833–1839.

(39) Casalegno, M.; Sello, G.; Benfenati, E. Top-priority fragment QSAR approach in predicting pesticide aquatic toxicity. *Chem. Res. Toxicol.* **2006**, *19*, 1533–1539.

(40) Khüne, R.; Ebert, R.-U.; Schüürmann, G. Model selection based on structural similarity-method description and application to water solubility prediction. *J. Chem. Inf. Model* **2006**, *46*, 636–641.

(41) Khüne, R.; Ebert, R.-U.; Schüürmann, G. Chemical domain of QSAR models from atom-centered fragments. *J. Chem. Inf. Model* **2009**, *49*, 2660–2669.

(42) Casalegno, M.; Benfenati, E.; Sello, G. An automated group contribution method in predicting aquatic toxicity: The diatomic fragment approach. *Chem. Res. Toxicol.* **2005**, *18*, 740–746.

(43) Fawcett, T. An introduction to ROC analysis. *Pattern Rec. Lett.* **2006**, *27*, 861–874.
(44) Young, S. S.; Gombar, V. K.; Emptage, M. R.; Cariello, N. F.; Lambert, C. Mixture-deconvolution and analysis of Ames mutagenicity data. *Chem. Int. Lab. Sys.* **2002**, *60*, 5–11.
(45) TOXNET. United States National Library of Medicine. http://toxnet.nlm.nih.gov/cgi-bin/sis/htmlgen?CCRIS (accessed May, 5, 2011).