SAR and QSAR in Environmental Research
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/gsar20

Automatic knowledge extraction from chemical structures: the case of mutagenicity prediction

T. Ferrari [a], D. Cattaneo [a], G. Gini [a], N. Golbamaki Bakhtyari [b], A. Manganaro [b] & E. Benfenati [b]

[a] Department of Electronics and Information, Politecnico di Milano, Milan, Italy

[b] Department of Environmental Health Sciences, Istituto di Ricerche Farmacologiche Mario Negri, Milan, Italy

Published online: 28 May 2013.

SAR and QSAR in Environmental Research, 2013
http://dx.doi.org/10.1080/1062936X.2013.773376

Automatic knowledge extraction from chemical structures: the case of mutagenicity prediction[£]

T. Ferrari[a], D. Cattaneo[a], G. Gini[a]*, N. Golbamaki Bakhtyari[b], A. Manganaro[b] and E. Benfenati[b]

[a]*Department of Electronics and Information, Politecnico di Milano, Milan, Italy;* [b]*Department of Environmental Health Sciences, Istituto di Ricerche Farmacologiche Mario Negri, Milan, Italy*

(*Received 17 June 2012; in final form 3 August 2012*)

This work proposes a new structure–activity relationship (SAR) approach to mine molecular fragments that act as structural alerts for biological activity. The entire process is designed to fit with human reasoning, not only to make the predictions more reliable but also to permit clear control by the user in order to meet customized requirements. This approach has been tested on the mutagenicity endpoint, showing marked prediction skills and, more interestingly, bringing to the surface much of the knowledge already collected in the literature as well as new evidence.

Keywords: SARpy; structural alerts; mutagenicity; data mining; SMILES; QSAR

1. Introduction

This paper deals with qualitative structure–activity relationships (SAR). SAR typically uses rules created by experts and expert systems to produce models that relate molecular or chemical substructures to a biological property – toxicity in our case. Here we show a new method to automatically develop such rules if a suitable set of results of biological experiments is available.

Data on the biological activity of chemical substances have triggered a proliferation of data mining approaches to predict the toxicity of unknown substances. In most cases, statistical tools search for a numerical correlation between chemical properties and biological activity. These models have significant prediction abilities on new compounds and can be profitably used for classification, but it is hard to extract the underlying rationale. Physicochemical properties or structural information on chemicals are numerically quantified into so-called molecular descriptors [1], whose chemical or biological meaning is not obvious. Sometimes, the equation that binds an instance to its prediction is not intelligible. This may be the case of neural networks, where often good performance is closely related to network complexity.

The structure of chemicals is explicitly taken into account by some graph-mining approaches, such as AGM [2], FSG [3] and MoFa [4], which mine large datasets for frequent substructures using ‘a priori’, an algorithm for rule induction designed for finding frequent item sets in a database [5].

Human experts usually estimate toxicity on the basis of detection of structural fragments already known to be responsible for the toxic property under investigation. Such fragments are

*Corresponding author. Email: gini@elet.polimi.it

[£]Presented at the 15th International Workshop on Quantitative Structure-Activity Relationships in Environmental and Health Sciences (QSAR2012), 18-22 June 2012, Tallinn, Estonia.

referred to as Structural Alerts (SAs) [6], toxicophores [7] or biophores [8], and human experts can obtain them from knowledge of the biochemical mechanism of action (such as the activation of an enzyme cascade or the opening of an ion channel, which leads to a biological response).

Only a few approaches have been developed to help experts extract this knowledge from data. Some are based on inductive logic programming (ILP) [9]; they cannot be directly applied to standard chemical formats for molecule representation and require extra computation. Others, including MCASE [10] and recently LAZAR [11], use a mixed approach. MCASE mines relevant fragments from a set of experimentally tested molecular structures (training set) by breaking down each structure into its constituent parts and selecting the ones with statistically significant non-random distribution among the active and inactive classes of compounds. The fragments that appear mostly in active molecules, and may therefore be responsible for the biological activity, are labelled biophores; additional features that seem to regulate a biophore's activity are called modulators and can influence the final prediction. LAZAR searches only for linear fragments selected with the chi-square statistical test; the final prediction is determined by a weighted majority vote from neighbours. In both cases, only simple substructures are taken into account on a purely statistical basis. It is worth mentioning that while LAZAR is open source and MCASE is commercial.

We developed and used SARpy (SAR in python), a new ad hoc SAR approach aimed at finding relevant fragments in a transparent way, to extract a set of rules directly from data without any 'a priori' knowledge. The algorithm generates substructures of arbitrary complexity, and the fragment candidates to become SAs are automatically selected on the basis of their prediction performance on a training set.

The output of SARpy consists in a set of rules in the form: 'IF contains <SA> THEN <apply label>', where the SA is expressed as Simplified Molecular Input Line Entry Specification (SMILES) [12], for use by human experts or other chemical software. Those rules can be used as a predictive model simply by calling a SMARTS matching program.

SMARTS (SMiles ARbitrary Target Specification) strings are a text representation of substructures [13]. While very similar to SMILES, SMARTS also allows specification of wildcard atoms and bonds, which can be used to formulate substructure queries for a chemical database. To be matched, the SMILES and the SMARTS strings are translated into graphs and the two graphs compared.

In SARpy, fragmentation is done directly on the SMILES notation of structures. A similar approach has been implemented in SMIREP [14], but there the SMILES strings were simply split into 'branching fragments' and 'cyclic fragments'. In other words, only entire branches or entire cycles are considered (that is to say, in the SMILES syntax, from parenthesis to parenthesis and/or from number to number). CORAL too [15] uses small SMILES fragments, but they are finally merged into a numerical molecular descriptor, so the whole structural information content of the SMILES string is never explicitly taken into account. In our method instead we explicitly consider each bond.

In the following sections we introduce the property we intend to model, the conceptual view of our approach, and its results on a large dataset of publicly available molecular structures tested with the Ames test for mutagenicity. Finally, the results and the knowledge extracted are discussed and compared with the present state-of-the-art of the mutagenicity domain.

1.1 *Tests and prediction methods for mutagenicity*

Mutagenic toxicity is the ability of a substance to cause genetic mutations. This property is of considerable public concern because of its close relationship with carcinogenicity and possible reproductive toxicity [16,17]. Today, regulators require the mutagenicity potency to correctly label mutagens/carcinogens and restrict exposure to them. Another important field is drug and pesticide discovery, where the development of potential mutagens/carcinogens should be stopped as early as possible.

Mutagenic toxicity can be experimentally assessed by various test systems. The most common is the Ames test for mutagenicity, using several strains of genetically engineered *Salmonella typhimurium*, sensitive to a large array of DNA-damaging agents [18,19]. As discussed by Piegorsch and Zeiger [20] the estimated inter-laboratory reproducibility of Salmonella test data is 85%.

In silico methods proposed for mutagenicity include quantitative structure–activity relationship (QSAR) and SAR. These can be generated using a wide variety of statistical methods and a large choice of molecular descriptors.

Usually, the first step in making a QSAR model is to calculate the molecular descriptors [1] or the other type of numerical values (such as fingerprints) used to encode the structural characteristics of a chemical compound into a fixed bit vector [21]. QSAR has been applied for predicting mutagenicity. One of the first attempts [22] used only four descriptors, namely the energy level of the lowest unoccupied molecular orbital (LUMO); the partition coefficient between octanol and water (log *P*); a structural indicator and a descriptor to exclude molecules considered outliers. This model was built using 230 nitro-aromatic compounds. One problem in the method was the use of a quantum computation descriptor (LUMO), which needed a long process time to be obtained, and the limitation to a single chemical class.

The SAR approach aims to identify particular structural fragments of a molecule known to be responsible for the toxic property under investigation. In the mutagenicity/carcinogenicity domain, the key contribution in the definition of such toxicophores was made by Ashby [19]. Basing his work on the electrophilicity theory of chemical carcinogenesis developed by Miller and Miller [23], which correlates the presence of electrophiles (like halogenated aliphatic or aromatic nitro substructures) to genotoxic carcinogenicity, Ashby compiled a list of 19 SAs for DNA reactivity. In a later study, Ashby and Tennant [24] mined a few hundred data from the US National Toxicology Program (NTP) manually to confirm their findings; however, the authors did not present numerical correlations between individual substructures and mutagenicity because their database was not large enough.

Every subsequent effort has started from knowledge collected by Ashby to derive more specific rules, such as a more recent work [7] where an understanding of the mechanism of action is combined with statistical criteria. The analysed dataset includes more than four thousand molecules with the respective Ames test binary results. A drawback is that molecules tested with different methods (with and without metabolic trial) are mixed; however, it is widely accepted and used in the scientific community. From this core data a few other papers have been published [25,26].

If the aim is to use mutagenicity as an indicator of carcinogenic substances, the correlation between mutagenicity and rat carcinogenicity is poor [27]. In particular, while mutagens correlate with carcinogens, non-mutagens do not correlate with non-carcinogens [28]; thus most of the SAs for genotoxicity are also present in the list of the SAs for carcinogenicity.

Practically SAs, in the context here, are rules which state the condition of mutagenicity depending on the presence or absence of specific chemical substructures.

The mutagenicity SAs are based on observations of chemicals with that moiety. These SAs derive from chemical properties and have a sort of mechanistic interpretation. However their presence alone does not give a definitive method to prove the mutagenicity of a new compound towards Salmonella, since there is a number of false positives in many cases, probably because other substituents can change the classification. For instance, Snyder and colleagues [29] reported the results of checking the main commercial systems built over rules in predicting the mutagenicity of pharmaceutical compounds; sensitivity of all systems was poor. Moreover, in many cases toxicity was present but no SAs.

In previous work [30,31] we showed the advantage of integrating QSAR and SAR to improve the accuracy of the classifier. Here we illustrate a deeper integration, since the SAs are automatically extracted from data. The SAs are combined using statistical tools.

2. Materials and methods

2.1 *The SARpy paradigm of knowledge extraction*

Given a training set of molecular structures, with their experimental activity binary labels, SARpy generates every substructure in the set and mines correlations between the incidence of a particular molecular substructure and the activity of the molecules that contain it. This is done in three steps starting just from the structural SMILES notation:

(1) Fragmentation: this novel, recursive algorithm considers every combination of bond breakages working directly on the SMILES string. This fast procedure is capable of computing every substructure of the molecular input set.
(2) Evaluation: each substructure is validated as potential SA on the training set. It is a complete match against the training structures, aimed at assessing the predictive power of each fragment.
(3) Rule set extraction: from the huge set of substructures collected, a reduced set of rules is extracted in the form: ‘IF contains <SA> THEN <apply activity label>’.

2.1.1 *Fragmentation*

The aim of this phase is to detect the chemical substructures present in the set of training chemicals. This challenging task is carried out in a straightforward manner: the totality of substructures is identified by recursively applying a very simple fragmentation algorithm. This performs only a rough fragmentation of the input structures, but by iterating each fragmentation step on the output of the previous one, it is possible to collect substructures of increasing complexity until the in-depth fragmentation of the original structures is complete.

In detail, each fragmentation step iterates over every bond in the input structures and collects the two fragments that would result if the bond were broken. Each step considers all the possible couples of fragments obtainable from each input structure. After the first step all the substructures derivable from every bond breakage (taken individually) are computed and collected. Applying the next fragmentation step to the output of the previous one, all the possibilities of a second bond breakage are explored, and so on, until no more new fragments can be extracted. For example, on a general A–B–C structure, the first fragmentation step will raise the two fragments A and B–C by breaking the first bond, then the A–B and C fragments by breaking the second bond; the B fragment will be found in the next step. See Figure 1 for a real example. The proposed breadth-first approach considers every combination of bond breakages, adding at each search level the possibility of further breaking, but only new substructures are added to the collection and propagated through the fragmentation.

The chemical structures are fragmented directly on their SMILES strings. The time complexity of algorithms working on strings is polynomial, an advantage with respect to

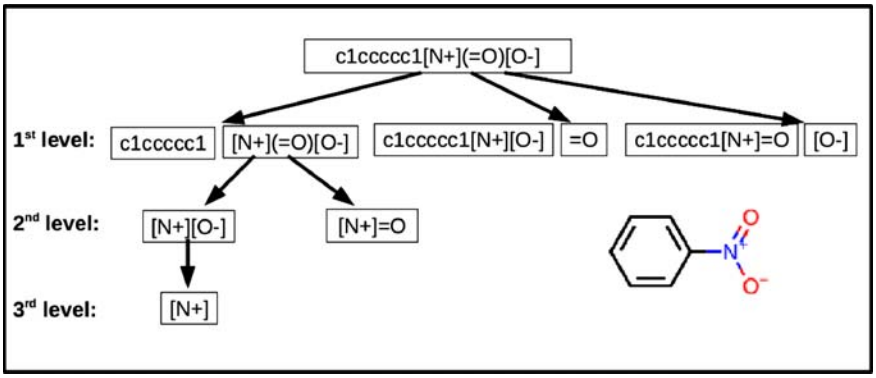


Figure 1. SMILES fragmentation. Duplicates are omitted. The SMILES of the starting structure is at the top.

algorithms working on graphs. In fact, considering that parsing with context-free grammars has a cubic worst-case time complexity, the complexity of SARpy fragmentation is polynomial too. Thus the problem of processing a two-dimensional molecular graph is reduced to fast processing of ASCII strings. Breaking a ring bond requires some 'workarounds' to correctly rearrange the SMILES structure. Breaking a ring bond does not mean splitting the structure, which should remain connected, with just the ring 'unrolled'. For simplicity, it was decided to consider rings as single entities during the fragmentation phase (i.e. ring bonds are not broken), the idea being that the same substructures contained in a ring might be found as an open skeleton in other compounds in the training data; otherwise, if always embedded in a ring, then only the whole ring itself has to be taken into account. However, fragments identified in other parts of the molecule in the fragmentation phase are anyhow identified in the evaluation phase, even in the ring, because more extensive analysis is done in this phase.

A further consideration concerns the length of relevant fragments in terms of number of atoms. If we are interested in general SAs, capable of identifying wide classes of chemicals, the largest substructures could be omitted without noticeable information loss, drastically reducing the time necessary to extract fragments. This is supported by several experiments carried out with the final implementation, where there is evidence that fragments longer than a certain length have no significant effect on the final model. The increase of the upper limit of atoms per fragment extends the required computational time, but over a certain threshold (18 atoms), the outcome of the computation remains the same. With the proposed approach it has been experimentally observed that, even on datasets containing thousands of molecules, the number of new fragments monotonically decreases after the first few fragmentation steps, dropping to zero in a reasonable time. This is because in large datasets, after processing the first thousands chemicals, the other remaining thousands chemicals contain in most of the cases fragments previously found.

2.1.2 *Evaluation*

Once all the existing substructures have been collected, the next phase consists in their individual evaluation as potential SAs on the training set of chemicals. For clarity, the binary case of a 'positive' or 'negative' experimental activity label associated with each structure is considered, with the focus on the search for SAs for positive activity.

First of all, each substructure is matched against every molecular structure in the training data; this huge task can be properly optimized, and time demand for this step is not an issue (see the section on implementation for details). Having the experimental labels, fragment matches can be divided: either against positive structures, called ‘true positives’ (TP), or against negative structures, called ‘false positives’ (FP). From these two, several indicators can be computed to assess the precision of each potential SA to predict the target activity label.

The likelihood ratio, which is a measure of precision intrinsic to the test (not depending on the prevalence of activity labels in the training set), is used:

$$\text{Likelihood ratio} = (\text{TP} / \text{FP}) \times (\text{negatives/positives})$$

Obviously the same procedure can be repeated for different labels, even not binary ones, just considering the target label ‘positive’ and the union of all the others ‘negative’.

The evaluation is aimed at identifying the substructures that best generalize the concept of biophores, with high precision and good sensitivity in the prediction of active chemicals.

We used the likelihood ratio in the next phase to dynamically extract the best set of SAs.

2.1.3 *Rule set extraction*

The final goal is to obtain a reduced set of rules from the huge list of potential alerts, with limited interferences, able to predict the target class with the best precision. This is done as follows:

(1) Order the list of potential alerts by likelihood ratio.
(2) Select the top ranked one, add it to the rule set and remove it from the list of potential alerts.
(3) Remove the TPs and FPs containing the alert just selected.
(4) Update TP and FP values of the remaining potential alerts.
(5) Update the likelihood ratios of potential alerts.
(6) Return to point 1.

With this procedure, which is partially similar to MCASE [32], we can select the next SA, keeping account of the effect of the SAs already in the rule set to minimize interferences and maximize efficiency. The definition of a termination condition markedly affects the behaviour of the rule set, making it more sensitive or more specific. Two basic approaches are described in the Implementation section.

The output can be presented to the user as an ordered set of rules in the form: ‘IF contains <SA> THEN <apply label>’.

2.2 *SARpy implementation*

The implementation is a Python script (about 500 lines of code) employing the open source OpenBabel 2.2.3 library via a set of bindings to the C++ code. The application is easily available through a graphic interface.

The input training structures and their experimental activity label can be submitted either as Structure Data File (SDF) or in a Comma Separated Values (CSV) table with structures expressed in SMILES notation. SARpy handles the molecular data by using Pybel [33], a set of convenience Python functions and classes that simplifies access to the OpenBabel module, converted into SMILES disregarding chirality information.

Figure 2 shows a flow chart of the model construction.

The fragmentation works directly on the SMILES strings, starting from the original training structures and recursively iterating on every resulting new SMILES fragment.

The fragments collected are uniquely identified by their canonical SMILES notation to prevent duplicates and are stored preserving the hierarchical relationships between structures and their substructures. The only parameterization required is the minimum and maximum number of atoms of fragments, set by default respectively at 2 and 18.

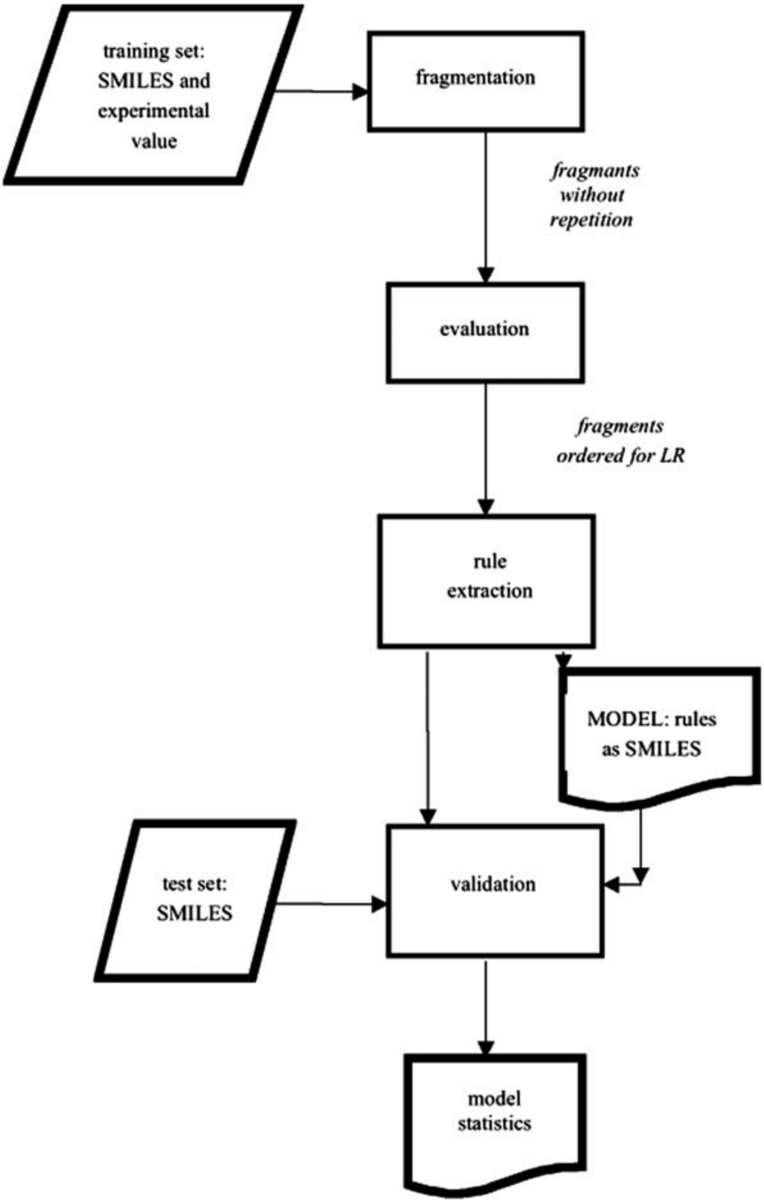


Figure 2. Flow chart of the model construction.

The evaluation phase matches each substructure against all the structures in the training data. The structural comparison is carried out by the OpenBabel SMARTS [13] matching function after being optimized, reducing the number of comparison to compute, by the use of fingerprints [21] and using the hierarchical organization of the fragmentation process. The search for structures potentially containing a given fragment can be restricted to the ones that contain a substructure of the fragment itself. Therefore, the evaluation is done backwards in the hierarchy of fragments; a given fragment is compared only with the structures already matched by one of its descendants. Fingerprints of training structures and potential alerts can be quickly computed and compared in order to discard obvious mismatches. After this step every fragment is paired with its TP and FP matches.

All the potential SAs are ranked according to their likelihood ratios; sensitivity is used as a secondary sorting key in case of equality. To avoid rules with irrelevant or unforeseeable behaviour, a lowest bound on the TP value is considered to exclude SAs with only limited information on their positive prediction ability, even if precise. Rules with a precision worse than the prevalence of the target class in the training set (i.e. likelihood ratio <1) are removed, since they predict worse than a ‘random rule’. Such rules are dynamically pruned again before the extraction of every rule, using the actual TP values and likelihood ratios.

The rule set extraction is driven by the ranking scheme based on likelihood ratios, explained in the previous section, and the termination condition of such extraction determines the behaviour of the model: an early stop could mean a specific but poorly sensitive rule set, with just a few but precise rules; the opposite is true for a late stop. Both the approaches have been explored and implemented to build flexible tools.

The resulting set of rules can be checked on an external test set or cross-validated many times.

2.3 *Experimental data processing*

The dataset employed [7] was retrieved from the European Commission funded CAESAR project [34] in which a mutagenicity model has been implemented [31]. The dataset in [7] originally contained 4337 molecular structures, but after a careful check of each chemical structure, some of them were corrected or removed to avoid inaccuracies. The resulting CAESAR mutagenicity dataset consists of 4204 compounds, 2348 classified as mutagenic and 1856 classified as non-mutagenic by the Ames test.

We used the original split into a training and a test set carried out for the CAESAR model, which was performed following a stratification criterion to make sure each subset approximately covered all major functional groups as well as all major features of the chemical domain of the total compound set. The training set consists of 80% of the data (3367 compounds); the other 20% (837 compounds) was left out for testing. The parameterisation is the default one, from 2 to 18 atoms per fragment. Using the training set, 112 rules were generated.

3. Results

Using SARpy we predicted the test set and obtained the statistics provided in Table 1. A five-fold cross-validation on the training set gave very similar statistics for accuracy. As a further check, we repeated the process on three other splits of the dataset, extracting randomly 20% of the chemicals as test set. As shown in Table 2, the resulting statistical values were

Table 1. SARpy: statistical evaluation on the test set.

	CAESAR mutagenicity dataset	
SARpy	*Training set*	*Test set*
Accuracy	83%	82%
Sensitivity	85%	85%
Specificity	80%	78%

Table 2. SARpy: statistical evaluation on three random test sets.

First split	*Training set*	*Test set*
rules: 114		
Accuracy	0.83	0.80
Sensitivity	0.84	0.79
Specificity	0.82	0.81
Second split	*Training set*	*Test set*
rules: 114		
Accuracy	0.83	0.81
Sensitivity	0.87	0.86
Specificity	0.77	0.74
Third split	*Training set*	*Test set*
rules: 115		
Accuracy	0.84	0.79
Sensitivity	0.86	0.83
Specificity	0.81	0.74

very similar. Accuracy was good on both the training and test sets, with balanced sensitivity and specificity, as illustrated in the confusion matrix of the predictions on the test set as in Table 1 and Table 3. In this case, the molecules not containing any SA are considered as non-toxic.

As a benchmark for the SARpy performance, we considered the collection of 33 SAs for mutagenicity obtained manually from literature sources and implemented in Toxtree 2.5.0 [35]. Table 4 reports the performance of this model on the same dataset.

In classification, the two approaches reach very similar accuracy. However, the specificity of the SARpy model is even better.

SARpy automatically identified most of the SAs listed in expert systems based on human knowledge, such as the Toxtree rule base.

Table 3. SARpy: confusion matrix on test set.

	Predictions	
Test set	*Active*	*Inactive*
Actual mutagens	393	72
Actual non-mutagens	82	290

Table 4. Toxtree: statistical evaluation.

	CAESAR mutagenicity dataset	
Toxtree v 2.5.0	*Training set*	*Test set*
Accuracy	80.8%	78%
Sensitivity	87.0%	86%
Specificity	72.5%	69%

In this study SARpy extracted 112 SAs from the training set (see below). Comparison of these alerts with the ones from Toxtree led to the conclusion that most of the Toxtree alerts had been detected. A few Toxtree fragments were not covered by SARpy. These include non-genotoxic carcinogens and SAs not occurring in the molecules of the dataset.

More interestingly, SARpy proved able to identify new fragments, not codified in well-known collections of SAs and not even present in the broad list of potentially genotoxic fragments recently defined by Marchant et al. [36].

4. Discussion

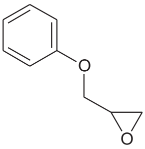
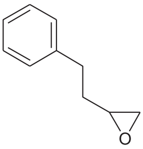
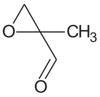
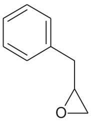
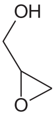
When comparing the SAs from SARpy with Toxtree we should bear in mind their different definition and that there are fragments which are more general or more specific. SARpy preferably identifies specific fragments which are more accurate than general ones. For instance, while Toxtree has one single fragment with the epoxide group, SARpy has five epoxide fragments, each containing additional features. Table 5 shows these fragments and the number of false positives. These fragments are more accurate than the Toxtree epoxide fragment. For this reason SARpy lists 89 fragments, which correspond to most of the Toxtree SAs. Table 6 shows the full list of fragments and the correspondence between the SARpy and Toxtree SAs.

The Toxtree fragments not covered by SARpy can be split into two categories. First the fragments associated with non-genotoxic carcinogenicity; the Toxtree list refers to SAs for carcinogenicity with genotoxic and non-genotoxic mechanisms of action. SARpy correctly did not identify the SAs for non-genotoxic carcinogenicity. Some of the compounds do not have genotoxic mechanisms so SARpy does not identify them because they are not mutagenic. This is the case of: SA17, thiocarbonyl; SA20, (poly) halogenated cycloalkanes (PAH); SA31a, halogenated benzene; SA31b, halogenated PAH (naphthalenes, biphenyls, diphenyls); and SA31c, halogenated dibenzodioxins.

Secondly, a few fragments of Toxtree are not found by SARpy: SA9, alkyl nitrite; SA15, isocyanate and isothiocyanate groups; SA23, aliphatic N-nitro; SA26, N-oxide; and SA30, coumarins and curocoumarins. This might be because there are not enough molecules with those fragments in the training set. In our case we set as threshold that at least three compounds had to contain a given fragment. If a chemical already has a certain fragment, it is excluded from further search. This reduces the importance of fragments that appear only in few cases.

Conversely, there are 23 SARpy fragments not present in Toxtree. Some are found in chemicals which anyway have another fragment listed by Toxtree. Thus, the finding of a certain fragment may not necessarily be related to the mutagenic effect. We checked all these 23 SARpy fragments and the related chemicals, and deleted the chemicals that could be

Table 5. Epoxide fragment and its variants found by SARpy.

ID	*SA*	*SMARTS*	*Variant*	*TP*	*FP*	**TM*	**FM*
14		O1C(C1)COc1ccccc1	[92,97]	18	0	18	0
22		O1C(C1)CCc1ccc(cc1)	[97]	11	0	16	0
27		C1(OC1)(C(=O))C	[97]	9	0	9	0
58		O1C(C1)Cc1ccc(cc1)	[97]	13	2	16	2
92		O1C(C1)CO	[97]	18	9	56	9
97		O1C(C1)	[]	24	17	123	28

*TM: true matches are all compounds in the training set containing the fragment. TP and FP refer to the presence of compounds in the remaining chemicals, not identified by a previous fragment. The opposite is true for FM (false matches).

labelled as mutagenic because of the presence of a Toxtree fragment. In this way we identified some fragments associated with mutagenicity but that do not show SAs listed by Toxtree. This was the case of the acenaphthylene fragment, associated with mutagenicity by SARpy, which indeed can be related to mutagenicity [37].

Another interesting fragment indentified by SARpy is 1,2-dichloroethene-sulphide, which is an S-halo alkenyl sulphide; the mutagenicity of these sulphides is supported in the literature [38].

Table 6. Rule sets extracted by SARpy from the training set and the corresponding expert evaluation.

SA_ID	*SMARTS*	*LR*	*Abs. LR*	*Variant of*	*TP*	*FP*	*True matches*	*False matches*	*SA for genotoxic carcinogenicity in Toxtree and further analysis*
1	[N+](=O)(c1c(c2c(cc1)cccc2))[O-]	inf	inf	[]	73	0	73	0	SA27
2	N(=O)N(C)C	inf	inf	[]	64	0	64	0	SA21
3	c12c(N)ccnc1cc(cc2)	inf	inf	[100]	47	0	47	0	SA28
4	c1oc([N+](=O)[O-])cc1	inf	inf	[]	46	0	46	0	SA27
5	c1cc([N+](=O)[O-])ccc1c1ccccc1	inf	inf	[60]	38	0	54	0	SA27
6	Nc3c(cc(N)cc3)N	inf	inf	[70,95,105]	26	0	26	0	SA28
7	C1c2c(C=CC1)cccc2	inf	inf	[]	25	0	33	0	No T-SA, DK: NR, LED
8	N1CC1	inf	inf	[]	24	0	24	0	SA7
9	c1(cc([N+](=O)[O-])sc1)	inf	inf	[]	21	0	21	0	SA27
10	c1(c2nccnc2ccc1N)	inf	inf	[90]	20	0	20	0	SA28
11	c12c(c3c(cc1ccc(c2)C)cccc3)	inf	inf	[81]	19	0	23	0	SA18
12	c1(c(ncn1))N	inf	inf	[]	19	0	21	0	SA28
13	n1c2c(n(c1)C)ccc1c2ncc(n1)	inf	inf	[90]	18	0	18	0	SA19
14	O1C(C1)COc1ccccc1	inf	inf	[92,97]	18	0	18	0	SA7
15	c12c(C(=O)c3c(C1=O)cccc3)c(c(cc2O))O	inf	inf	[59]	17	0	17	0	SA12
16	N(c1ccc(C=C)cc1)O	inf	inf	[104]	15	0	15	0	SA28
17	c12c3c(ccc1ccc(c2)N)ccc(c3)	inf	inf	[49,80]	14	0	54	0	SA28 and SA18
18	COCc1c2c(ccc1)cccc2	inf	inf	[39,82]	13	0	19	0	No T-SA, DK: NR, LED
19	C(=O)(c1ccc(NO)cc1)	inf	inf	[104]	13	0	13	0	SA11 and SA28
20	c1c(C(=O)Cl)cccc1	inf	inf	[]	12	0	13	0	SA1
21	C1(=Cc2c3c1cccc3ccc2)	inf	inf	[82]	11	0	16	0	No T-SA, DK: NR, experimentally mutagenic
22	O1C(C1)CCc1ccc(cc1)	inf	inf	[97]	11	0	16	0	SA7
23	C(=CCl)CO	inf	inf	[]	11	0	11	0	SA4
24	C(CBr)CO	inf	inf	[93]	10	0	10	0	SA8
25	[N+]=[NH-]	inf	inf	[]	10	0	15	0	SA29 OR SA14
26	c12c(cc(c(c2))C)c(cc(n1))	inf	inf	[100]	9	0	13	0	No T-SA, DK Alert 016: Quinoline
27	C1(OC1)(C(=O))C	inf	inf	[97]	9	0	9	0	SA7 and SA11
28	c1(nc2c(n1)c1c(nccc1)cc2)	inf	inf	[100]	8	0	8	0	SA19 and SA24
29	c1nc2c(C(=O)C=CC2=O)cc1	inf	inf	[]	8	0	10	0	SA12
30	N(=O)c1ccc(cc1)OC	inf	inf	[]	8	0	11	0	SA25

(*Continued*)

Table 6. (*Continued*).

SA_ID	*SMARTS*	*LR*	*Abs. LR*	*Variant of*	*TP*	*FP*	*True matches*	*False matches*	*SA for genotoxic carcinogenicity in Toxtree and further analysis*
31	SC(=CCl)Cl	inf	inf	[]	8	0	8	0	No T-SA, DK Alert 027: Alkylating agent
32	C1(=O)c2c(C(=O)c3c1cccc3N)cccc2	19.2	13.4	[]	16	1	17	1	SA28
33	n1(c2c(c3c1cccc3)cc(cc2)N)	inf	9.46	[38,49]	12	0	12	1	SA28
34	c12c3c(ccc1ccc(c2))ccc(c3)O	17.2	21.3	[80]	14	1	27	1	SA18
35	C=COCC(c1ccccc1)	17.3	11	[]	14	1	14	1	SA24
36	c12c3c4c(c1ccc(c2))cccc4ccc3	16.3	15.8	[]	26	2	40	2	SA18
37	N(c1ccc(Oc2ccccc2)cc1)O	16.6	10.2	[54]	13	1	13	1	SA28
38	n1(c2c(c3c1cccc3)cccc2)	15.5	9.46	[]	12	1	24	2	SA19
39	OCc1c2c(ccc1)cccc2	14.4	14.6	[82]	11	1	37	2	No T-SA, DK: NR, LED
40	c1cc2c(c3c(CC2)cccc3)cc1	17.1	16.9	[]	13	1	43	2	No T-SA, DK: NR, LED
41	c12sncc1cccc2	14.7	13.4	[]	11	1	17	1	No T-SA, DK: NR, LED
42	P(=O)(N(C)CC)(N)	inf	inf	[]	7	0	7	0	No T-SA, DK: NR, LED
43	C(N)Cl	inf	inf	[]	7	0	7	0	SA8
44	c1ccc(C=Cc2ccc(N)cc2)cc1	13.6	15	[104]	10	1	19	1	SA28
45	c1(NCCCl)ccc(cc1)	13.7	11	[106]	10	1	14	1	SA8
46	N(c1ccc(N=Nc2ccccc2)cc1)C	13.2	7.88	[105]	19	2	20	2	SA29 and SA28
47	C(=O)(NCc1ccccc1)C	12.7	7.88	[]	9	1	10	1	No T-SA, DK: NR, LED
48	c12c3cc(cc1ccc1c2c(cc3)ccc1)	12.3	22.9	[80]	26	3	87	3	SA18
49	c1(cc(N)ccc1)c1ccccc1	12.4	26.5	[]	17	2	101	3	SA28
50	[N+](=O)(c1c2ccccc2ccc1)[O-]	13.3	28.8	[]	9	1	73	2	SA27
51	C(=O)(Nc1ccc(c2ccccc2)cc1)	11.9	6.3	[60]	16	2	16	2	SA28
52	[N+](c1cc(C=O)ccc1)	12.1	17.3	[]	8	1	22	1	SA11
53	[N+](=O)(c1cc(c(c(c1)N))N)[O-]	12.2	9.46	[95]	8	1	12	1	SA27 and SA28
54	c1ccc(Oc2ccc(N)cc2)cc1	12.3	8.28	[]	8	1	21	2	SA28
55	C(=CC)C=COC	12.4	7.09	[]	8	1	9	1	SA24
56	N(=N)NC	12.5	7.09	[]	8	1	9	1	SA22
57	S(c1ccc(NO)cc1)	11	5.52	[]	7	1	7	1	SA28
58	O1C(C1)Cc1ccc(cc1)	10.3	6.3	[97]	13	2	16	2	SA7
59	C(=O)(c1c(cc(cc1))O)c1ccccc1	10.2	7.68	[]	19	3	39	4	No T-SA, DK: NR, LED
60	c1(c2ccccc2)ccc(N)cc1	8.45	10.9	[]	31	6	111	8	SA28
61	c12c(c(c3c(c1)cccc3))ccc1c2cccc1	8.1	7.59	[80,81]	24	5	77	8	SA18
62	c12c(c3c(cc1ccc(c2))cccc3)C	13.8	6.04	[81,82]	8	1	46	6	SA18

(*Continued*)

Table 6. (*Continued*).

SA_ID	*SMARTS*	*LR*	*Abs. LR*	*Variant of*	*TP*	*FP*	*True matches*	*False matches*	*SA for genotoxic carcinogenicity in Toxtree and further analysis*
63	c12c(c3c(cc2)cccc3)ccc2c1ccc(c2)	10.5	6.96	[80]	12	2	53	6	SA18
64	c2c3c(nc4c2cccc4)cccc3	7.96	14	[100]	18	4	71	4	SA19
65	C(=C(C(=O)c1ccc(cc1)))C(=O)	8.12	3.55	[]	9	2	9	2	SA10
66	n1c(n(c2c1ncc(c2)))	8.2	3.55	[]	9	2	9	2	No T-SA, DK: NR, LED
67	c1(n(ccn1)C)N	9.2	7.29	[]	10	2	37	4	SA28
68	C1(C(C1)C=C(C))(C)C	8.38	3.55	[]	9	2	9	2	No T-SA, DK: NR, LED
69	Nc1ccc([N+](=O)[O-])cc1	8.28	5.39	[105]	22	5	41	6	SA28 and SA27
70	c1(c(ccc(c1))N)N	7.72	9.14	[]	12	3	58	5	SA28
71	N=CC=C	7.83	4.33	[]	8	2	11	2	No T-SA, DK: NR, LED
72	OCc1ccc([N+](=O)[O-])cc1	7.91	4.2	[104]	8	2	16	3	SA27
73	CCNCCCl	14	9.98	[106]	7	1	38	3	SA8 and SA5
74	S(=O)(=O)(OCC)	8.06	3.15	[]	8	2	8	2	SA2
75	c12c(c3c(C1)cccc3)cccc2	7.47	5.25	[]	11	3	40	6	SA18
76	c1(c(n(c2c1cccc2)C))	7.23	5.12	[]	7	2	13	2	No T-SA, DK: NR, LED
77	C(Br)CBr	7.29	5.12	[93]	7	2	13	2	SA8
78	Nc1c(F)cccc1	7.36	4.33	[]	7	2	11	2	SA28
79	c1(c(cccc1N)N)C	7.42	3.68	[95]	7	2	14	3	SA28
80	c12c(cc(cc2))ccc2c1cccc2	7.28	10.8	[]	17	5	233	17	SA18
81	c1(c2c(cc3c1cccc3)cccc2)	7.3	7.52	[]	10	3	124	13	SA18
82	c1c(c2c(cc1)cccc2)C	8.88	6.61	[]	8	2	151	18	No T-SA, DK: NR, LED
83	c1(N)ccc(Cc2ccccc2)cc1	7.49	3.15	[104]	10	3	16	4	SA28
84	c12cc3c(oc1cc(cc2)O)ccc(c3)	6.83	3.35	[]	9	3	17	4	No T-SA, DK: NR, LED
85	C(Cl)(Cl)Cl	6.92	2.63	[]	9	3	10	3	SA8
86	c1(ccccc1OC)N	6.43	4.41	[]	11	4	28	5	SA28
87	N(c1ccccc1)(N)	5.94	3.15	[]	10	4	16	4	SA13
88	c1(nc2c(n1)cccc2)N	5.62	6.15	[]	7	3	39	5	SA28
89	N(C(=O)C)O	5.68	4.99	[]	7	3	19	3	No T-SA, DK: NR, experimentally mutagenic
90	c1(c2nccnc2ccc1)	5.53	9.46	[]	9	4	48	4	No T-SA, DK: NR, experimentally mutagenic
91	c1c(cc(c(c1)C)N)N	5.48	3.15	[95,104]	11	5	28	7	SA28
92	O1C(C1)CO	5.07	4.9	[97]	18	9	56	9	SA7

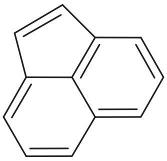
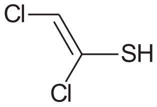
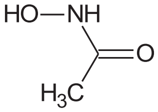
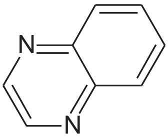
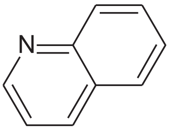
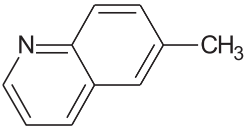
(*Continued*)

Table 6. (*Continued*).

SA_ID	*SMARTS*	*LR*	*Abs. LR*	*Variant of*	*TP*	*FP*	*True matches*	*False matches*	*SA for genotoxic carcinogenicity in Toxtree and further analysis*
93	C(Br)C	4.81	2.52	[]	24	13	48	15	SA8
94	C(N)OCC	4.87	2.76	[99]	9	5	21	6	SA16
95	Nc1cc(N)ccc1	4.81	5.64	[]	7	4	93	13	SA28
96	c1(c(n(nc1)))	4.44	1.42	[]	8	5	9	5	No T-SA, DK: NR
97	O1C(C1)	3.97	3.46	[]	24	17	123	28	SA7
98	N(N)CC	4.1	8.33	[]	7	5	74	7	SA13
99	OCN	4.15	2.48	[]	7	5	44	14	SA3
100	c12cc(ccc2cccn1)	3.89	3.3	[]	35	27	138	33	No T-SA, DK Alert 016: Quinoline
101	N=NC	19.2	5.52	[]	6	1	14	2	SA14
102	C(=C)(C=O)Cl	6.49	4.47	[106]	6	3	17	3	SA10
103	n1(c2c(nc1)c(ncn2)N)	3.94	1.02	[]	12	10	13	10	SA28
104	Cc2ccc(N)cc2	3.85	2.63	[]	24	21	137	41	SA28
105	Nc2ccc(N)cc2	3.83	3.9	[]	13	12	104	21	SA28
106	CCCl	3.86	2.08	[]	35	33	124	47	SA8
107	c12c(nnn(c1=O)N=C)c1c([nH]2)cccc1	inf	inf	[]	4	0	4	0	SA13
108	[N+](C(N))	inf	inf	[]	4	0	4	0	No T-SA, DK: NR, LED
109	[s+]1c2c(nc3c1cc(N)cc3)cc(c(c2)N)	inf	inf	[]	3	0	3	0	SA19 and SA28
110	n1(cc(c2c1cccc2)CC)N=O	inf	inf	[]	3	0	3	0	SA25 and SA21
111	C(=CC(=CC)C(=O))C	inf	inf	[]	3	0	4	0	SA10
112	C1(=O)OC(C1)	inf	inf	[]	3	0	3	0	SA6

Abs. LR: absolute LR refers to compounds not predicted by a previous alert. The LR ranges from 1 to infinite (inf). When LR is inf it indicates that the SA were found only in the experimentally positive compounds and non-mutagen compounds did not contain the target SA; DK: NR: a further evaluation with Derek Nexus provided 'nothing to report'; FP: false positive; LED: lack of experimental data; LR: likelihood ratio is the likelihood that a given prediction would be expected for a compound with the target SA compared with the likelihood that the same prediction would be expected for a compound without the target SA; No T-SA: no SA in Toxtree matches; SA: structural alert; TP: true positive.
The SAs was tested on the training set.

Table 7. Some SAs found by SARpy.

SA_ID	*Name*	*Fragments*	*Literature supporting mutagenicity*
21	Acenaphthylene		(35)
31	1,2-Dichloroethene-sulphide		(36)
89	*N*-Hydroxyacetamide		(37)
90	Quinoxaline		(38)
26 and 100	Quinoline	 	(39)

N-hydroxyacetamide [39], quinoxaline [40] and quinoline [41] are other fragments found by SARpy that are not detected as mutagens by Toxtree, although their mutagenicity is shown in the literature. We checked whether these fragments were included in the Derek Nexus [36] list of mutagenic fragments; Derek detects fragment 1,2-dichloroethene-sulphide and quinoline as mutagens but has nothing to report for the others.

Let us remember here that there are several lists of genotoxic fragments defined by human experts, but they overlap only partially [42].

Thus, this study clearly shows that it is possible to mimic human knowledge with good results and identify new rules that were not detected by human experts. Table 7 summarizes the comparison of some of the alerts extracted by SARpy with similar evidence in the literature.

Other results on the dataset used by Kazius and colleagues [7] have been published. In [25] a classifier using support vector machine (SVM) with radial basic function kernel obtained high accuracy in the training (92%) and test (83%) sets, but used 27 calculated molecular descriptors, thus increasing the risk of random correlations and making interpretation very difficult.

Alongside classical methods such as *in vivo* and *in vitro* experiments, computational tools are attracting more and more interest in the scientific community and in the industrial world to accompany or replace existing techniques.

For regulatory purposes it is important to obtain satisfactory classification accuracy on new chemical families that have not been fully studied. In this area models are needed that use statistical analysis on large numbers and can be further refined using cooperative methods to improve or confirm the results and give more information.

We have developed SARpy, a system focusing on the important structural features hidden in the database. SARpy differs from other (Q)SAR approaches in its ability to extract relevant knowledge in the form of SAs during the learning stage. Other approaches rely on pre-calculated descriptors or fingerprints, calculated by specialized software. Another advantage of SARpy over most of the similar data mining systems lies in the small set of rules produced. While approaches such as those described by Inokuchi et al. [2] and Deshpande et al. [3] typically find a large set of patterns satisfying a minimum frequency threshold which are not necessarily predictive, SARpy builds a small set of predictive rules. This rule set can be used to make expert predictions, or can be read by human experts, finding support in literature, or detecting new clues in the domain.

The model here obtained is implemented in the Virtual Models for Evaluating the Properties of Chemicals within a Global Architecture (VEGA) platform and freely available at the website, http://www.insilico.eu/use-qsar.html.

Acknowledgements

This work was partially supported by the European Commission contract ANTARES and the Italian *Ministero della Salute*.

References

[1] D.J. Livingstone, *The characterization of chemical structures using molecular properties: A survey*, J. Chem. Inform. Comput. Sci. 40 (2000), pp. 195–209.

[2] A. Inokuchi, T. Washio, and H. Motoda, *An a priori-based algorithm for mining frequent substructures from graph data*, in *Principles of Data Mining and Knowledge Discovery, Proceedings of 4th European Conference, PKDD 2000, 13–16 September 2000, Lyon, France*, D.A. Zighed, J. Komorowski, and J. Zytkow, eds., Springer, Berlin, 2000, pp. 13–23.

[3] M. Deshpande, M. Kuramochi, N. Wale, and G. Karypis, *Frequent substructure based approaches for classifying chemical compounds*, IEEE Trans. Knowl. Data Eng. 17 (2005), pp. 1036–1050.

[4] C. Borgelt and M.R. Berthold, *Mining molecular fragments: Finding relevant substructures of molecules, in Proceedings of the 2012 IEEE International Conference on Data Mining (ICDM 2002), 9–12 December 2002, Maebashi City, Japan*, IEEE Computer Society, 2002 (2002), pp. 51–58.

[5] R. Agrawal and R. Srikant, *Fast algorithms for mining association rules in large databases*, in *Proceedings of the 20th international conference on Very Large Data Bases, VLDB, Santiago, Chile*, B. Bocca, Jr., M. Jarke, and C. Zaniolo, eds., Morgan Kaufmann Publishers, San Francisco, 1994, p. 487–489.

[6] R. Benigni and C. Bossa, *Structural alerts for carcinogenicity, and the* Salmonella *assay system: A novel insight through the chemical relational databases technology*, Mutat. Res.-Rev. Mutat. 659 (2008), pp. 248–261.

[7] J. Kazius, R. Mcguire, and R. Bursi, *Derivation and validation of toxicophores for mutagenicity prediction*, J. Med. Chem. 48 (2005), pp. 312–320.

[8] H.S. Rosenkranz, Y.P. Zhang, and G. Klopman, *Studies on the potential for genotoxic carcinogenicity of fragrances and other chemicals*, Food Chem. Toxicol. 36 (1998), pp. 687–696.

[9] L. Dehaspe, H. Toivonen, and R.D. King, *Finding frequent substructures in chemical compounds, on Predictive Toxicology of Chemicals: Experiences and Impact of AI Tools, AAAI.SS.99*, Gini, G. and Katrizky, A., eds., AAAI Press, Menlo Park, CA, 1999, pp. 78–81.

[10] G. Klopman, *MULTICASE: A hierarchical computer automated structure evaluation program*, Quant. Struct.-Act. Rel. 11 (1992), 176–184.

[11] C. Helma, *Lazy structure–activity relationships (LAZAR) for the prediction of rodent carcinogenicity and* Salmonella *mutagenicity*, Mol. Divers. 10 (2006), pp. 147–158.

[12] D. Weininger, *SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules*, J. Chem. Inf. Comput. Sci. 28 (1988), pp. 31–36.

[13] R. Sayle, *1st-class SMARTS patterns*, EuroMUG 97, Bioinformatics Group, Research I.T., Glaxo Wellcome Research & Development, Stevenage, UK, 1997.

[14] A. Karwath and L. De Raedt, *SMIREP: predicting chemical activity from SMILES*, J. Chem. Inf. Model. 46 (2006), pp. 2432–44.

[15] A.A. Toropov, A.P. Toropova, and E. Benfenati, *Additive SMILES-based carcinogenicity models: probabilistic principles in the search for robust predictions*, Int. J. Mol. Sci. 10 (2009), pp. 3106–3127.

[16] R. Benigni, T.I. Netzeva, E. Benfenati, C. Bossa, R. Franke, C. Helma, E. Hulzebos, C. Marchant, A. Richard, Y.T. Woo, and C. Yang, *The expanding role of predictive toxicology: An update on the (Q)SAR models for mutagens and carcinogens*, J. Environ. Sci. Health C, 25 (2007), pp. 53–97.

[17] E. Benfenati, R. Benigni, D.M. Demarini, C. Helma, D. Kirkland, T.M. Martin, P. Mazzatorta, G. Ouedraogo-Arras, A.M. Richard, B. Schilter, W.G. Schoonen, R.D. Snyder, and C. Yang, *Predictive models of carcinogenicity and mutagenicity: Frameworks, state-of-the-art and perspectives*, J. Environ. Sci. Health C, 27 (2009), pp. 57–90.

[18] B.N. Ames, *The detection of environmental mutagens and potential*, Cancer 53 (1984), pp. 2030–2040.

[19] J. Ashby, *Fundamental structural alerts to potential carcinogenicity or noncarcinogenicity*, Environ. Mutagen. 7 (1985), pp. 919–921.

[20] W.W. Piegorsch and E. Zeiger, *Measuring intra-assay agreement for the Ames* Salmonella *assay assay*, in *Statistical Methods in Toxicology, Lecture Notes in Medical Informatics*, L. Hotorn, ed., Springer-Verlag, Berlin, 1991, pp. 35–41.

[21] J.L. Durant, B.A. Leland, D.R. Henry, and J.G. Nourse, *Reoptimization of MDL keys for use in drug discovery*, J. Chem. Inf. Comput. Sci 42 (2002), pp. 1273–1280.

[22] C. Hansch, P.P. Malony, T. Fujita, and R.M. Muir, *Correlation of biological activity of phenoxyacetic acids with Hammett substituent constants with partition coefficients*, Nature, 194 (1962), pp. 178–180.

[23] J.A. Miller and E.C. Miller, *Searches for ultimate chemical carcinogens and their reactions with cellular macromolecules*, Cancer 47 (1981), pp. 2327–2345.

[24] J. Ashby and R.W. Tennant, *Chemical structure,* Salmonella *mutagenicity and extent of carcinogenicity as indicators of genotoxic carcinogenesis among 222 chemicals tested by the U.S. NCI/NTP*, Mutat. Res. 204 (1988), pp. 17–115.

[25] Q. Liao, J. Yao, and S. Yuan, *Prediction of mutagenic toxicity by combination of recursive partitioning and support vector machines*, Mol. Divers. 11 (2007), pp. 59–72.

[26] M. Zheng, Z. Liu, C. Xue, W. Zhu, K. Chen, X. Luo, and H. Jiang, *Mutagenic probability estimation of chemical compounds by a novel molecular electrophilicity vector and support vector machine*, Bioinformatics, 22 (2006), pp. 2099–2106.

[27] A. Perrotta, D. Malacarne, M. Taningher, R. Pesenti, M. Paolucci, and S. Parodi, *A computerized connectivity approach for analyzing the structural basis of mutagenicity in* Salmonella *and its relationship with rodent carcinogenicity*, Mol. Mutagen. 28 (1996), pp. 31–50.

[28] R. Benigni, C. Bossa, O. Tcheremenskaia, and A. Giuliani, *Alternatives to the carcinogenicity bioassay: in silico methods, and the in vitro and in vivo mutagenicity assays*, Exp. Opin. Drug Metab. Toxicol. 6 (2010), pp. 1–11.

[29] R.D. Snyder, G.S. Pearl, G. Mandakes, W.N. Choy, F. Goodsaid, and I.Y. Rosenblum, *Assessment of the sensitivity of the computational programs DEREK, TOPKAT and MCASE in the prediction of the genotoxicity of pharmaceutical molecules*, Environ. Mol. Mutagen. 43 (2004), pp. 143–158.

[30] T. Ferrari, G. Gini, and E. Benfenati, *Support vector machines in the prediction of mutagenicity of chemical compounds*, Proceedings NAFIPS, Cincinnati, 2009.

[31] T. Ferrari, and G. Gini, *An open source multistep model to predict mutagenicity from statistical analysis and relevant structural alerts. CAESAR workshop on QSAR Models for REACH*. Chem. Cent. J. 4(Suppl. 1) (2010).

[32] *MCASE*, MultiCASE Inc., Beachwood, OH, USA; software available at http://www.multicase.com.

[33] N.M. O'Boyle, C. Morley, and G.R. Hutchison, *Pybel: A Python wrapper for the OpenBabel cheminformatics toolkit*, Chem. Cent. J. 2 (2008).

[34] Available at http://www.caesar-project.eu. Web site of the CAESAR project and QSAR platform.

[35] R. Benigni, C. Bossa, N.G. Jeliazkova, T.I. Netzeva, and A.P. Worth, *The Benigni/Bossa rulebase for mutagenicity and carcinogenicity – a module of Toxtree*, EUR 23241 EN, EUR-Scientific and Technical Report Series Office for the Official Publications of the European Communities, Luxembourg, 2008.

[36] C.A. Marchant, K.A. Briggs, and A. Long, *In silico tools for sharing data and knowledge on toxicity and metabolism: DEREK for windows, METEOR and VITIC*, Toxicol. Mech. Methods, 18 (2008), pp. 177–187.

[37] D.A. Kaden, R.A. Hites, and W.J. Thilly, *Mutagenicity of soot and associated polycyclic hydrocarbons to* Salmonella typhimurium, Cancer Res. 39 (1979), pp. 4152–4159.

[38] S. Vamvakas, W. Dekant, and M.W. Anders, *Mutagenicity of benzyl S-haloalkyl and S-haloalkenyl sulphides in the Ames-test*, Biochem. Pharmacol. 38 (1989), pp. 935–939.

[39] E. Dybing, E.J. Søderlund, W.P. Gordon, J.A. Holme, T. Christensen, G. Becher, E. Rivedal, and S.S. Thorgeirsson, *Studies on the mechanism of acetamide hepatocarcinogenicity*, Pharmacol. Toxicol. 60 (1987), pp. 9–16.

[40] Hu. Aeschbacher, U. Wolleb, J. Loliger, J.C. Spadone, and R. Liardon, *Contribution of coffee aroma constituents to the mutagenicity of coffee*, Food Chem. Toxicol. 27 (1989), pp. 227–232.

[41] C.J. Smith, C. Hansch, and M.J. Morton, *QSAR treatment of multiple toxicities: the mutagenicity and cytotoxicity of quinolines*, Mutat. Res.-Fund. Mol. M 379 (1997), pp. 167–175.

[42] A Worth, M. Fuart-Gatnik, S. Lapenna, E. Lo Piparo, A. Mostrag-Szlichtyng, and R. Serafimova, *The use of computational methods in the toxicological assessment of chemicals in food: current status and future prospects*, EUR 24748 EN, Joint Research Centre Scientific and Technical Report Series, Office for the Official Publications of the European Communities, Luxembourg, 2011.