Mining Toxicity Structural Alerts from SMILES: A New Way to Derive Structure Activity Relationships

Thomas Ferrari, Giuseppina Gini
Department of Electronics and Information
Politecnico di Milano
Milan, Italy

Nazanin Golbamaki Bakhtyari, Emilio Benfenati
Lab. of Environmental Chemistry and Toxicology
Mario Negri Institute
Milan, Italy

Abstract— **Encouraged by recent legislations all over the world, aimed to protect human health and environment, *in silico* techniques have proved their ability to assess the toxicity of chemicals. However, they act often like a black-box, without giving a clear contribution to the scientific insight; such over-optimized methods may be beyond understanding, behaving more like competitors of human experts' knowledge, rather than assistants. In this work, a new Structure-Activity Relationship (SAR) approach is proposed to mine molecular fragments that act like structural alerts for biological activity. The entire process is designed to fit with human reasoning, not only to make its predictions more reliable, but also to enable a clear control by the user, in order to match customized requirements. Such an approach has been implemented and tested on the mutagenicity endpoint, showing marked prediction skills and, more interestingly, discovering much of the knowledge already collected in literature as well as new evidences. The achieved tool is a powerful instrument for both SAR knowledge discovery and for activity prediction on untested compounds.**

Keywords - Structure Activity Relationships; structural alerts; SMILES; fragments; mutagenicity; knowledge discovery

I. Introduction

This paper deals with qualitative SAR (Structure-Activity Relationships). While SAR makes use of rules created by experts to produce models that relates subgroups of the molecule atoms to a biological property, as toxicity in our case, we show how to automatically develop such rules if a suitable set of results of biological experiments is available.

The recent availability of biological activity data on chemical substances has triggered a proliferation of data mining approaches for toxicity assessing. In most cases, classical statistical tools are used to search for a numerical correlation between chemical properties and biological activity. Such models are able to exhibit significant prediction abilities on new compounds, and can be profitably used for classification tasks; but it's hard to extract the underlying rationale. In fact, physicochemical properties or structural information of chemicals are numerically quantified into the so called molecular descriptors [1], whose chemical or biological meaning is not obvious. Moreover, the equation that binds an instance to its prediction could be not intelligible. It is the case of neural networks, where often good performance is tightly related to network complexity. On the other hand, the structural nature of chemicals is explicitly taken into account by several graph-mining approaches, as AGM [2], FSG [3] and MoFa [4], which mine large datasets for frequent substructures. All of the cited implementations are based on Apriori algorithm, an association rule induction method designed for *market basket analysis*; hence, the selection of relevant fragments is driven only by statistical criteria.

Human experts usually estimate toxicity through the detection of particular structural fragments, already known to be responsible for the toxic property under investigation. In the literature such fragments are usually referred to as *structural alerts* (SAs) [5], toxicophores [6] or biophores [7] and can be derived by human-experts, 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); these mechanisms are still poorly understood and largely unknown.

To assist experts in the extraction of such knowledge from data, by providing predictive and understandable models based on molecular fragments, a few approaches have been developed. Some are based on techniques from inductive logic programming (ILP) [8]. Whereas ILP techniques are theoretically appealing, they exhibit significant efficiency problems, and moreover cannot be directly applied to a standard chemical formats for molecule representation. Other, including MCASE [9], and more recently LAZAR [10] 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 that exhibit a 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 observed biological activity, are labeled biophores; additional features that seem able to regulate a biophore

978-1-4244-9925-0/11/$26.00 ©2011 IEEE

activity, such as molecular descriptors and/or other fragments, are called modulators and can influence the final prediction. It is hard to find detailed information about the way the fragments are generated (mainly linear fragments with branches around the backbones) and the scoring scheme which determines the final prediction is obscure (measured in "CASE unit") and not decisive (may contain "marginally active" predictions).

LAZAR, on the other hand, searches only for linear fragments selected, again, with a statistical test (chi-square); the final prediction of a molecular structure is determined by a weighted majority vote from neighbours (i.e., fragments above a predefined threshold of similarity). In both cases, only simple substructures are taken into account on pure statistical basis. It is worthy mentioning that while LAZAR is open source, MCASE is commercial, and its basic version (CASETOX) doesn't allow the user to extract knowledge from his own data, but only to use the application as an expert system with prepackaged knowledge.

In this work we present a new, ad hoc, SAR approach, capable of finding relevant fragments in a transparent way, and to extract a set of rules directly from data without using a priori knowledge. The fragmentation algorithm can generate substructures of arbitrary complexity, not only the simple ones, and the fragments candidates to become SAs are automatically selected on their actual prediction performance on a training set. Modulators are occasionally taken into account as particular classes of structural variants of the SA itself, considered harmful and therefore exceptions to the rule.

Both the input and the output are expressed as Simplified Molecular Input Line Entry Specification (SMILES) [11], ASCII strings obtained by printing the symbol nodes encountered in a depth-first tree visit of the chemical graph. SMILES are expressions of a context free language. The SMILES notation of a chemical compound is a string of atoms (represented by their atomic symbols), bonds, parentheses, and numbers. The four basic bond types are represented by the symbols '-', '=', '#', and ':' (single and aromatic bonds may always be omitted), while ionic bonds are represented by a '.'. Branches are specified by enclosing brackets. Cyclic structures are represented by breaking one bond in each ring; the atoms adjacent to the bond obtain the same number. Hydrogen is not included in a SMILES representation, but can be inferred from the available valences. Typically, a number of equally valid SMILES can be written for a molecule. For example, CCO, OCC and C(O)C all specify the structure of ethanol. Algorithms have been developed to ensure the same SMILES is generated for a molecule, and this is termed the canonical SMILES. Canonical SMILES are used in our data sets.

The output consists in a set of rules in the form

"*IF contains <SA> AND NOT <SA's exceptions> THEN active*"

where both SA and modulator structures are expressed as SMILES, then apt to be used by human experts or other chemical software.

Moreover, the user can easily drive the rule selection procedure to match particular purposes, by customizing the lower admissible precision of the rules, and by deciding the maximum length of the substructures to consider.

In the actual implementation, called SARpy (SAR in python), the fragmentation process is carried out directly on the SMILES notation of structures. A similar approach has been implemented in SMIREP [12], but in such work the SMILES strings are 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). Also CORAL [13] makes use of small SMILES fragments, but they are finally merged into a numerical molecular descriptor, therefore 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 section a conceptual view of the proposed approach is given. Then its prototypical implementation is detailed and validated on a large dataset of publicly available molecular structures. At the end, the results achieved and the extracted knowledge are discussed and compared to the present state of the art of the domain.

II. SARPY PARADIGM OF KNOWLEDGE EXTRACTION

Given a training set of molecular structures, along 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.

In the proposed paradigm, such task is carried out in four subsequent steps:

Fragmentation: it's a novel recursive algorithm that considers every combination of bond breakages, and computes every substructure of the molecular input set.

Evaluation: each fragment is validated as SA for positive activity on the training set; such matching produces a result in terms of *true positives* (TP), which are actual active compounds individuated by the SA, and *false positives* (FP), the inactive compounds incorrectly matched.

Exceptions induction: a screening on every SA is carried out to check how to refine its ability by excluding some of its structural variants.

Rule set extraction: from the large set of candidates SA, each one equipped with the TP and the FP values on the training set, the best subset is extracted and all explicit redundancies removed. After learning, the final set of rules is used for predicting the activity of new molecules. Compounds without any SA will be considered to be inactive.

A. Fragmentation

The total fragmentation of molecular structures is achieved by an iterative process; each iteration has the plain task of taking the input structures and computes all the substructures obtainable by individually considering every bond breakage. That is to say that fragmentation iteration simulates a split on each bond of every input molecule, and collects every time the resultant couple of fragments. The maximum number of output structures for any iteration is two fold the number of total bonds in the input set, since every split produces, at most, two fragments. The case of ring-bond splits will be later discussed.

If this process is applied on the training structures, the result of the first iteration is the collection of *1st level fragments*: all the substructures extracted by alternatively considering every possible single split. Now, applying the next iteration to the output of the previous, every *1st level fragment* is further split: this means to consider every possible pair of splits on the original structures. Substructures with exactly two broken bonds are the *2nd level fragments* and are added to the collection. And so on, until no more new fragments could be extracted.

Since we are interested in general SAs capable of identifying large classes of active compounds, we may decide a predefined maximum number of atoms; hence, the maximum number of bonds to be broken is known. Furthermore, the number of the newly produced fragments decreases with the depth of the fragmentation recursive call, and the number of output fragments goes to zero in a very few tens of steps. This observation can be supported by noticing that a fragment found in a molecule in a inner level could have been found, with good probability, in other molecules in an outer level, if not so rare to be useless.

When one or more cycles are present in a molecular structure, the first iteration produces not only the collection of every couple of fragments resulting from the split of all nonring-bonds, but also every possible way of unrolling each ring by breaking ring-bonds while keeping the structure undivided. Therefore the fragmentation of the ring will take place in the next iteration. So the presence of rings (even if fused) simply delays the appearance of some structural fragments due to the high number of bonds to be broken to split the structure in correspondence of rings.

Moreover, the search for structural alerts for positive activity restricts the fragmentation only to positive structures of the training set; the iterative procedures terminates when no more fragments are found or when the maximum depth (set by the user) has been reached.

Obviously the whole process can be performed even to identify “safe” substructures, simply considering as “positive” the inactive structures.

B. Evaluation

Once all the substructures have been produced, the next step consists in evaluating them as potential SAs for positive activity. Since the problem is a binary classification problem, and since we have the experimental activity value of the original structures in the training set, we evaluate positive correlations in terms of TP and FP predictions generated by their individual application on the training structures.

After computing TP and FP values, a candidate rule is generated and characterized by its precision rate, called Positive Predictive Value (*PPV*), and its sensitivity, which measure, respectively, the proportion of truly active compounds in the subset matched by the SA, and the rate of actual positives correctly identified as such by the single rule.

The task of associating each SA to its TP and FP molecules could be optimized in several ways, however, in the implementation such computation is explicitly carried out by a complete matching procedure, where each fragment is matched against every molecule in the training data, starting from the deepest level. In fact, the search for potential molecules containing a given structure can be restricted to those containing one of its descendants, being every descendant a substructure of its ancestor. Consequently, TP and FP of a SA can be found by searching in the subset of molecules already individuated by a related fragment from the deeper level (for instance, the one with less matches).

In addition, to avoid rules with irrelevant behaviour, a lowest TP threshold can be considered to exclude SAs with little information on their positive prediction ability. SAs with a number of TP molecules below the threshold can be pre-emptively excluded (without the computation of FP), and the ancestor branches pruned, since TP values can only decrease (being superstructures).

The evaluation is aimed at identifying the substructures that better generalize the concept of biophores, by showing high precision rate and good sensitivity for active chemicals. Several indicators can be derived, from the information just calculated, to assess the prediction ability, and to rank every fragment.

We propose two simple *score* assignments, which wraps *PPV* and sensitivity respectively in "OR" or “AND” fashion. Such *score* will be determinant in the following steps as a relative measure to compare the ability of two redundant SAs. The *score* is computed either as (1) or (2), according to the user choice:

$$score = PPV + sensitivity \tag{1}$$

$$score = PPV \times sensitivity \tag{2}$$

Each score uses *PPV* and is biased by sensitivity; it considers sensitivity just in case of similar precision. In fact, *PPV* has high values, and the sensitivity of such a precise SA might be of a different order of magnitude, since it refers the performance of a single rule to the entire dataset. The two measures can be seen as indicators for local and global ability to predict positivity, since one relates TP value to the subset of compounds matched by the rule, while the second refers to the total number of positives in the training set. To achieve a more concise model with fewer rules, the second definition of the *score* is used.

C. Exceptions Induction

So far, all the relevant substructures in the training data have been collected, along with the information about their individual prediction ability as potential SAs. To achieve a predictive model, generalization skills should be rewarded; however, some of the top ranked fragments, despite their good precision rate and high sensitivity, may represent a valid but still too general alert for positivity, by involving some harmless structural variants of the alert. A well known method to make these alerts more specific, preserving their generality, consists in searching for potential modulators [7]: factors that can regulate, and eventually deactivate, the action of the biophore.

In SARpy such idea is put into practice in the case of potential SAs which have a significant number of FP ascribable to the same few harmless superstructures of the SA itself. In other words, SARpy seeks for each alert the existence of variants, namely structural extensions that are present in its FP molecules, then candidates to become exceptions to the SA. The search and the evaluation of potential exceptions is simple, considering that all substructures have been computed and associated to the relative TP and FP molecules. Exceptions to an alert can be found just within the fragments of the FP molecules of the SA itself; furthermore, only its superstructures have to be considered. FP molecules associated with the exception structure in the previous evaluation on the entire dataset now indicate FP predictions removed by the rising of the exception (becoming *true negatives*); on the other hand, TP molecules indicate correct predictions become errors (*false negatives*).

At this point, the algorithm recalculates the new precision rate and sensitivity of the rule; if the exception enhances the performance on the training set, the resulting rule is expressed in the form "*IF contains <SA> AND NOT <SA's superstructures> THEN positive*".

D. Rule Set Extraction

At this point, a huge set of rules for positive activity has been collected, ranked and refined. The selection of the best rule set follows two steps. The first consists in constraining the precision of every rule (PPV). Then the second screening removes all the explicit redundancies: if two rules are found to be one implicit to the other, only the one with the best score is kept.

The minimum *PPV* threshold of every rule is user defined in the interval 0-1. A low threshold allows for more and not necessarily precise rules to be selected, and the final classification will be highly sensitive but poorly specific. A high value is more restrictive by admitting only rules with good precision; this makes the classification more accurate but maybe less sensitive. The default value is set to 0.8. The user can perform a fast trial and error procedure to find a custom-tailored set of rules, aimed at sensitivity, specificity, accuracy or simplicity of the rule set. Furthermore, by setting the *PPV* threshold at the highest value 1, the user can achieve the most sharp set of rules, the ones with (at the limit) no FP predictions, that represent extremely reliable alerts to take into account for further investigations.

Before testing new chemicals, with the aim to ensure that all the testing structures are inside the applicability domain of the training set, it is worth to check if some of the new structures contain fragments never found in the training phase. A priori unreliable predictions like these could be intercepted and the exclusion of such outliers from the testing task will guarantee higher prediction accuracy.

III. Implementation

The prototype implementation has been driven by a golden rule: fast and clear coding. It's a Python script (about 400 lines of code) employing the open source *Open Babel* 2.2.3 library [14] via a set of bindings to the C++ code. The aim of the prototype is to sample the efficacy of the proposed paradigm, either for classification purposes (the rule set used to classify untested compounds like in SAR) or for knowledge extraction (the rule set as source of information for further investigation).

The iterative fragmentation of chemical structures is implemented directly on their SMILES notation as a string fragmentation task (Fig. 1). In detail, each iteration alternatively splits all the SMILES strings in correspondence to each bond, and rearranges the resultant two substrings (keeping into account the parenthesis nesting level) to obtain the two SMILES that identify the relative couple of fragments.

The drawback in the current SMILES fragmentation is that ring breaking produces invalid fragments that will be rejected; thus, rings are considered as single entities. This doesn't necessarily means there is any other information loss: in fact, during the fragmentation of training data, the same toxic fragment involved in a ring may be still found as open skeleton in other compounds; then, during the evaluation phase, all the structures containing such fragment will be matched, regardless of them belonging to a ring or not. Otherwise, if the toxic fragment is always embedded in a ring, the ring itself is taken as the possible cause of toxicity.

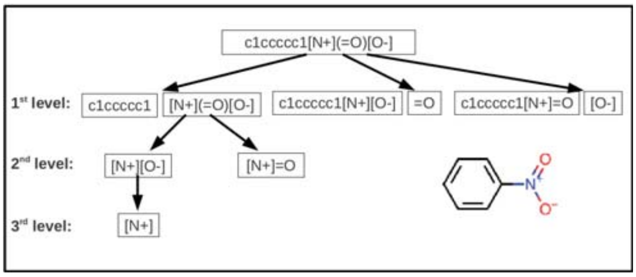


Figure 1- SMILES fragmentation: duplicates are omitted. At the top the starting structure.

Once found the 1st level fragments, every substructure is evaluated and potential exceptions are considered. The structural comparison is carried out by the Open Babel SMARTS [15] matching function. From this point only the substructures inclined to positive predictions are kept for the next iteration. In the default settings, to prevent imprecise SAs to be developed further, the minimum PPV is set to the trivial one (0.5, random rule); to avoid unexpected behaviors, rules with a statistically not significant number of matches on the training set are excluded, by constraining the minimum TP value in function of the total number of active compounds in the training set (natural logarithm is used).

For the remaining substructures the process is iterated until no more new fragments are found, or a user-defined fragmentation depth level is reached. Then, to remove explicit redundancies from the resulting subset, SAs are sorted by *score* and compared two by two, and if a rule implies the other (i.e., one SA is a substructure of the other), only the top ranked one is kept. Finally the selected rules are applied on the training set and the performance statistics (accuracy, sensitivity, specificity, and confusion matrix) are prompted to the user, who, if satisfied, can save the rule set, or can search for a better tuning by defining a new *PPV* threshold. As an additional feature, the search direction for the optimal *PPV* is suggested on the basis of the difference between sensitivity and specificity: with low specificity, higher precision rate should be requested, vice versa for the low sensitivity case. The resulting set of rules can be checked on an external test set or many-folds cross-validated.

IV. EXPERIMENTAL

To test the predictive capabilities of the proposed approach, the implementation has been deeply validated on the mutagenicity endpoint, for which large datasets of molecular structures are available along with the correspondent experimental outcome.

Mutagenicity is the capability of a substance to cause genetic mutations. It is a property of high public concern because it has a close relationship with carcinogenicity, in the case of genetic mutations, and with reproductive toxicity, in the case of germ cell mutations [16]. The mutagenic potential of chemical compounds is experimentally assessed by Ames test [17], which provides a cheap and short-term alternative to the rodent bioassay by means of a series of genetically engineered Salmonella Typhimurium bacterial strains. As discussed in [18], the estimated inter-laboratory reproducibility of this in vitro test is about 85%.

The employed dataset, usually referred to as *Bursi Mutagenicity Dataset* [6], was provided by its authors to the EC funded CAESAR project [19] as SDF (Structure Data Format) file. The dataset originally contained 4337 molecular structures, but after a minute check of each chemical structure some of them were corrected or removed, to avoid inaccuracies. The resulting CAESAR mutagenicity data set consists of 4204 compounds, 2348 classified as mutagenic and 1856 classified as non-mutagenic by Ames test.

For modeling, the data set was split into a training set and a test set following a stratification criterion in order to make sure that each subset would approximately cover 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), while the other 20% (837 compounds) is for testing.

The SDF input was read with *Pybel* [20], a set of convenience Python functions and classes that simplifies access to the *Open Babel* I/O module, and converted into SMILES disregarding chirality information.

The parameterization used in the experiment is: the minimum and maximum number of atoms in fragments, set respectively to 2 and 18, and the PPV threshold, set to 0.8, to get an high, yet enough accurate, set of rules. Using score (1) on the training set, 77 rules are generated.

We predicted the test set and obtained the statistics in Table 1. A 5-fold cross validation on the training set gives accuracy measures with very similar statistics. Good accuracy is achieved on both training and test set, with balanced sensitivity and specificity, as illustrated in the confusion matrix of the predictions on the test set (Table 2); in this case, the molecules not containing any SA are considered as non-toxic. The computation time on a laptop PC is in Table 3.

TABLE I. SARPY: STATISTICAL EVALUATION

SARpy	***CAESAR Mutagenicity Dataset***	
	Training set	***Test set***
Accuracy:	79.3 %	77.7 %
Sensitivity:	80.1 %	80.1 %
Specificity:	78.2 %	74.6 %

TABLE II. SARPY: CONFUSION MATRIX ON TEST SET

Test set	**Predictions**	
	Active	***Inactive***
Actual *mutagens*	407	58
Actual *nonmutagens*	86	286

TABLE III. SARPY: TIMING FOR MUTAGENICITY

Phase	Seconds	% of total time
Fragmentation:	93 s	12%
Evaluation:	221 s	29 %
Exceptions Induction:	440 s	58 %
Ruleset Extraction:	1 s	0.1%
Total time: 755 s		

Using score (2), a reduced set of 38 rules is generated, and the accuracy is similar, since the remaining rules are the most general and cover a large number of cases. However, here we prefer to discuss the relevance of the extracted rules more than possible improvements in accuracy performance that are the subject of the ongoing work.

V. DISCUSSION

In the mutagenicity/carcinogenicity domain, the key contribution in the definition of known SAs comes from Ashby's studies in the 80s [21]. Grounding his work on the electrophilicity theory of chemical carcinogenesis developed by [22], which correlates the electrophiles presence (like halogenated aliphatic or aromatic nitro substructures) to genotoxic carcinogenicity, Ashby compiled a list of 19 SAs for DNA reactivity. In a following paper [23], a few hundred data of NTP (National Toxicology Program of US) have been manually mined to confirm their findings. Every subsequent effort starts from knowledge collected by Ashby to derive more specific rules, like the already mentioned work by Kazius et al [6], where the cognition of the mechanism of action is joined to statistical criteria.

As a benchmark for the SARpy performance, we considered the collection of 30 SAs for mutagenicity manually derived from literature sources [21, 23, 24, and 7] and implemented in Toxtree [25]. Its performance on the same dataset is reported in Table 4. In classification, the two approaches reach a very similar accuracy, and the specificity of the SARpy model is even higher; moreover the *Bursi Mutagenicity Dataset* itself was entirely used to derive the Benigni/Bossa rule base coded in ToxTree. Furthermore, while a SARpy's fragment is just a SMILES, Toxtree's one can be a complex SMARTS, or even not definable by the SMARTS language only.

TABLE IV. TOXTREE: STATISTICAL EVALUATION

Toxtree	*CAESAR Mutagenicity Dataset*
Accuracy:	78.9 %
Sensitivity:	86.3 %
Specificity:	69.6 %

About knowledge discovery, SARpy was capable to automatically identify SAs which are listed in expert systems based on human knowledge, such as the Benigni/Bossa rule base. This is the case, for instance, of the aromatic amines and the azoderivatives (respectively SA_28 and SA_14 in Toxtree).

More interestingly, SARpy proved to be capable to identify new fragments, not codified into well-known collections of SAs, and even not present in the wide list of potentially genotoxic fragments recently defined in [26]. This is the case of the vinyl fragment, associated to mutagenicity by SARpy. Indeed, styrene, which is the smallest chemical containing this fragment, is mutagen in the Ames test [27].

Another fragment that SARpy discovered as related to mutagenicity is the 7-chloroquinoline. Interestingly, laboratory experiments have shown that, when chlorine is on the nitrogen ring, it does not cause mutagenicity; conversely, when chloro is in position 5, 6 or 8, the chloroquinoline is mutagenic, as reported in the CCRIS database[1] and in [28, 29]. Unfortunately, this database doesn't have experimental results for the mutagenicity of the 7-chloroquinoline.

Another interesting fragment individuated by SARpy is 1,2-dichloroethene-sulfides, that is an S-halo alkenyl sulfide, and the mutagenicity of S-halo alkenyl sulfides is supported in literature by [30].

A comparison between some of the alerts extracted by SARpy and similar evidences in literature is summarized in Table 5.

On the same dataset other results have been published. In [31] a machine learning approach mines the data set with Support Vector Machines (SVM) algorithm and RBF kernel, with high accuracy both in training (92%) and in test (83%) set, but the equations behind are extremely complex and the input requires 27 calculated molecular descriptors, thus increasing the risk of random correlations and making difficult the interpretation. Also LAZAR has been validated on a subset of the *Bursi Mutagenicity Dataset* and reached a prediction accuracy of 69% [10].

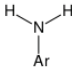
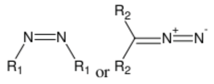
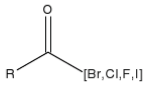
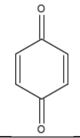
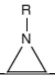
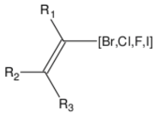
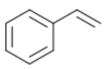
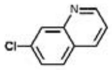
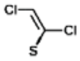
VI. CONCLUSION AND FUTURE WORK

There is an argument that, if the main aim of SAR and QSAR (Quantitative SAR) is simply prediction, the attention should be focused on model quality and not on its interpretation. Another argument is that it is dangerous to attempt to interpret models, since correlation does not imply causality. Regarding the interpretability of QSAR models, Livingstone [32] states: “The need for interpretability depends on the application, since a validated mathematical model relating a target property to chemical features may, in some cases, be all that is necessary, though it is obviously desirable to attempt some

[1] Toxnet: http://toxnet.nlm.nih.gov/cgi-bin/sis/htmlgen?CCRIS (accessed Jul 2010)

explanation of the mechanism in chemical terms, but it is often not necessary, per se".

TABLE V. SOME STRUCTURAL ALERTS FOUND BY SARPY

Name	Structure	Reference
Aromatic amine		Toxtree [26] (SA_28)
Aliphatic azo		Toxtree [26] (SA_14)
Acyl halide		Toxtree [26] (SA_1)
Quinones		Toxtree [26] (SA_12)
Aziridines		Toxtree [26] (SA_7)
Monohaloalkene		Toxtree [26] (SA_4)
Vinyl benzene		Styrene is mutagenic according to [28]
7-chloroquinoline		5,6 and 8 chloroquinoline are mutagenic according to [29, 30]
1,2-dichloroethene-sulfides		S-halo alkenyl sulfides are mutagenic according to [31]

On this basis, we can differentiate predictive (Q)SARs, focused on prediction accuracy, from descriptive (Q)SARs, focused on descriptor interpretability. The usage of predictive QSAR models is growing, since they provide fast, reliable and quite accurate estimates of the chemicals activity. These features make them suitable for legislative purposes, as envisaged in the European legislation REACH (Registration, Evaluation, Authorisation and Restriction of Chemicals).

Descriptive (Q)SAR however is highly appreciated by stakeholders to characterize the toxic risk of chemicals. Structural rules are expressions that correlate local characteristics of the molecule to a risk, and usually can be explained in terms of reactivity or activation of biological pathways. Toxicology has induced some of those rules from experiments on a few endpoints, notably mutagenicity and carcinogenicity.

Even though statistical (Q)SARs provide predictive models [33] using global characteristics of the molecule, there is a need to integrate the two approaches. To this end we have developed SARpy, a system able to focus on the important structural features hidden in the database.

The difference of SARpy with respect to other (Q)SAR approaches is its ability to extract relevant knowledge in the form of structural alerts during the learning stage. Other approaches rely on precalculated 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 FSG, and AGM 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. The resulting rule set can be used to carry out expert predictions, or can be read by human-experts, finding support in literature, or revealing new clues in the domain.

Furthermore, the same approach can produce new models of any property of interest. We tested SARpy to other relevant endpoints against the models freely provided in the CAESAR web site (http://www.caesar-project.eu) obtaining the same statistical performances and a set of possible SA still under investigation (since no structural alerts are known for bioaccumulation and developmental toxicity).

The limitation of SARpy of being a binary classifier is not a problem considering that the legislation indicates a threshold even for dose-related endpoints.

The theoretical limits and the time complexity will be demonstrated in future work. Considering that parsing with context-free grammars is decidable; and the worst-case time complexity is cubic, we expect that the complexity of SARpy in enumerating the substrings is still polynomial. Moreover, the worst case time complexity of the algorithm to break the aromatic bonds is polynomial in the number of atoms plus the number of bonds. In fact the algorithm behaves like a Depth First Search for each sub graph consisting of aromatic atoms. Because the total number of atoms and bonds in the aromatic sub graph does not exceed their total number in the molecule, the estimated time complexity is a worst case estimate.

Work is under way to make SARpy running on GPU to reduce computation times in case of very large data sets. Other future improvements will redefine the score in statistical terms as likelihood ratio. We will implement a graphical interface and use ROC curves for visualizing the substructures performance.

ACKNOWLEDGEMENTS

This work has been partially supported by the EU contracts CAESAR, ORCHESTRA and ANTARES. The

source code will be freely available through their web portals.

We acknowledge NVIDIA Academic partnership for providing us the GPU nVidia Tesla™ C1060 board.

REFERENCES

[1] R. Todeschini, and V. Consonni, "*Handbook of Molecular Descriptors*", Wiley-VCH, 2000.

[2] A. Inokuchi, T. Washio, and H. Motoda, “An apriori-based algorithm for mining frequent substructures from graph data,” *Proceedings PKDD'00*, pp 13–23, 2000.

[3] M. Deshpande, M. Kuramochi, N. Wale, and G. Karypis, “Frequent Substructure-Based Approaches for Classifying Chemical Compounds,” *IEEE Trans. on Knowl. and Data Eng*. 17, 8, pp 1036-1050, 2005.

[4] C. Borgelt, and M. R. Berthold, "Mining Molecular Fragments: Finding Relevant Substructures of Molecules," *Proceedings of the 2002 IEEE International Conference on Data Mining* (ICDM'02), pp 51, 2002.

[5] R. Benigni, and C. Bossa, “Structure alerts for carcinogenicity, and the Salmonella assay system: A novel insight through the chemical relational databases technology,” *Mutation Research/Reviews in Mutation Research*, Volume 659, Issue 3, 2008, pp 248-261, 2008.

[6] J. Kazius, R. Mcguire, and R. Bursi, "Derivation and validation of toxicophores for mutagenicity prediction," *Journal of Medicinal Chemistry*, 48(1), pp 312-320, 2005.

[7] H. S. Rosenkranz, Y. P. Zhang, G. Klopman, “Studies on the Potential for Genotoxic Carcinogenicity of Fragrances and Other Chemicals,” *Food and Chemical Toxicology*, 36 (8), pp 687-696, 1998.

[8] L. Dehaspe ,H. Toivonen ,R. D. King , “Finding Frequent Substructures in Chemical Compounds”, in (Gini G and A. Katrizky Eds) *Predictive Toxicology of Chemicals: Experiences and Impact of AI Tools*, AAAI-SS-99, AAAI Press, Menlo Park, CA, 1999.

[9] G. Klopman, “MULTICASE: A hierarchical computer automated structure evaluation program, *QSAR* 11, pp. 176–184, 1992.

[10] C. Helma, "Lazy Structure-Activity Relationships (lazar) for the Prediction of Rodent Carcinogenicity and Salmonella Mutagenicity," *Molecular Diversity* 10, pp 147-158, 2006.

[11] D. Weininger, "SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules," *Journal of Chemical Information and Computer Sciences* 28(1): 31-36, 1988.

[12] A. Karwath, and L. De Raedt, "SMIREP: predicting chemical activity from SMILES," *Journal of Chemical Information and Modelling*, 46(6):2432-44, 2006.

[13] 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, pp 3106-3127, 2009.

[14] R. Guha, M. T. Howard, G. R. Hutchison, P. Murray-Rust, H. Rzepa, C. Steinbeck, J. K. Wegner, and E. Willighagen. "The Blue Obelisk -- Interoperability in Chemical Informatics." *Journal of Chemical Information and Modelling* 46(3) pp 991-998, 2006.

[15] R. Sayle, "1st-class SMARTS patterns," presented at EuroMUG 97, Verona, Italy, 1997.

[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, pp 53-97, 2007.

[17] B. N. Ames, "The detection of environmental mutagens and potential," *Cancer,* 53, pp 2030-2040, 1984.

[18] W. W. Piegorsch, and E. Zeiger, "Measuring intra-assay agreement for the Ames salmonella assay," in (Hotorn L Ed) *Statistical Methods in Toxicology*, Lecture Notes in Medical Informatics, Springer-Verlag; pp 35-41, 1991.

[19] T. Ferrari, and G. Gini, “An open source multistep model to predict mutagenicity from statistical analysis and relevant structural alerts”, CAESAR QSAR Models for REACH, *Chemistry. Central Journa*l, Volume 4, Supplement 1, 2010.

[20] N. M. O'Boyle, C. Morley, and G. R. Hutchison, “Pybel: a Python wrapper for the OpenBabel cheminformatics toolkit,” *Chemistry Central Journal*, 2, 5, 2008.

[21] J. Ashby, "Fundamental structural alerts to potential carcinogenicity or noncarcinogenicity," *Environ Mutagen*, 7:919-921, 1985.

[22] J. A. Miller, and E. C. Miller, "Searches for ultimate chemical carcinogens and their reactions with cellular macromolecules," *Cancer* , 47, pp 2327-45, 1981.

[23] 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," *Mutation Research*, 204, pp 17-115, 1988.

[24] A. B. Bailey, N. Chanderbhan, N. Collazo-Braier, M. A. Cheeseman, and M. L. Twaroski, "The use of structure-activity relationship analysis in the food contact notification program," *Regulat Pharmacol Toxicol*, 42, pp 225-235, 2005.

[25] 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," Technical Report EUR 23241 EN, European Commission - Joint Research Centre 2008.

[26] S. J. Enoch, and M. T. D. Cronin, “A review of the electrophilic reaction chemistry involved in covalent DNA binding,” *Critical Reviews in Toxicology*, 40, 8, pp 728-748, 2010.

[27] C. de Meester, M. Duverger-van Bogaert, M. Lamboptte-Vandepear, M. Mercier, and F. Poncelet, “Mutagenicty of styrene in the Salmonella typhimurium test system,” *Mutation Research*, 90, pp 443-450, 1981.

[28] M. Kamiya, Y. Sengoku, K. Takahashi, K. Kohda, and Y. Kawazoe, “Antimutagenic structure modification of quinoline: fluorine-substitution at position-3”, *Basic Life Sciences*;52, pp 441-446, 1990.

[29] C. J. Smith, C. Hansch, and M. J. Morton, “QSAR treatment of multiple toxicities: the mutagenicity and cytotoxicity of quinolines,” *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis*, Volume 379, Issue 2, pp 167-175, 1997.

[30] S. Vamvakas, W. Dekant, and M.W. Anders, “Mutagenicity of benzyl s-haloalkyl and s-haloalkenyl sulfides in the ames-test,” *Biochemical Pharmacology* 38(6), pp 935-9, 1989.

[31] T. Ferrari, G. Gini, and E. Benfenati, “Support Vector Machines in the Prediction of Mutagenicity of Chemical Compounds,” *Proceedings NAFIPS*, June 14-17, Cincinnati, USA, 2009.

[32] D. J. Livingstone, “The characterization of chemical structures using molecular properties: a survey”, *Journal of Chemical Information and Computer Sciences*, 40(2), pp195-209, 2000.

[33] G. Gini, E. Benfenati, “e-modelling: foundations and cases for applying AI to life sciences", *International Journal on Artificial Intelligence Tools*, Vol 16, N 2, pp 243-268, 2007.