HAL
open science

New clues on carcinogenicity-related substructures derived from mining two large datasets of chemical compounds

Azadi Golbamaki, Emilio Benfenati, Nazanin Golbamaki, Alberto Manganaro, Erinc Merdivan, Alessandra Roncaglioni, Giuseppina Gini

To cite this version:

Azadi Golbamaki, Emilio Benfenati, Nazanin Golbamaki, Alberto Manganaro, Erinc Merdivan, et al.. New clues on carcinogenicity-related substructures derived from mining two large datasets of chemical compounds. Journal of Environmental Science and Health, Part C, 2016, 34 (2), pp.97-113. 10.1080/10590501.2016.1166879 . ineris-01863016

HAL Id: ineris-01863016

https://ineris.hal.science/ineris-01863016v1

Submitted on 28 Aug 2018

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

New clues on carcinogenicity-related substructures derived from mining two large datasets of chemical compounds

Azadi Golbamaki[1,*], Emilio Benfenati[1], Nazanin Golbamaki[2], Alberto Manganaro[1], Erinc Merdivan[3], Alessandra Roncaglioni[1],Giuseppina Gini[4]

[1]IRCCS - Istituto di Ricerche Farmacologiche Mario Negri, Milano, Italy

[2]Institut National de l'Environnement Industriel et des Risques (INERIS) , Verneuil en Halatte , France

[3]Sabancı University, Tuzla/Istanbul, Turkey; [4] Politecnico di Milano, Italy

[*]Corresponding author: Azadi Golbamaki, IRCCS - Istituto di Ricerche Farmacologiche Mario Negri, Department of Environmental Health Sciences, Laboratory of Environmental Chemistry and Toxicology. Via Giuseppe La Masa 19, 20156, Milan, Italy. Email address: azadi.golbamaki@marionegri.it, Phone number: +39 0239014595

Abstract

In this study, new molecular fragments associated with genotoxic and non-genotoxic carcinogens are introduced to estimate the carcinogenic potential of compounds. Two rule-based carcinogenesis models were developed with the aid of SARpy: model R (from rodents experimental data) and model E (from human carcinogenicity data). Structural alert extraction method of SARpy uses a completely automated and unbiased manner with statistical significance. The carcinogenicity models developed in this study are collections of carcinogenic potential fragments that were extracted from two carcinogenicity databases: the ANTARES

carcinogenicity dataset with information from bioassay on rats and the combination of ISSCAN and CGX datasets, which take into accounts human-based assessment. The performance of these two models was evaluated in terms of cross-validation and external validation using a 258 compound case study dataset. Combining R and H predictions and scoring a positive or negative result when both models are concordant on a prediction, increased accuracy to 72% and specificity to 79% on the external test set. The carcinogenic fragments present in the two models were compared and analysed from the point of view of chemical class. The results of this study show that the developed rule sets will be a useful tool to identify some new structural alerts of carcinogenicity and provide effective information on the molecular structures of carcinogenic chemicals.

Introduction

Identification, classification and risk assessment of carcinogenic chemicals by international organizations and national agencies of health and safety have made remarkable progress in recent years. The European Commission (EC) substantially modified and replaced the Directive 67/548/EEC and 93/101/EEC with Regulation (EC) 1272/2008 on risks and hazards of carcinogens and mutagens(1). The new regulation introduced the globally harmonised system of classification and labelling of chemicals (GHS). Under these directives experimental data studies on chemical carcinogens have been digitally collected with the aim of harmonizing national measures on classification, packaging and labelling of dangerous substances, to facilitate the establishment of a single market and to provide protection for public health and the environment. The new regulation complements the REACH regulation on the registration, evaluation, authorisation and restriction of chemicals.

Research has provided evidence that chemicals may cause cancer in animals and humans by one of several general mechanisms of action (MoA), generally classified into genotoxic and non-genotoxic. Genotoxic carcinogens cause damage to DNA, thus, many known mutagens are in this category, and often mutation is one of the first steps in the development of cancer (2). Epigenetic or non-genotoxic carcinogens do not bind covalently to DNA, and are usually negative in the standard mutagenicity assays (3). The unifying feature of all genotoxic carcinogens is that they are either electrophiles or can be activated to electrophilic reactive intermediates. On the contrary, non-genotoxic carcinogens act through a large variety of different and specific mechanisms.

For over 35 years, many chemicals have been tested by government agencies, private companies and research institutes using the two-year rodent carcinogenesis bioassay. Most of the chemicals or processes that have been associated with human carcinogenicity, as studied by epidemiological investigations, are shown to cause tumours in rats and mice (4-6). However, all compounds shown to induce cancer in laboratory rats and mice are not necessarily human carcinogens (7).

In the past ten years, research into the MoA and carcinogenesis has increased and the relevance of the carcinogenicity findings in rodents to human risk has been investigated in many publications (8-10). The results of research demonstrated that doses used in the bioassays may do not develop toxicity in humans exposed to same levels of these chemicals; in addition, rats and mice tumours occur in a sex, age and strain or stock dependent manner. In consequence of these points, the regulatory agencies consider that the high occurrence of tumours in the standard two-year rodent carcinogenesis bioassay is often not relevant to risk evaluation of human carcinogenesis (11). Variability of the tumours in rodents is another problem of this assay. To deal with the problems of two-year rodent carcinogenesis bioassay alternative methods are suggested by scientists and regulatory agencies. These methods include use of the toxicity level (LD50) in rodents (12), *in vitro* cell transformation and other assays, *in silico* methods or computerized prediction of carcinogenicity based on structure and chemical class (13). Each method has its own strengths and weaknesses, and analysis of carcinogenicity of a specific chemical and its MoA in human is better to be assessed based on the weight of evidence.

Among the *in silico* methods, the use of various computational techniques such as (quantitative) structure-activity relationship ((Q)SAR) modelling is supported by several legislative authorities

(14-16). (Q)SAR models consist of mathematical relationships between physicochemical properties of chemicals and their biological activity, thus being able to calculate a quantitative value (for the activity) given the structure of a chemical. These mathematical relationships can be simple linear regression equations, or more complex non-linear algorithms, and can be developed using several approaches such as neural networks, support vector machines, decision trees and many others. Conversely, SAR identifies the differences of compounds in two categories (e.g. active or inactive) and predicts an untested compound as "toxic" in case it has a toxic potential or "non-toxic" if not. Overall, (Q)SAR models are useful for the prediction of toxicity of untested chemicals saving costs and the need for testing on animals (17, 18).

Following the theory of eletrophilic reactivity of (many) carcinogens of James and Elizabeth Millers (19, 20), the advancement of the knowledge of carcinogenic chemicals have received distinguished contributions from many scientists. The *salmonella typhimurium mutagenicity* assay by Bruce Ames (7) and the compilation of the lists of carcinogenic and mutagenic structural alerts (SA) by John Ashby (21) were two fundamental contributions to this field. SAs identified and collected by John Ashby's are indeed reactive functional groups responsible for the induction of mutation or cancer, and are so-called genotoxic carcinogens. On the other hand, the *Salmonella* assay is the most predictive assay for genotoxic carcinogens and no other non-genotoxic mutagenicity test exists (22). Despite the extensive knowledge of genotoxic SAs, the use of SAs for identifying non-genotoxic carcinogens is restricted. Non-genotoxic carcinogens use many different MoA and they lack an apparent unifying mechanism. According to this diversity, different (Q)SAR models have been developed and made available for analysis and

 ACCEPTED MANUSCRIPT

identification of SAs. A number of non-genotoxic SAs and their characteristics have been published in (3).

One of the most recent rule sets defined by human expert for mutagenic carcinogenicity has been developed by Benigni and Bossa (23, 24). The updated version of this rule set (24) is implemented in Toxtree version 2.6.13 (25), a software application that investigates the presence of the genotoxic and non-genotoxic SAs in the chemical structures of the compounds. Alongside the rule-based (Q)SAR software that check the presence of human expert SAs in the chemical structures, there are statistically-based (Q)SARs which create models by using categorized active and inactive chemicals in a learning set to identify SAs that are associated with a particular toxicological activity. The high accuracy of the predictions performed by data mining and artificial intelligence has made these methods important tools to be used for preliminary research and for discovery of the mechanism of action that are still unknown. These methods however, comparing to rule-based models are less transparent to the end user. Historically, the Computer Automated Structure Evaluation (CASE/MultiCASE) (26) program is a SAR expert system that identifies two-dimensional structural features or biophores which can be used for the prediction of unknown compounds as potential toxins. This statistically-based program does not use the knowledge on the mechanisms of action, but reanalyse the dataset of chemicals trying to link the structures of chemicals into their toxic activity. On the other hand, SAs developed by human experts were integrated in software such as OncoLogic (27) and DEREK (28).

In this study we used SARpy (29), a commercially free statistically-based program, for the extraction of potential carcinogenic SAs from two different learning sets. The approach that we have taken in developing the two new carcinogenicity models is mainly based on statistical

 ACCEPTED MANUSCRIPT

evaluation of the chemicals in our learning sets categorized in two groups of carcinogens and non-carcinogens. The SARpy's method of identification of the SAs that are associated with a particular biological or toxicological activity does not demand *a priori* knowledge about MoA of the compounds and performs purely on a statistical basis. Two different carcinogenicity datasets have been prepared as learning sets and SARpy extracted two different models from these two datasets. The internal and external evaluation of the models have been assessed thoroughly. The choice of taking into consideration two substantially different learning sets and developing two models is due to different characterization of these data. The first dataset contains exclusively rodent carcinogenicity data based on presence of carcinogenic effects in male or female rats, while the second dataset takes into account human-based assessments and data retrieved from different assays. This suggests to obtain two different carcinogenicity models.

Finally, the SAs in the two rule sets are analyzed from the point of view of chemical class and the same SAs present in both rule sets are revised. The two developed models have been made available inside VEGA (http://www.vega-qsar.eu/) (30), an open source platform that already offers several (Q)SAR models.

Material and Methods

Carcinogenesis data sources

ANTARES carcinogenicity dataset: Rat carcinogenesis learning set

Compounds for the first model's learning set were obtained from the carcinogenicity database of EU-funded project ANTARES (31). The ANTARES' carcinogenicity database is a collection of chemical rat carcinogenesis data (presence of carcinogenic effects in male or female rats)

 ACCEPTED MANUSCRIPT

obtained from the EU-funded project CAESAR (32) dataset and the "FDA 2009 SAR Carcinogenicity - SAR Structures" database. The CAESAR toxicity values were originated from the Distributed Structure-Searchable Toxicity DSSTox database, which was built from the Lois Gold's Carcinogenic Potency Database (CPDB) (33). The compounds with a definite TD50 (which is the dose that produces an increase of 50% of the tumours in animals) value for rat in this dataset were labeled as carcinogenic, while the remaining were labeled as non-carcinogenic. Additional 738 chemicals different from the 805 CAESAR compounds were added. The added chemicals are from the "FDA 2009 SAR Carcinogenicity - SAR Structures" database using the Leadscope database (34). Here a categorical label for carcinogenicity was already contained in the original dataset and again the compound was labeled as carcinogenic if a positive outcome was detected in male or female rats. So a total number of 1543 compounds constituted the ANTARES dataset.

ISS Carcinogenicity database and Carcinogenicity Genotoxicity eXperience dataset: Different species carcinogenesis learning set

The ISS Carcinogenicity (ISSCAN) database (35) is provided by the Istituto Superiore di Sanità (ISS). It is originally aimed at developing predictive models for carcinogenicity of chemicals. The great part of the chemicals in this database are classified as carcinogens by various regulatory agencies and scientific bodies. The database has been specifically designed as an expert decision support tool and contains information on chemicals tested with the long-term carcinogenicity bioassay on rodents (presence of carcinogenic effects in male or female rats and mice). This carcinogenicity dataset contains 622 carcinogens, 210 non-carcinogens and 58 equivocals.

Compounds for the second model's learning set were obtained by merging the ISSCAN database and the Carcinogenicity Genotoxicity eXperience (CGX) database. More information on the CGX database can be found in (36). In this study, compounds used for development of the new models had to be either positive or negative, thus, compounds with equivocal results in the databases have been removed. In particular, from the original ISSCAN dataset with 890 compounds, we removed 58 compounds, while the CGX database did not contain any equivocal result.

All compounds in the combined dataset have been checked for their consistency between the two sources. We found 651 compounds in common, 15 of them with inconsistent carcinogenicity values. These compounds have been removed from the combined dataset.

Comparison with the ANTARES dataset

We compared the final list of compounds with the ANTARES carcinogenicity dataset prepared for the development of the first model. We found 105 compounds with conflicting values when compared with the compounds in the ANTARES dataset. In order to develop a more conservative model, we opted to remove only 15 compounds which had positive result in the ANTARES dataset and negative results in the combined second dataset, and left as carcinogenic those that had carcinogenicity result the opposite way. Consequently, there are 90 positive compounds in the combined database which are negative in the ANTARES dataset. Afterwards, we checked and cleaned the structures manually, and by the help of the istMolBase (37) and InstantJChem (38) software formed the final dataset.

In addition, the compounds have been checked for their molecular structure. We adopted only the substances with connected molecular structure; those which had unconnected structures have been removed from the dataset. The overall dataset consisted of 986 compounds with 734 carcinogens and 252 non-carcinogens. Each compound in the list had a chemical name, a CAS number, a Simplified Molecular Input Line Entry Specification (SMILES) (39), and its categorical designation (i.e. carcinogen or non-carcinogen). In the present study, this combined dataset is conventionally called ISSCAN-CGX.

Data for model validation

ECHA database

We prepared an external test set for the validation of the developed models from carcinogenicity the eChemPortal inventory (40). For this purpose, we made two queries on this database. The first query contained the following restrictions:

i) Study result type: experimental result

ii) Reliability: 1 and 2

iii) Species: mouse and rat

iv) Maximum number of studies: 4

The second query consisted of:

i) Study result type: experimental result

ii) Reliability: 1 and 2

 ACCEPTED MANUSCRIPT

iii) Species: mouse and rat

iv) Sources: any guideline and exposure route

The list resulted from the first query comprised 308 compounds, whereas, the second query returned a list of 166 compounds, which were mostly in common with the results of the first query. The studies conducted for the first list of compounds have been manually evaluated. Afterwards, we looked into the Classification Labelling and Packaging (CLP) inventory (41) for the positive (i.e. carcinogenic) chemicals collected by the above mentioned queries. Inside the CLP inventory we found 68 compounds, which were already present in our data collection. The latter search confirmed the carcinogenic property of these compounds.

The dataset consisted of 64 positive compounds, 169 negative compounds, and 90 equivocal compounds. The equivocal results are due to the presence of conflicting information in different sources or different studies in the same source.

It should be noticed that, for already classified compounds (no conflicting information), the level of uncertainty in the assignment is not homogeneous, because some of the compounds were classified on the basis of a single study (i.e. data present in one single source).

From the reliability point of view, in the data collected in our dataset, 49 positive compounds have positive carcinogenic effect in at least two sources. 57 negative compounds are non-carcinogenic in both lists, and they are not present in the list of compounds retrieved from the CLP inventory. 64 compounds are considered as non-carcinogens because of the presence of only one single study in the two lists.

SARpy

 ACCEPTED MANUSCRIPT

The SARpy (SAR in Python) program is a Python script based on the OpenBabel chemical library. SARpy creates classification models by using categorized active and inactive chemicals in a learning set to identify molecular fragments that are associated with a particular biological, pharmaceutical or toxicological activity. The algorithm generates molecular substructures of arbitrary complexity, and the fragments candidates to become SAs are automatically selected on the basis of their prediction performance in a learning 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 a SMARTS string, for use by human experts or other chemical software. SMARTS notations are text representations of substructures (36) that allow specification of wildcard atoms and bonds, which can be used to formulate substructure queries for a chemical database. Those rules can be used as a predictive model simply by calling a SMARTS matching program. For the matching phase, SMILES and the SMARTS strings are translated into graphs and the two graphs are compared to each other (42).

Extracting active fragments

R (rat) model

To obtain a more comprehensive collection of potential carcinogenic fragments, five learning sets were randomly created from the ANTARES carcinogenicity dataset with 1543 compounds, preserving 80% for the learning set and 20% for the evaluation set. In other words, for each model a random set of 20% of chemicals in the learning set was removed, with the remaining 80% of the compounds a model was developed and the activity of the compounds left out was

predicted with the same model. We combined the five models and put together the lists of the potential active fragments, removed the duplicates and eliminated the SAs with likelihood ratio lower than 2. We opted for the likelihood ratio threshold of 2 in order to retain the SAs which are statistically more significant. A measure of each fragment's association with biological activity is determined by SARpy as "training likelihood ratio" and it is given along with the list of the potential fragments or the rule set in the output. The likelihood ratio can be taken into account to determine the goodness of a SA identified by SARpy. Even if a SA that is associated with activity (e.i. carcinogenicity) is present in a molecular structure, the molecule may contain other fragments that make it inactive (e.i. non-carcinogen), thus the specific SA might not be expected to be found only in active compounds. This evidence is the basis of the determination of the likelihood ratio.

Using the SARpy software, each chemical in the learning set was fragmented *in silico* into all possible fragments meeting user-specified criteria. For this study we extracted only the "ACTIVE" fragments (or SAs) and the default values for the minimum and maximum number of atoms in a fragment were set for the fragment extractions of each model (minimum=2; maximum=18). Another configuration to establish by the user is the minimum number of compounds in the learning set in which an active (or inactive) fragment is found. In our analysis, the minimum number of compounds that contain a potential active fragment was set to 3. Conventionally, in this study we call this model R.

E (expert) model

SARpy was used for model development and statistical analysis using the ISSCAN-CGX dataset.

 ACCEPTED MANUSCRIPT

The extraction settings are as follows: the minimum number of atoms in a fragment is equal to 4, whereas, the maximum number of atoms is equal to 10, and the minimum number of compounds containing the active fragment is 6. These configurations have been set in favour of a model with a more balanced sensitivity and specificity values. In order to assess the predicitivity of the model, statistical analysis have been conducted in terms of accuracy, sensitivity and specificity using cross-validation routine as an internal evaluation, in addition to an external evaluation using an external test set. In this paper, we name this model E.

Internal evaluation of the models

Accuracy, sensitivity and specificity have been determined for the internal evaluation of each model using the SARpy program. For the internal validation, 5-fold cross-validation routine was conducted for each model. In the 5-fold cross-validation the learning set is randomly partitioned into five equal sized subsets. For each iteration, a single subset of chemicals was retained as the validation data for testing the model, and the remaining subsets were used as training data. The cross-validation process was repeated five times (the folds). The evaluation results of five iterations were then averaged to produce a single estimation. Accuracy, sensitivity and specificity of the internal evaluation are assessed in addition to the Matthews correlation coefficient (MCC).

External evaluation of the models

The predictability of the models has been evaluated on two external test sets: the first external set is the dataset used as the learning set of the opposite model (e.g. for the R model we used ISSCAN-CGX dataset and vice versa), and the second dataset is a collection of 258 compounds

collected from the eChemPortal inventory. Accuracy, sensitivity, specificity and the MCC for the external evaluation are determined using SARpy. Although the external evaluation is considered the best mean for the assessment of the predictive ability of a (Q)SAR model (43, 44), the results of the external evaluation of any model are highly related to the relative similarity of the external evaluation set in relation to the learning set.

Results and discussions

R model

Each learning set produced its own model, which is a collection of active SAs with their likelihood ratios. The final model merging all sets of SAs consisted of 127 active SAs. Table 1 shows the predictive performance of five models developed based on five different splits of the ANTARES database. The performance of each model has been evaluated on its own learning set using cross-validation analysis. Further, an external evaluation using the corresponding test set is performed on each model. To have an overview of the statistical analysis of the performance of the models, we calculated the average of the predictive values of all the five models, and reported in Table 1 as well. The averages of accuracy, sensitivity and specificity for the 778 compound internal cross-validation using five rule sets extracted from the ANTARES dataset were 71%, 73% and 69%, respectively. The average of accuracy, sensitivity and specificity for 337 compounds in the test set as an external validation of these models, were 63%, 63% and 62%, respectively.

Using the R model, the results of cross-validation on the whole training set were 66% accuracy, 83% sensitivity, 48% specificity and 0.34 the MCC (Table 2). Analysis of the external validation

for the R model demonstrated that the concordance between experimental and predicted value on the ECHA dataset is higher than using the ISSCAN-CGX dataset. The accuracy of the R model on the ECHA dataset was 67%, compared to 58% of accuracy for the ISSCAN-CGX dataset. The complete list of these alerts are presented in the VEGA platform.

E model

With the configuration set as mentioned above, SARpy extracted 43 active rules from the ISSCAN-CGX learning set. Analysis of the cross-validation for the E model demonstrated that the second model produced an accuracy of 73%, with a sensitivity of 77% and a specificity of 62% (Table 2). The MCC value for this analysis is 0.36. The accuracy values for the external evaluation of the E model on the ANTARES dataset and the ECHA database were 59% and 64%, respectively. Analysis of the external validations for the E model demonstrated that the model produced a higher sensitivity (77%) compared with the specificity (41%) of the R model. On the contrary, the specificity of the external evaluation on the chemicals from the ECHA database was higher (72%) compared to its sensitivity (48%) (Table 2). The complete list of the SAs present in this model is accessible through VEGA.

Analysis of the combination of the prediction results of the R and the E models

Another analyses has been done on the prediction results of the R model and the E model. In this new approach, we considered the final results as correctly predicted only in case both models have predicted them consistently. Table 3 summarizes the results of combining the R and E model external validation predictions on the chemicals from the ECHA database.

 ACCEPTED MANUSCRIPT

The results suggested that when both models are concordant on a negative prediction for a compound the reliability of the result is much higher than in case a positive prediction is done.

We observe an improvement of the results compared to the use of the individual models, for accuracy (72%) and specificity (79%). In fact, combining the predictions of the two models the MCC is increased to 0.37, compared to 0.31 for the R model and 0.20 for the E model. Only sensitivity is higher using the R model (62%). Thus, users may choose a solution or another depending if they prefer a conservative or a realistic assessment.

Fragments analysis

Comparison of the SAs in the R and E models

The SAs present in the R and E models have been compared and those that are in common between the two rule sets categorized into chemical classes and listed as follows. The SAs in the R model are presented with their ID number and written in order of their correspondence to the identical SAs in the E model.

1) Aromatic amine (R model: 6, 41, 36, 22, 10 / E model: 27, 31, 33, 38, 104)
2) Aromatic heterocyclic (R model: 12, 19, 2 / E model: 75, 108, 117)
3) Hydrazide (R model: 28, 27 / E model: 2, 50)
4) N-Nitroso (R model: 1 / E model: 8)
5) Phenyl-Hydrazine (R model: 32 / E model: 48)
6) α,β- Haloalkanes (R model: 25 / E model: 56)
7) Sulfite (R model: 8 / E model: 68)
8) Nitrogen Mustard like (R model: 11 / E model: 73)

 ACCEPTED MANUSCRIPT

9) Phosphonite (R model: 15 / E model: 98)

Categorization of the SAs in the R and E models

The SAs present in the models R and E are categorized from a chemical class point of view. The substructures within each category are presented with their ID number in their original rule set and are as follows:

Nitrogen containing substructures (Azo type):

1) Aromatic amine (R model: 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 33, 34, 35, 36, 37, 38, 40, 42, 83, 104, 110, 113 / E model: 6, 10, 22, 31, 35, 36, 41, 42)
2) Aromatic heterocycles containing Nitrogen (R model: 74, 75, 80, 81, 83, 95, 113, 122 / E model:12, 17, 43)
3) Azine (Hydrazine) (R model:46, 47, 49, 50, 51, 53, 54, 55, 101 / E model: 27, 32)
4) Azide (Hydrazide) (R model: 2, 3, 44, 45, 52 / E model: 3, 28)
5) Nitrosamine (R model:4, 5, 7, 9, 10 / E model: not found (NF))
6) Nitrogen or sulfur mustard (R model: 72, 73, 115 / E model: 11, 34)
7) Aromatic methylamine (R model: 30, 34, 36 / E model: NF)
8) Aliphatic N-Nitroso (R model: 62, 63/ E model: NF)
9) Aromatic Nitro (R model: 90, 123 / E model: NF)
10) 1 aryl 2 monoalkyl hydrazine (R model: 48 / E model: NF)
11) Aziridine (R model:120 / E model: NF)
12) Aromatic hydroxylamine (R model: 32 / E model: NF)
13) Diazo (R model:92 / E model: NF)

14) Aromatic Azo (R model: 71 / E model: NF)

15) Aromatic Nitroso (R and E models: NF)

Other substructures:

1) (1,2, and 3 membered) Aromatic Heterocycles (R model: 74, 75, 80, 81, 83, 90, 95, 103, 108, 113, 117, 121, 122, 123 / E model: 2, 12, 17, 19, 43)
2) Aliphatic halide (R model: 57, 58, 59, 70, 125 / E model: 18, 25)
3) Heterocyclic Alkane (R model: 84, 105, 109, 120 / E model: 23)
4) Polycyclic aromatic systems (R model: 39, 43, 60, 61 / E model: 30)
5) Sulfonate bonded carbon (R model: 67, 68 / E model: 8)
6) Epoxide (R model: 105 / E model: 23)
7) B propiolactone (R model: 114 / E model: NF)

Not only SARpy was able to find the already known carcinogen substructures that were represented by the SAs of Kazius et al. (45), but a number of SAs have been identified for the first time. Table 4 demonstrates the new identified SAs that have been classified into six chemical classes. The substructures within each category are listed with their ID number and are as follows:

1) Nitrosurea (R model: 12, 13, 14, 19 / E model: NF)

2) Nitrogen or sulfur mustard like (R model: 72, 115 / E model: 34)

3) Benzodioxole and Benzendiol (R model: 17, 18 / E model: 9)

4) Teritiary amine substituted by a Sulfur atom (E model: 24)

5) α,β-oxy and carboxy substitutions (R model: 20, 21, 76 / E model: NF)

6) α,β-haloalkanes (R model: 56, 69 / E model: 25)

7) Oximes (R model: 78 / E model: NF)

For the sake of example, we illustrated the chemicals from which the SA 24 (form the chemical class teritiary amine substituted by a Sulfur atom) in the E model has been extracted (Table 5). All the chemicals that contain the above mentioned SA in the ISSCAN-CGX data set are carcinogenic.

Discussion

Automated extraction of SAs has been implemented by the statistically-based program SARpy on two learning sets. The ANTARES learning set collects rodent bioassay carcinogenicity data on 1543 chemicals, while ISSCAN-CGX database containing 986 chemicals takes into account human-based assessments and data retrieved from different assays. The predictive performance of the developed models were evaluated internally, as well as using a 258 compound external validation dataset collected from the ECHA inventory. The two developed models for carcinogenicity have been implemented in the VEGA platform and are indeed freely available for end users.

Recent progresses in data mining provide effective competence in the automated discovery of SAs associated to toxicological endpoints. An important contribution of the statistically-based methods to the carcinogenicity field is identification of new SAs which help us in refining the existing rule sets. While the most known carcinogenicity rule sets (23) are composed on the basis of human expert judgement, the SAs identified in our study are extracted in an unbiased manner

by SARpy with no *a priori* knowledge about the MoA of the chemicals. This approach shed light to the new clues about genotoxic and non-genotoxic SAs. Some primary analyses have been provided on the SA lists; chemical classes of the identified SAs have been evaluated, however, further study for the new SAs should be performed considering other collections of alerts (45). SARpy SAs resulting from the current analysis on the ANTARES and ISSCAN-CGX data sets follow the SAs presented by Kazius et al. (46).

Furthermore, the models are developed on the basis of two learning sets with different carcinogenicity data from the point of view of origin and provenance. Concerning the learning sets with substantially variant carcinogenicity data assessed within different properties, each set of the extracted SAs constituted a purpose oriented model. The user may consider the results of the model with more realistic predictions or the one with more conservative assessments.

Generally, the best approach in making a conclusion to estimate the reliability of a prediction is combining evidence from different information sources such as (Q)SAR model predictions, *in vitro* and *in vivo* test results. This is reflected in the general trend of developing ensemble models and/or combining the output of different existing models. An example of the latter approach has been done on a similar endpoint, mutagenicity (Ames test), by the integration of the different models available on the VEGA platform (47). The advantage of having the two presented models available on the VEGA platform, where other models for the same endpoint are available, is also the possibility of performing a similar activity to make a conclusion.

Finally, the results of the presented models will be exploited for the improvement of ToxRead (http://www.toxgate.eu), a recent platform that uses set of rules for different endpoints to filter and select similar compounds and assist the user in performing read-across studies (48, 49).

Also, these rules can be compared and possibly explained considering reasoning about mechanisms, including adverse outcome pathways (50).

Acknowledgement

The research for this paper was financially supported by the Life PROSIL project and Ministero della Salute of Italy / Istituto Superiore di Sanità (ISS) within the project “Messa a punto di strategie integrate di testing -basate su metodi alternativi- per l’identificazione di sostanze cancerogene per il REACH”.

 ACCEPTED MANUSCRIPT

References

1. Regulations R: No. 1272/2008 of the European Parliament and of the Council, on Classification, Labeling and Packaging of Substances and Mixtures, Amending and Repealing Directives 67/548/EEC and 1999/45/EC, and Amending Regulation (EC) No. 1907/2006, Official J. *Eur Union* 2008:L353.

2. Arcos J.C., 1995. Chemical Induction of Cancer: Modulation and Combination Effects. An Inventory of the Many Factors Which Influence Carcinogenesis: Springer Science & Business Media.

3. Woo Y., Lai D., 2003. Mechanisms of action of chemical carcinogens, and their role in Structure-Activity Relationships (SAR) analysis and risk assessment. Quantitative Structure-Activity Relationship (QSAR) models of mutagens and carcinogens; 41-80.

4. Huff J., 1999. Long-Term Chemical Carcinogenesis Bioassays Predict Human Cancer Hazards: Issues, Controversies, and Uncertainties. Annals of the New York Academy of Sciences, 895(1); 56-79.

5. Tomatis L., 2006. Identification of carcinogenic agents and primary prevention of cancer. Annals of the New York Academy of Sciences,1076(1); 1-14.

6. Doll R., 2001. The causes of cancer. Revue d'épidémiologie et de santé publique, 49(2); 193.

7. Ames B.N., Gold L.S., 1990. Chemical carcinogenesis: too many rodent carcinogens. Proceedings of the National Academy of Sciences, 87(19); 7772-7776.

8. Anisimov V.N., Ukraintseva S.V., Yashin A.I., 2005. Cancer in rodents: does it tell us about cancer in humans? Nature Reviews Cancer, 5(10); 807-819.

9. Knight A., Bailey J., Balcombe J., 2006. Animal carcinogenicity studies: 1. Poor human predictivity.

10. Knight A., Bailey J., Balcombe J., 2004. Animal carcinogenicity studies: 2. Obstacles to extrapolation of data to humans. System.

11. Ward J.M., 2007. The two-year rodent carcinogenesis bioassay-Will it survive? Journal of toxicologic pathology, 20(1); 13-19.

12. Ashby J., Paton D., 1993. The influence of chemical structure on the extent and sites of carcinogenesis for 522 rodent carcinogens and 55 different human carcinogen exposures. Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis, 286(1); 3-74.

13. Ward J.M., 1996. Rat or Mouse Cancer Bioassay—or None of the Above? Toxicologic pathology, 24(6); 734-735.

14. Louekari K., Sihvonen K., Kuittinen M., Sømnes V., 2006. In vitro tests within the REACH information strategies. Alternatives to laboratory animals: ATLA, 34(4); 377-386.

15. Combes R., Grindon C., Cronin M., Roberts D.W., 2008. Garrod JF: Integrated decision-tree testing strategies for mutagenicity and carcinogenicity with respect to the requirements of the EU REACH legislation. Alternatives to laboratory animals: ATLA, 36; 43-63.

16. Wells M.Y., Williams E.S., 2009. The transgenic mouse assay as an alternative test method for regulatory carcinogenicity studies—Implications for REACH. Regulatory Toxicology and Pharmacology, 53(2); 150-155.

17. Benigni R., Bossa C., 2008. Predictivity of QSAR. Journal of chemical information and modeling, 48(5); 971-980.

18. Benigni R., Bossa C., 2008. Predictivity and reliability of QSAR models: The case of mutagens and carcinogens. Toxicology mechanisms and methods, 18(2-3); 137-147.

19. Miller J., Miller E., 1977. Origins of human cancer. Cold Spring Harbor Laboratory, Cold Spring Harbor, NY; 605-627.

20. Miller E.C., Miller J.A., 1981. Searches for ultimate chemical carcinogens and their reactions with cellular macromolecules. Cancer, 47(10); 2327-2345.

21. Ashby J., Tennant R.W., 1988. Chemical structure, Salmonella mutagenicity and extent of carcinogenicity as indicators of genotoxic carcinogenesis among 222 chemicals tested in rodents by the U.S. NCI/NTP. Mutat Res, 204(1); 17-115.

22. Benigni R., Bossa C., Tcheremenskaia O., Giuliani A., 2010. Alternatives to the carcinogenicity bioassay: in silico methods, and the in vitro and in vivo mutagenicity assays. Expert Opin Drug Metab Toxicol, 6(7); 809-819.

23. Benigni R., Bossa C., 2008. Structure alerts for carcinogenicity, and the Salmonella assay system: a novel insight through the chemical relational databases technology. Mutat Res, 659(3); 248-261.

24. Benigni R., Bossa C., Tcheremenskaia O., 2013. Nongenotoxic carcinogenicity of chemicals: mechanisms of action and early recognition through a new set of structural alerts. Chemical reviews, 113(5); 2940-2957.

25. Toxtree, Ideaconsult Ltd, Sofia, Bulgaria. http://toxtree.sourceforge.net/.

26. Rosenkranz H., 2003. Quantitative Structure-Activity Relationship (QSAR) models of chemical mutagens and carcinogens. In.: CRC Press: Boca Raton.

27. OncoLogic, EPA. http://www2.epa.gov/tsca-screening-tools/oncologictm-computer-system-evaluate-carcinogenic-potential-chemicals.

28. Derek Nexus, Lhasa Limited, Leeds, UK. http://www.lhasalimited.org.

29. Ferrari, T., Cattaneo, D., Gini, G., Bakhtyari, N. G., Manganaro, A., Benfenati, E., 2013. Automatic knowledge extraction from chemical structures: the case of mutagenicity prediction. SAR and QSAR in Environmental Research, 24(5).

30. VEGA. Istituto di Ricerche Farmacologiche Mario Negri Milano. http://www.vega-qsar.eu.

31. ANTARES project, Grant Agreement LIFE08 ENV/IT/00435. http://www.antares-life.eu/.

32. CAESAR project, no. 022674 – SSPI. http://www.caesar-project.eu/.

33. Gold L.S., 2011. The Carcinogenic Potency Project and Database (CPDB). University of California, Berkeley; Lawrence Berkeley, National Laboratory; National Library of Medicine's (NLM®). http://potency.berkeley.edu.

34. Leadscope. http://www.leadscope.com/.

35. Benigni R., Cecilia B.. ISSCAN: Istituto Superiore di Sanita, "CHEMICAL CARCINOGENS: STRUCTURES AND EXPERIMENTAL DATA". https://w3.iss.it/site/BancaDatiCancerogeni/.

36. Kirkland D., Aardema M., Henderson L., Müller L., 2005. Evaluation of the ability of a battery of three in vitro genotoxicity tests to discriminate rodent carcinogens and non-carcinogens: I. Sensitivity, specificity and relative predictivity. Mutation Research/Genetic Toxicology and Environmental Mutagenesis, 584(1); 1-256.

37. istMolBase. Kode. http://chm.kode-solutions.net/products_istmolbase.php.

38. InstantJChem.ChemAxon. https://www.chemaxon.com/products/instant-jchem-suite/instant-jchem/.

39. Weininger D., 1988. 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.

40. eChemPortal. In. OECD. http://www.echemportal.org/echemportal/index?pageID=0&request_locale=en.

41. CLP inventory. In.: ECHA. http://echa.europa.eu/it/regulations/clp/cl-inventory.

42. Ferrari T., Gini G., Bakhtyari N.G., Benfenati E., 2011. Mining toxicity structural alerts from SMILES: A new way to derive Structure Activity Relationships. In: Computational Intelligence and Data Mining (CIDM), IEEE Symposium on: 11-15 April 2011; 120-127.

43. Tropsha A., Gramatica P., Gombar V.K., 2003. The importance of being earnest: validation is the absolute essential for successful application and interpretation of QSPR models. QSAR & Combinatorial Science, 22(1); 69-77.

44. Perkins R., Fang H., Tong W., Welsh W.J., 2003. Quantitative structure-activity relationship methods: Perspectives on drug discovery and toxicology. Environmental Toxicology and Chemistry, 22(8); 1666-1679.

45. Ceccaroli C., Pulliero A., Geretto M., Izzotti A., 2015. Molecular Fingerprints of Environmental Carcinogens in Human Cancer. Journal of Environmental Science and Health, Part C, 33(2); 188-228.

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

47. Cassano A., Raitano G., Mombelli E., Fernández A., Cester J., Roncaglioni A., Benfenati E., 2014. Evaluation of QSAR Models for the prediction of Ames genotoxicity: A retrospective exercise on the chemical substances registered under the EU REACH regulation. Journal of Environmental Science and Health, Part C, 32(3); 273-298.

48. Gini G., Franchi A.M., Manganaro A., Golbamaki A., Benfenati E., 2014. ToxRead: a tool to assist in read across and its use to assess mutagenicity of chemicals. SAR QSAR Environ Res, 25(12); 999-1011.

49. Benfenati E., Manganelli S., Giordano S., Raitano G., Manganaro A., 2015. Hierarchical Rules for Read-Across and In Silico Models of Mutagenicity. Journal of Environmental Science and Health, Part C, 33(4); 385-403.

50. Benigni R., Battistelli C. L., Bossa C., Giuliani A., Tcheremenskaia, O., 2015. Alternative toxicity testing: analyses on Skin sensitization, Toxcast Phases I and II, and Carcinogenicity provide indications on how to model mechanisms linked to adverse outcome pathways. Journal of Environmental Science and Health, Part C, 33(4); 422-443.

Table 1. R model internal and external validation for five different splits and the average of the model performance

		1° split (59 active rules)	2° split (65 active rules)	3° split (61 active rules)	4° split (58 active rules)	5° split (57 active rules)	Average
Learning set (778 compounds)	Accuracy	71 %	72 %	71 %	70 %	71 %	71 %
	Sensitivity	75 %	75 %	71 %	73 %	70 %	73 %
	Specificity	65 %	69 %	71 %	66 %	72 %	69 %
Test set (337 compounds)	Accuracy	63 %	60 %	64 %	65 %	62 %	63 %
	Sensitivity	68 %	58 %	62 %	67 %	61 %	63 %
	Specificity	56 %	63 %	66%	61 %	64 %	62 %

 ACCEPTED MANUSCRIPT

Table 2. R model and E model internal and external validation

	R model (127 active rules)			E model (43 active rules)		
	Cross-validation	external validation on ISSCAN and CGX data	external validation on ECHA data	Cross-validation	external validation on ANTARES data	external validation on ECHA data
Accuracy	66%	58%	67%	73%	59%	64%
Sensitivity	83%	76%	62%	77%	77%	48%
Specificity	48%	40%	70%	62%	41%	72%
TP[a]	651/783	593/735	55/89	562/735	599/783	43/89
TN[b]	367/760	142/254	119/169	157/254	315/760	121/169
FP[c]	393/760	112/254	50/169	95/254	445/760	48/169
FN[d]	132/783	142/735	34/89	172/735	184/738	46/89
MCC[e]	0.34	0.35	0.31	0.36	0.19	0.20

[a] True positive

[b] True negative

[c] False positive

[d] False negative

[e] Matthews Correlation Coefficient

 ACCEPTED MANUSCRIPT

Table 3. The combination of the predictions of the R and E models on the ECHA external validation set

Combined model	
TP[a]	33/89
TN[b]	96/169
FP[c]	25/169
FN[d]	24/89
Accuracy	72%
Sensitivity	58%
Specificity	79%
MCC[e]	0.37
Coverage	178/258

[a] True positive

[b] True negative

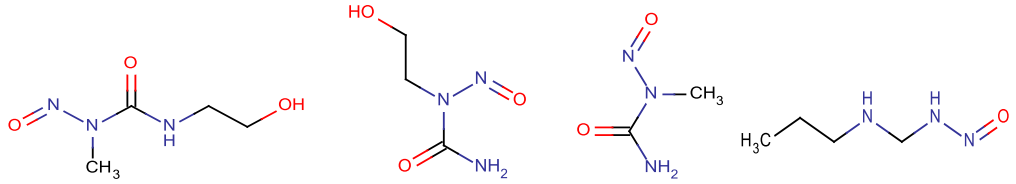
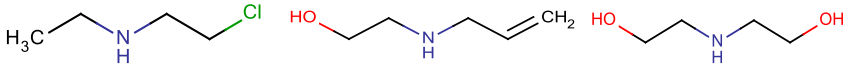
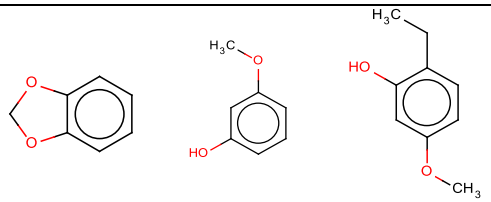
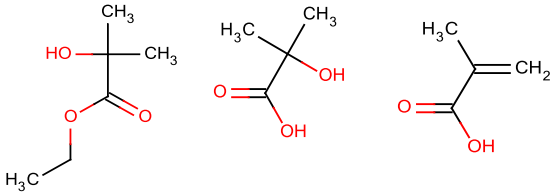
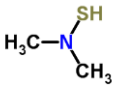
[c] False positive

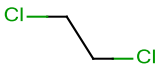
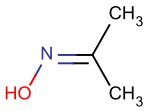
[d] False negative

[e] Matthews correlation coefficient

 ACCEPTED MANUSCRIPT

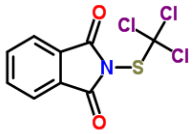
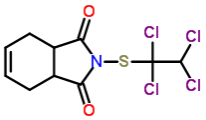
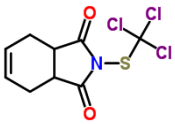
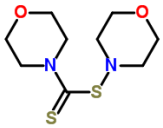
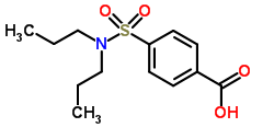
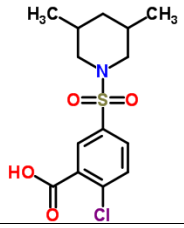
Table 4. New carcinogenic structural alerts identified by SARpy in the R and E models

Nitrosurea:

Nitrogen or sulfur mustard like:

Benzodioxole and Benzendiol:

α,β-oxy and carboxy substitutions:

Teritiary amine substituted by a Sulfur atom:


α,β-haloalkanes:

Oximes:


 ACCEPTED MANUSCRIPT

Table 5. Chemicals structures in the ISSCAN-CGX data set from which structural alert 24 has been extracted

		
O=C1c2ccccc2C(=O)N1SC(Cl)(Cl)Cl	O=C1N(C(=O)C2CC=CCC12)SC(C(Cl)Cl)(Cl)Cl	O=C1N(C(=O)C2CC=CCC12)SC(Cl)(Cl)Cl
		
O1CCN(C(=S)SN2CCOCC2)CC1	O=C(O)c1ccc(cc1)S(=O)(=O)N(CCC)CCC	O=C(O)c1cc(ccc1Cl)S(=O)(=O)N1CC(C)CC(C)C1