Carcinogenicity Prediction of Noncongeneric Chemicals by Augmented Top Priority Fragment Classification.

Mosè Casalegno,[a] Guido Sello[b,*]

[a] Department of Chemistry, Materials, and Chemical Engineering "Giulio Natta", Via Mancinelli 7, I-20131 Milano, Italy. E-mail: mose1.casalegno@polimi.it

[b] Dipartimento di Chimica, Università degli Studi di Milano, via Golgi 19, I-20133 Milano, Italy. E-mail: guido.sello@unimi.it

Abstract.

Carcinogenicity prediction is an important process that can be performed to cut down experimental costs and save animal lives. The current reliability of the results is however disputed. Here, a blind exercise in carcinogenicity category assessment is performed using augmented top priority fragment classification. The procedure analyses the applicability domain of the dataset, allocates in clusters the compounds using a leading molecular fragment, and a similarity measure. The exercise is applied to three compound datasets derived from the Lois Gold Carcinogenic Database. The results, showing good agreement with experimental data, are compared with published ones. A final discussion on our viewpoint on the possibilities that the carcinogenicity modelling of chemical compounds offers is presented.

Keywords. Carcinogen classes; functional groups; molecular fragments; structural alerts; structure-activity relationships; carcinogenicity prediction.

1. INTRODUCTION

Carcinogenicity in humans is the most alarming characteristic of chemicals for citizens and it is

frequently cited as the biggest barrier to the commercialization of chemicals.(Benigni, 2005) Regulatory agencies are deeply involved in assessing compound safety and, in this regard, are making any sensible effort to control that the cancer risk is minimized; in this perspective, several controls are required to companies before a new compound can be introduced into the market.(OECD, 2002; US EPA, 2005) Considering the cost of the experimental tests and the ethical aspects of the animal tests, the possibility to predict the carcinogenicity of compounds is appealing. However, the mechanisms that operate in cancer development are numerous and not yet fully understood.(Benigni and Bossa, 2011; Cimino, 2006) In addition, the variability of the chemical structure is so wide that sometimes it is difficult to locate the moiety that is responsible for the activity. To complicate the problem, it should be mentioned that, although the presence of a particular moiety can be defined as the responsible for a harmful action, it is much more difficult to determine the molecular part that can partly or fully prevent the cancer development.(Maurici et al., 2005)

Chemicals' carcinogenicity prediction has been discussed and studied for long time.(Guyton et al., 2009; Benigni et al., 2007) Several reports present both quantitative and qualitative models with results at variable level.(Fjodorova et al., 2010; Zhong et al., 2013; Patlewicz et al. 2003; Helguera et al., 2005a; Helguera et al., 2005b; Helguera et al., 2006; Passerini, 2003) Referring to a recent review by Benigni and Bossa (Benigni and Bossa, 2011) it appears clear that the modelling of carcinogenicity is difficult because the problem is complex and not well understood. Experiments are often scarce in number and scope, thus limiting a consistent rationalization. Benigni and Bossa (2011) clearly show that: a) the characterization of experimental mechanisms of action of chemicals in cancer promotion is inadequate; b) the relation of structure characteristics with cancer Mode of Action (MOA) is puzzling; c) the current models, though using approximate theoretical principles, are comparable to experimental determinations. The last sentence does not imply that the problem of predicting compound carcinogenicity is solved, it only highlights that the still limited models

available give predictions that are correct at the same level as the experimental determinations. As already pointed in the previous lines, the understanding of cancer development is still limited and, as a consequence, the developed models still need improvements.

Some recent studies use the common procedure to classify compounds into carcinogens and non carcinogens.(Fjodorova et al., 2010; Zhong et al., 2013) The models therein developed follow the conventional statistical approach: 1) choosing a set of experimental results; 2) calculating several molecular descriptors; 3) dividing the experimental data into a training and a test set; 4) statistically validating the models. The results are very similar, as expected, and their discussion points to the model predictability and to the search for a molecular rationalization of the model descriptors.

In this paper we pursue a different goal: we test the possibility to reach a result comparable with that present in the literature using a classification based exclusively on the molecular structure. In our study we do not use the experimental results to drive the model development; in contrast, we will only check a posteriori if our classification can be used to predict carcinogenicity with the same confidence. In addition, we will briefly discuss the current reliability of our prediction.

Augmented Top Priority Fragments were recently described.(Casalegno and Sello, 2013) They are the result of the combination of two different procedures: the first groups compounds using a mixture of molecular similarity and atomic group composition;(Casalegno et al., 2008) the second analyses and validates the groups using functional groups calculated by means of the electronic energy.(Sello, 1992) The application of this procedure to a set of compounds comprises the following steps: 1) a check of the applicability domain to identify the outliers in the chemical space associated with a set of structural descriptors (i.e. molecular fragments); 2) the exclusion of all the molecules that cannot be inserted in a subset of at least four components; 3) the assignment of the remaining compounds to one or more sets; 4) the selection of the most significant set for each compound. No use of the biological activity is required to perform this procedure.

The aim of the current study is to: a) present the results, that can be used in many applications, in

the context of carcinogenicity prediction; b) discuss the results in comparison with recently published ones; c) eventually, assess the feasibility of the use of models for carcinogenicity prediction.

2. MATERIALS AND METHODS

2.1. Chemical Data

In this study, we considered three datasets. The first set (hereafter called SET1) derived from the Carcinogenic Potency Database (CPDBAS) and was used by Zhong et al. (2013) It contains 852 non congeneric compounds with well-defined chemical structures, including 449 carcinogens and 403 non carcinogens. The second set (hereafter called SET2) contains 802 chemicals, including 420 carcinogens and 382 non carcinogens, which were extracted from the Distributed Structure-Searchable Toxicity (DSSTox) Public Database Network (http://www.epa.gov/ncct/dsstox/sdf_cpdbas.html)(CPDBAS) derived from the Lois Gold Carcinogenic Database (http://potency.berkeley.edu/) (CPDB); the set was used by Fiodorova et al.(2010) The third set (hereafter called SET3) contains 1118 compounds extracted from the Istituto Superiore della Sanità (ISS) database (ISSCANv3a),(Benigni et al., 2008) including 574 carcinogens, 354 non carcinogens and 190 undetermined compounds. It should be noted that the three data sets share several molecules (90% overlap between SET1 and SET2; 83% overlap between SET2 and SET3). The carcinogenicity is referred to studies on rats. The total number of unique compounds is 1359 (of which 887 were analysed, see Table 1). The use of three overlapping datasets allows to highlight the differences that can originate by the application of the model, thus permitting the result comparison and the model study. We stress that no experimental data was used to train our model; as a consequence, all the three sets should be considered external test sets. In contrast with more conventional training/test methodology, our model assigns compounds to subsets ignoring any kind of activity information. Hence, carcinogenicity data are only used with

the purpose to test the model predictive capabilities.

2.2. Method Overview

The method used in the current study has been already described.(Casalegno and Sello, 2013) As a consequence, we are not going to provide here an exhaustive description; however, we would like to provide a qualitative view of its fundamentals. The procedure is based on the application of two methods; the first generates some clusters that contains a selection of molecules;(Casalegno et al., 2011) the second extends and validates the clusters.(Sello, 1992) Cluster generation is based on the mapping of a fragment-based representation onto a cluster-based one.(Casalegno et al., 2008) Using structural fragments (SFs) we generate a chemical space that is used to build the clusters. As in our previous works, we adopted here the Atomic Centered Units (ACUs) as SFs. Clusters are groups of compounds sharing a common SF. Thus, each compound belong to as many clusters as the number of its constituent fragments. This permits the description of a molecule by means of membership in clusters (also referred to as groups); at the same time, a function called affinity is used to determine cluster memberships. At the end of the calculation, groups containing some compounds are formed; a compound can be present in more than one group. Compounds not belonging to any group are collected in a special group, called the outlier group.

In this method the fragments do not have any special reactivity meaning; so, to complete the molecule description we use a second method, developed to define and locate functional groups. The method uses the electronic description of atoms to define their importance and to collect interacting atoms into functional groups (FGs, hereafter).(Sello, 1992)

As described previously (Casalegno and Sello, 2013), the two above strategies can be combined with the aim at grouping compounds sharing similar structures and activity. To this end, the representation provided by the SFs is mapped onto that based on the FGs. This process comprises four sequential steps, namely: 1) SF-FG mapping, 2) cluster merging, 3) cluster selection, and 4)

cluster splitting. For better clarity, hereafter, we quickly resume these steps. SF-FG mapping is carried out by cross checking the atoms that belong to the SF and to the FG: all the SF atoms should belong to one FG, otherwise the molecule is removed from the cluster. This process aims at establishing a one-to-one correspondence between SFs and FGs, for each molecule in a specific cluster. Cluster merging is performed to merge SFs that are closely similar, and reduce the total number of SFs. When two SFs are merged, also the molecules assigned to the corresponding clusters become part of the same cluster. Cluster selection is then performed to select for each compound belonging to many clusters, only the cluster associated with the highest affinity value. It should be noted that the affinity values were obtained during the initial SF-based clustering process. Finally, we inspect each cluster, looking for molecules characterized by FGs of different lengths. Since these FGs may show different reactivities, these molecules are assigned to different sub-clusters within the main cluster (i.e., cluster splitting).

In the following we will often use the terms selectivity, specificity and accuracy with reference to statistical significance of the results. These are defined as:

Selectivity = TP / (TP + FN),

Specificity = TN / (TN + FP)

Accuracy = (TP+TN)/(TP+FP+TN+FN),

where TP, FP, TN, and FN, refer to the number of True Positives, False Positives, True Negatives, and False Negatives, respectively. The total number of experimental carcinogenic compounds is TP + FN, and of non-carcinogenic compounds is TN + FP.

3. RESULTS

Carcinogenicity prediction is an important target because it can reduce both costs and animal experiments. The use of either QSAR or SAR is of great importance because the modelling of scientific instances not only helps the solution of a problem but also requires its rationalization. It is

clear enough that we cannot currently apply QSAR to carcinogenicity prediction with confidence and that the assessment of carcinogenicity categories can be nevertheless an important achievement.(Benigni and Bossa, 2011) However, even this limited objective is not completely established. The question is: can we accept the theoretical categories and use them with confidence? In the following we present the results of the application of our model to the three selected datasets and in the Discussion section we will try to answer the question. As already mentioned the three datasets that we used in our analysis are derived from the same source: the Lois Gold Carcinogenic Database (CPDB). As a consequence, there is some overlap between the sets; more exactly, 90% overlap between SET1 and SET2; 83% overlap between SET2 and SET3, considering the smallest set. We can thus expect similar results in the model application. However, it should be noted that our model generates classes using all the compounds of a set; this has the consequence that classes can be different even using sets with similar member composition. Another difference between SET3 and SET1 and SET2 concerns the compound activity; because both SET1 and SET2 were selected to train Zhong et al. (2013) and Fiodorova et al. (2010) models using the compound activity they cannot include in the analysis compounds whose activity is missing. By contrast, our model has not such limitation and can be also applied to SET3 that contains compounds whose activity has not been measured.

3.1. Applicability Domain

The first outcome of the application of the model is the selection of those compounds that are outside the applicability domain. These compounds exclusively contain SFs that either are singly represented or are found in too many molecules (thus being useless for the selection procedure). The applicability domain is strictly dependent on the set composition; this fact is evident looking at the detected outliers. The number of outliers is 85, 88, and 96, for SET1, SET2, and SET3,

respectively. Even in the presence of an important set overlap the outlier overlap is nevertheless quite limited: 33 common compounds between SET1 and SET2, 54 common compounds between SET1 and SET3, 27 common compounds between SET2 and SET3. This result is explained by the rigid rule used by the procedure to select the outliers that asks for a complete uniqueness of the compound. It is thus sufficient that two structure share a common fragment to create a group.

Some examples of outliers are reported in Figure 1.

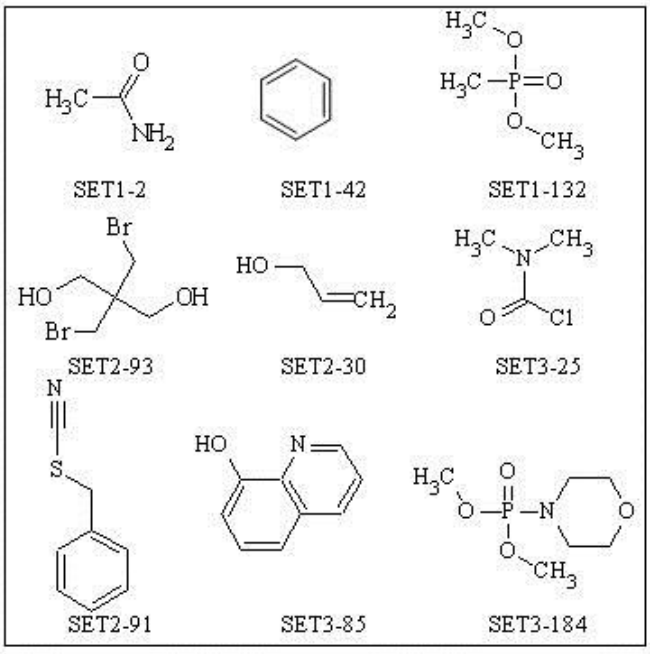


Figure 1. Examples of outliers detected in SET1, SET2 and SET3.

Table 1. Dataset composition

Dataset	Compounds	Positive	Negative	Outliers	Rare	Analysed	SF
SET1	852	449	403	85	151	616	72
SET2	802	421	381	88	187	527	58
SET3	1118	574[a]	354	96	213	809	74

[a] In SET3 there are 190 compounds whose activity has not been experimentally determined

In these examples it easy to see the presence of very small and simple compounds (SET1-2, SET1-132, SET2-30, SET3-25), of uncommon compounds (SET2-93, SET3-184), and of

compounds containing only ordinary and unusual SFs (SET1-42, SET2-91, SET3-85).

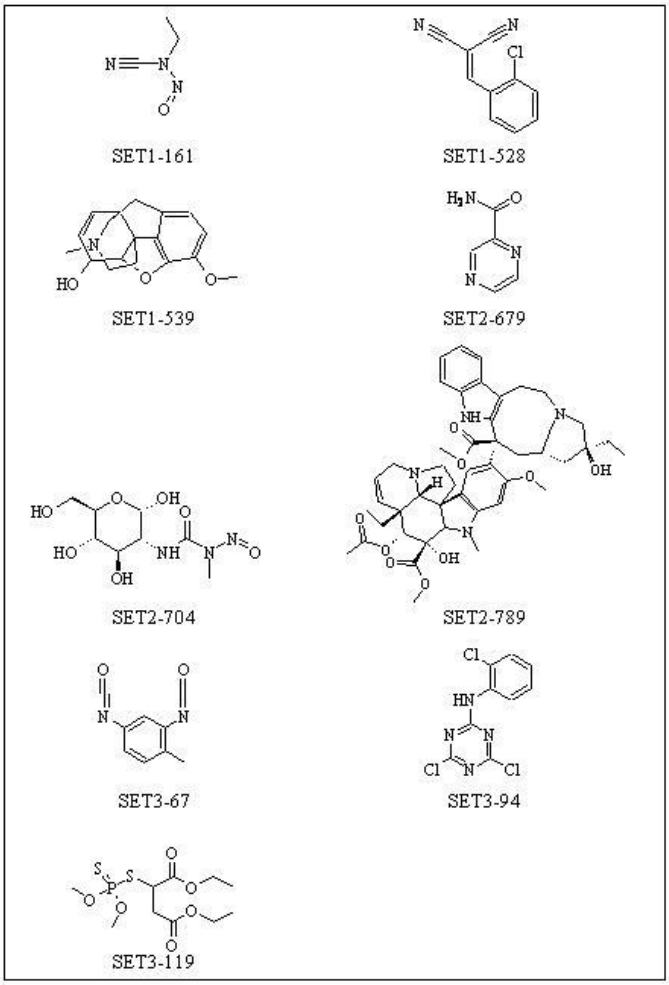


Figure 2. Examples of compounds excluded from SET1, SET2, and SET3, because members of too small subsets.

The consequence of the formation of very small groups is considered by the successive group formation procedure. In fact, groups formed by less than four molecules are not considered. This generates a second subset (for each main set) of compounds that are not outliers but are not considered significant in the analysis of the compound activity. These subgroups contain 151 compounds for SET1, 187 compounds for SET2, 213 compounds for SET3. The total number of excluded molecules is thus 236, 275, and 309, for SET1, SET2, and SET3, respectively; the analysed compounds are therefore 616, 527, and 809. In addition, in SET3 there are 125 compounds

that have not an assigned experimental activity (Table 1). Some examples of excluded compounds are reported in Figure 2. In this case, the eliminated compounds usually contain some unusual SFs in a complex structure.

At the end of this section a comment is necessary. It is sufficient the preliminary analysis performed by our model to highlight that the three datasets cannot be used without considering their composition. The applicability domain strongly influences the reliability of the results; the use in a model of compounds that are not well inside the domain should be considered with care, in particular when the dataset is highly varied. There is sometimes a problem with the definition of the applicability domain; in fact, using continuous variables as descriptors the domain can be easily defined by the variable ranges. In contrast, using structural descriptors the domain is defined by the substructure chemical space. In this case, the domain definition is not straightforward because it depends on the current interpretation of the chemical space. In our model two attributes are checked: the presence of the molecule fragments inside the space and the possibility of assigning the molecule to a group of similar compounds.

3.2. Group Activity Assessment

Our procedure divides compounds into subsets using a reference SF and a variable called affinity that measures the similarity of a compound to the other group components. The number of the subsets depends on two principal factors: the number of compounds and the variance of the set. All the examined datasets generate subsets containing less than 10 members on average, even after the exclusion of outliers and rare compounds. Obviously, there are subsets with more members (20-50) and subsets with less members (4, that is the minimum to have a subset). In this situation it is clear that the subsets have also a different reliability. It should be also considered that the same SF can generate more than one subgroup because the affinity of the compounds can function as an

additional distinctive feature.

As mentioned the subsets are formed without taking into account the activity data. When the division is terminated we assign to each subset a reference activity based on the most frequent activity in the subset; in case of equal activity distribution the positive attribute is assigned (i.e. we privileged the false positive result). To help understanding this point we can consider a subset containing 10 members; if more than 4 members (i.e. between 5 and 10) are active than the subset is defined active. As reported in Table 2, it is now possible to calculate the usual variables, accuracy, specificity, and selectivity, in order to verify the reliability of the predicted activities. Of course, the contribution of each subset to the total variables is weighted against the subset dimension. Concerning the accuracy, i.e. the overall correct prediction, the result is very similar near the 75% for all the sets. This was expected because the three sets are overlapping; however, the presence in SET3 of a consistent number of undetermined compounds could have implied a different result.

The result is different when we examine the selectivity and the specificity. In fact, whilst SET1 and SET2 show similar values, SET3 has a high selectivity and a low specificity. This outcome can be explained by two factors: the above mentioned choice towards false positive assignments and the presence of the undetermined compounds that allows the formation of sets containing some compounds that do not contribute to the activity determination. In addition, the ratio between active and inactive compounds is greater in SET3.

Table 2. Significant statistical variables concerning the dataset analysis.

Dataset	Positive	Negative	FP	FN	Total errors	Accuracy	Selectivity	Specificity
SET1	325	291	81	61	142	76.9	81.2	72.2
SET2	276	251	77	58	135	74.4	79.0	69.3
SET3	405[a]	279	144	47	191	72.1	88.4	48.4

[a] In SET3 there are 125 compounds whose activity has not been experimentally determined

3.3. Subset Analysis

The datasets used in the analysis contain compounds with very different structures. As a consequence, the identification of the moiety responsible for the activity is not always straightforward. Nevertheless, in the following we will report the main known group located by the procedure.

3.3.1. Nitro furans

This moiety (Figure 3) is present in all the datasets. It is mainly found in carcinogenic compounds; nitro aromatic moieties are considered structure alerts by the Toxtree program.(Toxtree) Looking at the subsets formed by the procedure we observe that:

In SET1 there is a subset containing 19 nitro furan derivatives, 18 of them are carcinogenic; the only exception is reported in Figure 4a. In addition compound SET1-692 is considered an outlier and compound SET1-264 is not analyzed (figure 4b). A second subset contains 4 more nitro furans that are all carcinogenic. Three other nitro furans (SET1-136, SET1-187 and SET1-196) are inserted in different subsets, because they also contain other moieties.

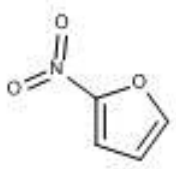


Figure 3. Nitro furan moiety.

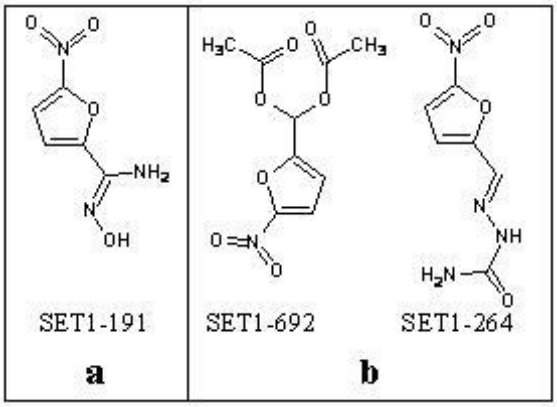
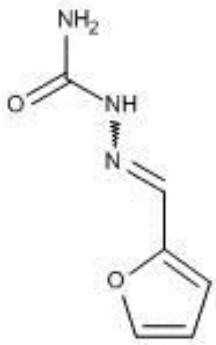


Figure 4. a) SET1-691 is a non carcinogenic compound. b) SET1-692 is an outlier and SET1-264 is not analyzed because its affinity is too low.

In SET2 there is a subset containing 16 compounds, all of them are carcinogenic. A second subset contains 6 compounds, 5 of them are also nitro furans. In the subset there is also the compound shown in Figure 5, that is not toxic; however, Toxtree (Toxtree) classifies the compound as carcinogenic. Three nitro furans cannot be inserted in any subset and are not classified. Finally, three more compounds are classified in different subsets.



SET2-338

Figure 5. SET2-338 is a non carcinogenic compound.

In SET3 there is a subset containing 18 compounds, all of them are carcinogenic. A second subset contains 12 compounds, one is not toxic (the same present in SET1, Figure 4a); one compound is not a nitro furan (the same present in SET2, Figure 5).

Nitro furans are thus recognized as highly toxic moieties and very often the procedure correctly classifies them. The few exceptions are easily detected when a different classification is possible; whilst the wrong predictions cannot be explained.

3.3.2. N-nitroso compounds

The N-NO moiety is another well known carcinogenic alert. In the datasets many compounds containing this group are present; thus we can expect that they will be divided into some subsets.

In SET1 118 compounds contain the N-NO group. They are representative of several molecular structures, as evidenced by their distribution into six main subsets containing 51, 6, 7, 5, 13, and 8, compounds; the remaining 28 compounds are: 2 outliers, 8 not analyzed, 18 in other groups. The main subset (51 compounds) contains 43 active and 8 inactive compounds (all the inactive compounds are found active by the Toxtree program). Considering all the classified compounds we observe 74 actives and 16 inactives. The error percentage is thus 17.8, showing that even in this particularly well defined case the error number can be significant. The 10 unassigned compounds have either few atoms or quite special structure; three examples are shown in Figure 6. Finally, 11 compounds of the 18 assigned to other groups have a correct activity assignment.

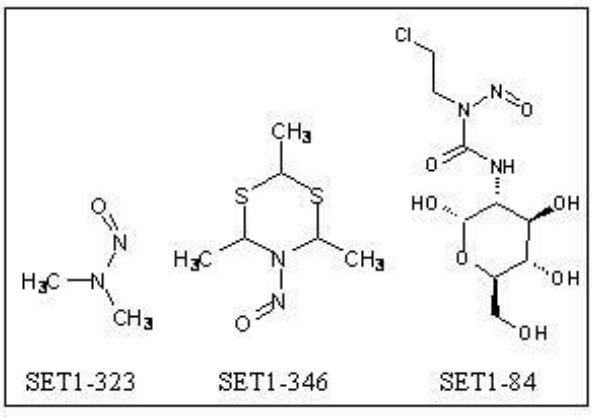


Figure 6. Unassigned compounds of SET1.

In SET2 104 compounds contain the N-NO group. They are distributed into five main subsets

containing 45, 12, 5, 6, and 7, compounds; the remaining 29 compounds are: 2 outliers, 6 not analyzed, 21 in other groups. The main subset (45 compounds) contains 37 active and 8 inactive compounds. Considering all the classified compounds we observe 59 actives and 16 inactives. The error percentage is thus 21.3.

In SET3 120 compounds contain the N-NO group. They are distributed into four main subsets containing 57, 12, 8, and 4, compounds; the remaining 39 compounds are: 2 outliers, 12 not analyzed, 25 in other groups. The main subset (57 compounds) contains 48 active, 8 inactive, and 1 undetermined compounds. Considering all the classified compounds we observe 69 actives, 11 inactives, and 1 undetermined. The error percentage is thus 13.8.

The N-NO moiety is certainly a reliable structural alerts; however, its presence is not sufficient to predict compound carcinogenicity. It could be interesting to analyze if there is some clear structural reason that detoxifies the group presence. However, looking at the structures of the 11 inactives present in SET3 (as an example, Figure 7) the reason for their lack of reactivity is not always clear; e.g., some can be considered water soluble (e.g. containing a carboxylic residue), others geometrically hindered.

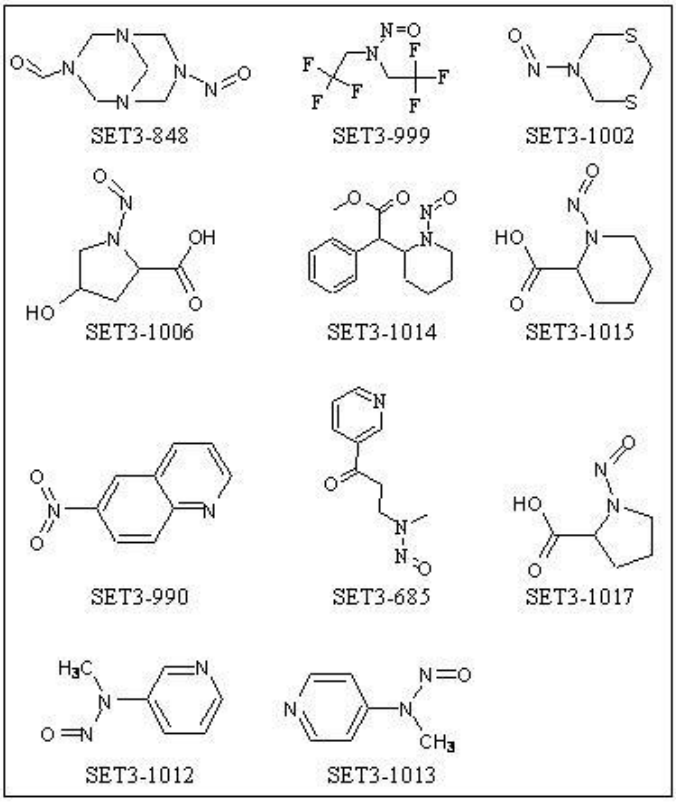


Figure 7. Non carcinogenic compounds in SET3 containing the N-NO moiety.

3.3.3. Nitro benzenes and alkyl halides

Aromatic compounds containing a nitro group are commonly considered carcinogenic and the group is considered a structural alert. However, this is not always the case. Analyzing a subgroup of this class in SET1, containing 21 compounds, we found 9 actives and 12 inactives. The structures are reported in Figure 8a and 8b. A subgroup of SET2 contains 18 nitro benzenes, of which 7 are actives and 11 are inactives. Similarly, the 42 nitro derivatives found in a subgroup of SET3 were classified as actives (19), inactives (17), and undetermined (6). It is almost impossible to find any reason to discriminate between the two molecule categories.

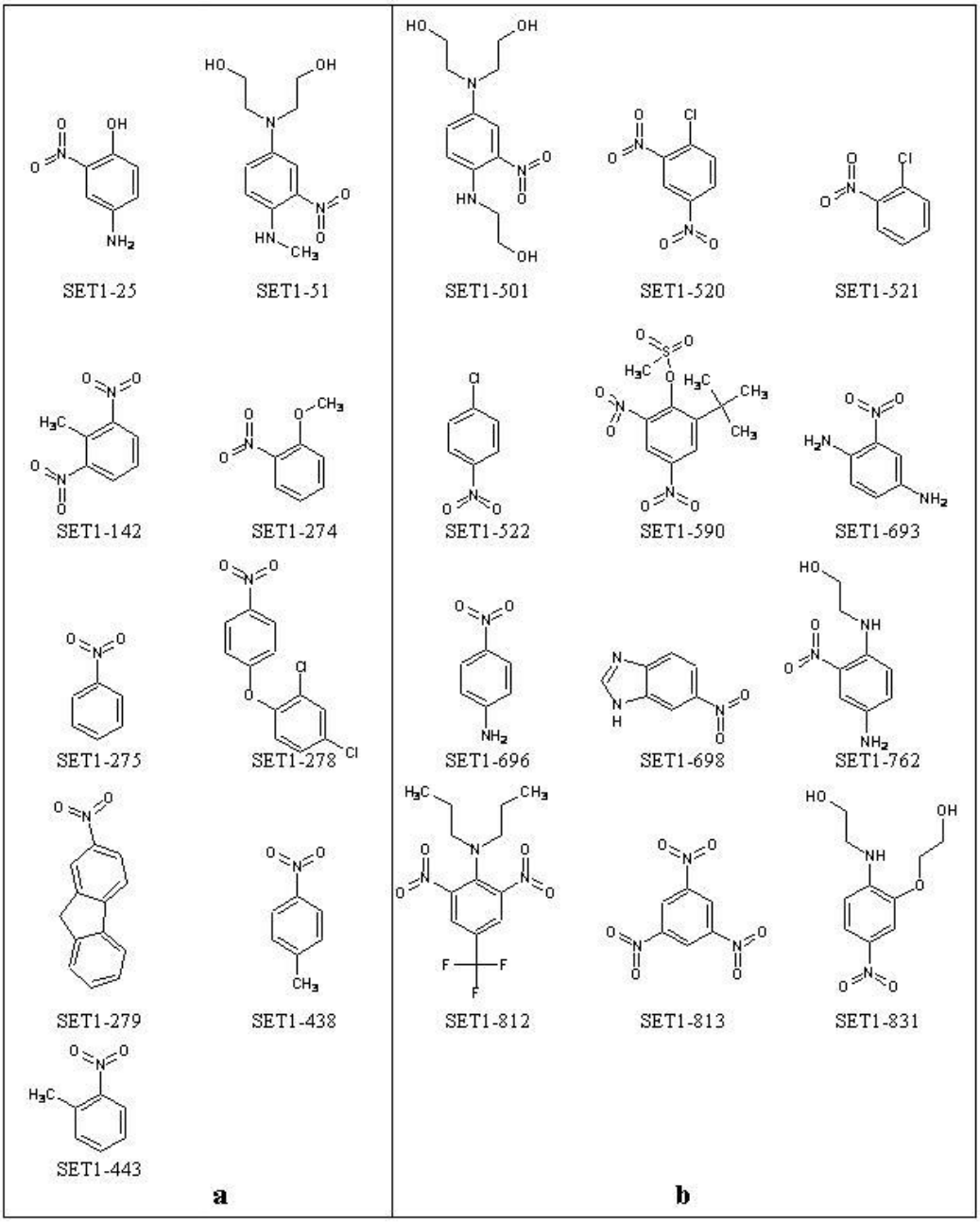


Figure 8. a) Carcinogenic compounds containing the aromatic NO_2 moiety. b) Non carcinogenic compounds containing the aromatic NO_2 moiety.

Also alkyl halides are commonly recognized as structural alerts; however, our analysis reveals that the percentage of compounds of this class that are classified non carcinogenic is not far from 50. Six examples (3 active, upper, and 3 inactive, lower) are reported in Figure 9.

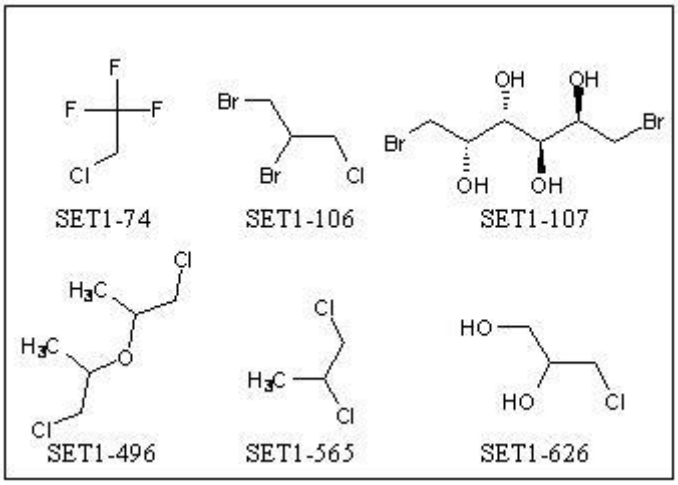


Figure 9. Examples of carcinogenic (upper) and non carcinogenic (lower) compounds containing the alkyl halide moiety.

3.3.4. Primary aromatic amines (anilines)

Another well known and deeply studied functional group whose presence is related to the carcinogenicity of chemical compounds is represented by anilines. Also in this case we selected one subgroup in each set to present the analysis result. In SET1 the subgroup contains 21 compounds with 13 actives and 8 inactives. In SET2 the subgroup contains 25 molecules with 17 actives and 8 inactives. In SET3 the subgroup contains 22 compounds with 12 actives, 4 inactives, and 6 undetermined. Again the reason of the different experimental results is not evident looking at the structures. Some examples are reported in Figure 10a (actives) and 10b (inactives).

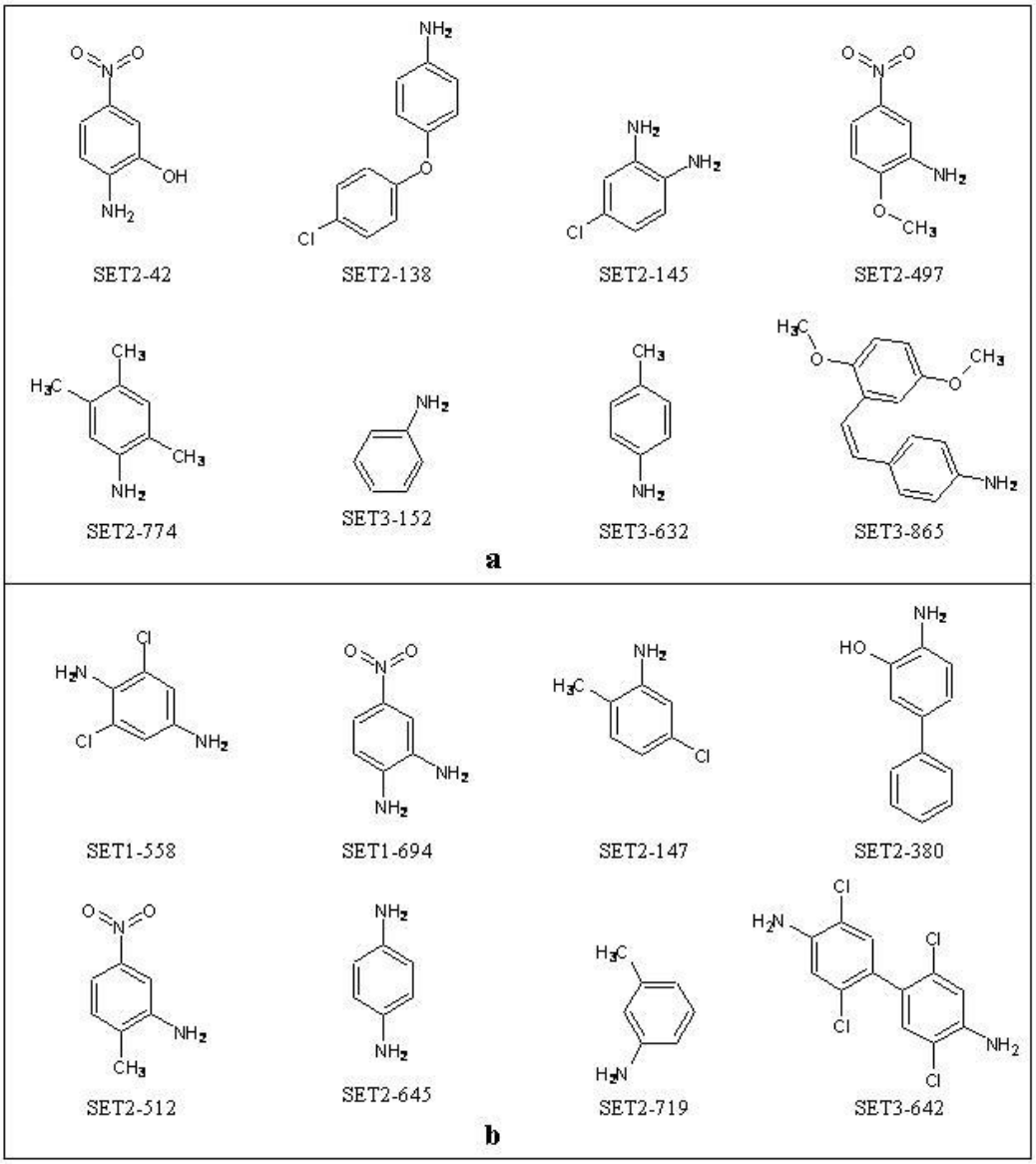


Figure 10. a) Carcinogenic compounds containing the aromatic NH_2 moiety. b) Non carcinogenic compounds containing the aromatic NH_2 moiety.

4. DISCUSSION

Cancer is still one disease of great concern for the developed world population. Life style and genetic heritage are certainly important factors for cancer development; however, at least part of the current occurrences can be associated to chemical toxicity. It is thus important to assert the hazard of compounds before, during, and after production and commercialization. In this perspective the possibility of using theoretical models can be of high interest for both economic and ethical reasons;

but it is important to assess the reliability level of the predictions to avoid excessive expectation followed by rapid disappointment.(Benigni and Bossa, 2011)

Cancer development is a complex and not well understood mechanism; its modelling is therefore not well settled yet. Consequently, QSAR models are still missing (excluding few examples with congeneric compound set)(Passerini, L., 2003; Benigni, 2003) and SAR models are only few. Since the seventies the notion of structural alerts is the main qualitative scheme to help in carcinogenicity prediction, giving suggestions on the compound toxicity.(Miller and Miller, 1977; Miller and Miller, 1981a; Miller and Miller, 1981b) Structural alerts introduce the idea of biological mechanisms driven by the presence of specific substructures; in addition, they can be used to divide compounds into classes. Their importance is well recognized by regulatory agencies that are currently asking for compound classification as a practical mean to assign compound toxicity.

In their deep and comprehensive recent review Benigni and Bossa (2011) discuss the state of the art and the reliability of the cancer prediction through models, highlighting the role of structural alerts and the contribution of mechanistic studies. Here, the attention is pointed to the importance that small substructures can have in driving specific interactions with cell elements; undoubtedly, the use of structural alerts had a noteworthy success in biological activity prediction.(Ashby and Tennant, 1988; Benigni and Bossa, 2006) A theoretically different approach uses the hypothesis that sufficiently similar moieties should have similar biological actions; as a consequence, molecular similarity is used to find the common structural characteristics.(Willett, 2006) Also in this case the straightforward use of the model takes to the formation of compound classes, that are also more well defined than those formed using alerts. Comparing the two approaches we can see that, whilst in the first approach the presence of a small part of a molecule could be overestimated, in the second approach the similarity between two compounds can be found even neglecting the presence of exactly that small part that is fundamental for the biological action. In addition, neither approach is explicitly forming subsets.

This work is a former attempt to demonstrate that compound carcinogenicity can be predicted by means of a fully blind analysis, with an accuracy comparable to that reached by models trained on experimental data. Our proposal focuses on the combination of the alert presence together with a significant molecular similarity. This way, the procedure is going to form subsets that implicitly comply with both requirements. The analysis of the applicability domain of the set is part of the procedure; the model analyses the set of compounds searching for outliers and for compounds that do not comply with the method rules. The identification of these compounds leads to their exclusion from the subsequent analysis. The procedure is also general and applicable to any molecular set, generating subsets that can be then used to assess any molecular property: It is the property that originates from the starting dataset that determines the application of the results.

Two recent studies on carcinogenicity prediction (by Zhong et al., 2013; and by Fiodorova et al., 2010) use very similar datasets to classify compounds: carcinogenic and non carcinogenic. Both models use a standard statistical approach where, through the activity information taken from the literature, the model is trained and optimized; then, one or more test sets are used for the validation step. The obtained result is sufficiently similar: high accuracy for the training sets (>80% and >90%) and good accuracy for the test sets (>70% and >65%). Also selectivity and specificity are notable, as can be appreciated looking at Table 3. The result interpretation is also discussed to some extent, mainly by Zhong et al., controlling the presence/absence of structural alerts, or of compound subgroups.

Table 3. Significant statistical variables concerning the dataset SET1 and SET2 from literature data (Fjodorova et al., 2010; Zhong et al., 2013).

Dataset	Accuracy	Selectivity	Specificity
Training SET1	82.3	n.a.	n.a.
Test SET1	72.0	72.4	71.3
Training SET2	98.9	87.5	92.2

Test SET2	65.2	70.7	68.3

Our results on SET1 and SET2 were obtained without partitioning the full sets in training and test subsets. Nonetheless, a comparison with the results above referenced can be informative. The statistical significance of our results is comparable to that reported in Table 3 (accuracy: 76.9, 74.4; selectivity: 81.2, 79.0; specificity: 72.2, 69.3; for SET1 and SET2, respectively). Some other benefits emerge from the application of our approach: 1) some compounds that are supposed to hinder the analysis are automatically excluded; 2) the reliability of the subgroups can be assessed from the number of their members; 3) the presence of possible structural alerts can be directly evidenced looking at the SF of the subgroup, or it can be verified checking the subgroup in a following action.

The main achievement is however the possibility to find subgroups of inactive compounds. Because the inactivity is not determined by the presence/absence of alerts, it is always very difficult to search for inactive compounds. Our procedure groups together those compounds that are similar and that possess all the characteristics to be inactive. This result is possible even if the compounds contain an alert. This is a fundamental aspect in activity prediction: whilst the activity is triggered by the presence of a particular moiety, the lack of activity is always the result of the contribution of several different features. Even the concept of detoxifying groups is not sufficient, because it does not consider other compound attributes like, for example, molecular dimension or solubility.

In the Result section the presence of special SF in some subgroups is presented. The unbiased formation of subgroups permits to easily uncover both expected and unexpected outcomes. For example, the well known toxicity of nitro furans is clear. In contrast, the presence of both nitro benzenes and alkyl halides, commonly considered toxic substructures, is questioned by our result.

It is now the time to answer to our initial question: is it reasonable to assess compound carcinogenicity via in-silico models ? Since cancer represents a major social and economical issue,

also in relation with the its financial and psychological impact, we believe we should positively answer to this question.

Unfortunately, the availability of reliable experimental data and the understanding of the cancer mechanisms are still insufficient to guarantee a meaningful quantification of the risk. It remains the possibility of achieving a qualitative compound classification. If we consider the statistical significance of model application to test compounds (accuracy always greater than 70%), our answer to the initial question can be affirmative. At least, in the limited target of qualifying compounds as potentially carcinogenic we can accept the application of models. However, great attention should be paid to the result analysis; we still have to live in a significant uncertainty. Whilst the predicted toxicity of a compound can be often accepted, the prediction of the absence of toxicity is uncertain; we should carefully check all the facets of the prediction report before accepting the result. The applicability domain is fundamental; an accurate control of the relation between the structures in the dataset and the membership of the query compound should be done. It is sometimes inadequate to check the density of the domain and the position of the query; we should also check the presence/absence of special features in the target structure and/or in the dataset. In case of a classification, it is necessary to control the composition of each class and to give a reliability index to the class of concern.

Eventually, if an uncertainty is still present, it is preferable to reject the prediction.

5. CONCLUSIONS

We presented this exercise with the aim at carefully testing the possibility of predicting carcinogenicity. In this perspective, we chose one of the most popular dataset (CPDB) as the source of three partitions for the exercise; we studied their applicability domains; we blindly applied a categorization procedure; we compared our result to two of the most recent examples. The final outcome does not fully support the implementation of theoretical models to solve the task.

Nevertheless, we believe that the development of the modelling activity should be pursued with conviction because this rationalization effort helps in understanding the experimental data and in making clearer the reliability of compound classification. In the future we expect more exciting results from both the experimental and the theoretical sides.

6. REFERENCES

Ashby, J. and 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./Genetic Toxicol. 204, 17-115.

Benigni, R., 2003. SARs and QSARs of mutagens and carcinogens: understanding action mechanisms and improving risk assessment, in: Benigni, R. (Ed.), Quantitative structure-activity relationship (QSAR) models of mutagens and carcinogens, CRC Press, Boca Raton, pp 259–282.

Benigni, R., 2005. Structure-activity relationship studies of chemical mutagens and carcinogens: mechanistic investigations and prediction approaches. Chem. Rev. 105, 1767–1800.

Benigni, R. and Bossa, C., 2006. Structural alerts of mutagens and carcinogens. Curr. Comput.-Aided Drug Des. 2, 169−176.

Benigni, R. and Bossa, C., 2011. Mechanisms of Chemical Carcinogenicity and Mutagenicity: A Review with Implications for Predictive Toxicology. Chem. Rev. 111, 2507–2536.

Benigni, R., Bossa, C., Netzeva, T., Worth, A., 2007. Collection and evaluation of (Q)SAR Models for Mutagenicity and Carcinogenicity. European Commission Directorate General Joint Research Centre 2007 EUR 22772EN © European Communities.

Benigni, R., Bossa, C., Richard, A.M., Yang, C., 2008. Ann. Ist. Super. Sanità, 44, 48. ISSCAN is a curated database on chemical carcinogens freely available at http://www.iss.it/ampp/dati/cont.php?id=233&lang=1&tipo=7 (accessed October 2012).

Casalegno, M. and Sello, G., 2013. Determination of Toxicant Mode of Action by Augmented Top Priority Fragment Class. J. Chem. Inf. Model. 53, 1113–1126.

Casalegno, M., Sello, G., Benfenati, E., 2008. Definition and Detection of Outliers in Chemical Space. J. Chem. Inf. Model. 48, 1592−1601.

Casalegno, M., Benfenati, E., Sello, G., 2011. Identification of Toxifying and Detoxifying Moieties for Mutagenicity Prediction by Priority Assessment. J. Chem. Inf. Model. 51, 1564−1574.

Cimino, M.C., 2006. Comparative overview of current international strategies and guidelines for genetic toxicology testing for regulatory purposes. Environ. Mol. Mutagen. 47, 362–390.

CPDB. http://potency.berkeley.edu/

CPDBAS: Carcinogenic Potency Database Summary Tables, http://www.epa.gov/ncct/dsstox/sdf_cpdbas.html.

Fjodorova, N., Vracko, M., Tušar, M., Jezierska, A., Novic, M., Kühne, R., Schüürmann, G., 2010. Quantitative and qualitative models for carcinogenicity prediction for non-congeneric chemicals using CP ANN method for regulatory uses. Mol. Divers. 14, 581–594.

Guyton, K.Z., Kyle, A.D., Aubrecht, J., Cogliano, V.J., Eastmond, D.A., Jackson, M., Keshava, N., Sandy, M.S., Sonawane, B., Zhang, L., Waters, M.D., Smith, M., 2009. Improving prediction of chemical carcinogenicity by considering multiple mechanisms and applying toxicogenomic approaches. Mutation Research 681, 230–240.

Helguera, A.M., Perez, M.A.C., Combes, R.D., González, M.P., 2005a. The prediction of carcinogenicity from molecular structure. Curr. Comput. Aided Drug Des. 1, 237–255.

Helguera, A.M., Perez, M.A.C., González, M.P., Ruiz, R.M., Gonzalez-Diaz, H., 2005b. A topological substructural approach applied to the computational prediction of rodent carcinogenicity. Bioorg. Med. Chem. 13, 2477–2488.

Helguera, A.M., Perez, M.A.C., Combes, R.D., González, M.P., 2006. Quantitative

structure-activity relationships for the computational prediction of nitrocompounds carcinogenicity. Toxicology 220, 51–62.

Maurici, D., Aardema, M., Corvi, R., Kleber, M., Krul, C., Laurent, C., Loprieno, N., Pasanen, M., Pfuhler, S., Phillips, B., Prentice, D., Sabbioni, E., Sanner, T., Vanparys, P., 2005. Carcinogenicity. ATLA 33, Suppl. 1, 177–182.

Miller, J.A. and Miller, E.C., 1977. Origins of Human Cancer; Cold Spring Harbor Laboratory: Cold Spring Harbor, p 605.

Miller, E.C. and Miller, J.A., 1981a. Cancer Searches for ultimate chemical carcinogens and their reactions with cellular macromolecules. Cancer, 47, 2327–2345.

Miller, E.C. and Miller, J.A., 1981b. Mechanisms of chemical carcinogenesis. Cancer 47, 1055–1064.

OECD, 2002. Guidance notes for analysis and evaluation of chronic toxicity and carcinogenicity studies. OECD Environment, Health and Safety Publications Series on Testing and Assessment No. 35, Paris, France.

Passerini, L., 2003. QSARs for individual classes of chemical mutagens and carcinogens, in: Benigni, R. (Ed.), Quantitative structure-activity relationship (QSARs). Models of mutagens and carcinogens, CRC Press, Boca Raton, pp 81–123.

Patlewicz, G., Rodford, R., Walker, J.D., 2003. Quantitative structure-activity relationships for predicting mutagenicity and carcinogenicity. Environ. Toxicol. Chem. 22, 1885–1893.

Sello, G., 1992. A New Definition of Functional Groups and a General Procedure for Their Identification in Organic Structures. J. Am. Chem. Soc. 114, 3306−3311.

Toxtree (Estimation of Toxic Hazard - A Decision Tree Approach), Ideaconsult Ltd., Version 2.5.0. http://toxtree.sourceforge.net (accessed July 2013).

US EPA, 2005. Guidelines for carcinogen risk assessment and supplemental guidance for assessing susceptibility from early-life exposure to carcinogens. Federal Register 70, 17765–17817.

Willett, P., 2006. Similarity-based virtual screening using 2D fingerprints. Drug Discovery Today, 11, 1046−1053.

Zhong, M., Nie, X., Yan, A., Yuan, Q., 2013. Carcinogenicity Prediction of Noncongeneric Chemicals by a Support Vector Machine. Chem. Res. Toxicol. 26, 741−749.