Computational Biology and Chemistry 61 (2016) 145–154

Contents lists available at ScienceDirect

Computational Biology and Chemistry

journal homepage: www.elsevier.com/locate/compbiolchem

Research article

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
[b] Dipartimento di Chimica, Università degli Studi di Milano, via Golgi 19, I-20133 Milano, Italy

ARTICLE INFO

Article history:
Received 5 March 2015
Received in revised form 15 December 2015
Accepted 26 January 2016
Available online 29 January 2016

Keywords:
Carcinogen classes
Functional groups
Molecular fragments
Structural alerts
Structure–activity relationships
Carcinogenicity prediction

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.

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,b, 2006; Passerini, 2003). Referring to a recent review by 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)

* Corresponding author.
E-mail addresses: mose1.casalegno@polimi.it (M. Casalegno), guido.sello@unimi.it (G. Sello).

http://dx.doi.org/10.1016/j.compbiolchem.2016.01.011
1476-9271/

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 Fjodorova 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.

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.

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).