Computational Toxicology 3 (2017) 33–43

Contents lists available at ScienceDirect

Computational Toxicology

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

Classification of carcinogenic and mutagenic properties using machine learning method

N.S Hari Narayana Moorthy [a,*,1], Surendra Kumar [b], Vasanthanathan Poongavanam [c,*]

[a] Departamento de Química e Bioquímica, Faculdade de Ciências, Universidade do Porto, s/n, Rua do Campo Alegre, 4169-007 Porto, Portugal
[b] Department of Pharmaceutical Chemistry, Faculty of Pharmacy, Babu Banarasi Das Northern India Institute of Technology, Lucknow, India
[c] Department of Physics, Chemistry and Pharmacy, University of Southern Denmark, DK-5230 Odense M, Denmark

ARTICLE INFO

Article history:
Received 17 March 2017
Received in revised form 23 June 2017
Accepted 5 July 2017
Available online 6 July 2017

Keywords:
Carcinogenicity
Machine learning
Fingerprints
Random forest
SRD

ABSTRACT

An accurate calculation of carcinogenicity of chemicals became a serious challenge for the health assessment authority around the globe because of not only increased cost for experiments but also various ethical issues exist using animal models. In this study, we provide machine learning-based classification models for the carcinogenicity and mutagenicity. The carcinogenic and mutagenic information of 1481 chemically diverse molecules in various species (e.g. dog, hamster, rat, single-cell and multi-cell) has been used for classification models, and these models include random forest method using physicochemical descriptors and structural fingerprints. In addition, the sum of ranking difference (SRD) method has been used to rank the developed models. The best models based on the random forest approach correctly classify more than 70% of compounds in the test set. Furthermore, the MACCS fingerprints were utilized to understand the structural features of the chemicals that cause mutagenicity or carcinogenicity. The results obtained from these studies along with the qualitative models could potentially be employed to screen a large number of chemicals for carcinogenicity and mutagenicity assessment.

Introduction

Recent studies suggest that majority of the drug candidate failure in the drug discovery project was due to poor pharmacokinetic properties, particularly toxicity which may vary from drug-drug to xenobiotic-gene interactions, for example, interaction with the biomolecules like DNA. In the latter interaction, xenobiotics could potentially be a drug or any environmental components, which affects the normal cellular or genetic functions and causes cancer or related malignant tumors. These carcinogens are categorized as genotoxic and non-genotoxic that based on their effect on the genetic materials [1–4]. A larger number of chemicals in the environment are prone to the cause of carcinogenic effects [5]. According to the new implementation of REACH (European's Registration, Evaluation, Authorization, and Registration of Chemicals Directive), the toxicological information will be submitted for the registration or authorization of new chemicals. As part of the approval process, the risk to human and environments is evaluated for all substances. The identification of chemical mutagens and carcinogens are of high priority within the EU and other countries. However, the evaluation of toxicity especially carcinogenicity in animal tests (rodent bioassays) is very laborious, costly, and require a significant number of animals for these experiments [6,7].

The prediction of toxicities (carcinogenicity, mutagenicity, and other toxicities such as skin-sensitisation) is not only necessary for chemical regulatory purposes but also essential in the drug discovery process. The primary reason for early prediction of carcinogenicity and related properties is to find and eliminate compounds with poor toxicological endpoints. With the availability of advanced computational techniques and software, the rapid prediction of toxicity has increased significantly in recent years especially, (quantitative) structure-activity relationships (QSARs), classification analysis and knowledge-based expert systems [4]. These approaches identify the key molecular functionalities (or structural features) that are known to cause these toxicities [8]. In this context, *in silico* techniques are efficient and accurate enough to be used to predict the hazards effect (toxicity) of many compounds [9].

In recent years, several studies have been published on the prediction of carcinogenicity using the computational methods, for instances, Pereira and Schmitz have reported support vector

* Corresponding authors.
E-mail addresses: hari.nmoorthy@gmail.com (N.S Hari Narayana Moorthy), nobelvasanth@gmail.com (V. Poongavanam).
[1] Present Address: Department of Pharmacy, Indira Gandhi National Tribal University, Amarkantak 484887, (MP), India.

http://dx.doi.org/10.1016/j.comtox.2017.07.002
2468-1113/

machine (SVM) models using the pharmacophore-based fingerprints on relatively a large dataset consisting of 1547 compounds obtained from the carcinogenic potency database (CPDB) [10]. Another research group has reported the counter propagation artificial neural network (CP-ANN) models using a set of 805 chemicals from the CPDB with an overall accuracy of ~69–73% in test set [11,12]. Similarly, Zhong et al. also have reported carcinogenicity models using the SVM method (accuracy of 71.96%) based on the extended connectivity fingerprint (ECFP) using a set of 852 chemicals obtained from the CPDB database [13]. Furthermore, a dataset of ~1500 chemicals with their carcinogenicity data obtained from the IARC (International Agency for Research on Cancer), EU (European Union), EPA (US Environmental Protection Agency), NTP, ACGIH (American Conference of Governmental Industrial Hygienists), and JSOH (Japan Society for Occupational Health) was used for the development of SVM-based classification model and the model showed an overall accuracy of ~70% [5,14,15]. Interestingly, Chen et al. have reported the prediction models based on a large dataset of 17233 chemicals from the STITCH database with an overall accuracy of 79.50% [16]. Recently, Zhang et al. have reported the Naive Bayes-based classification models on 1042 non-congeneric carcinogenic compounds from the CPDB with a prediction accuracy of 68% [17]. Similarly, binary and ternary classification models were also developed on 829 carcinogenic compounds with a predictive accuracy of 83% and 80% in the test set, respectively [18].

Although, several carcinogenic or mutagenic studies particularly based on the machine learning methods have been reported and often SVM is shown to be the best method. However, none of the previous studies have investigated the carcinogenicity data of different species. Hence, in the present study, we have used Random Forest (RF)-based classification method as an alternative to the SVM method, because of its efficient characteristic (i.e. capable of handling thousands of input variables) without overfitting the input data. To this end, six different mutagenic or carcinogenic endpoints (e.g. dog, hamster, rat, single-cell or multi-cell and mutagenic) for approximately 1500 compounds from the DSSTox database (also known as CPDB) [19] have been evaluated and based on the structural analysis, it suggested the key substructural features that are frequently seen in these carcinogenic or mutagenic compounds. We hope that current classification models could be useful as the screening tool for the identification of compounds or chemicals that potentially causes cancers.

Computational methods and materials

Dataset collection and preprocessing

A set of 1481 structurally diverse compounds includes natural products and drugs, was retrieved from the Carcinogenic Potency Database (CPDB) [19] and CPDB often used as a benchmark dataset for carcinogenicity prediction because of its structural diversity and high-quality toxicological endpoints. In the current study, compounds having different carcinogenic information were used, for instance, (1) carcinogenic studies on rat, dog and hamster, (2) carcinogenicity screened in single and multi-cell and (3) chemicals classified as mutagenic or non-mutagenic in the *Salmonella* assay. Categorical carcinogenic activity score of all these types based on the TD_{50} (drugs or chemicals dose at which toxicity occurs in 50% of cases), and activity score 100 (active) and 0 (inactive) are assigned for all screened compounds. Here, single and multi-cell endpoints are based on the minimal and multi-cell evidence for or against activity, respectively. Compounds with more than one TD_{50} or tumor site listed for carcinogenicity experiments on specific species cell (e.g. rat, mouse, hamster etc.) are annotated as active compounds and compounds with no TD_{50} are annotated as inactive (0).

Table 1
Number of active and inactive compounds used in the study.

S. No.	Species	Active	Inactive	Total
1	Dog	13	12	25
2	Hamster	41	31	72
3	Mutagenicity	317	341	658
4	Multi-cell	459	359	818
5	Single-cell	698	483	1121
6	Rat	473	424	897

The 2D coordinates of all compounds with its experimental activities (a binary data consisting of 1 for active (carcinogenic or mutagenic) and 0 for inactive (non-carcinogenic or non-mutagenic)) were extracted and used for further ligand preprocessing. The number of chemicals used for each species in this study is provided in Table 1. To avoid uncertainty in the model building, initially the data set was cleaned by excluding the mixtures, polymers, inorganic compounds, organometallic compounds, salts and chemicals with undefined activity or missing stereochemical information using the ChemAxon tool [20]. The 2D coordinates of the remaining 1370 compounds (528 active and 842 inactive) were imported into the Molecular Operating Environment (MOE) software (Version 2014.13) [21] for 3D conversion followed by energy minimization using the MMFF94x force field (with Generalized Born solvation model). Subsequently, the energy-minimized structures were used to compute molecular descriptors [22], which include 272 physicochemical descriptors, 76 vol Surface descriptors (VolSurf). These descriptors include physical properties, surface area, atom and bond counts, shape indices, pharmacophore descriptors, etc. Furthermore, a set of 166 MACCS fingerprints was also calculated using the PaDEL software [23]. The MACCS structural keys are particularly useful to analyze substructure of a large dataset and often used to explore the structural pattern. Finally, a descriptor matrix (x-variables) was merged with the binary experimental activity (y-variable) for the model developments.

Subsequently, the descriptor pool was reduced by eliminating those descriptors which possessed either 0 or no variance in the values. The best possible descriptors for the classification analysis were selected through principal component analysis (PCA) and stepwise feature selection procedure as implemented in the STATISTICA software (Version 12) [24]. For model building, the dataset was further divided into training (2/3) and test set (1/3) for classification using the WEKA (Version 3.6.4) (random number generator option) with a seed value 1 and care was taken to avoid imbalance between active and inactive ratios in the dataset. In addition, the software package SIMCA (version 11.0; Umetrics, Umeå, Sweden) was used for multivariate analysis, for instance, the dataset was characterized using the PCA method. A list of selected descriptors used for PCA is provided in the Supporting Information (SI-Table 1). A VIF value >10 is an indication of potential multicollinearity problems (inflated standard errors of regression coefficients), and in this analysis, all the models have the VIF values <2.5. Hence, the descriptors involved in these models do not possess any serious multicollinearity problem.

Machine learning method

The choice of machine learning method used in this study is Random Forest (RF) method. Briefly, the RF method was developed by Breiman [25] as an extension of Decision tree method [26]. The idea behind, it provides a high predictive ability by averaging the predictions of a large number of individual decision trees. This