Food and Chemical Toxicology 97 (2016) 141–149

Contents lists available at ScienceDirect

Food and Chemical Toxicology

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

Novel naïve Bayes classification models for predicting the carcinogenicity of chemicals

Hui Zhang [a, c, *], Zhi-Xing Cao [b, c], Meng Li [a], Yu-Zhi Li [b], Cheng Peng [b]

[a] College of Life Science, Northwest Normal University, Lanzhou, Gansu, 730070, PR China
[b] Pharmacy College, Chengdu University of Traditional Chinese Medicine, Key Laboratory of Systematic Research, Development and Utilization of Chinese Medicine Resources in Sichuan Province-key Laboratory Breeding Base of Co-founded by Sichuan Province and MOST, Chendu, Sichuan, PR China
[c] State Key Laboratory of Biotherapy and Cancer Center, West China Hospital, West China Medical School, Sichuan University, Chengdu, Sichuan, 610041, PR China

ARTICLE INFO

Article history:
Received 18 January 2016
Received in revised form
2 August 2016
Accepted 1 September 2016
Available online 3 September 2016

Keywords:
Carcinogenicity
In silico prediction
Naïve Bayes classifier
Molecular descriptors
Extended connectivity fingerprints
(ECFP_14)

ABSTRACT

The carcinogenicity prediction has become a significant issue for the pharmaceutical industry. The purpose of this investigation was to develop a novel prediction model of carcinogenicity of chemicals by using a naïve Bayes classifier. The established model was validated by the internal 5-fold cross validation and external test set. The naïve Bayes classifier gave an average overall prediction accuracy of 90 ± 0.8% for the training set and 68 ± 1.9% for the external test set. Moreover, five simple molecular descriptors (e.g., AlogP, Molecular weight (M_W), No. of H donors, Apol and Wiener) considered as important for the carcinogenicity of chemicals were identified, and some substructures related to the carcinogenicity were achieved. Thus, we hope the established naïve Bayes prediction model could be applied to filter early-stage molecules for this potential carcinogenicity adverse effect; and the identified five simple molecular descriptors and substructures of carcinogens would give a better understanding of the carcinogenicity of chemicals, and further provide guidance for medicinal chemists in the design of new candidate drugs and lead optimization, ultimately reducing the attrition rate in later stages of drug development.

© 2016 Published by Elsevier Ltd.

1. Introduction

Toxicity of drugs, as a significant issue for the pharmaceutical industry, most frequently lead to increased attrition and cost, late-stage failures and even market withdrawals (Lasser et al., 2002; Paul et al., 2010; Segall and Barber, 2014). In order to reduce attrition of drug candidates as a result of adverse drug reactions (ADRs), extensive studies, including chemical structure, genetic, biological systems and clinical perspectives, have been applied in the drug discovery process (Kennedy, 1997). Carcinogenicity is among the toxicological endpoints that pose the highest public concern (Fjodorova et al., 2010a). Any substance that damages the genome or disrupts the cellular metabolic processes might induce tumors, increase tumor incidence, or short the time to tumor occurrence, called carcinogen (Fjodorova et al., 2010a, 2010b). Presently, various factors, such as lifestyle, diet, smoking, pollution, indoor and workplace exposure, are found to have the potential risk of cancer causation, and exposure to chemicals is an essential carcinogen (Belpomme et al., 2007; Benigni and Bossa, 2011; Marone et al., 2014). In order to reduce the risk of chemical induction of carcinogenicity in drug development, different strategies, ranging from biochemical investigations, to the use of assay systems, to *in silico* methodologies, have been extensively implemented throughout the pharmaceutical industry (Marone et al., 2014). However, owing to the experimental approaches for the carcinogenicity testing of chemicals is very expensive, time consuming, and even unethical, the *in silico* methodologies for predicting the carcinogenicity of chemicals has become a research focus in recent years.

Computational techniques for the prediction of toxicity is a rapidly growing field, and the major driving force for which is the implementation of the European Union REACH (Registration, Evaluation and Authorization of Chemicals) legislation (Lagunin et al., 2009) and ICH (International Conference on Harmonization) guideline (Jena et al., 2005a,b). These regulations explicitly encourage the use and development of alternative methods, such as quantitative structure-activity relationships (QSARs), for *in vivo*

[*] Corresponding author. College of Life Science, Northwest Normal University, Lanzhou, Gansu, 730070, PR China.
E-mail address: zhanghui123gansu@163.com (H. Zhang).

http://dx.doi.org/10.1016/j.fct.2016.09.005
0278-6915/© 2016 Published by Elsevier Ltd.

toxicological assessment (FDA, 2015; REACH, 2011). For example, The ICH M7 guidance (FDA, 2015) suggests the computational toxicology assessment should be performed using two complementary QSAR methodologies: “expert rule-based” and “statistically-based”, and points out the QSAR models utilizing these prediction methodologies should adhere to the general validation principles set forth by the Organization for Economic Cooperation and Development (OECD, 2007). Presently, many computational prediction approaches for the carcinogenicity of chemicals have been reported (Benigni and Zito, 2003; Contrera et al., 2003, 2007; Fjodorova et al., 2010a; Helguera et al., 2005; Hulzebos et al., 2005; Kar and Roy, 2011; Klopman et al., 2004; Lagunin et al., 2005; Matthews and Contrera, 1998; Singh et al., 2013; Tan et al., 2009; Valerio et al., 2007; Woo and Lai, 2005; Zhu et al., 2008; Zhong et al., 2013). Among these prediction approaches, the QSARs methods have been broadly applied in prediction of carcinogenicity of chemicals. The QSARs attempt to find relationships between the chemical structure (structural and physicochemical features) with an endpoint (e.g., toxic effect) using a statistically derived mathematical equations (Dearden, 2016; Roy et al., 2015). For example, Fjodorova et al. (2010a) performed a classification model of chemicals of carcinogenic potency based on 27 two-dimensional MDL descriptors and propagation artificial neural network (CP ANN) technique, which gave an accuracy of 92% for the training set and 68% for the test set. Zhong et al. (2013) constructed support vector machine (SVM) model with using 24 molecular descriptors, and the prediction model gave over 80% for the test set. Singh et al. (2013) extracted 834 structurally diverse chemicals of rat data to construct probabilistic neural network (PNN) prediction model using five descriptors, and gave classification accuracy of 92.09% in complete rat data. Although these machine learning methods could give satisfactory accuracies in the forecast of carcinogenicity, the end-points of previous studies were considered as receptorial or in general fine-tuned process. Moreover, we found a large number of molecular descriptors were used in previous researches, which may limit the ability to interpret the mechanisms of carcinogenicity. In this research, the naïve Bayes classifier was considered to assess the carcinogenicity of chemical compounds. The naïve Bayes classification model employs the versatile machine learning algorithms based on the Bayes' theorem with the conditional independence assumptions (Box and Tiao, 2011; Berger, 2013), in which each variable can be independently estimated as a one dimensional variable. Because of the conditional independence assumption is rarely true in real-world applications, the naïve Bayes classifier often performs surprisingly well.

The objective of this study is to build a novel prediction model to discriminate chemicals as carcinogens and non-carcinogens with using naïve Bayes method, and identify some important molecular descriptors and substructures of carcinogenic compounds. The established prediction models will be validated by the internal 5-fold cross validation and external test set. We hope the established naïve Bayes prediction model of the carcinogenicity of chemicals could be applied to filter early-stage molecules for this potential carcinogenicity adverse effect. Furthermore, the identified simple molecular descriptors and substructures of carcinogens would give a better understanding of the carcinogenicity of chemicals, and provide guidance for medicinal chemists in the design of new candidate drugs and lead optimization, ultimately reducing the attrition rate in later stages of drug development.

2. Materials and methods

2.1. Dataset

Presently, animal experiments are the major source of information on chemical carcinogens, and several on-line database of rodent carcinogenicity are available, such as the Carcinogenic Potency Database (CPDB) (http://potency.berkeley.edu/cpdb.html), the US National Toxicology Program (NTP) database (http://ntp-apps.niehs.nih.gov/ntp_tox/index.cfm), Istituto Superiore di Sanita, Chemical Carcinogens: “Structures and Experimental Data” (ISSCAN) (http://www.epa.gov/ncct/dsstox/sdf_isscan_external.html), and Pesticides Action Network (PAN) database (http://www.pesticideinfo.org). Among these databases, the Carcinogenic Potency Database (CPDB), containing a large diversity of chemical structures (1547 substances), is considered as a single standardized resource of information on many chronic long term bioassays (Singh et al., 2013). The rat driven data of carcinogenicity is considered more suitable for human carcinogenicity prediction than those of other rodents (Huff et al., 1991; Huff, 1999; Fung et al., 1995). Thus, in this study, we only considered the rat data of carcinogenicity that extracted from the Carcinogenic Potency Database (CPDB). After deletion of some inorganic compounds and complex compounds, 1042 compounds of rat carcinogenicity, including 506 carcinogens (positives) and 536 non-carcinogens (negatives), were remained. These selected compounds were then randomly separated into five equal-sized subsets. Of the five subsets, four subsets were used as training set (834 compounds, 80% of the data), and the remaining one subset was employed for the test set (208 compounds, 20% of the data) (Table 1). This process was repeated five times in such a way that each subset was used exactly once as the external test set. Finally, five datasets (Dataset 1–5) were obtained.

2.2. Molecular descriptors

All the molecular descriptors were calculated by Discovery Studio 3.1 software (http://accelrys.com/products/discovery-studio/). In this investigation, seventeen descriptors that widely used in the ADME/T prediction were selected (Wang et al., 2012; Hou and Wang, 2008; Zhang et al., 2015, 2016). The descriptors include the number of N atom, the number of O atom, ALogP, Apol, logD, molecular solubility, molecular weight, the number of aromatic rings, the number of H acceptors, the number of H donors, the number of rings, the number of rotatable bonds, molecular fractional polar surface area, molecular polar surface area, molecular surface area, Wiener and Zagreb.

2.3. ECFP_14 fingerprints

The extended-connectivity fingerprints (ECFPs), a class of topological fingerprints for molecular characterization, are derived using a variant of the Morgan algorithm (Morgan, 1965). The ECFPs are designed to capture molecular features relevant to molecular activity, and recently applied in substructure searching, drug activity predicting, similarity searching, clustering, and virtual screening (Rogers and Hahn, 2010). In this study, the ECFP_14 fingerprints were used to analyze the structure features of carcinogenic/non-carcinogenic compounds.

Table 1
The number of compounds used in each of the training set and test set.

	Training set	Test set	Total
Carcinogenic agents	405	101	506
Non-carcinogenic agents	429	107	536
Total	834	208	1042

Download English Version:

https://daneshyari.com/en/article/2584832

Download Persian Version:

https://daneshyari.com/article/2584832

Daneshyari.com