Food and Chemical Toxicology 110 (2017) 122–129

Contents lists available at ScienceDirect

Food and Chemical Toxicology

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

Development of novel prediction model for drug-induced mitochondrial toxicity by using naïve Bayes classifier method

Hui Zhang[a,b,*], Peng Yu[a], Ji-Xia Ren[b,c], Xi-Bo Li[a], He-Li Wang[a], Lan Ding[a,**], Wei-Bao Kong[a]

[a] College of Life Science, Northwest Normal University, Lanzhou, Gansu 730070, PR China
[b] State Key Laboratory of Biotherapy and Cancer Center, West China Hospital, West China Medical School, Sichuan University, Chengdu, Sichuan 610041, PR China
[c] College of Life Science, Liaocheng University, Liaocheng, Shandong 252059, PR China

ARTICLE INFO

Keywords:
Mitochondrial toxicity
Naïve Bayes classifier
Recursive partitioning classifier
Molecular descriptors
Extended connectivity fingerprints (ECFP_6)

ABSTRACT

Mitochondrial dysfunction has been considered as an important contributing factor in the etiology of drug-induced organ toxicity, and even plays an important role in the pathogenesis of some diseases. The objective of this investigation was to develop a novel prediction model of drug-induced mitochondrial toxicity by using a naïve Bayes classifier. For comparison, the recursive partitioning classifier prediction model was also constructed. Among these methods, the prediction performance of naïve Bayes classifier established here showed best, which yielded average overall prediction accuracies for the internal 5-fold cross validation of the training set and external test set were 95 ± 0.6% and 81 ± 1.1%, respectively. In addition, four important molecular descriptors and some representative substructures of toxicants produced by ECFP_6 fingerprints were identified. We hope the established naïve Bayes prediction model can be employed for the mitochondrial toxicity assessment, and these obtained important information of mitochondrial toxicants can provide guidance for medicinal chemists working in drug discovery and lead optimization.

1. Introduction

Mitochondria are essential organelles and play a wide range of diverse functions in the cell, such as energy production, cellular respiration, regulation of intracellular homeostasis, reduction/oxidation balance, proliferation, and apoptosis (Malty et al., 2015; Nunnari and Suomalainen, 2012; Vlasblom et al., 2014). Mitochondrial impairment has been considered as an important contributing factor in the etiology of drug-induced organ toxicity, such as hepatotoxicity (Liu et al., 2016), cardiotoxicity (Varga et al., 2015), and acute kidney injury (AKI) (Wallace and Starkov, 2000). Moreover, mitochondrial dysfunction has also been considered a potential unifying factor in the pathogenesis of some diseases, including cancer (Guaragnella et al., 2014), neurodegenerative disorders (Bhat et al., 2015; Moon and Paek, 2015), obesity (Bournat and Brown, 2010; Nasrallah and Horvath, 2014) and diabetes (Sivitz and Yorek, 2010). Presently, numerous environmental chemicals and drugs have been reported to impair mitochondrial function (Chan et al., 2005; Dykens and Will, 2007; Dykens et al., 2007; Hynes et al., 2013; Wills et al., 2013). For example, the Food and Drug Administration has issued 528 Black Box Warnings for patients through 2010, and approximately 35% of those were associated with mitochondrial toxicity (O'Connor, 2010). Notably, some drugs, such as troglitazone and cerivastatin, were withdrawn from the market due to they directly/indirectly impair mitochondrial function. (Golomb and Evans, 2008; Kaufmann et al., 2006; Okuda et al., 2010; Rachek et al., 2009). Mitochondrial toxicity has been considered as a major cause for preclinical therapeutic candidate failures as well as post market drug withdrawals (Dykens and Will, 2007). Thus, establishing a simple, fast and high-throughput screening approach alternative to experimental mitochondrial toxicity evaluation has become an important and urgent task in the early stages of drug development, which would avoid candidate failures, allowing resources to be focused on those compounds with the highest chance of success to the market.

Computational techniques for hazard identification and risk assessment have become a research focus in recent years. Developing and using alternative approaches to the experimental toxicological assessment has been explicitly encouraged by some legislations (OECD; REACH, 2011). Presently, various computational methods for organ-level toxicities have been extensively developed and reported. However, the reports of *in silico* prediction methods for drug-induced mitochondrial toxicity were very few. For example, to date, only Zhang et al. (2009) developed a SVM prediction model based on 27 molecular

* Corresponding author. College of Life Science, Northwest Normal University, Lanzhou, Gansu 730070, PR China.
** Corresponding author.
E-mail addresses: zhanghui123gansu@163.com (H. Zhang), dinglan@nwnu.edu.cn (L. Ding).

http://dx.doi.org/10.1016/j.fct.2017.10.021
Received 1 September 2017; Received in revised form 10 October 2017; Accepted 13 October 2017
Available online 16 October 2017
0278-6915/

descriptors, which gave concordance of 84.59% for the training set, and 77.14% for external test set. Naven et al. (2013) developed structure-activity relationships for mitochondrial dysfunction, and identified 11 toxicophores associated with the mechanism of potent uncoupling activity. Nelms et al. (2015) developed an *in silico* profiler based on the different mechanistic information, and obtained mechanism based eight structural alerts for mitochondrial toxicity. All of these suggested that creating and developing new computational method for mitochondrial toxicity prediction with an reasonable accuracy is important and necessary. Thus, in this investigation, the naïve Bayes classifier approach was considered to assess drug-induced mitochondrial toxicity. The naïve Bayes classification model based on the Bayes' theorem with the conditional independence assumptions, in which each variable can be independently estimated as a one dimensional variable (Berger, 2013; Box and Tiao, 2011). Presently, the naïve Bayes classifier has been widely applied for the ADMET related properties prediction and drug-likeness analysis (Chen et al., 2011; Li et al., 2014; Shi et al., 2015; Tian et al., 2012, 2014; Wang et al., 2016; Zhang et al., 2016a, 2017).

The aim of this investigation is to build a novel prediction model of drug-induced mitochondrial toxicity with using naïve Bayes method, and identify some important molecular descriptors and substructures of mitochondrial toxicants. 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 mitochondrial toxicity could be applied to filter early-stage molecules for this potential adverse effect. Furthermore, the identified important molecular descriptors and substructures of toxicants would give a better understanding of the mitochondrial toxicants, 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. Mitochondrial toxicity dataset collection

The dataset, containing 288 agents, was extracted from Zhang et al. (2009). Within this dataset, 171 chemicals were designed as mitochondrial toxicants according to the literature reported, and 117 agents were considered as negatives according to FDA-approved drug. In this investigation, five compounds were deleted because of four compounds were duplicate, and one compound was not found. Finally, the 283 compounds, containing 168 mitochondrial toxicants and 115 non-toxicants, were remained. The chemical structures and categories of these compounds have been given in the supplementary material (Table S1). Then, these selected compounds were randomly separated into five equal-sized subsets. Of the five subsets, four subsets were used as training set (226 compounds, 80% of the data), and the remaining one subset was employed for the test set (57 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. Calculations of molecular descriptors and fingerprints

All the molecular descriptors were calculated by Discovery Studio 3.5 software (http://accelrys.com/products/discovery-studio/). In this investigation, 23 descriptors that widely applied in the ADME/T prediction were selected (Hou and Wang, 2008; Wang et al., 2012; Zhang et al., 2015). The descriptors include the number of C atom, number of H atom, number of O atom, ALogP, Apol, logD, molecular solubility, molecular weight, number of rings, number of rings 6, number of aromatic rings, number of H acceptors, number of H donors, number of rotatable bonds, molecular fractional polar SASA, molecular fractional polar surface area, molecular polar SASA, molecular polar surface area, molecular SASA, molecular SAVol, molecular surface area, Wiener and Zagreb.

In addition, the topological fingerprints 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). Previous researches have proved the topological fingerprints significantly influence the prediction performance of naïve Bayes classifier (Zhang et al., 2016a, 2017). In this investigation, the ECFP_6 fingerprints (extended connectivity fingerprints (ECFPs)) were used to analyze the structural features of toxic/non-toxic compounds because of it could give the highest prediction accuracy.

Table 1
The number of compounds applied in each of the training sets and test sets.

	training set	test set	sum
toxicants	134	34	168
non-toxicants	92	23	115
total	226	57	283

2.3. Naïve Bayes (NB) classifiers

The naive Bayesian categorization technique was used to develop the classifiers to discriminate between toxicants and non-toxicants. The naïve Bayes classifier is a probabilistic classification method based on Bayes's theorem with strong independence assumptions between the features. The mathematical details of naïve Bayes classifier were described previously (Berger, 2013; Box and Tiao, 2011; Zhang et al., 2016b). In this investigation, the naïve Bayes classifiers were developed by using Discovery Studio (DS) version 3.5 (http://accelrys.com/products/discovery-studio/). The cross validation of the training set was assigned to 5. In addition, selection of the number of bins appeared in the histogram, which was used to divide the entire range of observed values for the variable into a series of intervals, and then count how many values fall into each interval (Shimazaki and Shinomoto, 2007). The bin size critically influences the performance of naïve Bayes model. In this study, the parameter of “number of Bins” was changed from 10 to 300 systematically in order to find a better naïve Bayes classifier model. The “Learn Options” was selected as Validate Models, Ignore uninformative Bins and Equipopulate Bins. The ECFP_6 was picked as “Model Domain Fingerprint".

2.4. Recursive partitioning (RP) classifier

As a statistical method for multivariable analysis, the recursive partitioning (RP) creates a decision tree that divides data by using a hierarchical set of yes/no questions per node to split a data set into smaller subsets (Cook and Goldman, 1984). The splitting process continues until no more significant nodes are obtained or when a minimum number of samples per node is reached. In the present investigation, the recursive partitioning classifiers were built using the Discovery Studio 3.5 software package. The 5-fold cross-validation was used, the split method was set as Gini index, the weighting method was defied as Uniform, and the minimum number of samples at each node was used as 10 to avoid excessive partitioning. In addition, the maximum tree depth was changed from 2 to 30 systematically in order to find a better RP model.

2.5. Validating the prediction accuracies of the classification models

The following parameters were used to assess the predictive performance of the classification models: the overall prediction accuracy (Q(Eq. (1))); sensitivity (SE (Eq. (2))), the prediction accuracy for the toxicants); specificity, (SP (Eq. (3))), the prediction accuracy for the non-toxicants); positive predictive value (PPV (Eq. (4))); negative