Neurocomputing 145 (2014) 68–74

Contents lists available at ScienceDirect

Neurocomputing

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

Interpretable prediction of non-genotoxic hepatocarcinogenic chemicals

Chun-Wei Tung [a,b,c,*], Jhao-Liang Jheng [b]

[a] School of Pharmacy, Kaohsiung Medical University, Kaohsiung 80708, Taiwan
[b] PhD Program in Toxicology, Kaohsiung Medical University, Kaohsiung 80708, Taiwan
[c] National Environmental Health Research Center, National Health Research Institutes, Miaoli County 35053, Taiwan

ARTICLE INFO

Article history:
Received 14 December 2013
Received in revised form
28 April 2014
Accepted 6 May 2014
Available online 26 June 2014

Keywords:
Non-genotoxic hepatocarcinogenicity
Decision tree
Chemical–chemical interaction
Chemical–protein interaction
Quantitative structure–activity relationship
Toxicogenomics

ABSTRACT

The assessment of non-genotoxic hepatocarcinogenicity of chemicals relies on time-consuming rodent bioassays. The development of alternative methods for non-genotoxic hepatocarcinogenicity could help the identification of potential hepatocarcinogenic chemicals. This study evaluated four types of features for the interpretable prediction of non-genotoxic hepatocarcinogenic chemicals including chemical–chemical interactions (CCI), chemical–protein interactions (CPI), chemical descriptors (QSAR) and gene expression profiles (TGx). Based on the results of decision tree classifiers, the CPI-based features perform best with independent test accuracies of 90% and 86% for interaction scores from combined scores and databases, respectively. Informative features were identified and analyzed to give insights into the non-genotoxic hepatocarcinogenicity of chemicals. The difference between CPI scores and gene expression profiles for the identified important proteins shows that CPI could play more important roles in non-genotoxic hepatocarcinogenicity.

1. Introduction

Chemical carcinogenesis can be classified into two main categories of genotoxic (mutagenic) and non-genotoxic (non-mutagenic) agents according to the mechanism of action [1,2]. The evaluation of chemical carcinogenesis is important for both the drug safety and regulation of environmental chemicals. Several short-term *in vitro* and *in vivo* assays have been developed to assess genotoxic agents by measuring DNA damage, mutagenic effects, and chromosomal aberrations [3]. However, due to the complex nature of non-genotoxic agents, the assessment of non-genotoxic hepatocarcinogenicity of chemical compounds is based on 2-year rodent bioassays that are labor-intensive, time-consuming and expensive. There are only 1500 chemicals studied by National Toxicology Program during the past 30 years [4]. It is desirable to develop alternative methods to efficiently prioritize potential non-genotoxic hepatocarcinogenicity of chemicals for further studies.

Numerous computational models have been developed to predict various toxicity endpoints with reasonably good prediction performance. For example, the quantitative structure–activity relationship (QSAR) models have been extensively used to analyze and predict carcinogenicity [5–8]. QSAR models aiming to correlate chemical structure information and toxicity endpoints could provide useful information of important structures for toxicity alerts. However, the application of QSAR models for predicting non-genotoxic hepatocarcinogenicity yields a low accuracy of 55% [9] showing the high complexity of non-genotoxic hepatocarcinogenicity.

Recently, toxicogenomics (TGx) correlating gene expression profiles and toxicity endpoints has emerged as important alternative methods. With the power of machine learning methods, TGx performs well in non-genotoxic hepatocarcinogenicity with a test accuracy of 80% [9,10]. In contrast to traditional 2-year rodent bioassays, TGx methods require less experimental effort. Generally, published TGx methods select 29–120 genes as important biomarkers and require short-term experiments with 5–28 days [9,11,10]. However, the classification methods for the above-mentioned methods are not interpretable. For practical uses, the number of biomarkers is expected to be as small as possible. It is desirable to develop interpretable classification methods with a high accuracy and a small number of features. Also, TGx methods utilize only gene expression information without incorporating protein-binding effects that are important mechanisms of non-genotoxic carcinogenicity.

The chemical–chemical (CCI) and chemical–protein interaction (CPI) information grows very fast in recent years. Benefit from the development of CCI and CPI databases, enormous interaction data obtained from databases, experiments and text-mining can be easily accessed from the structured databases of STITCH [12–14].

[*] Corresponding author at: School of Pharmacy, Kaohsiung Medical University, Kaohsiung 80708, Taiwan. Tel.: +886 73121101x2316.
E-mail addresses: cwtung@kmu.edu.tw, cwtung@livemail.tw (C.-W. Tung).
URL: http://cwtung.kmu.edu.tw (C.-W. Tung).

http://dx.doi.org/10.1016/j.neucom.2014.05.073
0925-2312/© 2014 Elsevier B.V. All rights reserved.

It provides a great opportunity to study the performances of CCI and CPI for analyzing and predicting non-genotoxic hepatocarcinogenicity. To study the effects of chemical–protein interactions on non-genotoxic hepatocarcinogenicity, we proposed a CPI-based method to identify protein biomarkers with interpretable rules for predicting non-genotoxic hepatocarcinogenic chemicals [15]. A protein biomarker ABCC3 was identified as a potential biomarker for further exploration with an accuracy of 86% outperforms the state-of-the-art methods of gene expression profile-based toxicogenomics using 90 gene biomarkers [15].

Moreover, the CCI feature was recently utilized to predict chemical toxicity effects based on the assumption that interactive chemicals are more likely to share similar toxicity [16]. The CCI-based feature was also successfully applied to the prediction of effective drug combinations [17]. The combination of CCI and CPI features has been applied to predict drug side effects [18]. Four features can be generated based on scores from databases, experiments, text-mining and combined scores for each of the CCI and CPI features.

In this study, 12 features of 4 CCI, 4 CPI, 1 QSAR and 3 TGx features were proposed to predict non-genotoxic hepatocarcinogenicity using interpretable decision tree classifiers. Important features and associated rules were analyzed to give insights into the mechanism of non-genotoxic hepatocarcinogenicity. The CPI-database feature performs well in both 5-fold cross-validation and independent test accuracies of 82% and 86%, respectively. This study stressed out that the importance of chemical–protein interaction effects on the prediction of non-genotoxic hepatocarcinogenicity of chemicals.

2. Materials and methods

2.1. Dataset

In order to demonstrate and compare prediction performances of various features of the four feature types including CCI, CPI, QSAR and TGx, this study utilized a dataset developed by Liu et al. [9] consisting of 62 chemicals with publicly available gene expression profiles in rat. There are 13 positive chemicals with non-genotoxic hepatocarcinogenicity and 49 negative chemicals without non-genotoxic hepatocarcinogenicity. The 62 chemicals are divided into a training dataset and an independent test dataset according to the previous study [9]. The training and independent test datasets consisting of 8 positive and 32 negative chemicals and 5 positive and 17 negative chemicals are utilized for training and testing models, respectively.

2.2. Chemical–chemical and chemical–protein interactions

Chemical–chemical (CCI) and chemical–protein interaction (CPI) data are obtained from STITCH 3.1 database [13,14,12]. STITCH database is an aggregated database of interactions connecting over 300,000 chemicals and 2.6 million proteins from 1133 organisms [19]. The interaction data are obtained from three major sources of experiments, databases and text-mining. The experiment part consists of direct chemical–chemical and chemical–protein interacting data with experimental evidences. The database part contains interaction data from pathway databases. The text-mining data is obtained by extracting information of interactions from literatures using text-mining techniques. Likelihood or relevance scores of interactions are available for each evidence type. An overall score for a given chemical–protein interaction is generated by combining the three scores of corresponding evidence types that is available at STITCH. For chemical–chemical interaction, the combined score is calculated from the three scores and a similarity score available at STITCH. The score is divided by 1000 and is ranging from 0 (no interaction) to 1 (strong interaction). Chemical–protein interactions are transferred between species based on the sequence similarity of the proteins [19].

2.3. Chemical descriptors

To generate chemical descriptors, chemical 2D structures were firstly extracted from PubChem database. Subsequently, PaDEL-Descriptor [20], a software for calculating molecular descriptors and fingerprints, was utilized to calculate 770 1D and 2D descriptors and PubChem fingerprints. The calculation of descriptors and fingerprints is mainly based on the Chemistry Development Kit [21] with some additional descriptors and fingerprints including atom type electrotopological state descriptors, McGowan volume, molecular linear free energy relation descriptors, ring counts, count of chemical substructures, binary fingerprints and count of chemical substructures. The final feature vector is a 1610-dimensional vector.

2.4. Gene expression profiles

The gene expression profiles associated with the chemicals were extracted from a publicly available dataset GSE8858 [22] of Gene Expression Omnibus database. The experiments were performed on three layouts of EXP5280X2-584, EXP5280X2-613 and EXP5280X2-648 of GE Healthcare/Amersham Biosciences Code-Link UniSet Rat I Bioarray. Three datasets are available for three time points of 1 day (TGx-1d), 3 days (TGx-3d) and 5 days (TGx-5d). The dose for each chemical was selected according to Liu et al. [9]. For each dataset of TGx-1d, TGx-3d and TGx-5d, there are 10,399 expression values associated with each chemical resulting in a 10,399-dimensional feature vector.

2.5. Decision tree algorithm

Decision tree algorithms capable of generating interpretable rules are widely used in various classification and regression problems such as immunogenic peptides [23], ubiquitylation sites [24], gamma-turn types [25], protein subnuclear localization [26], promoters [27] and esophageal squamous cell carcinoma [28]. In this study, the decision tree method C5.0 is applied to construct decision tree classifiers and derive interpretable rules for classifying non-genotoxic hepatocarcinogenicity. C5.0 is an improved version of C4.5 with smaller trees and less computation time [29]. The implementation of C5.0 used in this study is the R package C50 [30].

The construction of a decision tree is described as follows. First, information gain is utilized to rank features. Second, the top-ranking features are iteratively appended as nodes to split data into subsets. The tree growing process stops when the data subset in each leaf node belongs to the same class. The fully grown tree is prone to over-fit the training data. Therefore, a pruning process is applied to reduce the tree size by replacing a subtree with a leaf node to avoid over-fitting problems. The pruning process is based on a default threshold value of 25% confidence. The samples in the leaf node are the covered samples of the rule. The class label of a leaf node is determined by using a majority rule. The samples with a relative small size in the leaf node are regarded as misclassified samples. The final decision tree can directly generate if-then rules where one leaf node corresponds to one rule.

2.6. Performance measurement

To evaluate classifiers for their prediction performance, the widely used 5-fold cross-validation method is applied. Four