Acquiring Decision Rules for Predicting Ames-Negative Hepatocarcinogens Using Chemical-Chemical Interactions

Chun-Wei Tung

School of Pharmacy, Kaohsiung Medical University, 80708, Taiwan
Ph.D. Program in Toxicology, Kaohsiung Medical University, 80708, Taiwan
National Environmental Health Research Center, National Health Research Institutes, Miaoli County 35053, Taiwan
cwtung@kmu.edu.tw
http://cwtung.kmu.edu.tw

Abstract. Chemical carcinogenicity is an important safety issue for the evaluation of drugs and environmental pollutants. The Ames test is useful for detecting genotoxic hepatocarcinogens. However, the assessment of Ames-negative hepatocarcinogens depends on 2-year rodent bioassays. Alternative methods are desirable for the efficient identification of Ames-negative hepatocarcinogens. This study proposed a decision tree-based method using chemical-chemical interaction information for predicting hepatocarcinogens. It performs much better than that using molecular descriptors with accuracies of 86% and 76% for validation and independent test, respectively. Four important interacting chemicals with interpretable decision rules were identified and analyzed. With the high prediction performances, the acquired decision rules based on chemical-chemical interactions provide a useful prediction method and better understanding of Ames-negative hepatocarcinogens.

Keywords: Ames-Negative Hepatocarcinogens, Decision Tree, Chemical-Chemical Interaction, Interpretable Rule, Toxicology.

1 Introduction

The assessment of carcinogenicity is crucial for drug development that is based on 2-year rodent bioassays. The bioassays are labor-intensive, time-consuming and expensive. Chemical carcinogens can be classified as either genotoxic (mutagenic) or non-genotoxic (non-mutagenic) agents according to the mechanism of action [1]. 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 [2]. Among the assays, the predictivity of Ames test has been extensively studied for carcinogenicity. The Ames test is useful for identifying mutagenic carcinogens with an accuracy of 80% [3,4]. However, 48% of Ames-negative chemicals are carcinogens [5] and additional bioassays do not help in detecting carcinogens from Ames-negative chemicals [6]. It is desirable

M. Comin et al. (Eds.): PRIB 2014, LNBI 8626, pp. 1–9, 2014.

to develop alternative methods for assessing carcinogenicity of Ames-negative chemicals.

A quantitative structure-activity relationship (QSAR) model has been evaluated for its prediction performance of non-genotoxic hepatocarcinogens. However, the accuracy is only slightly better than random (55%) [7]. Recently, toxicogenomics (TGx) correlating gene expression profiles and toxicity endpoints has emerged as important alternative methods. The TGx methods performed well in non-genotoxic hepatocarcinogenicity with a test accuracy of 80% [7,8]. However, gene expression profiles are only available for a small number of chemicals. It is highly expensive to conduct a large-scale TGx study for hepatocarcinogens.

Chemical-protein interaction (CPI) information has been proposed to predict non-genotoxic hepatocarcinogens with a high accuracy of 86% using only one protein biomarker [9]. Notably, both the aforementioned TGx and CPI methods were performed on a small dataset with less than 62 chemicals. Although the CPI information is useful for analyzing and predicting hepatocarcinogens, the information is incomplete that many chemical-protein pairs have not been studied yet. The development of computational methods for a large number of chemicals is desirable.

This study constructed a relatively large dataset consisting of 166 chemicals by extracting information of Ames-negative chemicals and corresponding hepatocarcinogenicity from NCTRlcdb [10]. The more complete chemical-chemical interaction (CCI) information from STITCH database [11] was proposed to predict hepatocarcinogens based on the assumption that interactive chemicals are more likely to share similar toxicity. The CCI information has been successfully applied to predict various chemical activities such as cancer drugs and chemical toxicity [12,13].

In order to acquire rule-based knowledge, interpretable decision tree classifiers were applied to predict hepatocarcinogenicity with accuracies of 85% and 76% for validation and independent test, respectively. The CCI-based method performs much better than a QSAR-based method with 12% and 6% improvements in terms of accuracy for validation and independent test, respectively. The decision rules were also analyzed to give insights into hepatocarcinogenicity.

2 Materials and Methods

2.1 Dataset

Ames-negative rodent hepatocarcinogens and noncarcinogens were extracted from a liver cancer database NCTRlcdb [10]. The annotations of organ-specific carcinogenicity and mutagenicity are available for 999 chemical compounds. Mutagenic chemicals (Ames-positive) were firstly removed. Subsequently, hepatocarcinogens and noncarcinogens were identified according to the field of OVERALL. Six noncarcinogens without corresponding chemical-chemical interaction data were also excluded. The final dataset consists of 73 hepatocarcinogens and

93 noncarcinogens. The dataset was randomly divided into three datasets with similar ratios between hepatocarcinogens and noncarcinogens for training (60%), validation (20%) and independent test (20%). The three datasets are available at http://cwtung.kmu.edu.tw/nghc.

2.2 Chemical Descriptors

The software of PaDEL-Descriptor [14] was utilized to generate chemical descriptors from chemical 2D structures extracted from PubChem database. The final feature vector is a 1610-dimensional vector consisting of 770 1D and 2D descriptors and 840 PubChem fingerprints.

2.3 Chemical-Chemical Interactions

Chemical-chemical interaction (CCI) data are obtained from STITCH 3.1 database [11], an aggregated database of interactions connecting over 300,000 chemicals and 2.6 million proteins from 1,133 organisms. For each CCI, there is a combined score calculated by combining four evidence sources of experiments, databases, text-mining and similarity. In this study, the scores divided by 1,000 are utilized to represent CCI features. The scores are ranging from 0 (low confident) to 1 (high confident).

2.4 Decision Tree Algorithm

Decision tree algorithms capable of generating interpretable rules are widely used in various biological problems such as immunogenic peptides [15], ubiquitylation sites [16] and esophageal squamous cell carcinoma [17]. In this study, the decision tree method C5.0 is applied to construct decision tree classifiers and derive interpretable rules based on CCI features for predicting hepatocarcinogenicity. C5.0 is an improved version of C4.5 with smaller trees and less computation time [18]. The implementation of R package C50 is utilized in this study [19].

The construction of a decision tree is briefly 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.

2.5 Feature Selection

The selection of important features can provide better insights into the biological problems and improve prediction performances [20,16,21,17]. This study utilized a two-step feature selection method. First, features with near zero variances were removed. Baseline models are constructed by using features whose variances are not near zero. Second, a wrapper-based feature selection method using a minimum redundancy-maximum relevancy (mRMR) method [22] is utilized to identify important CCI features for analyses and development of prediction methods. The mRMR selection process is described as follows. First, mRMR is utilized to rank the importance of CCI features. Subsequently, a sequential backward feature elimination algorithm is applied to iteratively remove CCI features with lowest ranks for selecting a subset of CCI features giving the highest 10-fold cross-validation (10-CV) accuracy. The selected feature subset is used to construct a decision tree model for predicting hepatocarcinogens.

2.6 Performance Measurement

To evaluate classifiers for their prediction performance, the widely used 10-fold cross-validation method is applied. Four measurements were used to evaluate prediction performances including sensitivity, specificity, precision and accuracy, defined as follows:

$$\text{Sensitivity} = \frac{TP}{TP + FN}, \tag{1}$$

$$\text{Specificity} = \frac{TN}{TN + FP}, \tag{2}$$

$$\text{Precision} = \frac{TP}{TP + FP}, \tag{3}$$

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + FN + TN}, \tag{4}$$

where TP, FP, FN and TN are the numbers of true positives, false positives, false negatives and true negatives, respectively. In this work, accuracy is used as the major indicator for estimating the performance of classifiers.

3 Results and Discussion

3.1 Selection of Informative Features

A baseline model using all 223 CCI features whose variances are not near zero is firstly evaluated for comparison. The accuracies of 10-CV and validation for the baseline model are 64% and 72.73% using training and validation datasets, respectively. To identify informative features for Ames-negative hepatocarcinogens, the sequential backward feature elimination algorithm was applied to the training dataset consisting of 45 hepatocarcinogens and 55 noncarcinogens.

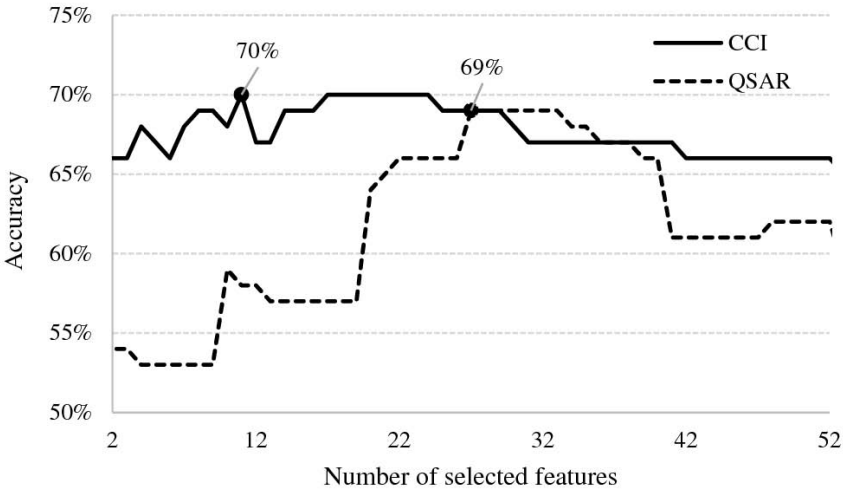

Fig. 1. The cross-validation performance for various numbers of selected features

The feature selection process and corresponding 10-CV accuracies are shown in Figure 1. Based on the training dataset, the algorithm selected a small subset of 11 CCI features giving a highest 10-CV accuracy of 70% that is 6% higher than the baseline model. The feature selected model performs well in training dataset with an accuracy of 82%. To evaluate the validation performance of the feature selected model, a decision tree model constructed by using the 11 CCI features and training dataset was utilized to classify chemicals in the validation dataset consisting of 14 hepatocarcinogens and 19 noncarcinogens. A high validation accuracy of 84.85% is obtained from the feature selected model that is 12% higher than the baseline model. Detailed performance is shown in Table 1. In addition to the mRMR method, three additional methods of chi-square test, variable importance of random forest, and relief were also evaluated with worse validation accuracies of 72.73%, 69.70% and 69.70%, respectively. The mRMR method aiming to select a feature subset of minimum redundancy and maximum relevancy might be able to avoid overfitting problems.

3.2 Independent Test Performance

To further evaluate the prediction performance of the proposed method, the decision tree model constructed by using the 11 selected CCI features was utilized to predict the chemicals in the independent test dataset consisting of 14 hepatocarcinogens and 19 noncarcinogens. The test performances are 75.76%, 50.00%, 94.74% and 87.50% for accuracy, sensitivity, specificity and precision, respectively. Compared to the test accuracy of the baseline model (66.67%), the constructed decision tree model performs well with 9% improvement. The CCI-based

Table 1. Prediction performance

	Validation		Test	
Method	CCI	QSAR	CCI	QSAR
Accuracy	84.85%	72.73%	75.76%	69.70%
Sensitivity	78.57%	57.14%	50.00%	71.43%
Specificity	89.47%	84.21%	94.74%	68.42%
Precision	84.62%	72.73%	87.50%	62.50%
AUC	0.8421	0.7030	0.7180	0.6880

model with high performances of precision and specificity is expected to be a useful tool for screening Ames-negative hepatocarcinogens. The detailed performance of the constructed model is shown in Table 1.

3.3 Comparison to Quantitative Structure-Activity Relationship (QSAR) Models

For comparison, a QSAR model was developed using the same feature selection algorithm and decision tree classifier. After feature selection, the QSAR model with a 10-CV accuracy of 69% is slightly worse than the CCI-based model (Figure 1). As shown in Table 1, the QSAR model with 27 selected features performs much worse than the CCI-based model in both validation and test dataset. The prediction accuracies of the CCI-based model are 12% and 6% higher than that of the QSAR model for validation and independent test, respectively. Due to different specificity levels of CCI-based and QSAR models, it is hard to conclude the superiority of the CCI-based model. An additional nonparametric measurement of area under receiver operating characteristic (ROC) curve (AUC) is applied to evaluate the CCI-based and QSAR models. As shown in Table 1, results show that the CCI-based model is better than the QSAR model with 14% and 3% improvement on validation and test datasets, respectively.

3.4 Decision Rules for Ames-Negative Hepatocarcinogenicity

To better understand the relationship between important CCI features and Ames-Negative Hepatocarcinogenicity, the decision tree model constructed by using the training dataset and 11 selected CCI features is shown in Figure 2. Five decision rules corresponding to five leaf nodes can be derived from the decision tree. In brief, a chemical interacting with one of the four chemicals is a hepatocarcinogen that correctly predict 27 hepatocarcinogens. Otherwise, it is a noncarcinogen that 55 noncarcinogens are correctly predicted with 18 miss-classified hepatocarcinogens. The four compounds are di-(4-aminophenyl)ether (CID000007579), ethane (CID000006324), 2-acetylaminofluorene (CID000005897), and deoxyguanosine (CID000187790). Among the four compounds, the di-(4-aminophenyl)ether and 2-acetylaminofluorene are Ames-positive carcinogens.

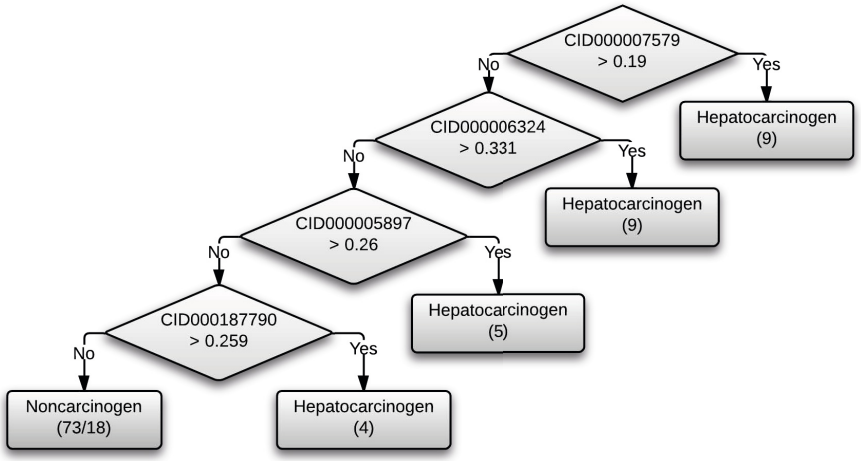

Fig. 2. Decision tree classifier for Ames-negative hepatocarcinogens

4 Conclusions

The development of computational methods for the assessment of hepatocarcinogenicity is important for efficient drug development compared to the traditional 2-year rodent bioassays. Most of the non-mutagenic hepatocarcinogens could be identified by the *in vitro* Ames test. However, it is desirable to develop alternative methods for assessing Ames-negative hepatocarcinogens. The acquisition of rules for efficient recognition of Ames-negative hepatocarcinogens is especially important for practical application. This study proposed a decision tree-based method using the CCI information and mRMR feature selection method for the acquisition of decision rules for predicting hepatocarcinogenicity of Ames-negative chemicals. The prediction model performs well with validation and test accuracies of 85% and 76%, respectively. The acquired simple decision rules are useful for identifying Ames-negative hepatocarcinogens with high specificity and precision. Future works can be the application and comparison of other machine learning methods to improve the prediction performance of Ames-negative hepatocarcinogens.

Acknowledgement. The author would like to acknowledge the financial support from National Science Council of Taiwan (NSC 101-2311-B-037-001-MY2), Kaohsiung Medical University Research Foundation (KMU-M103009), NSYSU-KMU Joint Research Project (NSYSUKMU103-P002), and National Health Research Institutes (EH-103-PP-09).

References

1. Hayashi, Y.: Overview of genotoxic carcinogens and non-genotoxic carcinogens. Exp. Toxicol. Pathol. 44, 465–471 (1992)
2. Weisburger, J.H., Williams, G.M.: The distinction between genotoxic and epigenetic carcinogens and implication for cancer risk. Toxicol. Sci. 57, 4–5 (2000)
3. Zeiger, E.: Identification of rodent carcinogens and noncarcinogens using genetic toxicity tests: premises, promises, and performance. Regul. Toxicol. Pharmacol. 28, 85–95 (1998)
4. Benigni, R., Bossa, C., Tcheremenskaia, O., Giuliani, A.: Alternatives to the carcinogenicity bioassay: in silico methods, and the in vitro and in vivo mutagenicity assays. Expert Opin. Drug Metab. Toxicol. 6, 809–819 (2010)
5. Cunningham, A.R., Carrasquer, C.A., Qamar, S., Maguire, J.M., Cunningham, S.L., Trent, J.O.: Global structure-activity relationship model for nonmutagenic carcinogens using virtual ligand-protein interactions as model descriptors. Carcinogenesis 33, 1940–1945 (2012)
6. Zeiger, E.: Historical perspective on the development of the genetic toxicity test battery in the united states. Environ. Mol. Mutagen. 51, 781–791 (2010)
7. Liu, Z., Kelly, R., Fang, H., Ding, D., Tong, W.: Comparative analysis of predictive models for nongenotoxic hepatocarcinogenicity using both toxicogenomics and quantitative structure-activity relationships. Chem. Res. Toxicol. 24, 1062–1070 (2011)
8. Yamada, F., Sumida, K., Uehara, T., Morikawa, Y., Yamada, H., Urushidani, T., Ohno, Y.: Toxicogenomics discrimination of potential hepatocarcinogenicity of nongenotoxic compounds in rat liver. J. Appl. Toxicol. (2012)
9. Tung, C.W.: Prediction of non-genotoxic hepatocarcinogenicity using chemical-protein interactions. In: Ngom, A., Formenti, E., Hao, J.-K., Zhao, X.-M., van Laarhoven, T. (eds.) PRIB 2013. LNCS, vol. 7986, pp. 231–241. Springer, Heidelberg (2013)
10. Young, J., Tong, W., Fang, H., Xie, Q., Pearce, B., Hashemi, R., Beger, R., Cheeseman, M., Chen, J., Chang, Y.C., Kodell, R.: Building an organ-specific carcinogenic database for sar analyses. J. Toxicol. Environ. Health A 67, 1363–1389 (2004)
11. Kuhn, M., Szklarczyk, D., Franceschini, A., von Mering, C., Jensen, L.J., Bork, P.: Stitch 3: zooming in on protein-chemical interactions. Nucleic Acids Res. 40, D876–D880 (2012)
12. Lu, J., Huang, G., Li, H.P., Feng, K.Y., Chen, L., Zheng, M.Y., Cai, Y.D.: Prediction of cancer drugs by chemical-chemical interactions. PLoS One 9, e87791 (2014)
13. Chen, L., Lu, J., Luo, X., Feng, K.Y.: Prediction of drug target groups based on chemical-chemical similarities and chemical-chemical/protein connections. Biochim. Biophys. Acta 1844, 207–213 (2014)
14. Yap, C.W.: Padel-descriptor: an open source software to calculate molecular descriptors and fingerprints. J. Comput. Chem. 32, 1466–1474 (2011)
15. Tung, C.W., Ziehm, M., Kämper, A., Kohlbacher, O., Ho, S.Y.: Popisk: T-cell reactivity prediction using support vector machines and string kernels. BMC Bioinformatics 12, 446 (2011)
16. Tung, C.W., Ho, S.Y.: Computational identification of ubiquitylation sites from protein sequences. BMC Bioinformatics 9, 310 (2008)

17. Tung, C.W., Wu, M.T., Chen, Y.K., Wu, C.C., Chen, W.C., Li, H.P., Chou, S.H., Wu, D.C., Wu, I.C.: Identification of biomarkers for esophageal squamous cell carcinoma using feature selection and decision tree methods. The Sci. World J. 2013, 782031 (2013)
18. Quinlan, J.: C4. 5: programs for machine learning (1993)
19. Kuhn, M., Weston, S.: Code for C5.0 by R. Quinlan, N.C.C.: C50: C5.0 Decision Trees and Rule-Based Models (2014); R package version 0.1.0-016
20. Tung, C.W.: Prediction of pupylation sites using the composition of k-spaced amino acid pairs. J. Theor. Biol. 336, 11–17 (2013)
21. Tung, C.W., Ho, S.Y.: Popi: predicting immunogenicity of mhc class i binding peptides by mining informative physicochemical properties. Bioinformatics 23, 942–949 (2007)
22. De Jay, N., Papillon-Cavanagh, S., Olsen, C., El-Hachem, N., Bontempi, G., Haibe-Kains, B.: Mrmre: an r package for parallelized mrmr ensemble feature selection. Bioinformatics 29, 2365–2368 (2013)