

Improvement of the Prediction for Genotoxicity and Carcinogenicity Based on Toxicogenomics Data and Modified Ensemble Algorithm

Authors: Woo Sun Lee¹, Eun Mi Lee¹, Euisik Han¹, Yhun Yong Sheen², Seung Hee Kim¹, Sue Nie Park¹

¹Korea Food and Drug Administration/Department of Toxicological Researches, Korea

²Ewha Women's University/College of Pharmacy, Korea

Keywords: genotoxicity, carcinogenicity, toxicogenomics, PCA, ensemble learning method

Prediction of toxicity related to genotoxicity and carcinogenicity of chemicals has been attempted by carrying out detection and analysis of candidate marker genes for them by toxicogenomics approaches using microarray method. However, the procurement of enough amount data set for prediction analysis is usually not so easy since the cost of tests could be too expensive. Therefore, it would be useful to improve the prediction power by applying proper analytical methods on the data set obtained for small number of chemicals. Thus, in this study, we attempted to apply and compare the classification approaches such *k*-NN (*k*-nearest Neithborhood), SVM (Support Vector Machine), PLS (Partial Least Square regression), PCR (Principle Component regression), ridge regression discriminant classification. Test materials chosen as model compounds were four chemicals each representing genotoxic carcinogen (glycidol), genotoxic noncarcinogen (8-hydroxyquinoline), nongenotoxic carcinogens (o-nitrotoluene), and nongenotoxic noncarcinogen (1,2-dichlorobenzene) to evaluate the classification methods we applied. These chemicals were administered to L5178Y mouse lymphoma cell to obtain toxicogenomic data using Affimatrix genechip and the data sets were obtained from three independently repeated tests for each chemical. Genes were selected after data manipulation. For normalization, mas5 & quantile combination methods were applied, and data filtered with fold change >1.3 and one-way ANOVA ($p < 0.001$). 3D plotting of Principle component of PCA (Principle component Analysis) using microarray whole profiles obtained for 4 chemicals resulted in nice separation of all chemicals into 4 group representing the homogeneity within replications. *k*-NN, SVM, PLS, PCR, ridge regression classified data with wrapping filtered data from smallest to largest to find optimal genes selection. Each minimal errors was *k*-NN(0.2), SVM(0.2), PLS(0.2), PCR(0.06), ridg(0.0) for training step. Finally, using weight from error, we aggregate result of each classifier, and we reached final error rate to 0%. Currently, we are attempting to test data sets obtained for 4 more chemicals to see if the prediction power changes or not. In addition, we are also testing other statistical methods such as Ensemble learning which aggregates each single classifier's result through selecting optimal candidate marker genes to see if it enhances toxicity class prediction accuracy on microarray data. Our study may offer the more effective method in evaluating toxicity of chemicals related to genotoxicity and carcinogenicity resulting in more effective prediction method for novel toxicants.

Point of Contact:

Sue Nie Park, Ph.D.

Director, Division of Genetic Toxicology

Korea Food & Drug Administration

5 Nokbeon-dong, Eunpyeong-gu

Seoul, Korea 122-704
82-02-380-1792
suenie@kfda.go.kr