Available online at www.sciencedirect.com
SCIENCE DIRECT®
Mutation Research 575 (2005) 17–33
Fundamental and Molecular Mechanisms of Mutagenesis
www.elsevier.com/locate/molmut
Community address: www.elsevier.com/locate/mutres

Comparison of supervised clustering methods to discriminate genotoxic from non-genotoxic carcinogens by gene expression profiling

J.H.M. van Delft*, E. van Agen, S.G.J. van Breda, M.H. Herwijnen, Y.C.M. Staal, J.C.S. Kleinjans

Department of Health Risk Analysis and Toxicology, Maastricht University, P.O. Box 616, 6200 MD Maastricht, The Netherlands

Received 4 October 2004; received in revised form 17 February 2005; accepted 23 February 2005
Available online 19 April 2005

Abstract

Prediction of the toxic properties of chemicals based on modulation of gene expression profiles in exposed cells or animals is one of the major applications of toxicogenomics. Previously, we demonstrated that by Pearson correlation analysis of gene expression profiles from treated HepG2 cells it is possible to correctly discriminate and predict genotoxic from non-genotoxic carcinogens. Since to date many different supervised clustering methods for discrimination and prediction tests are available, we investigated whether application of the methods provided by the Whitehead Institute and Stanford University improved our initial prediction. Four different supervised clustering methods were applied for this comparison, namely Pearson correlation analysis (Pearson), nearest shrunken centroids analysis (NSC), *K*-nearest neighbour analysis (KNN) and Weighted voting (WV). For each supervised clustering method, three different approaches were followed: (1) using all the data points for all treatments, (2) exclusion of the samples with marginally affected gene expression profiles and (3) filtering out the gene expression signals that were hardly altered. On the complete data set, NSC, KNN and WV outperformed the Pearson test, but on the reduced data sets no clear difference was observed. Exclusion of samples with marginally affected profiles improved the prediction by all methods. For the various prediction models, gene sets of different compositions were selected; in these 27 genes appeared three times or more. These 27 genes are involved in many different biological processes and molecular functions, such as apoptosis, cell cycle control, regulation of transcription, and transporter activity, many of them related to the carcinogenic process. One gene, *BAX*, was selected in all 10 models, while *ZFP36* was selected in 9, and *AHR*, *MT1E* and *TTR* in 8. Summarising, this study demonstrates that several supervised clustering methods can be used to discriminate certain genotoxic from non-genotoxic carcinogens by gene expression profiling in vitro in HepG2 cells. None of the methods clearly outperforms the others.

Keywords: Gene expression profiling; Clustering methods; Genotoxic and non-genotoxic carcinogens

Abbreviations: GTX, genotoxic; KNN, *K*-nearest neighbour; NGTX, non-genotoxic; NSC, nearest shrunken centroids; Pearson, Pearson correlation; WV, weighted voting

* Corresponding author. Tel.: +31 43 3881092; fax: +31 43 3884146.
E-mail address: j.vandelft@grat.unimaas.nl (J.H.M. van Delft).

0027-5107/$ – see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.mrfmmm.2005.02.006

1. Introduction

Application of functional genomics technologies—meaning profiling of differential expression for many genes simultaneously by DNA microarrays—in toxicology, including genetic toxicology and chemical carcinogenesis, has proven to be very promising over recent years. Soon after the introduction in the late 90s, it was already clear that by toxicogenomics a wealth of information can be obtained on the mode-of-action of toxic compounds by identifying genes with deregulated mRNA levels following exposure [1,2]. Furthermore, it became clear that compounds which cause similar toxic effects, also modulate to a certain extent similar gene expression profiles [3–6]. Therefore, in principle these similar gene expression profile changes can serve as fingerprints for the specific toxic effects, and be used to predict the toxic properties of other compounds [7]. Clustering methods that group samples with similar gene expression profiles and discriminate clusters of samples with more dissimilarities in those profiles, are frequently used for this.

Clustering methods based on gene expression profiles can be divided in two groups, namely unsupervised and supervised clustering [8,9]. Unsupervised clustering methods, such as the classical hierarchical clustering method developed by Eisen et al. [10] (currently also called heat maps) and *K*-means clustering methods [11], perform the clustering without taking into account any a priori knowledge about the samples to be clustered. In order to determine whether a specific sample belongs to one or another specific group of samples (e.g. samples belonging to different classes of toxicants), supervised clustering methods are used. In these methods, a priori knowledge about which clusters of samples are present is added and used to identify genes whose expression changes best discriminate between the various clusters. Thereafter, this information can be used to determine for other samples to which cluster they belong.

To date, the ability of discriminating toxic compounds based on expression patterns, has been demonstrated frequently, both by in vitro studies using cell lines or primary hepatocytes [5,12] as well as by in vivo studies with rat [13–16]. That toxicogenomics cannot only be used to discriminate classes of toxicants but also to accurately establish the class of a toxicant, has recently been shown for hepatoxicants in in vivo rat studies [17,18]. Furthermore, we recently demonstrated that expression profiling in HepG2 cells following exposure to chemicals carcinogen in rodents, can discriminate and predict certain genotoxic (GTX) from non-genotoxic (NGTX) carcinogenic agents [19]. HepG2 cells are metabolically competent with respect to biotransformation of mutagens and carcinogens, frequently used in toxicology and gene expression studies and carry no p53 mutations [3,20–24].

In clinical settings, such as in cancer diagnostics and therapy, many different supervised clustering methods for classification of disease state or prediction of therapy efficacy have been applied, such as Pearson correlation analysis, *K*-nearest neighbour analysis, neural networks, support vector machines, decision tree classifiers, weighted voting and nearest centroids analysis [25–30]. None of them is clearly preferred, and each has specific advantages or disadvantages [31,32]. Basically, all these methods consist of two steps: first identification of the genes—the classifiers—that best discriminate between the two classes based on known samples (a training set), and then to use the expression profiles of those classifiers to establish the class of for an unknown sample. Many of these methods have been made available to the scientific community, among others by the pioneers in this field at the Whitehead Institute (MIT, Boston; *K*-nearest neighbours and weighted voting methods as part of GeneCluster 2; www.broad.mit.edu/cancer/software; [33]) and Stanford University (Stanford, USA; nearest shrunken centroids method; http://www.stat.stanford.edu/∼tibs/pam; [25]).

The method applied previously by us to discriminate GTX from NGTX carcinogens, was by Pearson correlation analysis [19]. We also demonstrated that the discrimination could be drastically improved by exclusion of weak data that possibly only increase noise and thus hide the available information. In that study, both omitting the samples with marginally affected gene expression profiles as well as excluding the gene expression signals that were hardly altered, improved correct class discrimination from about 70 to 90% or more.

Here we investigated whether the application of the prediction methods provided by the Whitehead Institute and Stanford University improves the prediction. Four different supervised clustering methods were applied for discrimination of GTX from NGTX carcinogens based on gene expression profiling. These meth-