Microchemical Journal xxx (2017) xxx–xxx

Contents lists available at ScienceDirect

Microchemical Journal

The use of omics-based approaches in regulatory toxicology: an alternative approach to assess the no observed transcriptional effect level

Daniele Quercioli[a], Andrea Roli[b], Elena Morandi[a], Stefania Perdichizzi[a], Laura Polacchini[a], Francesca Rotondo[a], Monica Vaccari[a], Marco Villani[c], Roberto Serra[c], Annamaria Colacci[a,*]

[a] Center for Environmental Toxicology, Regional Agency for Prevention, Environment and Energy – Emilia Romagna (Arpae), c/o DIMES, University of Bologna, Italy
[b] Department of Computer Science and Engineering (DISI), University of Bologna, Italy
[c] Department of Physics, Informatics and Mathematics, University of Modena and Reggio Emilia, Italy

ARTICLE INFO

Article history:
Received 15 September 2016
Received in revised form 30 January 2017
Accepted 31 January 2017
Available online xxxx

Keywords:
NOTEL
Transcriptomics
Anomaly detection
Gene expression

ABSTRACT

The evaluation of chemical hazard is based on the identification of the quality and the quantity of adverse effects as a consequence of exposure. The adverse effects that do not involve genetic damage are often related to the chemical dose or concentration. The adverse outcome is the consequence of a row of key events, each targeting a different biological trait. The identification of these key events at molecular and cellular level would provide novel biomarkers of exposure and risk. The application of toxicogenomics approaches to experimental models of chemical exposure allows the detection of gene pathways involved in response to low doses of the chemical as an early endpoint of adversity. The use of toxicogenomics would improve the knowledge on the dose-response relationship, linking the environmental exposure to the effect on the population and allowing a better refinement of the quantitative risk assessment. In this context, the gene modulation data can be used to calculate a No Observed Transcriptional Effect Level (NOTEL).
In this paper we present a method for evaluating the NOTEL based on anomaly detection: a classifier is built that discriminates between target class instances, i.e., normal cases, and anomalies, i.e., samples with significant transcriptional effects. The strength of this method is that (*i*) it can be applied to cases in which few samples are available and the space dimension is high and (*ii*) it makes use of the complete gene expression profiles.

1. Introduction

The evaluation of chemical hazard is based on the identification of the quality and the quantity of adverse effects as a consequence of exposure. The adverse effects that do not involve genetic damage are often related to the chemical dose or concentration. This implies that for non-genotoxic compounds it is possible to identify a level of dose, below which no adverse effects are observed. The identification of NOAEL (No Observed Adverse Effect Level) or NOEC (No Observed Effect Concentration) is of paramount importance in the field of regulatory toxicology. If the exposure is below the NOAEL, then it is assumed that the risk for human exposure, or environment impact, in case of NOEC, is negligible. However, NOAEL and NOEC are often calculated on experimental data from animal studies, with high experimental doses, which do not reflect the low and very low concentrations of chemicals in the environment. For chemicals that do not show a linear dose-response, it is possible that adverse effects are induced at doses below the NOAEL. When it is not possible to highlight a NOAEL in a study, the lowest observed adverse effect level (LOAEL) is considered, which increases the level of uncertainty in the risk calculation.

During the last 30 years other approaches have been developed to overcome the limitations of the NOAEL approach. The benchmark dose (BMD) is calculated on a predetermined level of effect. The lower confidence limit of the BMD (BMDL) is used as the point of departure for the calculation of allowable exposures [29]. Even if the BMD offers some advantages with respect to NOAEL, the majority of the chemical risk assessment is still based on the NOAEL approach [4]. The lower bound on the benchmark dose or the NOAEL serves as the point of departure to calculate the "safe" dose for human exposure to the single chemical (e.g. Acceptable Daily Intake, Occupational Exposure Level, and Reference Dose) [11].

As the role of toxicogenomics in the prediction of chemical hazard and risk has become more and more central in the last few years, the

* Corresponding author.
E-mail address: annamaria.colacci@unibo.it (A. Colacci).

http://dx.doi.org/10.1016/j.microc.2017.01.029
0026-265X/

approach to the identification of the adverse outcome has changed consequently. The adverse outcome is the consequence of a row of key events, each targeting a different biological trait. The identification of these key events at molecular and cellular level would provide novel biomarkers of exposure and risk [10]. The application of toxicogenomics approaches to experimental models of chemical exposure allows the detection of gene pathways involved in the response to the chemical as an early endpoint of adversity, which is often related to the exposure to low doses of chemicals. The biological interpretation of omics results sustains the identification of the key events in the pathway leading to the adverse outcome [25]. The modulation of genes related to the adverse outcome is an early step towards the process leading to the phenotypic evidence. The key genes in a modulated pathway may be regarded as early biomarkers of exposure at low doses. Moreover, the use of toxicogenomics data in the risk assessment can reduce the uncertainty related to the interspecies differences and variability, which is one of the main limitations in the use of animal studies [7,21]. Toxicogenomics, however, is one of the main high throughput methods included in the predictive toxicology approach, which aims at reducing the use of animals, supporting the application of the 3Rs principles of replacement, refinement and reduction of animal studies [25]. As Chiu et al. have recently noted [7], the use of toxicogenomics, besides supporting the weight-of-evidence of alternative methods within the integrated testing strategies and assessment approaches, would improve the knowledge on the dose-response relationship, linking the environmental exposure to the effect on the population and allowing a better refinement of the quantitative risk assessment. In this context, the gene modulation data can be used to calculate a No Observed Transcriptional Effect Level (NOTEL), which could replace the NOAEL as a better point of departure to model the exposure at low doses.

Current methods to determine the NOTEL [13,23,31] rely on the estimation of the proportion of genes differentially expressed between treated samples and controls. A weakness of this approach is that the distribution of gene expression levels is not taken into account in the dose-response analysis, which might result in an information loss for NOTEL assessment. As far as we know, there are no NOTEL assessment methods that explicitly make use of the activation profile of genes. In our viewpoint, a computational method for the NOTEL assessment that is based on gene expression profiles would be more robust and more informative. To this aim, we developed a method based on a one-class classification anomaly detection technique. Anomaly detection [5] is defined as the problem of finding those instances in a data set that do not conform to the expected behaviour. Hence, a treated sample not conforming to the control cases can be viewed as an anomaly. Specifically, in the problem of one-class classification, a model is learnt from a set of target data instances—i.e., the *training set*—obtaining a so-called one-class classifier, which is used to classify new instances into either the target class or as an anomaly. In our case, the target instances composing the training set are samples of non-treated cases.

The aim of this work is to illustrate an alternative way for assessing transcriptional significant response to given dose levels. The structure of this contribution is as follows. We introduce the notion of anomaly detection and we describe our approach in Section 2. The approach is validated on data extracted from a database of experiments on rats, undertaken in a collaborative Japanese project [30], and results are presented in Section 3 and then discussed in Section 4. Finally, we summarize the results achieved and provide an outlook to further improvements in Section 5.

2. Methods

In this section we illustrate our approach. We first provide some details on the selection of significant gene probes; subsequently, we describe the technique for detecting significant treated cases as anomalies of the one-class model.

2.1. Modulated genes selection

The number of gene probes in a standard microarray is relatively large and, in a typical experiment, most probes are not significantly altered by the experimental treatment. Therefore, an initial probe selection process has been performed to eliminate probes that are likely to introduce noise in data analysis. The probe selection process consists in a one-way analysis of variance (ANOVA) combined with a false discovery rate correction for multiple comparisons [3].

Before the training phase, the data are also preprocessed by a feature scaling step [14], i.e., probe gene values are scaled so as to vary in the same range.

2.2. Anomaly detection for NOTEL assessment

In our approach, the target class is composed of the instances that do not show any significant variation in the gene expression profile. A model for this class has to be learnt by examples, which are provided by non-treated cases and their collection is called the *training set*. A model for the target class is defined by means of a training procedure, which will be described in the following. Once learnt, the model can be used to test whether a given treated sample is in the class or not.

Let X be the training set in an N-dimensional vector space drawn from the target distribution. X is composed of vectors $\mathbf{x}_i \in \mathbb{R}^N$, corresponding to untreated samples. The training process returns a one-class classifier that characterises this target class. In general, one-class classifiers can be expressed in terms of a vector $\mathbf{x}$ and the set X as follows:

$$h(\mathbf{x}, X) = \mathcal{I}(d(\mathbf{x}, X) \leq \theta) = \begin{cases} 1 & \text{if } \mathbf{x} \text{ is classified as a target} \\ 0 & \text{otherwise} \end{cases}$$

where $\mathcal{I}(\cdot)$ is an indicator function which returns 1 when the distance between $\mathbf{x}$ and the set X is less than or equal to θ, parameter of the model. Therefore, $h(\cdot)$ relies on the definition of a distance between an N-dimensional vector $\mathbf{x}$ and the set X, and on a parameter θ. Besides θ, the classifier may also depend on other parameters, depending on the kind of classifier. The objective of the training process is to properly tune θ—and other parameters, if any—given the target instances.

Among the available methods for anomaly detection, we chose one that is particularly suited for situations in which few samples are available and the space dimension is high, which is exactly our case. This method, named the *Minimum Spanning Tree Class Descriptor* (MST_CD), is non-parametric and it is based on a graph representation of the target training data, aiming to capture the underlining data structure [15]. Informally, the method constructs a tree T in $\mathbb{R}^N$ connecting the points in X, which represents the skeleton of target class model. We assume that the tree is composed of the points in X and the segments connecting them, i.e., $X \subset T \subset \mathbb{R}^N$. A point $\mathbf{x}$ is then said to belong to the target class if its distance to T is not greater than θ. A common choice for the distance $d(\cdot)$ is $d(\mathbf{x}, T) := \min\{d_{Eucl}(\mathbf{x}, \mathbf{y}), \mathbf{y} \in T\}$, where $d_{Eucl}(\mathbf{x}, \mathbf{y})$ is the Euclidean distance[1] between $\mathbf{x}$ and $\mathbf{y}$. In other words, $d(\mathbf{x}, T)$ is the length of the segment connecting $\mathbf{x}$ with its closest point in T.

The graph defining the reference for the target class is constructed as follows. Let X be a training set composed of n instances $\mathbf{x_i} \in \mathbb{R}^N$, $i = 1, \ldots, n$. The points in X are the vertices of a complete

[1] The Euclidean distance is the most common choice; however, other distance measures can of course be chosen.

Please cite this article as: D. Quercioli et al., The use of omics-based approaches in regulatory toxicology: an alternative approach to assess the no observed transcriptional effect level, Microchemical Journal (2017), http://dx.doi.org/10.1016/j.microc.2017.01.029