Bioinformatics, 31(13), 2015, 2115–2122
doi: 10.1093/bioinformatics/btv108
Advance Access Publication Date: 19 February 2015
Original Paper

OXFORD

Gene expression

Pattern recognition methods to relate time profiles of gene expression with phenotypic data: a comparative study

Diana M. Hendrickx*, Danyel G. J. Jennen, Jacob J. Briedé, Rachel Cavill, Theo M. de Kok and Jos C. S. Kleinjans

Department of Toxicogenomics, School of Oncology and Developmental Biology (GROW), Maastricht University, 6200 MD Maastricht, The Netherlands

*To whom correspondence should be addressed.

Associate Editor: Janet Kelso

Received on August 11, 2014; revised on January 29, 2015; accepted on February 16, 2015

Abstract

Motivation: Comparing time courses of gene expression with time courses of phenotypic data may provide new insights in cellular mechanisms. In this study, we compared the performance of five pattern recognition methods with respect to their ability to relate genes and phenotypic data: one classical method (k-means) and four methods especially developed for time series [Short Time-series Expression Miner (STEM), Linear Mixed Model mixtures, Dynamic Time Warping for -Omics and linear modeling with R/Bioconductor limma package]. The methods were evaluated using data available from toxicological studies that had the aim to relate gene expression with phenotypic endpoints (i.e. to develop biomarkers for adverse outcomes). Additionally, technical aspects (influence of noise, number of time points and number of replicates) were evaluated on simulated data.

Results: None of the methods outperforms the others in terms of biology. Linear modeling with limma is mostly influenced by noise. STEM is mostly influenced by the number of biological replicates in the dataset, whereas k-means and linear modeling with limma are mostly influenced by the number of time points. In most cases, the results of the methods complement each other. We therefore provide recommendations to integrate the five methods.

Availability: The Matlab code for the simulations performed in this research is available in the Supplementary Data (Word file). The microarray data analysed in this paper are available at ArrayExpress (E-TOXM-22 and E-TOXM-23) and Gene Expression Omnibus (GSE39291). The phenotypic data are available in the Supplementary Data (Excel file). Links to the pattern recognition tools compared in this paper are provided in the main text.

Contact: d.hendrickx@maastrichtuniversity.nl

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Comparing time courses from different types of measurements is an important topic in biological research (Cavill *et al.*, 2013; Smith *et al.*, 2009; Tan *et al.*, 2006), because it contributes toward understanding the response of complex cellular systems to perturbations. In particular in toxicogenomics (which stands for the application of -omics technologies to toxicology), comparing time courses of gene expression with time courses of phenotypic endpoints (i.e. biomarkers for the adverse outcomes induced by a chemical compound) may help to distinguish gene expression changes related to toxicity from changes not related to toxicity (Powell *et al.*, 2006). The

present study focuses on comparing time courses of gene expression with time courses of phenotypic data, also called phenotypic anchoring of gene expression data. Phenotypic anchoring may provide additional insight in the response to perturbations that cannot be generated by analyzing gene expression alone. Various statistical tools can be used for phenotypic anchoring (Ganter *et al.*, 2008), including pattern recognition.

Pattern recognition includes clustering (e.g. k-means), classification (e.g. support vector machines) and model reduction (e.g. principal component analysis) (de Ridder *et al.*, 2013). Biological entities belonging to the same cluster are assumed to be functionally related (Bar-Joseph, 2004; Liao, 2005).

In this study, five representative pattern recognition methods were chosen. First, we included a clustering method that is able to remove random patterns, called Short Time-series Expression Miner (STEM). Removing random patterns is important for datasets with a high number of variables compared with the number of time points (Ernst and Bar-Joseph, 2006). Secondly, Dynamic Time Warping for -Omics (DTW4omics), a method that takes into account time delays in biological processes (Cavill *et al.*, 2013), is considered. In contrast to clustering methods (which treat all variables in the dataset as equivalent entities), DTW4omics treats the phenotypic data as predefined profiles and searches for genes matching these profiles using a distance metric. Thirdly, Linear Mixed Model (LMM) mixtures, a clustering method correcting for variability between replicates (Celeux *et al.*, 2005), is discussed. Furthermore, we applied k-means, a classical clustering method for static data. Finally, we used a regression method [linear modeling with the R/Bioconductor package limma (Ritchie *et al.*, 2015)], to find relationships between gene expression and phenotypic endpoints.

The five methods were compared with respect to their ability to extract functionally related groups of genes, their sensitivity to measurement noise, the influence of the number of time points and the influence of the number of biological replicates. The technical aspects (influence of noise, number of time points and number of replicates) were evaluated on simulated data, mimicking real expression and phenotypic data. The biological outcome was evaluated exploiting two public datasets from toxicogenomics.

2 Materials and methods

2.1 Pattern recognition methods

In this study, both simulated and real datasets were analyzed by means of STEM, DTW4Omics, LMM mixtures, k-means and linear modeling with limma. Table 1 gives an overview of the properties of the four methods. Calculations were performed in R (http://www.r-project.org/), except for STEM, which is a Java application.

2.1.1 Short time-series expression miner

STEM (Ernst and Bar-Joseph, 2006) is a freely available Java tool (http://www.cs.cmu.edu/~jernst/stem/) for clustering time profiles that run in parallel. STEM distinguishes between real and random patterns in time series by identifying time profiles that are significantly present in the dataset. Significance is determined by permutation testing. Phenotypic endpoints were treated as genes and added to the study. Opposite time profiles of phenotypic endpoints were also added to the study in order to determine negative relationships between genes and phenotypic endpoints. The variables in the dataset (genes, phenotypic endpoints) are assigned to the significant profiles based on similarity (positive correlation). Genes assigned to the same profile as a phenotypic endpoint are assumed to be related to that endpoint. The STEM clustering algorithm is described in detail by Ernst and coworkers (Ernst *et al.*, 2005). STEM has been previously applied to cluster toxicogenomics time profiles (Alm *et al.*, 2009; Briede *et al.*, 2010; Hebels *et al.*, 2013). In this study parameters for STEM were taken as described in previous work (Briede *et al.*, 2010).

Table 1. Properties of the four pattern recognition methods discussed in this article

Takes into account	STEM	DTW	LMM	k-Means	Limma
Time dependencies	Yes	Yes	Yes	No	Yes
Correction for random patterns	Yes	NA	No	No	NA
Delays	No	Yes	No	No	No
Variability between replicates	No	No	Yes	No	Yes

Note: NA, not applicable.

2.1.2 Dynamic time warping for –omics

DTW4Omics is publicly available in-house R code (http://web.tgx.unimaas.nl/svn/public/dtw/) to detect similar patterns on different time scales, due to delays or differences in speed (Cavill *et al.*, 2013). For each endpoint, a list of significant genes (false discovery rate <0.05) is produced based on the calculation of the optimal distance between time profiles and permutation testing. All genes in this list are supposed to be positively related to that endpoint. To determine negative relationships, we repeat the previous steps for the opposite time profiles of the phenotypic endpoints. DTW4Omics uses the distance function from the DTW package described by Giorgino and coworkers (Giorgino, 2009).

2.1.3 LMM mixtures

LMM mixtures is a model-based method for clustering time series that run in parallel, taking into account variability between replicates (Celeux *et al.*, 2002). Phenotypic endpoints were treated as genes and added to the study. Opposite time profiles of phenotypic endpoints were also added to the study in order to determine negative relationships between genes and phenotypic endpoints. LMM mixtures divide the dataset in an optimal number of clusters. For each cluster, the average profile for the cluster is described by a LMM, with the times as fixed effects and variability between replicates as random effect. Genes assigned to the same cluster as a phenotypic endpoint (respectively, the opposite profile of a phenotypic endpoint) are assumed to be positively (respectively, negatively) related to that endpoint. The optimal number of clusters is determined using the Bayesian Information Criterion (BIC) and maximum-likelihood parameters are iteratively determined by an optimization algorithm (Celeux *et al.*, 2005). To determine a starting value for the optimization algorithm for LMM mixtures, spectral clustering (Ng *et al.*, 2001) was applied. Using the cluster solution from spectral clustering has previously been reported as an efficient initialization strategy for LMM mixtures (Scharl *et al.*, 2010). Spectral clustering was performed in R using the 'specc' function of the 'kernlab' package (Zeileis *et al.*, 2004). The mixture of LMM was fitted with the 'FLXMRlmm' interface of the 'flexmix' package (Grün and Leisch, 2008; Leisch, 2004). Both R packages are available on the CRAN website (http://cran.r-project.org).

2.1.4 k-Means clustering

Phenotypic endpoints were treated as genes and added to the study. Opposite time profiles of phenotypic endpoints were also added to the study in order to determine negative relationships between genes

and phenotypic endpoints. k-Means then clusters profiles with an iterative optimization algorithm that minimizes within-cluster variability, while maximizing between-cluster variability (Kintigh and Ammerman, 1982; Liao, 2005). Each cluster is represented by its cluster center. Genes assigned to the same cluster as a phenotypic endpoint (respectively, the opposite profile of a phenotypic endpoint) are assumed to be positively (respectively, negatively) related to that endpoint. The optimal number of clusters is selected based on the sum of squared error (SSE), the sum of squared distances between cluster members and center (Kintigh and Ammerman, 1982). The optimal solution was calculated in R using the code available at http://www.mattpeeples.net/kmeans.html and is the number of clusters for which the SSE for the original data differs the most from the average SSE for 250 random datasets. k-Means was performed in R using the source code available at http://www.mattpeeples.net/kmeans.html, which uses the 'kmeans' function of the 'stats' package (http://cran.r-project.org).

2.1.5 Linear modeling with limma

For each phenotypic endpoint, genes that vary with that endpoint were determined by fitting a LMM to the gene expression data using the R/Bioconductor package limma (Ritchie *et al.*, 2015*)* (www.bioconductor.org/packages/release/bioc/html/limma.html). Time, treatment (exposed versus control) and the phenotypic endpoint were taken as fixed effect, and replicates as random effect in the model. Using the phenotypic endpoint as a contrast (coefficient in the linear model), limma tests which genes vary significantly with the phenotypic endpoint. A cut-off value of 0.05 was taken for the false discovery rate. The regression coefficients for the endpoints determine whether the relationships are positive or negative.

2.2 Simulated data

Performance of statistical methods can be evaluated by checking whether the results correspond with prior biological knowledge. However, often our knowledge about the underlying biology is incomplete. In this case, computer simulations are very useful to assess how well a method works (Mendes *et al.*, 2003). Here, the influence on the results of making changes, for example adding noise, to a reference dataset is studied.

2.2.1 Simulation method

We simulated datasets with properties of microarray data in Matlab® Version 8.1.0. (r2013a) (copyright©, 1984–2013, The Mathworks Inc.) applying the method described by Mendes and coworkers (Mendes *et al.*, 2003) and on the companion website of Nykter *et al.* (2006) (http://www.cs.tut.fi/sgn/csb/mamodel/). Thousand variables were generated for 10 time points. Each variable in the dataset is described by a differential equation including rate constants, affinity constants and Hill coefficients as parameters. After simulating control samples, a perturbation was simulated by changing one of the rate constants. Three biological replicates were generated for both control and perturbation by adding small random variables to the affinity constants and Hill coefficients. The reference dataset consisted of the log2-ratios for the three biological replicates. Clusters of variables relating to variables 996–1000 were determined applying the four pattern recognition methods, so variables 996–1000 are considered to be the endpoints. The influence of measurement noise was studied by successively adding 15%, 20% and 25% Gaussian noise to the reference data and comparing the solutions for the pattern recognition methods with the solutions for the reference dataset. Results of the reference dataset were also compared with pattern recognition results when using less time points (the first eight and six time points) and when using only two biological replicates instead of three. Table 2 presents an overview of the simulated data. More details about the simulation and the Matlab code are provided in the Supplementary Data, Section 1.

Table 2. Overview of the simulated datasets

	Measurement noise	Number of time points	Number of biological replicates
Reference data	0%	10	3
15% Noise	15%	10	3
20% Noise	20%	10	3
25% Noise	25%	10	3
8 Time points	0%	8	3
6 Time points	0%	6	3
2 Biological replicates	0%	10	2

2.2.2 Evaluation of pattern recognition methods

The results for the different simulated datasets were compared with the results obtained from the reference dataset. The accuracy describes how close the results of the simulation were to the results for the reference dataset.

2.3 Real datasets

The five pattern recognition methods were applied to two datasets obtained from the human hepatoma cancer cell line HepG2: after exposure to either benzo(a)pyrene (B(a)P), a human carcinogen, or menadione (vitamin K3), an agent producing reactive oxygen species (ROS). Both positive and negative associations with particular endpoints were determined.

2.3.1 Response of HepG2 to B(a)P

B(a)P is a carcinogenic polycyclic aromatic hydrocarbon, having genotoxic and non-genotoxic properties. B(a)P exposure causes, among others, oxidative stress, DNA adduct formation and apoptosis (van Delft *et al.*, 2010). Sources of exposure are (among others) wood burning, coal tar, vehicle exhaust and cigarette smoke (Bostrom *et al.*, 2002). Gene expression and phenotypic endpoints (DNA adducts, cell cycle, apoptosis indicative of molecular responses to DNA damage induced by the carcinogen) were measured at 12 time points after exposure to B(a)P (3, 6, 9, 12, 15, 18, 24, 30, 36, 48, 54 and 60 h). HepG2 cells were treated with 3 μM B(a)P or vehicle control. Cell cycle profiles (G1, G2/M and S phase) and apoptotic cell levels were determined by flow cytometry. DNA adduct levels were determined by ^{32}P post-labeling. Gene expression was determined with Agilent microarrays, labeled with cyanine 3 (Cy3) and cyanine 5 (Cy5). Two biological experiments were conducted, with two hybridizations per time point for each experiment (by swapping Cy3 and Cy5). Microarray data are available at ArrayExpress (E-TOXM-22 and E-TOXM-23). Phenotypic data are available in the Supplementary Data (Excel file). Detailed information about this dataset was described earlier in van Delft *et al.* (2010).

2.3.2 Response of HepG2 to menadione

Menadione (vitamin K3) is a compound causing oxidative stress through the formation of ROS. Oxidative stress has been related to chemically induced liver injury, chronic liver diseases and

hepatocellular carcinoma (Deferme *et al.*, 2013). Menadione has been used as a drug therapy and sometimes also as nutritional supplement (Truong and Booth, 2011). Gene expression and phenotypic endpoints (oxidative DNA damage, protein oxidation, cell cycle, apoptosis, which are all toxic responses inflicted by ROS) were measured at seven time points after exposure to menadione (0.5, 1, 2, 4, 6, 8 and 24 h). HepG2 cells were treated with 100 µM menadione or vehicle control. Cell cycle profiles (G1, G2/M and S phase) and apoptotic cell levels were determined by flow cytometry. Protein oxidation was determined by protein carbonyl assay, measuring protein carbonyl formation by spectrophotometry. Oxidative DNA damage was determined using the FPG-comet assay. Gene expression was determined using Affymetrix microarrays. Three biological experiments were conducted. Microarray data are available at Gene Expression Omnibus (GSE39291). Phenotypic data are available in the Supplementary Data (Excel file). Detailed information about this dataset was described earlier in Deferme *et al.* (2013).

2.3.3 Normalization, selecting differentially expressed genes and scaling

Normalization was conducted as described earlier in van Delft *et al.* (2010) and Deferme *et al.* (2013) for the B(a)P and menadione dataset, respectively.

Differentially expressed genes (DEGs) were determined using the R/Bioconductor package limma (Ritchie *et al.*, 2015), controlling for dye swaps in case of two-color arrays. 'Treatment—time-matched control' contrasts were fitted at each time point and with each treatment (exposure, control). Genes with false discovery rate lower or equal than 0.01 and absolute log fold change higher than 1.5 were selected.

Log2 ratios were taken to transform the data to be normally distributed. Data were centered and unit variance scaled to make the values of all variables (genes and endpoints) comparable.

2.3.4 Evaluation of pattern recognition methods

Overrepresentation analyses (pathway and GO analysis) were performed on the lists of genes related to the endpoints using ConsensusPathDB, an interaction database containing pathways from 32 databases (Kamburov *et al.*, 2013). Pathways and GO terms with at least two genes in common with the gene lists were selected. To correct for multiple testing, only pathways and GO terms with false discovery rate lower or equal than 0.05 were selected. Overlap between the genes, pathways or GO terms was determined using the online tool Venn Diagrams (http://bioinformatics.psb.ugent.be/webtools/Venn/). Lists of overrepresented pathways and GO terms were compared with prior biological knowledge, among others from the Comparative Toxicogenomics Database (Davis *et al.*, 2013).

3 Results

3.1 Simulated data

LMM mixtures divided the dataset into 19 clusters (based on the BIC), while k-means only distinguished 5 clusters (based on the SSE). Because of their different properties and underlying assumptions, there is minor overlap among all of the five methods (Fig. 1 and Supplementary Fig. S1). Which methods show the largest overlap is different for different variables (e.g. for variable 996 DTW4omics and k-means show the largest overlap, while for variable 1000 STEM and DTW4omics show the largest overlap). The overall biggest overlap is found between DTW4omics and k-means,

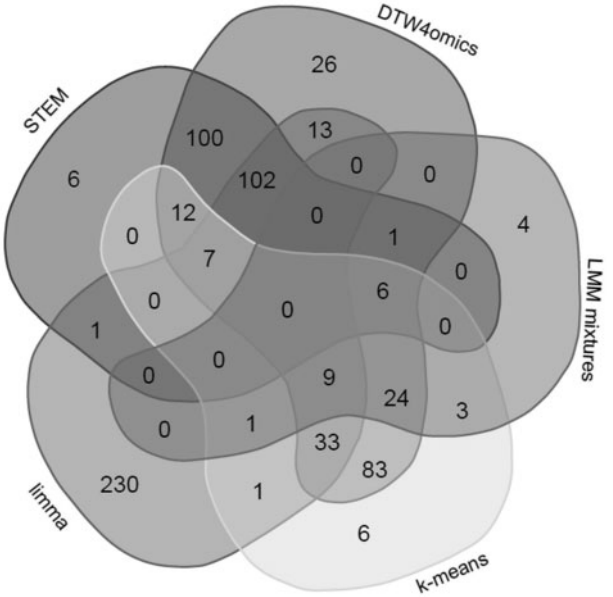

Fig. 1. Simulated data (reference dataset)—variables related to variable 1000 for each of the five pattern recognition methods

followed by STEM and DTW4omics. The accuracy of the five methods for the simulated datasets is shown in Table 3. Accuracies above 80% were found for all methods after noise had been added to the data, except for limma, where accuracies for datasets with 25% noise were between 74% and 81%. Limma and k-means appeared the most influenced by the number of time points. Lowering the number of time points lowered the accuracy significantly. STEM was most influenced by the number of biological replicates available.

3.2 Real datasets

3.2.1 Response of HepG2 to B(a)P

Lists of genes, pathways and GO terms for each endpoint are provided in the Supplementary Excel file for each of the five methods. Applying STEM, DTW4omics and limma resulted in a separate gene list for each phenotypic endpoint. For LMM mixtures and k-means, some of the clusters contained more than one phenotypic endpoint:

- DNA adducts and G1 were assigned to the same cluster by both LMM mixtures and k-means;
- Apoptosis and S are assigned to the same cluster by LMM mixtures;
- Apoptosis, G2 and S are assigned to the same cluster by k-means.

For DTW4omics and limma, some of the gene lists had genes in common (see Supplementary Excel file, tab 'B(a)P intersections gene lists'). The other methods by definition divide a dataset into clusters that do not have genes in common.

LMM mixtures divided the dataset into 11 clusters (based on the BIC), while k-means only distinguished 7 clusters (based on the SSE). Apart from a few exceptions, there was minor overlap across the five methods when comparing lists of genes, pathways and GO terms (Supplementary Data, Supplementary Figs. S2–S16). Figures 2 and 3 present a summary of the molecular response pathways related to DNA adducts for the five methods. Detailed tables with pathway lists for all endpoints, including references to prior biological knowledge, are available in the Supplementary Data (Supplementary Tables S2–S11). The pathway 'direct p53 effectors'

Table 3. Accuracy of the five methods for the simulated datasets

	STEM	DTW	LMM	k-Means	Limma
15% Noise	0.83–0.96	0.93–0.94	0.91–0.93	0.80–0.90	0.80–0.81
20% Noise	0.88–0.95	0.90–0.92	0.92–0.93	0.89–0.98	0.80–0.81
25% Noise	0.82–0.94	0.87–0.93	0.90–0.94	0.89–0.98	0.74–0.81
8 Time points	0.87	0.87	0.95	0.80	0.76
6 Time points	0.87	0.79	0.91	0.66	0.68
2 Biological replicates	0.77	0.88	0.93	0.90	0.89

Note: The range of the accuracy for the influence of noise is based on three experiments per noise level.

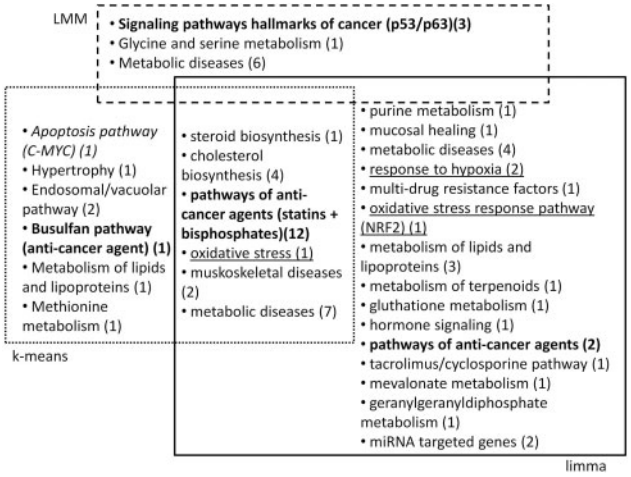

Fig. 2. Response to B(a)P (dataset van Delft *et al.*, 2010)—pathways positively related to DNA adducts (ConsensusPathDB, FDR ≤ 0.05) for LMM mixtures, k-means and limma. Number of pathways of each kind between brackets. No pathways were found by STEM and DTW4omics. Bold, related to cancer; italic, apoptotic pathway; underlined, related to oxidative stress

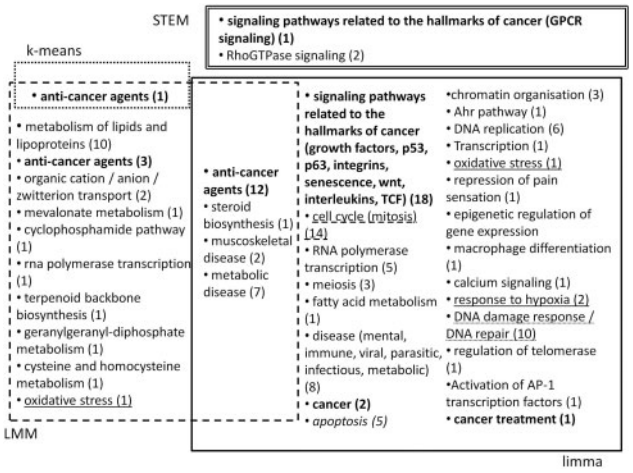

Fig. 3. Response to B(a)P (dataset van Delft *et al.*, 2010)—pathways negatively related to DNA adducts (ConsensusPathDB, FDR ≤ 0.05) for STEM, LMM mixtures, k-means and limma. Number of pathways of each kind between brackets. No pathways were found for DTW4omics. Bold, related to cancer; italic, apoptotic pathway; underlined, related to oxidative stress; underlined (dashed), (mitotic) cell cycle; underlined (dotted), DNA damage response/ DNA repair

was found by both LMM mixtures (positive relationship, see Fig. 2 and Supplementary Table S2) and limma (negative relationship, see Fig. 3 and Supplementary Table S3). Transcription factor p53 is important in signaling DNA damage (Hanahan and Weinberg, 2000). Limma also found 10 pathways related to DNA damage response (Fig. 3 and Supplementary Table S3).

3.2.2 Response of HepG2 to menadione

Lists of genes, pathways and GO terms for each endpoint are provided in the Supplementary Excel file for each of the five methods. Apart from a few exceptions, there was minor overlap among all of the five methods when comparing lists of genes, pathways and GO terms (see Supplementary Data, Supplementary Figs. S17–S34). Applying DTW4omics and limma resulted in a separate gene list for each phenotypic endpoint, except for the G1 and S phases wherefore DTW4omics did not find genes that were positively and negatively related, respectively. For STEM, LMM mixtures and k-means, some of the clusters contained more than one phenotypic endpoint:

- apoptosis and protein oxidation were assigned to the same cluster by both STEM and k-means;
- protein oxidation, G1 and S were assigned to the same cluster by LMM mixtures;
- G1 and S were assigned to the same cluster by k-means.

For DTW4omics and limma, some of the gene lists had genes in common (see Supplementary Excel file, tab 'MEN intersections gene lists'). The other methods by definition divide a dataset into clusters that do not have genes in common.

LMM mixtures divided the dataset into 10 clusters (based on the BIC), while k-means only distinguished four clusters (based on the SSE). Figures 4 and 5 present a summary of the molecular response pathways related to oxidative DNA damage for the five methods. Detailed tables with pathway lists for all endpoints, including references to prior biological knowledge, are available in the Supplementary Data (Supplementary Tables S12–S23). The pathway 'direct p53 effectors', an important pathway in signaling DNA damage (Hanahan and Weinberg, 2000), was found by LMM mixtures (positive relationship, see Fig. 2 and Supplementary Table S12). Furthermore, LMM mixtures found two DNA damage response pathways (Fig. 2 and Supplementary Table S12).

3.3 Computational time

The computational time for running the analyses was about 10 s for STEM and limma, about 30–40 min for DTW4omics and about 30 min for k-means (including the algorithm for determining the number of clusters) on a 64 bit Windows 7 computer, 64 GB memory, intel i5-3320 M processor. Running LMM mixtures takes 1 day for a given number of clusters, so it would, e.g., take 24 days to generate the cluster solutions with 2, 3, . . . , 25 clusters. Therefore, the calculations for different numbers of clusters were run in parallel on a 64-bit Windows 7 server, 64 GB memory, intel i7-3930 K(hexacore).

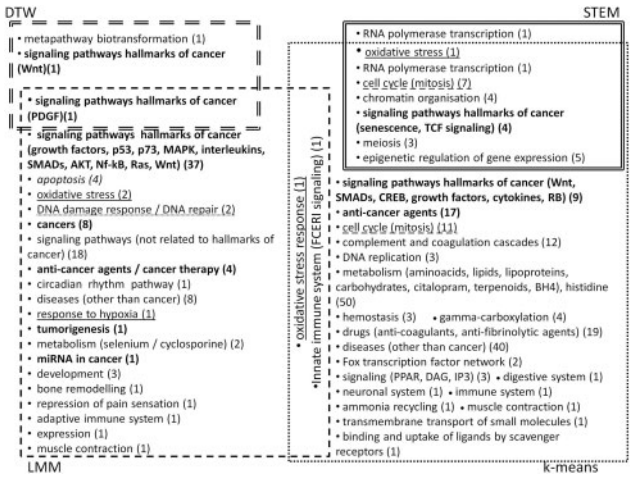

Fig. 4. Response to menadione (dataset Deferme *et al.*, 2013)—pathways positively related to oxidative DNA damage (ConsensusPathDB, FDR ≤ 0.05) for STEM, DTW4omics, LMM mixtures and k-means. Number of pathways of each kind between brackets. No pathways were found by limma. Bold, related to cancer; italic, apoptotic pathway; underlined, related to oxidative stress; underlined (dashed), (mitotic) cell cycle; underlined (dotted), DNA damage response/DNA repair

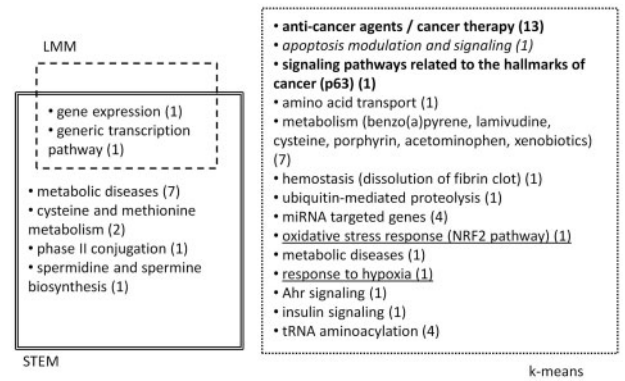

Fig. 5. Response to menadione (dataset Deferme *et al.*, 2013)—pathways negatively related to oxidative DNA damage (ConsensusPathDB, FDR ≤ 0.05) for STEM, LMM and k-means. Number of pathways of each kind between brackets. No pathways were found by DTW4omics and limma. Bold, related to cancer; underlined, related to oxidative stress

4 Discussion

Applying pattern recognition methods to toxicogenomics time series of both microarrays and phenotypic endpoints can provide new hypotheses about interactions of the genome with the adverse outcome of exposure to a toxicant. These new hypotheses will guide wet laboratory experiments (e.g. knock down experiments) that aim to verify such new relationships.

In this study, we evaluated five pattern recognition methods. The main focus is on methods developed for time series methods, except k-means, which we added to the study in order to assess the impact of time dependencies on pattern recognition. Another method for static data, that has been frequently applied to time series, is Weighted Correlation Network Analysis (WGCNA) (Elo and Schwikowski, 2013), which combines correlation analysis with hierarchical clustering (Langfelder and Horvath, 2008). Clusters obtained by applying static-based methods on time series are often less biologically coherent than clusters determined with time-series methods (Bar-Joseph *et al.*, 2012). However, this could not be observed from this study.

STEM, LMM mixtures and k-means treat all the genes and endpoints as equivalent entities and cluster them all together. As a consequence several endpoints may end up in the same cluster and some clusters will have no endpoints. In our study, having several endpoints within the same cluster was found upon applying STEM, LMM mixtures and k-means. Having two or more endpoints within the same cluster has the advantage that similarity between the two endpoints is explicitly known. A drawback of having two endpoints in the same cluster is that one cannot really assess whether any particular gene in the cluster is more strongly associated with one endpoint than the other. Another drawback is that one can have only certainty that the gene expressions really match the profile of the endpoint if the endpoint is at the center of the cluster.

DTW4omics treats the endpoints as profiles and matches each of the genes to these profiles. Limma uses the endpoints as fixed effects in a regression model. As a consequence, for DTW4omics and limma, each cluster has only one endpoint and a gene can match more than one endpoint or none.

Clustering methods that do not correct for random patterns, like LMM mixtures and k-means divide the datasets into large clusters. This mostly leads to a large amount of pathways related to a large number of biological processes. Correcting for random patterns (like STEM) results in smaller gene lists, which has the advantage that these lists are mostly easier to interpret. A disadvantage is that if two variables are highly correlated to one of the model profiles by coincidence, the two variables would also be in the same cluster, which leads to false positives. Another drawback is that in cases when the variable is not correlated to one of the model profiles, it is not taken into account for further analysis. In this way information may be lost. While STEM, LMM mixtures, k-means and limma find gene expression modifications that covary simultaneously with the investigated endpoints, DTW4omics allows time lags and delays. Because of this property results from applying the DTW4omics approach are more influenced by the number of time points than the two other time series methods (Table 3). Limma is more sensitive to noise than the other methods (Table 3).

LMM mixtures and limma need biological replicates (for modeling the random effects), while STEM, DTW4omics and k-means in principle can be applied without having replicates. However, this is not recommended because it decreases the accuracy of STEM, and to a lesser extent, of DTW4omics and limma (Table 3).

STEM and DTW4omics both use a permutation test for calculating significance. Permutation tests assume independence between subsequent time points, which can be disadvantageous in case clear trends are observed in the data (Xia *et al.*, 2013).

LMM mixtures clustering and limma take into account variability between replicates. The consequence of having a low number (2–3) of replicates is that variability (the random effect in the model) can be overestimated or underestimated. For LMM mixtures, underestimation (respectively, overestimation) of the variability between replicates leads to clustering too strictly (respectively, not strict enough), and as a consequence, may generate false negatives (respectively, false positives). Another disadvantage of LMM mixtures, and of model-based clustering in general, is that the algorithm is slow compared with other methods. For limma, underestimation (respectively, overestimation) of the variability between replicates has influence on variance shrinkage and leads to decreased (respectively, increased) sensitivity to detect relationships between genes and phenotypic endpoints, which in its turn leads to false negatives (respectively, false positives).

For all five methods studied, pathway and GO analysis of these gene lists provided pathways and GO terms that have been previously related with exposure to the particular toxicant or with one of its adverse outcomes (see Supplementary Material).

The hallmarks of cancer (Hanahan and Weinberg, 2000, 2011) review biological processes related to all types of cancer, and the signaling pathways involved in those biological processes. Those hallmarks are: (1) sustaining proliferative signaling; (2) evading growth suppressors; (3) avoiding immune destruction; (4) enabling replicative immortality; (5) tumor-promoting inflammation; (6) activating invasion and metastasis; (7) inducing angiogenesis; (8) genome instability and mutation; (9) resisting cell death and (10) deregulating cellular energetics. Because B(a)P is a carcinogen and menadione causes hepatocellular carcinoma, pathways known to be involved in the hallmarks of cancer provide biologically relevant hypotheses to test in follow-up experiments. For B(a)P, in particular signaling pathways related to DNA damage response are relevant (Bolotina *et al.*, 2007). Examples are signaling pathways involving p53, NF-kB and MYC (Bolotina *et al.*, 2007; Hanahan and Weinberg, 2000). For menadione, pathways related to oxidative stress provide relevant hypotheses. Signaling pathways influenced by oxidative stress given in Martindale and Holbrook (2002) include pathways involved in growth arrest, cell proliferation, senescence and apoptosis. All five methods provide biologically relevant pathways and GO terms that are related to the exposure to the toxicant or its adverse outcomes (e.g. cancer, oxidative stress) (see Figs. 2–5 and Supplementary Material, Section 3). Therefore, in terms of retrieving biologically relevant information, there is no method that outperforms the others.

If we compare the results of limma, the only regression-based model in the study, with the results for the other (correlation- and distance-based) methods, we observe the following. LMM mixtures positively relate the pathway 'direct p53 effectors' to DNA adducts, while limma finds a negative relationship. Four genes of 'direct p53 effectors' are in both the genes lists resulting from LMM mixtures and limma analysis (PCNA, TNFRSF10B, LIF and BAX), which means that these four genes have a positive correlation with DNA adducts, but a negative regression coefficient. This can be explained as follows. The regression coefficient of the limma model gives the relationship between gene expression and DNA adducts, when all other variables in the model (treatment, time) are held constant. This means that the positive relationship observed by determining similarities between the time profiles was due to another (confounding) variable. This shows an advantageous property of limma, namely correcting for confounding factors.

For the menadione dataset (dataset Deferme *et al.*, 2013), limma cannot find any pathways related to oxidative DNA damage, while all other methods generate a pathway list. For this dataset, applying DTW4omics results in detecting some pathways that cannot be found without time warping. One of these pathways is related to the hallmarks of cancer (Fig. 4). This shows the relevance of having a method taking into account time delays.

Applying the methods for phenotypic anchoring described in this article results in poorly overlapping gene lists. This is a general problem of clustering algorithms. There are several reasons for the lack of consistency between methods. First, parameters (optimal number of clusters, thresholds) are determined by statistical means and not on biological properties (Swift *et al.*, 2004). Second, differences in output are due to different intrinsic properties of these methods. Limma is a regression method, while the other methods are based on similarity measures (correlation, distance). Other differences in intrinsic properties (taking into account time dependencies, delays, variance; correcting for random patterns) are shown in Table 1. For static-based methods, several attempts have already been undertaken to address these issues. WGCNA selects thresholds based on scale-free topology, a network property previously observed in biological networks (Zhang and Horvath, 2005). Consensus clustering attempts to combine the different approaches, capturing the advantageous properties of each methods (Swift *et al.*, 2004). However, to our knowledge, similar attempts have not been undertaken for methods developed for time series, taking into account time dependencies and delays.

In summary, we conclude that all five methods are suitable for extracting new hypotheses concerning gene-phenotypic endpoint relationships and none of the five methods outperforms the others in terms of biology. Furthermore, all methods have their limitations. Because these methods provide complementary results, we recommend developing a method that integrates the results from the different methods. A possible way to do this is to calculate a weighted score for the probability of each gene–endpoint relationship based on the accuracy of the methods. Simulation experiments can guide the choice of the weight for each method. For example, because STEM was highly influenced by the number of biological replicates, we decrease the weight for STEM when having less replicates. In a similar way, we can adapt the weight of k-means and limma according to the number of time points. We can then multiply the false discovery rate for each pathway with this probability score in order to correct for inaccuracies due to the low number of replicates and/or time points.

Acknowledgements

The authors like to thank Joost van Delft and Lize Deferme (Maastricht University) for providing additional information on the B(a)P and menadione datasets.

Funding

This work was supported by the Research Data Alliance (RDA) Europe, a part of the EU Seventh Framework Programme, under grant agreement number 312424.

Conflict of Interest: none declared.

References

Alm,H. *et al.* (2009) In vitro neurotoxicity of PBDE-99: immediate and concentration-dependent effects on protein expression in cerebral cortex cells. *J. Proteome Res.*, **9**, 1226–1235.

Bar-Joseph,Z. (2004) Analyzing time series gene expression data. *Bioinformatics*, **20**, 2493–2503.

Bar-Joseph,Z. *et al.* (2012) Studying and modelling dynamic biological processes using time-series gene expression data. *Nat. Rev. Genet.*, **13**, 552–564.

Bolotina,N.A. *et al.* (2007) Benzo[a]pyrene-dependent activation of transcription factors NF-kappaB and AP-1 related to tumor promotion in hepatoma cell cultures. *Biochemistry*, **72**, 552–557.

Bostrom,C.E. *et al.* (2002) Cancer risk assessment, indicators, and guidelines for polycyclic aromatic hydrocarbons in the ambient air. *Environ. Health Perspect.*, **110**, 451–488.

Briede,J.J. *et al.* (2010) Global gene expression analysis reveals differences in cellular responses to hydroxyl- and superoxide anion radical-induced oxidative stress in caco-2 cells. *Toxicol. Sci.*, **114**, 193–203.

Cavill,R. *et al.* (2013) DTW4Omics: comparing patterns in biological time series. *PLoS one*, **8**, e71823.

Celeux,G. *et al.* (2002) Mixture of linear mixed models. Application to repeated data clustering. *Inria Research Report 4566*. https://hal.inria.fr/inria-00072022/document (9 March 2015, date last accessed)

Celeux,G. *et al.* (2005) Mixture of linear mixed models for clustering gene expression profiles from repeated microarray experiments. *Stat. Model.*, **5**, 243–267.

Davis,A.P. *et al.* (2013) The comparative toxicogenomics database: update 2013. *Nucleic Acids Res.*, **41**, D1104–D1114.

de Ridder,D. *et al.* (2013) Pattern recognition in bioinformatics. *Brief Bioinform.*, **14**, 633–647.

Deferme,L. *et al.* (2013) Time series analysis of oxidative stress response patterns in HepG2: a toxicogenomics approach. *Toxicology*, **306**, 24–34.

Elo,L.L. and Schwikowski,B. (2013) Analysis of time-resolved gene expression measurements across individuals. *PLoS one*, **8**, e82340.

Ernst,J. and Bar-Joseph,Z. (2006) STEM: a tool for the analysis of short time series gene expression data. *BMC bioinformatics*, **7**, 191.

Ernst,J. *et al.* (2005) Clustering short time series gene expression data. *Bioinformatics*, **21**(Suppl. 1), i159–i168.

Ganter,B. *et al.* (2008) Pathway analysis tools and toxicogenomics reference databases for risk assessment. *Pharmacogenomics*, **9**, 35–54.

Giorgino,T. (2009) Computing and visualizing dynamic time warping alignments in R: the dtw package. *J. Stat. Softw.*, **31**, 1–24.

Grün,B. and Leisch,F. (2008) FlexMix version 2: finite mixtures with concomitant variables and varying and constant parameters. *J. Stat. Softw.*, **28**, 1–35.

Hanahan,D. and Weinberg,R.A. (2000) The hallmarks of cancer. *Cell*, **100**, 57–70.

Hanahan,D. and Weinberg,R.A. (2011) Hallmarks of cancer: the next generation. *Cell*, **144**, 646–674.

Hebels,D.G. *et al.* (2013) Performance in omics analyses of blood samples in long-term storage: opportunities for the exploitation of existing biobanks in environmental health research. *Environ. Health Perspect*, **121**, 480–487.

Kamburov,A. *et al.* (2013) The ConsensusPathDB interaction database: 2013 update. *Nucleic Acids Res.*, **41**, D793–D800.

Kintigh,K.W. and Ammerman,A.J. (1982) Heuristic approaches to spatial analysis in archaeology. *Am. Antiq.*, **47**, 31–63.

Langfelder,P. and Horvath,F. (2008) WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics*, **9**, 559.

Leisch,F. (2004) FlexMix: a general framework for finite mixture models and latent class regression in R. *J. Stat. Softw.*, **11**, 1–18.

Liao,T.W. (2005) Clustering of time series data—a survey. *Pattern Recogn.*, **38**, 1857–1874.

Martindale,J.L. and Holbrook,N.J. (2002) Cellular response to oxidative stress: signaling for suicide and survival. *J. Cell Physiol.*, **192**, 1–15.

Mendes,P. *et al.* (2003) Artificial gene networks for objective comparison of analysis algorithms. *Bioinformatics*, **19**(Suppl. 2), ii122–ii129.

Ng,A.Y. *et al.* (2001) On spectral clustering: analysis and an algorithm. In: *Proceedings of Advances in Neural Information Processing Systems*, Vol. 14, MIT Press, Cambridge, MA, pp. 849–856.

Nykter,M. *et al.* (2006) Simulation of microarray data with realistic characteristics. *BMC Bioinformatics*, **7**, 349.

Powell,C.L. *et al.* (2006) Phenotypic anchoring of acetominophen-induced oxidative stress with gene expression profiles in rat liver. *Toxicol. Sci.*, **93**, 213–222.

Ritchie,M.E. *et al.* (2015) Limma powers differential expression analysis for RNA-sequencing and microarray studies. *Nucleic Acids Res.*, [Epub ahead of print, doi:10.1093/nar/gkv007].

Scharl,T. *et al.* (2010) Mixtures of regression models for time course gene expression data: evaluation of initialization and random effects. *Bioinformatics*, **26**, 370–377.

Smith,A.A. *et al.* (2009) Clustered alignments of gene-expression time series data. *Bioinformatics*, **25**, i119–i127.

Swift, *et al.* (2004) Consensus clustering and functional interpretation of gene-expression data. *Genome Biol.*, **5**, R94.

Tan,Y. *et al.* (2006) Integrating time-course microarray gene expression profiles with cytotoxicity for identification of biomarkers in primary rat hepatocytes exposed to cadmium. *Bioinformatics*, **22**, 77–87.

Truong,J.T. and Booth,S.L. (2011) Emerging issues in vitamin K research. *J. Evid. Based Complementary Altern. Med.*, **16**, 73–79.

van Delft,J.H. *et al.* (2010) Time series analysis of benzo[A]pyrene-induced transcriptome changes suggests that a network of transcription factors regulates the effects on functional gene sets. *Toxicol. Sci.*, **117**, 381–392.

Xia,L.C. *et al.* (2013) Efficient statistical significance approximation for local similarity analysis of high-throughput time series data. *Bioinformatics*, **29**, 230–237.

Zeileis,A. *et al.* (2004) Kernlab-an S4 package for kernel methods in R. *J. Stat. Softw.*, **11**, 1–20.

Zhang,B. and Horvath,S. (2005) A general framework for weighted gene co-expression network analysis. *Stat. Appl. Genet. Mol.*, **4**, 1.