Computational Toxicology 30 (2024) 100313

Contents lists available at ScienceDirect

Computational Toxicology

journal homepage: www.sciencedirect.com/journal/computational-toxicology

Full Length Article

MoS-TEC: A toxicogenomics database based on model selection for time-expression curves

Franziska Kappenberg *, Benedikt Küthe, Jörg Rahnenführer

Department of Statistics, TU Dortmund University, Vogelpothsweg 87, 44227 Dortmund, Germany

ARTICLE INFO

Edited by Dr. Daniel Dietrich

Keywords:
Time-response modelling
Gene expression data
MCP-Mod
Open TG-GATEs

ABSTRACT

MoS-TEC is a newly developed toxicogenomics database for time-expression curves fitted with a statistical model selection approach. Toxicogenomic data provide information on the response of the genome to compounds, often measured in terms of gene expression values. When such experimental data are available for different exposure times, the functional relationships between the exposure time and the expression values of genes might be of interest. The TG-GATEs (Open Toxicogenomics Project-Genomics Assisted Toxicity Evaluation System) database provides such information for genomewide gene expression data for 170 compounds. We performed extensive model selection using MCP-Mod on these data. Specifically, gene expression data measured for eight time points from in vivo experiments on rat liver for 120 compounds with complete datasets were considered. MCP-Mod is a two-step approach, including a multiple comparison procedure (MCP) and a modelling (Mod) approach. The results are estimated time-expression curves that model the relationship between exposure time and gene expression values for all combinations of genes and compounds. We present an appropriate data normalization approach and report which models were selected per compound and in total. For high-quality model fits with a large value for the explained variance, the sigEmax model was most frequently selected. The new R Shiny application MoS-TEC provides easy access for researchers to the best curve fit for all genes individually for all compounds. It can be used online without installing additional software.

1. Introduction

In preclinical and toxicological research, knowledge about the properties of compounds is often derived from data obtained from animal or cell experiments. In the field of toxicogenomics, it is investigated how the genome responds to compounds, often measured in terms of gene expression or protein expression [25]. Molecular technologies such as Microarray or RNA-Seq technology [7, e.g.] allow to measure the gene expression in cells simultaneously genomewide for tens of thousands of genes.

A frequent goal is to obtain information about the onset of a possibly adverse effect of a specific compound, depending on the value of a certain condition of interest [15]. As condition, in toxicology, mostly the concentration or the dose of the compound is used, but also other conditions are of interest, such as the exposure time of the cells to the compound. The standard approach then is to measure several values of the condition of interest and to model the condition-response relationship. From a statistical point of view, the type of condition does not influence the curve modelling approach. Therefore, in this paper terms such as 'dose–response analysis' or 'alert concentration' can also refer to other condition types, especially exposure time, i.e. we will also use the terms 'time-response analysis' or 'alert time'.

The Open TG-GATEs[1] database [13] contains genomewide gene expression data for 170 different compounds, measured in vivo in rats and additionally in vitro in human and rat hepatocytes. For the in vivo measurements, for all compounds, three different doses and eight (exposure) times are given, such that the gene expression profiles can be modelled as a function of time.

* Corresponding author.
E-mail address: kappenberg@statistik.tu-dortmund.de (F. Kappenberg).

[1] Abbreviations: MCP-Mod: Multiple comparison procedure and modelling, MCP: multiple comparison procedure, Mod: modelling, Open TG-GATEs: Open toxicogenomics project-genomics assisted toxicity evaluation system, RMA: Robust multiarray average, PCA: Principal component analysis, PC: principal component, PoC: proof of concept, signifMCP: Significant result of the multiple comparison procedure, MoS-TEC: Model Selection for Time-Expression Curves applied to compounds from the TG-GATEs database', or short 'Model Selection Time-Expression Curves', AOP: Adverse outcome pathway, OECD: Organisation for Economic Cooperation and Development, AIC: Akaike information criterion, FDA: food and drug administration, ADHD: attention deficit hyperactivity disorder.

https://doi.org/10.1016/j.comtox.2024.100313

Available online 8 May 2024
2468-1113/© 2024 The Author(s). Published by Elsevier B.V. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

Data from the TG-GATEs database have recently been used in several applications, e.g. for validation of a newly developed multi-dose model to predict drug-induced hepatotoxicity based on gene expression and toxicity data [33], or for identifying a set of 1000 genes that are highly responsive to chemical exposures [31]. While these works make use of the large set of compounds available in the TG-GATEs database, data from selected individual compounds can also be used as an example dataset to demonstrate the application of a new method, as done in Serra et al. [30].

The approach to model concentration–response curves has many advantages, such as being able to interpolate between actually measured concentrations and making use of the entire concentration–response profile instead of considering response values only concentration-wise. However, for gene expression data, the approach to consider the concentrations individually is still often preferred in the published literature [17]. When using the modelling approach, a major challenge is the choice of an appropriate model. A combination of relevance testing of some concentration–response effect and model selection is provided by the MCP-Mod (Multiple comparison procedure and modeling) approach, originally introduced by Bretz et al. [6]. It has been shown that this approach is also suitable for high-dimensional gene expression data [11].

In a two-step approach, it is first determined whether the gene exhibits a non-flat profile, and then suitable models are fitted and compared. In more detail, the MCP-Mod approach is a two-step procedure originally designed for Phase II clinical trials. In the first step, for a pre-specified set of candidate models (for modelling the dose–response relationship), the proof of concept (PoC) is established via a multiple comparison procedure. This means that it is identified which statistical models for describing the dose–response relationship better fit the data than a flat profile. Only for the candidate models that pass this PoC-step, the respective model parameters are estimated, and either model selection based on an information criterion or model averaging is performed. Since in the analysis discussed here, genomic data is considered, the terminology PoC does not match the common usage. Thus, we refer to a significant result of the multiple comparison procedure as 'signifMCP', in line with Duda et al. [11].

In this paper, we provide a detailed analysis of the gene expression profiles for a large number of compounds in the TG-Gates database, where the profile is a function of the (exposure) time. Our goal is to provide comprehensive insights for a huge number of genes, treated with different compounds, over several time points. We proceed in three steps. We explain the results for specific selected compounds, give an overview of results for all compounds, and present our new publicly freely available database MoS-TEC ('Model Selection for Time-Expression Curves applied to compounds from the TG-GATEs database', or short 'Model Selection Time-Expression Curves'). In MoS-TEC, for each compound-gene combination, fitted time-expression curves can be viewed and downloaded. We also discuss meaningful and necessary pre-processing of the data from the TG-GATEs database.

The database MoS-TEC can be assessed via a user-friendly R Shiny application [9], which is also called MoS-TEC.[2] It illustrates the compound-wise results via PCA plots and presents the results of the MCP-Mod procedure in a summarized way. The main application of the app is to view and download resulting time-expression curves for individual genes of interest. As output for a combination of a gene and a compound, only the model selected by the MCP-Mod methodology or all models that passed the signifMCP step can be plotted.

Based on the fitted time-response models, it is possible to calculate a variety of alerts based on a pre-defined effect level. Examples are different versions of the minimum effective dose introduced in the context of MCP-Mod [6], effective times [28], (absolute) lowest effective times [16; 18; 24], or benchmark times [15]. However, in this work, the focus is on the appropriate pre-processing and on the modelling of the time-response relationship itself, not on the derivation of alerts.

The MoS-TEC application provides detailed information on the time-response profile for individual genes. Its use is thus mostly given in applications where some genes are pre-selected based on biological considerations, and their properties after treatment with different compounds are of interest. One such area of application are adverse outcome pathways (AOP) [2]. An AOP is a concept, where within a sequence of events, beginning with a molecular initiating event, and progressing via key events, an adverse outcome is reached. A main purpose of AOPs is to gain insight on the connection between mechanisms of toxicity and the adverse outcome [19]. A summary of all currently characterized AOPs is given in the AOP knowledge base maintained by the OECD (Organisation for Economic Co-operation and Development) (https://aopkb.oecd.org/). Rooney et al. [29] and Corton et al. [10] analysed gene expression measurements for several compounds from the TG-GATEs database to identify rat liver tumorigens in short-term assays. The authors considered specific molecular initiating events that were only based on gene expression biomarkers and key events.

The remainder of the paper is structured as follows. First, the TG-GATEs database and the corresponding datasets used for modelling curves are introduced in detail. The MCP-Mod methodology is explained, and the specific parameter choices and adaptions required for the given datasets are discussed. The results of the analysis are first presented in terms of the necessary normalization steps. Then, some general results of the MCP-Mod procedure on 120 compounds from the database are presented, and for four specifically chosen compounds, the results are discussed in detail. Finally, the newly developed Shiny application 'MoS-TEC' that allows to individually perform the analyses for all compounds is introduced.

2. Materials and methods

This section presents the Open TG-GATEs database and, in particular, the in vivo gene expression datasets that are the basis of the MoS-TEC database. The initial analysis of the dataset, which consists of preprocessing and some descriptive analyses, is explained, as well as the MCP-Mod method along with the selected candidate models.

2.1. The Open TG-GATEs database

The Open TG-GATEs (Open Toxicogenomics Project-Genomics Assisted Toxicity Evaluation System) is a freely available large-scale and high-quality toxicogenomics database [13]. It contains gene expression profiles derived from in vivo (rat) and in vitro (primary rat hepatocytes, primary human hepatocytes) exposure to 170 compounds at multiple doses and (exposure) time points. The compounds include pharmaceuticals and chemicals demonstrated to produce liver and kidney damage. For the in vivo studies, only performed in rats, the gene expression data for a test compound were obtained from single compound administration at up to four doses (control, low, middle, high) and eight time points. The time points correspond to four single-dose studies (where the animals were treated only once and then sacrificed after 3, 6, 9, or 24 h) and four repeated-dose studies (where the animals received one dose per day and were sacrificed after 4, 8, 15, or 29 days). For the in vivo studies, the concentrations for the low, middle, and high doses were chosen in a ratio of 1:3:10. For the in vitro studies, the ratio of the concentrations was set as 1:5:25. The gene expression (microarray) analysis was performed using the Rat Genome 230 2.0 Array and the Human Genome U133 Plus 2.0 Array.

2.2. Initial analysis of the datasets

In the following, only data from in vivo experiments on rat liver and only compounds that are complete in terms of doses and time points are

[2] http://shiny.statistik.tu-dortmund.de:8080/app/MoS-TEC.

considered. The number of complete datasets is 120. The handling, analysis, and visualization of the data was performed in R, version 4.1.0 [26]. The microarray data were preprocessed with the RMA (robust multiarray average) algorithm, which applies a background correction, quantile normalization, and a $\log_2$-transformation [14]. This was done using the additional software packages `affy` [12] and `rat2302.db` [8].

The considered datasets contain 31,099 probe sets, respectively. For each probe set (in the following referred to as gene for simplicity), $N = 96$ measurements are available, consisting of three measurements for each of the four doses and eight time points. In this work, only the control and the high dose are considered for the time-response modelling, reducing the dataset to $N = 48$ samples per compound.

For the descriptive analysis of the data, principal component analysis (PCA) is performed in four scenarios, which are distinguished by different filters and transformations according to specific criteria. For the first scenario, all 31,099 genes are used for the calculation of the PCA. These data are referred to as untreated data in the following. For the second scenario, the variance is calculated for each gene across all replicates, doses, and time points, and instead of all 31,099 genes only the 1000 genes with the greatest variance are considered. For the third PCA, the gene expression data are normalized with respect to the control values. For each time point separately, normalization is done by calculating per gene the arithmetic mean of the expression values for the control dose, and subtracting genewise this mean from the single values for the high dose. The resulting mean values thus are (normalized) high dose values of the respective compound for each time point. The control values are then not considered further and are removed from the data set. These data are referred to as normalized data in the following. In the fourth scenario, again, only the 1000 genes with the greatest variance are considered, but now for the normalized data. For all four scenarios, the first two principal components are used for plotting and interpreting the PCA results.

2.3. MCP-Mod procedure

In this section, the Multiple Comparison Procedure and Modelling (MCP-Mod) approach is presented, for details see also Bornkamp et al. [4]. In this work, the effect is not modeled as a function of concentration or dose, but of exposure time. Therefore, in the following the term time-response model is used instead of dose–response model.

MCP-Mod is an approach for the investigation of time-response relationships. The method consists of a multiple comparison procedure (MCP) and a modeling (Mod) approach. In the MCP step, an optimal contrast test is used to determine whether a gene has a statistically significant signal for a predefined set of candidate models. If a significant signal is detected for at least one model, a so-called signifMCP (significant result of the multiple comparison procedure) is present and the Mod step is performed. For the set of models for which a signifMCP is available, either the 'best' model is selected based on the Akaike information criterion (AIC), or model averaging over the set of significant models is performed, both resulting in an estimate of the time-response profile.

Given are k time points $t_1 < \ldots < t_k$ with n_i replicates. A model equation of the preprocessed gene expression value for time point $i = 1, \ldots, k$ and replicate $j = 1, \ldots, n_i$ is given by

$$Y_{ij} = \mu_{t_i} + \epsilon_{ij}, \ \epsilon_{ij} \sim \mathcal{N}(0, \sigma^2), \tag{1}$$

with homogeneous variance $\sigma^2 > 0$. The mean response $E(Y_{ij}) = \mu_{t_i} = f(t_i, \boldsymbol{\theta})$ at time point t_i is assumed to follow a time-response model with parameter vector $\boldsymbol{\theta}$ and ϵ_{ij} as independent errors. The total sample size is $N = \sum_{i=1}^{k} n_i$. For the MCP step, a set $\mathcal{M}$ of M models is specified. An overview of the models and their parameters that are used in this work is given in Table 1. The models can be reformulated as

Table 1
Time-response models $f(t, \boldsymbol{\theta})$, their standardized versions $f^0(t, \boldsymbol{\theta}^0)$, and the guesstimates for $\boldsymbol{\theta}^0$ for the analysis. For the beta model, B is defined as $B(\delta_1, \delta_2) = (\delta_1 + \delta_2)^{\delta_1+\delta_2}/\delta_1^{\delta_1}\delta_2^{\delta_2}$ and $D = 696$. For the linlog model, the offset c is not estimated but set to $0.01 \times$ maximal time point, which is 6.96 in our case, to avoid taking the logarithm of 0.

Model	$f(t, \boldsymbol{\theta})$	$f^0(t, \boldsymbol{\theta}^0)$	$\boldsymbol{\theta}^0$
linear	$E_0 + \delta t$	t	$\varnothing$
linlog	$E_0 + \delta \log(t + c)$	$\log(t + c)$	$\varnothing$
quadratic	$E_0 + \beta_1 t + \beta_2 t^2$	$t + (\beta_2/\lvert\beta_1\rvert)t^2$	$\delta = \beta_2/\lvert\beta_1\rvert = -0.0014$
exponential	$E_0 + E_1(\exp(t/a) - 1)$	$\exp(t/a) - 1$	$a = 1391.9999$
emax	$E_0 + E_{max} t/(ET_{50} + t)$	$t/(ET_{50} + t)$	$ET_{50} = 9$
Beta	$E_0 + E_{max} B(a_1, a_2) \cdot (t/D)^{a_1}(1 - t/D)^{a_2}$	$B(a_1, a_2)(t/D)^{a_1}(1 - t/D)^{a_2}$	$a_1 = 2, a_2 = 1$
sigEmax	$E_0 + E_{max} t^h/(ET_{50}^h + t^h)$	$t^h/(ET_{50}^h + t^h)$	$ET_{50} = 9, h = 0.8043$

$$f(t, \boldsymbol{\theta}) = \theta_0 + \theta_1 f^0(t, \boldsymbol{\theta}^0) \tag{2}$$

(see second column of Table 1), where $f^0(t, \boldsymbol{\theta}^0)$ is the standardized version of a model, parameterized by the vector $\boldsymbol{\theta}^0$. Here θ_0 is a location and θ_1 is a scale parameter, so only the parameter vector $\boldsymbol{\theta}^0$ determines the shape of the model function. The concept of standardized models is important for the choice of optimal contrasts.

In addition, initial estimates, so-called guesstimates, are determined for the parameter vector $\boldsymbol{\theta}^0$. In practice, these are translated into estimates for $\boldsymbol{\theta}^0$ through careful consideration and prior knowledge of the expected percentages of maximal effects at specific time points. Because of the large number of genes, an individual choice of guesstimates per gene is practically infeasible. Therefore, the same guesstimates are used for all genes. For some genes the guesstimates will correspond more closely to their actual shapes, and for some genes, candidate models and actual profiles might deviate further. The estimates used are listed in the last column of Table 1. They are chosen in a way that a broad range of possible time-response shapes is covered, following the general idea from Duda et al. [11]. For the sigEmax model, it is assumed that at time 9 h 50% and at time 350 h 95% of the maximum effect is achieved. For the Emax model, 50% of the maximum effect is assumed to be reached at time 9 h. These assumptions are based on some explorative analyses. In the exponential model, at time 9, 9/700% of the maximum effect is assumed to be reached, yielding a pronounced exponential shape. In the quadratic model, the guesstimate is calculated based on the assumption that 100% of the maximum effect is reached at a time of 350 h, which corresponds to approximately half of the maximal tested time. A graphical display of the resulting candidate models is given in Figures A.1 (log-scale) and A.2 (original scale).

Given are $M = |\mathcal{M}|$ candidate models with model functions $f_m(t, \boldsymbol{\theta})$ and guesstimates $\boldsymbol{\theta}^0$ of the standardized models f_m^0. For each candidate model an optimal contrast test is calculated, where the hypothesis $H_m^0 : \mathbf{c}_m^T \boldsymbol{\mu}_m = 0$ versus $H_m^1 : \mathbf{c}_m^T \boldsymbol{\mu}_m \neq 0$ is tested. Here $\mathbf{c}_m = (c_{m1}, \ldots, c_{mk})^T$ is the optimal contrast vector of model m and $\boldsymbol{\mu}_m = (\mu_{m1}, \ldots, \mu_{mk})^T$ defines a mean response vector for the m-th candidate shape. The test statistics for the contrasts are given by

$$T_m = \frac{\sum_{i=1}^{k} c_{mi} \overline{Y}_i}{S\sqrt{\sum_{i=1}^{k} c_{mi}^2 \Big/ n_i}}, \ m = 1, \ldots M, \tag{3}$$

where $\overline{Y}_i$ is the mean at time point t_i and $S^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij} - \overline{Y}_i)^2/(N-k)$ is the pooled variance. The values $\mathbf{c}_m$ are chosen to maximize the power of the assumed mean response vector

$\boldsymbol{\mu}_m$. Under $H_0, (T_1, \ldots, T_M)^T$ follows a central multivariate t-distribution. A signifMCP is established if $T_{max} = \max(T_1, \ldots, T_M) > q_{1-\alpha}$, where $q_{1-\alpha}$ is the multiplicity-adjusted critical value at level $1-\alpha$. Models with contrast test statistics greater than the $1-\alpha$ quantile can be considered statistically significant, controlling the family-wise error rate. The models with $T_m > q_{1-\alpha}$ form the set $\mathscr{M}^* = \{M_1, \ldots, M_L\} \subseteq \mathscr{M}$ of L significant models. Only if a signifMCP is present for at least one gene, next the modeling step is performed. In the modeling step, the 'best' model is chosen according to the AIC.

For a gene with expression values y_{ij} $(i = 1, \ldots, k, j = 1, \ldots, n_i)$ and associated model fit $f(\cdot, \widehat{\boldsymbol{\theta}})$, the adjusted R^2 is given by

$$R^2_{\text{adj}} = 1 - \frac{(1-R^2)(N-1)}{N-p}, \quad R^2 = 1 - \frac{SSE}{SST},$$

with the sum of squared errors $SSE = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - f(t_i, \widehat{\boldsymbol{\theta}}))^2$ and the total sum of squares $SST = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y})^2$. The number of parameters is given by p.

The MCPMod procedure is implemented in the software package `DoseFinding` [4]. In this implementation, for the calculation of the multiple testing procedure, the lowest condition is required to take a value of 0. However, in the TG-GATEs data, the lowest considered condition is the time of 3 h. Thus, for the calculation of the candidate models and the subsequent calculation of the optimal contrasts, this condition of 3 h is set to 0. This only affects the calculation of the signifMCP via a slight adjustment of the optimal contrasts. The modelling step is again performed on the original data, also the model selection based on the AIC is not affected by this approach. In particular, this means that the final displayed model, i.e., the fitted curve, is shown for the original data with the lowest time point of 3 h.

3. Results

In this section, the results of the data analysis are presented. First, the necessity to perform appropriate normalization is illustrated with an example. A general overview of results when applying the MCP-Mod procedure to compounds from the TG-GATEs database is given, followed by a detailed description of the analysis for four chosen sample compounds. Finally, the newly developed Shiny application for performing the here presented analysis with the help of an easily comprehensible graphic interface is briefly introduced.

For displaying data, the `R` packages `ggplot2` [34] and `dplyr` [35] were used.

3.1. Normalization of gene expression data

When only considering compounds with complete measurements for the in vivo experiments on rat liver, 120 datasets are considered in total. After pre-processing via the RMA algorithm, the data of the high dose are normalized with respect to the respective control values of the same time point. The benefit of performing such a normalization is illustrated in Fig. 1. Here, for the compound propylthiouracil, PCA plots for different situations are shown. In contrast to the later MCPMod analyses, here all dose levels are considered.

In the top left corner, the PCA plot for the complete data, after pre-processing, but without any normalization or gene selection is shown. Four dose levels (Control, Low, Middle, High) for the 8 different time points and three replicates for each dose-time combination are displayed. The time points corresponding to a single dose (3, 6, 9, 24 h) are clearly separated on the horizontal axis (i.e. for the first principal component (PC)) from the time points corresponding to the repeated doses (4, 8, 15, 29 days), as indicated by the black line. This separation

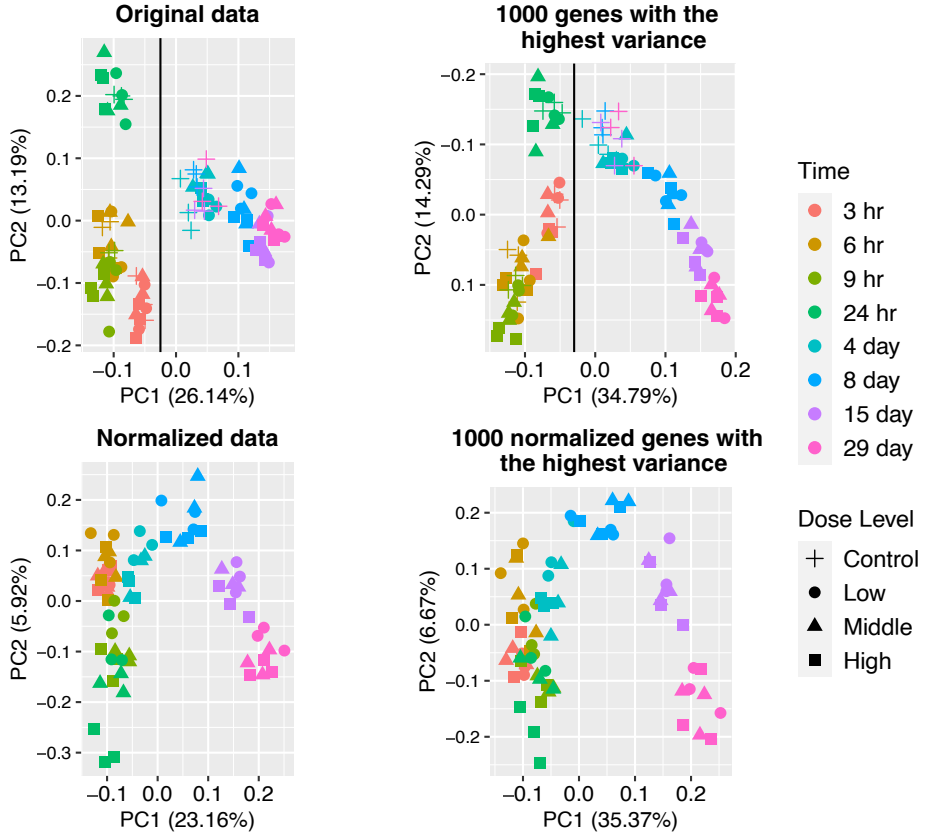


Fig. 1. PCA plots for the compound propylthiouracil for all dose levels and all time points. The plots in the top row show the original data and the plots in the bottom row show the data after normalizing with respect to the corresponding control values. The plots on the left consider all genes, while the plots on the right are based on the respective 1000 genes with the highest variance across all samples, respectively.

between single and repeat doses is far more pronounced than any separation between the individual time points or the dose levels, suggesting some kind of batch effect. To avoid blurring the results with primarily noisy genes with no real effect between conditions, the next step is to consider only the top 1000 genes with the highest variance across all samples. This is a common default procedure in several standard implementations of PCA, e.g. in the `plotMDS` function from the R-package `limma` [27], the `plotPCA` function from the R-package `DESeq2` [21] or in the R package `pcaExplorer` [22]. The resulting PCA plot is shown in the top right corner. Still, a clear separation between the single doses and the repeated doses can be seen, but overall, the values for the different times are pulled apart and overlap less. For the repeated doses, all the samples corresponding to the control dose level cluster very closely together, while for the single doses, the control dose levels cluster more closely with the other samples of the respective same time point.

To avoid these batch effects, data of the dose levels Low, Middle, and High are normalized with respect to the corresponding control replicates. For each time point and each gene, the mean expression value of the three control replicates is calculated, and this value is subtracted from all expression values for the other doses. The corresponding PCA plots are shown in the bottom row of Fig. 1. The bottom left plot shows the entire dataset, accordingly normalized, and the bottom right plot again only based on the 1000 genes with the highest variances in their expression values, after normalization. Here a notably higher percentage of the overall variability is explained than when all genes are considered, as indicated by the percentages given on the labels, while the overall profile in the PCA plot looks very similar. From all four PCA plots shown here, it becomes clear that time has a greater impact on the resulting gene expression level than the dose.

In the following, only the normalized expression values are used for the calculation of PCA plots and the application of the MCP-Mod methodology. In contrast to the results shown here, from now on only the high dose level will be considered.

3.2. Overview of results for the entire database

The MCP-Mod methodology is applied to the normalized gene expression values for all 31,099 genes and each of the 120 compounds separately. The set of candidate models with the respective guesstimate values is chosen as explained before, see Table 1. In the first step of the MCP-Mod procedure, for each gene, it is statistically tested whether a time-dependent signal is present, and if so, it is said that a signifMCP is established for that gene. As significance level $\alpha = 0.05$ is chosen for this statistical test. Fig. 2 shows a violin plot of the number of signifMCP per compound. While there are some values higher than 20,000 (corresponding to the compounds methylene dianiline, methapyrilene, and acetamidofluorene), the median (and the quartiles) of the numbers of signifMCP are given by 11029.50 (and 8499.00; 12862.75).

Only for the genes with a signifMCP, the significant candidate models are fitted to the expression data, and the best model is chosen according to the AIC criterion. In total, a model is fitted to 1,340,745 genes across the 120 compounds. Overall, the sigEmax model is chosen most often, followed by the quadratic model.

Fig. 3 shows the number of times each of the candidate models is the selected model across all compounds. The display is stratified according to the goodness-of-fit as determined by R^2_{adj}, the adjusted R^2. These numbers are additionally summarized in Table A.1. Moreover, the reverse plot, i.e. a display of the value of R^2_{adj} of the winner model, stratified by the winner model, is shown in Figure A.3.

It can be seen that the sigEmax model is most often chosen in situations with medium to high values of the R^2_{adj}, i.e. for high-quality fits. Particularly for very high values with $R^2_{\text{adj}} > 0.75$, the sigEmax model is the best model according to the AIC in almost half of the cases. If sigEmax is not the winner, then betaMod, emax, linlog, and quadratic

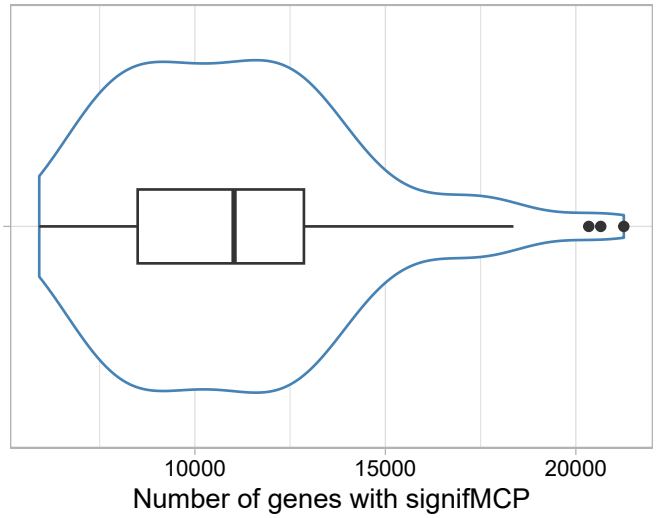


Fig. 2. Number of genes with signifMCP per compound. This display summarizes the number of signifMCPs, i.e. the genes with a statistically significant signal for a predefined set of candidate models, for all 120 compounds.

models are the best model similarly often. In a situation where only very little of the variation is explained by the resulting best model ($R^2_{\text{adj}} < 0.25$), the linear or quadratic model is the respective best model in the vast majority of cases. For medium low values of R^2_{adj}, the quadratic model is most often the best model, but overall, there is less difference in the frequency of winning between the different models than in the other situations. When less variation is explained by the resulting best model, comparatively simpler models (linear, linlog, quadratic) are more often the winner model. The exponential model is a clear exception to this observation, as it is the winner model in the least number of cases. The relatively flexible betaMod model is far less often the best model than the sigEmax model, which is equally complex in terms of the number of parameters, but more restricted in terms of the shape.

3.3. Detailed analysis for four chosen compounds

Four compounds were chosen for a detailed analysis of the results of the MCP-Mod procedure applied to the normalized expression values of the high dose. Corresponding results for all 120 compounds are summarized in Supplement B. The selected compounds represent a wide range of results that are observed in the entire dataset. The choice is based both on a visual assessment of PCA plots and by considering the total number of signifMCPs for the compounds:

- Compound 1: methapyrilene:
  The PCA plot shows a clear progression of the different time points, and many genes lead to a signifMCP.
- Compound 2: pemoline:
  Some progression along the time points can be seen, but with more random noise, and the number of signifMCPs is smaller.
- Compound 3: quinidine:
  The random noise in the data almost overlays the structures in the PCA plot, and only a medium number of signifMCPs are obtained.
- Compound 4: famotidine:
  Almost no structure can be seen in the PCA plots, and only a low number of signifMCPs are obtained.

While the selection of compounds was based on statistical considerations, their biological and therapeutic properties are briefly introduced here as well:

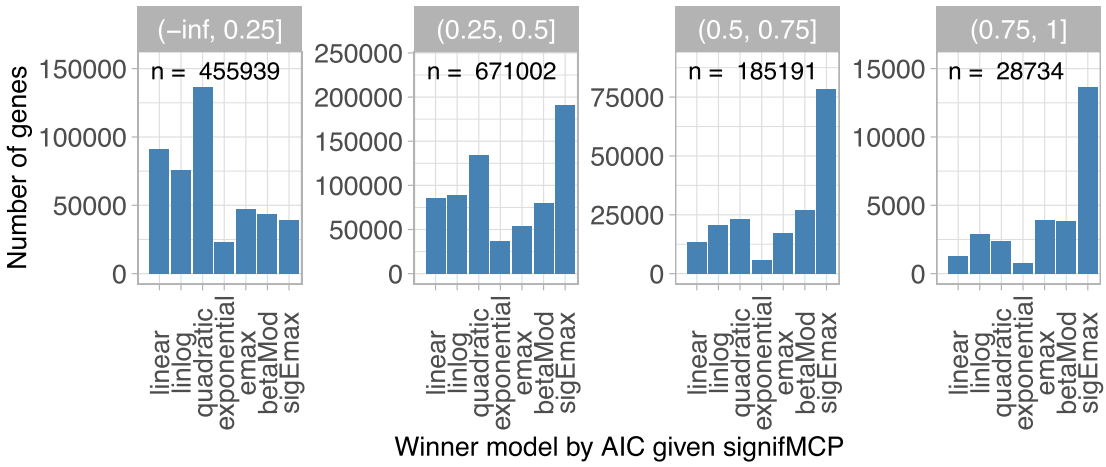


Fig. 3. Number of times each of the candidate models is the selected model according to the AIC across all compounds. The display is stratified according to the goodness-of-fit (R^2_{adj}) of the respective selected model.

- Compound 1: methapyrilene:

  Methapyrilene was used until the 1970s to treat sleep disorders due to its sedative properties. However, the discovery of a possible hepatocarcinogenic potential after continuous administration in high doses to rats led to the drug being withdrawn from the market [20]. Since mutagenicity and genotoxicity tests generally showed negative results, methapyrilene is not considered a genotoxic carcinogen [23]. It is often used as a model hepatotoxin in basic research, especially for toxicogenomic approaches to distinguish genotoxic and non-genotoxic carcinogens or to categorize different hepatotoxic substances based on their mechanisms of action [5].
- Compound 2: pemoline:

  Pemoline was approved by the Food and Drug Administration (FDA) in 1975 under the name Cylert® as a drug for the treatment of attention deficit hyperactivity disorder (ADHD). Subsequent publications have described cases of serious pemoline-induced liver

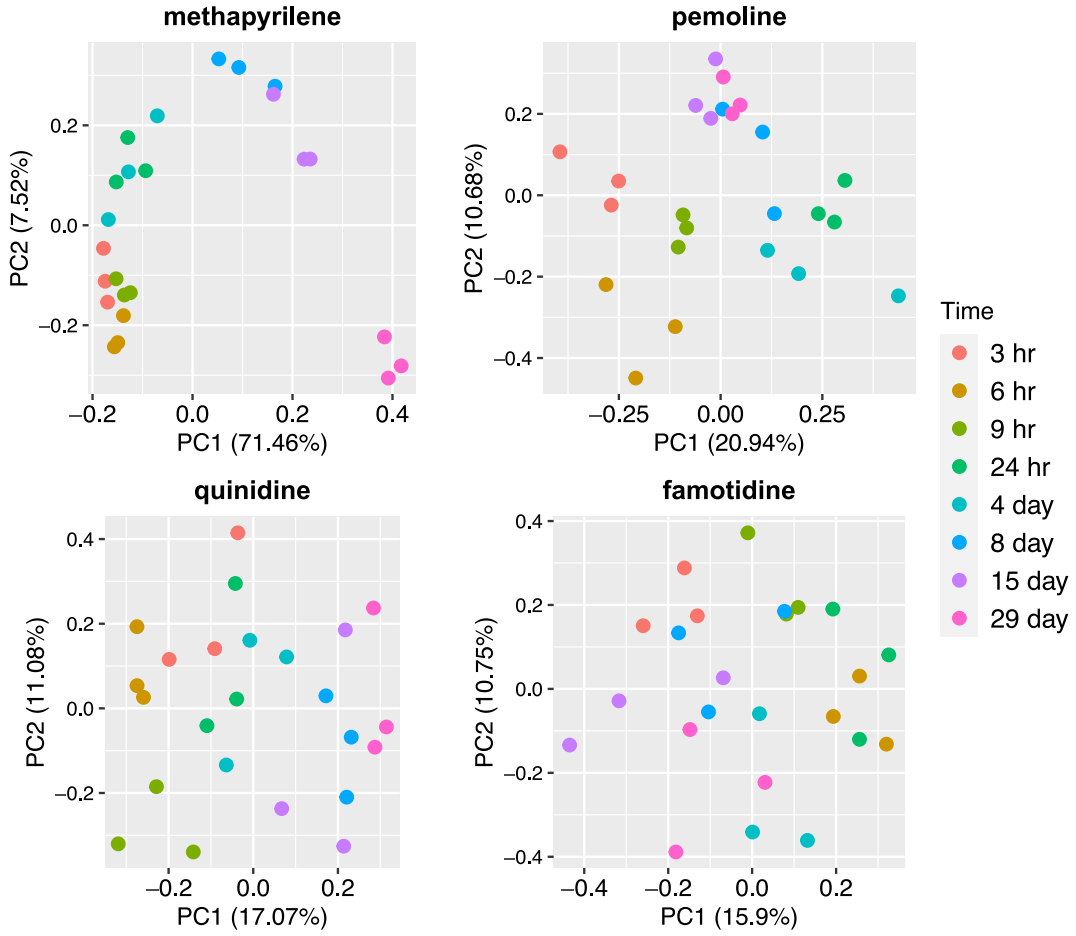


Fig. 4. PCA plot for four selected compounds, displaying all time points for the high dose. Data are normalized with respect to the corresponding control samples, and only the 1000 genes with the largest variance are used for the calculation of the plots.

injury, including acute liver failure. Due to ongoing reports of liver toxicity, the FDA changed the labelling of pemoline in December 1996. In particular, the risk of liver failure associated with the use of pemoline was noted, and the drug was reclassified from first-line to second-line treatment for ADHD [36].

- Compound 3: quinidine:

  Quinidine was one of the first antiarrhythmic drugs developed and is useful in the treatment of a variety of arrhythmias. Following concerns about an increased risk of ventricular arrhythmias and deaths with quinidine in the 1950s, its use declined dramatically. However, recent studies have sparked renewed interest in quinidine, particularly in combination with verapamil for the treatment of atrial fibrillation [37].
- Compound 4: famotidine:

  Famotidine belongs to a class of medications that block H2 receptors, functioning as an antihistamine. It is frequently prescribed to treat gastric and duodenal ulcers, as well as gastroesophageal reflux disease [1].

This shows how these compounds have relevance in several real-world applications.

PCA plots for the four selected compounds methapyrilene, pemoline, quinidine, and famotidine are displayed in Fig. 4. Methapyrilene is chosen as the compound with clear progression along the time points in the PCA plot. Moving from left to right (first PC, explaining more than 70% of the variation), with an additional vertical separation (second PC), an increase in the time can be observed. The three replicates corresponding to the same time, respectively, always cluster relatively close together.

For the second compound, pemoline, the first PC explains a far smaller proportion of the total variation. However, some progression of the times can still be seen: Smaller times are located in the lower left corner, medium times in the lower right corner, and the highest times in the upper part of the PCA plot. Although they are more spread in comparison to the first compound, the three replicates of each time still cluster together.

The spread of the three replicates is even larger for the third compound, quinidine. The first and second PC explain a comparatively similar ratio of the total variation as for pemoline, but the time-dependent structure is less clear.

Finally, for the fourth compound famotidine, the spread of the replicates is generally much larger than the differences between the times. Several times overlap and even less structure than for the previous plot can be seen, and the ratio of the explained variation is also even smaller.

As a second step of the analysis, the MCPMod procedure as described above is applied to the normalized gene expression values for all genes in the high dose. The resulting winner models for the four selected compounds, for genes with signifMCP, stratified by the value of R^2_{adj} for the respective winner model, are summarized in Fig. 5. The total number of signifMCPs is 20,654 for methapyrilene, 7,283 for pemoline, 7,268 for quinidine, and 6,665 for famotidine.

All four compounds have in common that the respective largest proportion of genes exhibits a resulting R^2_{adj} of the winner model in the interval $(0.25, 0.5]$. However, for methapyrilene, a very large number of genes also result in R^2_{adj} values of the winner model in higher intervals,

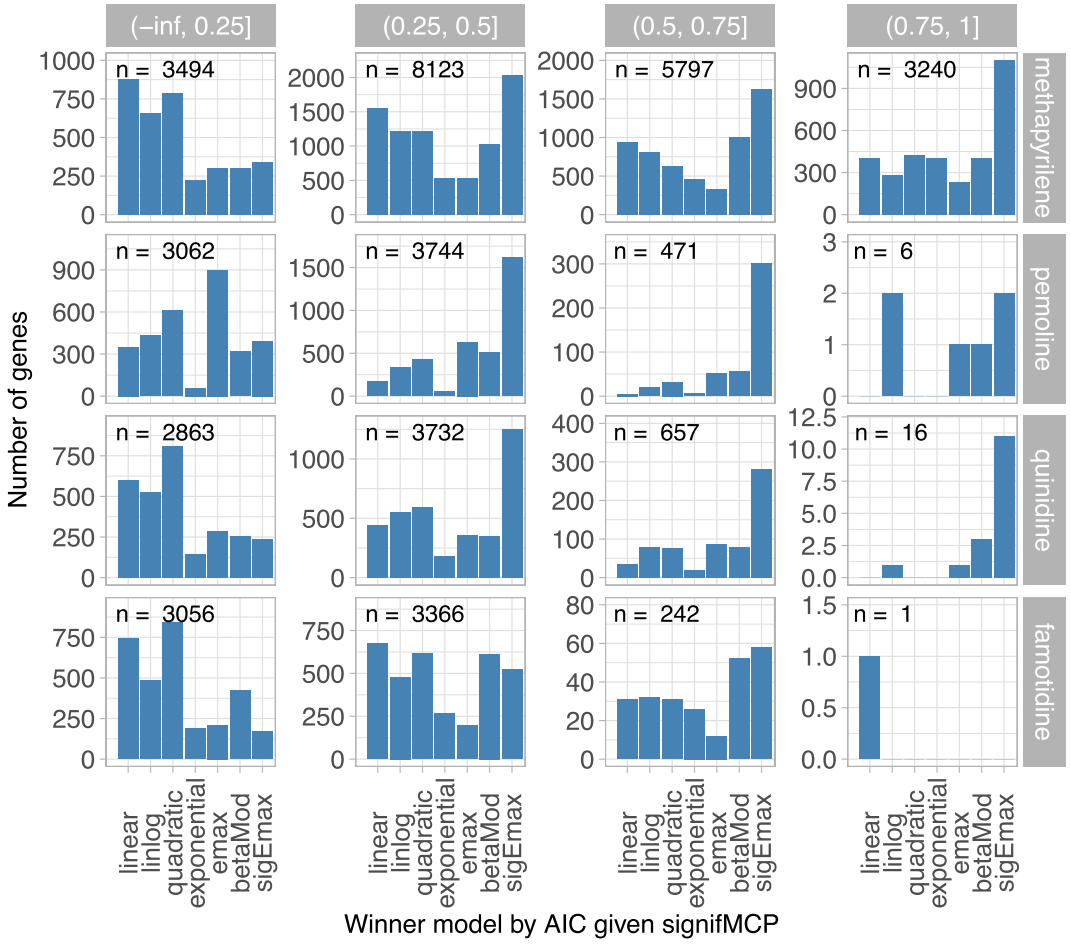


Fig. 5. Winner model by AIC given signifMCP, for four selected compounds. Results are stratified by the value of R^2_{adj} of the respective selected model and displayed with individual y-axes.

indicating more clearly pronounced shapes and less noisy data, respectively. For low values of R^2_{adj}, the linear, linlog, and quadratic models are the winner models for the largest number of genes for all compounds except for pemoline, where the emax model is the winner model most often. For higher values of the R^2_{adj}, the sigEmax model is the winner model most often, and especially for the highest values of R^2_{adj}, with quite a large difference to the second most frequent winner model. For famotidine and R^2_{adj} in the interval $(0.25, 0.5]$, the betaMod is the winning model even more often than the sigEmax model.

Fig. 6 displays the distribution of the values of R^2_{adj} for the four selected compounds. From the histograms, the much higher number of signifMCP for the compound methapyrilene than for the other compounds can be seen. While for the three compounds pemoline, quinidine, and famotidine, the histograms are centered around a value of 0.2 and have a similar shape, larger values can be observed for methapyrilene. Here, values up to 1.0 are observed, and notably many values of R^2_{adj} are in the range between 0.6 and 1.0. For the other compounds, in contrast, only very few values larger than 0.6 are observed.

Generally, the results for the compounds shown here are in line with the findings presented in Duda et al. [11]: Genes with a clearer signal (i.e. higher value of the R^2_{adj}) often lead to more complex models such as the sigEmax model as winner models, while genes with more noise (i.e. low value of R^2_{adj}) are best fitted with a linear model or a linear model on a log scale (linlog model).

3.4. Example genes

Examples of the resulting fitted curves are shown in Fig. 7. Here, for each of the four selected compounds and each of the seven candidate models, the fitted time-response curve resulting in the highest value of the R^2_{adj} is shown. Corresponding values of R^2_{adj} and the gene names or, if these are not available, the probe set IDs are summarized in Table A.2. As for the analyses before, the fits are calculated for the normalized data in the high dose only.

A logarithmic scale is chosen for the display of the time in hours. Thus, the linear model appears to have an exponential shape, since the linearity refers to the original scale. In contrast, the linlog model shows a linear time-response relationship on this logarithmic scale.

3.5. Shiny application 'Mos-TEC'

The Shiny application MoS-TEC was developed to present the results of the model fits for the 120 datasets, i.e. the 120 different compounds, used. As a first step, this application makes it possible to perform a principal component analysis for the investigated compounds, as described in Chapter 2.2. For a more detailed analysis of the results obtained with the MCP-Mod method for a specific compound, the application offers barplots of the winner model according to the AIC for a given signifMCP. These barplots are then stratified according to the R^2_{adj}, similar to the approach outlined in Chapter 3.3. It is also possible to generate time-response curves for individual genes with signifMCP.

The analysis area of the application then provides results referring to model fits. The selection of a compound of interest is facilitated by a drop-down menu. The compound can be selected either manually by scrolling and performing a targeted search or by entering the initial letters using the auto-complete function. Below the drop-down menu, three tabs provide access to the previously mentioned PCA plots, barplots, and time-response curves. An exemplary menu view is shown in Fig. 8.

In the section Time-Response-Curve, another drop-down menu allows the selection of specific genes. If at least one signifMCP is available for the selected gene, scatterplots of the high-dose level data are generated with all models that have a signifMCP. In addition, a table with information on the respective model, the AIC, the R^2_{adj}, and the parameter estimates is displayed next to each plot. If not signifMCP is available for the selected gene, only the scatterplot without any fitted curve is shown. Scatterplots for the low and middle dosages are also generated. Note that fitting the models for a gene can take up to 30 s. An example is provided in Fig. 9.

4. Conclusion

In this work, we applied the multiple comparison and modelling procedure (MCP-Mod) to time-response relationships for 120 compounds from the openly available TG-GATEs database [13]. We used the in vivo gene expression data, where the compounds were given to rats in four different doses (Control, Low, Middle, and High), and at eight different time points. Rats were sacrificed and expression values were measured genome-wide for cells taken from the livers.

The MCP-Mod methodology was originally developed for Phase II clinical studies [6; 3], but has been shown to lead to meaningful results also when applied to high-dimensional gene expression data [11]. Briefly, the methodology combines a multiple comparison procedure to detect genes with some meaningful signal ('signifMCP') with a model selection approach, to fit one of several candidate models.

MCP-Mod analyses and descriptive analyses via principal component analyses (PCA) are shown and discussed in this work for four specific selected compounds that cover a broad spectrum of the observed results. Results for all 120 compounds are summarized in an additional supplement and are made available via the Shiny application 'MoS-TEC' (Model Selection for Time-Expression Curves applied to compounds from the TG-GATEs database). This application provides an intuitive and easy-to-use graphical user interface to select the compound of interest and obtain overall results such as PCA plots and summaries of the MCP-Mod results, and in addition, also allows the plotting of individual genes with the respective best-fitting model, e.g. after biological pre-selection

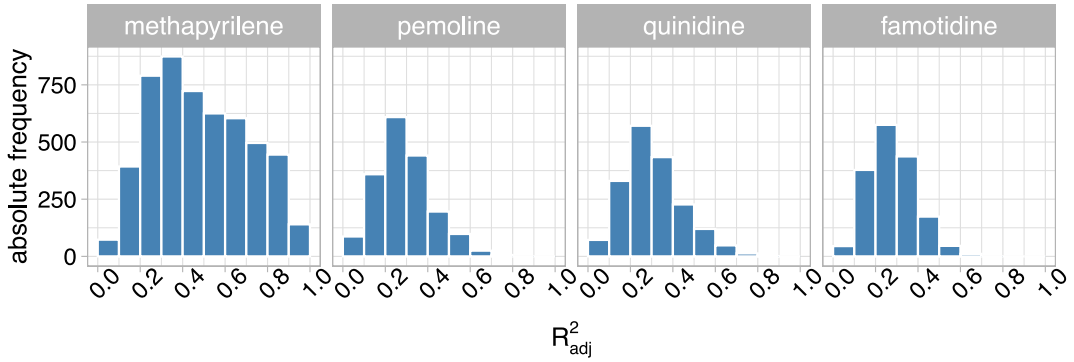


Fig. 6. Values of R^2_{adj} for four selected compounds. Values are shown for all genes with a signifMCP and the respective winner model, stratified by the compound.

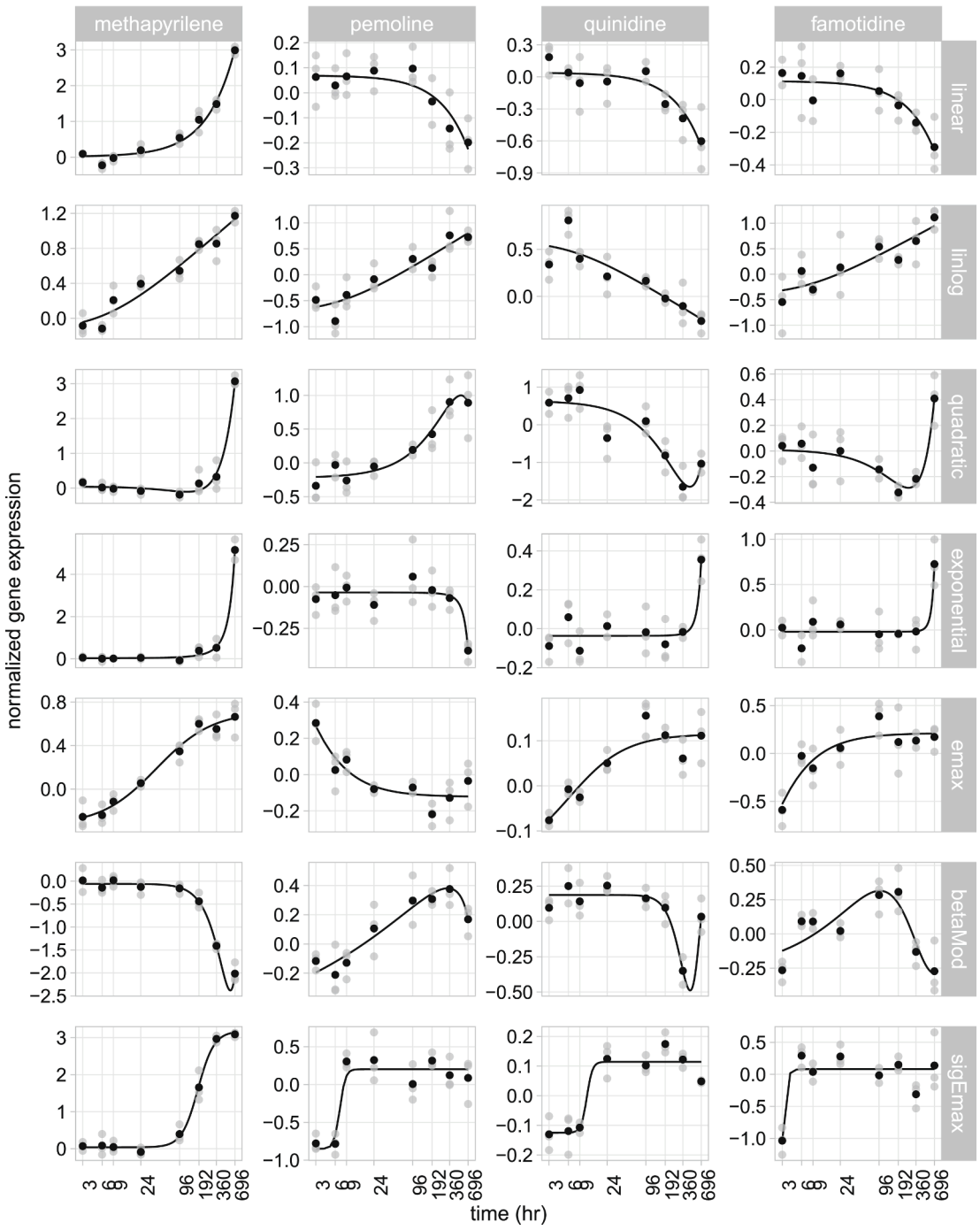


Fig. 7. Examples of fitted time-response curves. For each of the four compounds and each candidate model, the respective curve with the highest value of R^2_{adj} is displayed, where grey dots indicate the individual measurements, and black dots the time-wise mean expression values. All profiles are shown with individual y-axes.

of potentially interesting genes. The application can be used online without installing additional software.

The general findings in this paper underline the results from Duda et al. [11]. In their paper, gene expression values for cells treated in different concentrations with only one compound were assessed, whereas the present work considers time-response data from in vivo experiments for 120 compounds. However, the main results are the same. All different candidate models are chosen via the model selection step in the MCP-Mod methodology, emphasizing the importance of considering a variety of models when fitting curves to gene expression as response data. When stratifying the winner model by the degree of noise in the data, reflected by the adjusted R^2 (R^2_{adj}), higher degrees of noise (lower values of R^2_{adj}) most often lead to simpler models with only few parameters. For a higher signal-to-noise ratio, associated with a high value of R^2_{adj}, the very popular sigEmax model is the winner model most often. However, it is also observed that for some compounds, only very few genes exhibit a high value of R^2_{adj} in the range between 0.75 and 1.

While the R^2_{adj} is a very popular measure for goodness-of-fit of a fitted model, its use in nonlinear model fits has been criticized [32] due to the property that it may become negative. This is also sometimes the case for the curves considered here, such that this point should be kept in mind when interpreting the results. For model selection purposes, in this work, only the Akaike information criterion was used, whereas the R^2_{adj}

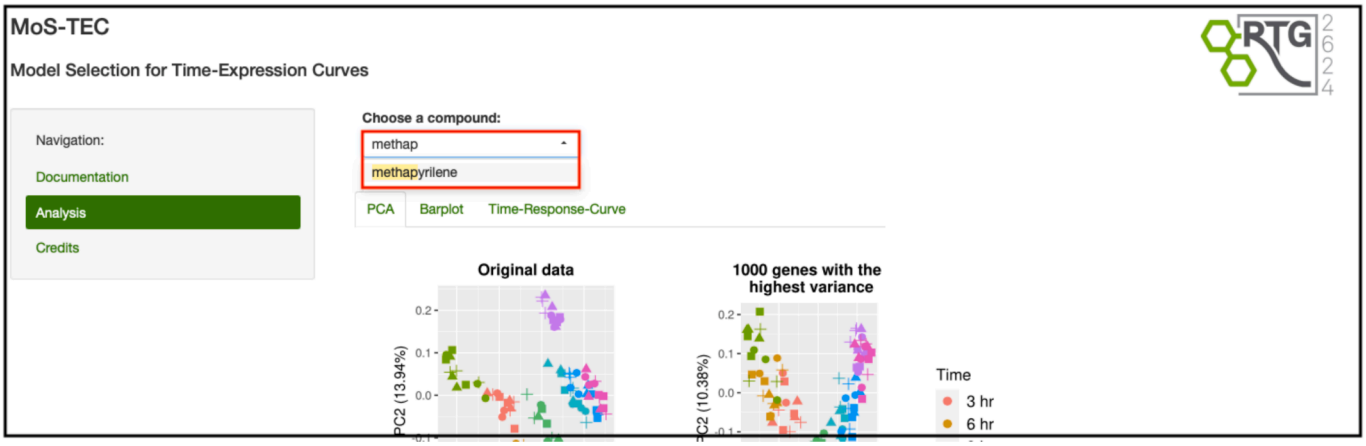


Fig. 8. Example of the menu view for the application MoS-TEC. Here, one compound is selected and the corresponding PCA plots are shown.

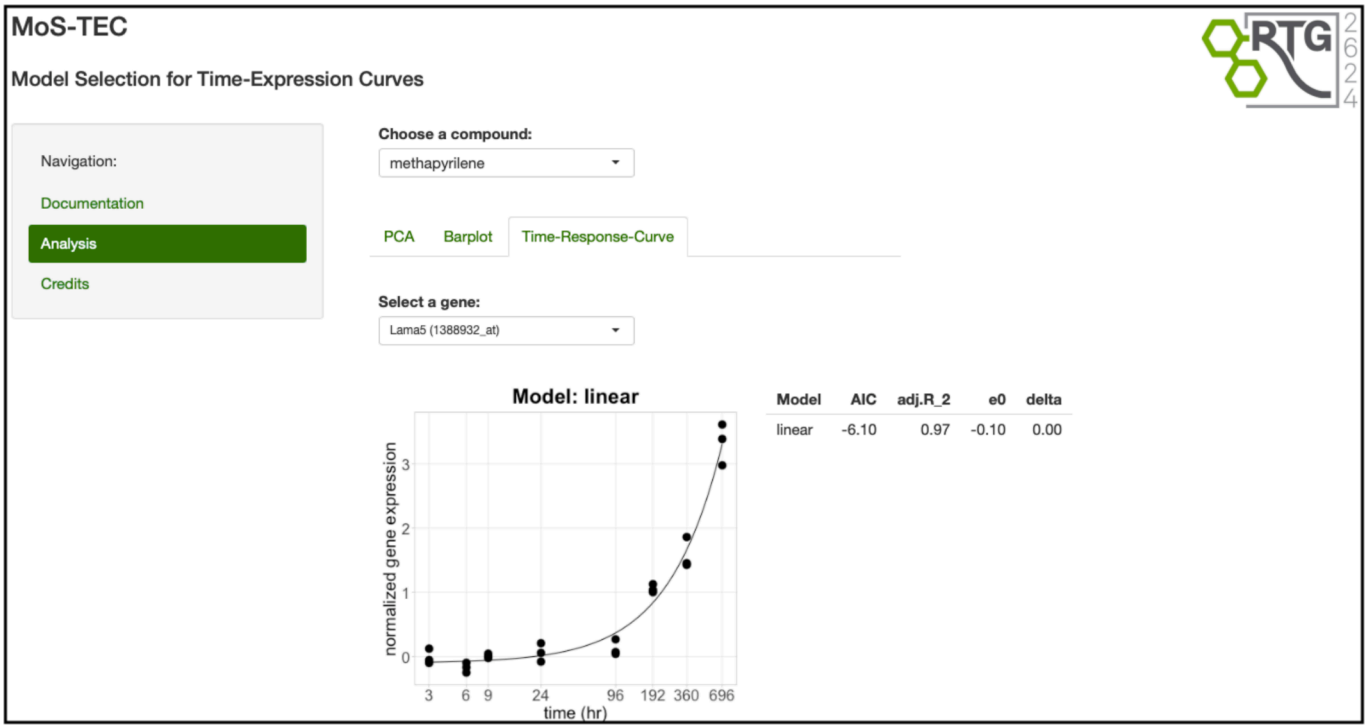


Fig. 9. Example of the menu view for the application MoS-TEC. Here, one compound and one gene are selected and a fitted curve for a model with signifMCP alongside the table with further information on the fitted model is shown.

was only used for stratification for display purposes.

In the Shiny application MoS-TEC, the resulting model fit is only displayed for the high dose. However, data points for the middle and the low dose (normalized in the same way as the high dose with respect to the control) are plotted as well. The corresponding time-response curves do not necessarily exhibit a signifMCP or lead to the same selected model according to the AIC. Since the quality of a fitted model is usually better when more data points are available, the possibility of fitting models with shared parameters between different doses for the same compound and gene could be of interest in a future extension of the application.

Funding information

This work has been supported by the Research Training Group 'Biostatistical Methods for High-Dimensional Data in Toxicology' (RTG 2624) funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation - Project Number 427806116).

CRediT authorship contribution statement

Franziska Kappenberg: Conceptualization, Methodology, Project administration, Supervision, Writing – original draft. **Benedikt Küthe:** Data curation, Formal analysis, Software, Visualization, Writing – original draft. **Jörg Rahnenführer:** Conceptualization, Funding acquisition, Project administration, Resources, Supervision, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data is provided in a Shiny-Application (http://shiny.statistik.tu-dortmund.de:8080/app/MoS-TEC)

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at https://doi.org/10.1016/j.comtox.2024.100313.

References

[1] S.Y. Al Samarrai, R. AlZubaidi, N. Al-Ansari, Charge transfer complex-based spectrophotometric analysis of famotidine in pure and pharmaceutical dosage forms, Scientific Rep. 14 (1) (2024).

[2] G. Ankley, R. Bennett, R. Erickson, D. Hoff, M. Hornung, R. Johnson, D. Mount, J. Nichols, C. Russom, P. Schmieder, J. Serrrano, J. Tietge, D. Villeneuve, Adverse outcome pathways: a conceptual framework to support ecotoxicology research and risk assessment, Environ. Toxicol. Chem. 29 (2010).

[3] B. Bornkamp, J. Pinheiro, F. Bretz, MCPMod: an R package for the design and analysis of dose-finding studies, J. Stat. Softw. 29 (7) (2009) 1–23.

[4] B. Bornkamp, J. Pinheiro, F. Bretz, L. Sandig, DoseFinding: Planning and Analyzing Dose Finding Experiments, R package version 1.0-2 (2018).

[5] A. Braeuning, A. Oberemm, T. Heise, U. Gundert-Remy, J.G. Hengstler, A. Lampen, In vitro proteomic analysis of methapyrilene toxicity in rat hepatocytes reveals effects on intermediary metabolism, Arch. Toxicol. 93 (2) (2018) 369–383.

[6] F. Bretz, J.C. Pinheiro, M. Branson, Combining multiple comparisons and modeling techniques in dose-response studies, Biometrics 61 (3) (2005) 738–748.

[7] P. Bushel, R. Paules, S. Auerbach, A comparison of the TempO-Seq S1500+ platform to RNA-Seq and microarray using rat liver mode of action samples, Frontiers (2018) in Genetics.

[8] M. Carlson, rat2302.db: Affymetrix Rat Genome 230 2.0 Array annotation data (chip rat2302), R package version 3.13.0 (2021).

[9] W. Chang, J. Cheng, J. Allaire, C. Sievert, B. Schloerke, Y. Xie, J. Allen, J. McPherson, A. Dipert, B. Borges, shiny: Web Application Framework for R, R package version 1 (7) (2022) 4.

[10] J.C. Corton, T. Hill, J.J. Sutherland, J.L. Stevens, J. Rooney, A set of six gene expression biomarkers identify rat liver tumorigens in short-term assays, Toxicol. Sci. 177 (2020) 11–26.

[11] J.C. Duda, F. Kappenberg, J. Rahnenführer, Model selection characteristics when using MCP-Mod for dose–response gene expression data, Biometrical J. (2022).

[12] L. Gautier, L. Cope, B.M. Bolstad, R.A. Irizarry, affy—analysis of Affymetrix GeneChip data at the probe level, Bioinformatics 20 (3) (2004) 307–315.

[13] Y. Igarashi, N. Nakatsu, T. Yamashita, A. Ono, Y. Ohno, T. Urushidani, H. Yamada, Open TG-GATEs: a large-scale toxicogenomics database, Nucl. Acids Res. 43 (Database issue) (2015) D921–D927.

[14] R.A. Irizarry, B.M. Bolstand, F. Collin, L.M. Cope, B. Hobbs, T.P. Speed, Summaries of affymetrix genechip probe level data, Nucl. Acids Res. 31 (4) (2003) 1–8.

[15] S.M. Jensen, F.M. Kluxen, C. Ritz, A review of recent advances in benchmark dose methodology, Risk Anal. 39 (19) (2019) 2295–2315.

[16] X. Jiang, Estimation of effective concentrations from in vitro dose-response data using the log-logistic model (Ph.D. thesis), Medical Faculty of Ruprecht-Karls-University in Heidelberg, 2013.

[17] F. Kappenberg, J. Duda, L. Schürmeyer, O. Gül, T. Brecklinghaus, J. Hengstler, K. Schorning, J. Rahnenführer, Guidance for statistical design and analysis of toxicological dose-response experiments, based on a comprehensive literature review, Arch. Toxicol. 97 (2023).

[18] F. Kappenberg, M. Grinberg, X. Jiang, K.-S.A., J.G. Hengstler, J. Rahnenführer, Comparison of observation-based and model-based identification of alert concentrations from concentration-expression data, Bioinformatics 37 (14) (2021) 1990–1996.

[19] M. Leist, A. Ghallab, R. Graepel, R. Marchan, R. Hassan, S.H. Bennekou, A. Limonciel, M. Vinken, S. Schildknecht, T. Waldmann, E. Danen, B. van Ravenzwaay, H. Kamp, I. Gardner, P. Godoy, F.Y. Bois, A. Braeuning, R. Reif, F. Oesch, D. Drasdo, S. Höhme, M. Schwarz, T. Hartung, T. Braunbeck, J. Beltman, H. Vrieling, F. Sanz, A. Forsby, D. Gadaleta, C. Fisher, J. Kelm, D. Fluri, G. Ecker, B. Zdrazil, A. Terron, P. Jennings, B. van der Burg, S. Dooley, A.H. Meijer, E. Willighagen, M. Martens, C. Evelo, E. Mombelli, O. Taboureau, A. Mantovani, B. Hardy, B. Koch, S. Escher, C. van Thriel, C. Cadenas, D. Kroese, B. van de Water, J.G. Hengstler, Adverse outcome pathways: opportunities, limitations and open questions, Arch. Toxicol. 91 (2017).

[20] W. Lijinsky, M.D. Reuber, B.-N. Blackwell, Liver tumors induced in rats by oral administration of the antihistaminic methapyrilene hydrochloride, Science 209 (4458) (1980) 817–819.

[21] M.I. Love, W. Huber, S. Anders, Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2, Genome Biol. 15 (2014) 550.

[22] A. Ludt, A. Ustjanzew, H. Binder, K. Strauch, F. Marini, Interactive and Reproducible Workflows for Exploring and Modeling RNA-seq Data with pcaExplorer, ideal and GeneTonic, Curr. Protocols 2 (4) (2022) e411.

[23] J. Mirsalis, Genotoxicity, toxicity and carcinogenicity of the antihistamine methapyrilene (MTR 07216), Mutation Res./Rev. Genet. Toxicol. 185 (3) (1987) 309–317.

[24] K. Möllenhoff, K. Schorning, F. Kappenberg, Identifying alert concentrations using a model-based bootstrap approach, Biometrics (2022) 1–13.

[25] National Research Council, Applications of Toxicogenomic Technologies to Predictive Toxicology and Risk Assessment, The National Academies Press, Washington, DC, 2007.

[26] R Core Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, 2021.

[27] M.E. Ritchie, B. Phipson, D. Wu, Y. Hu, C.W. Law, W. Shi, G.K. Smyth, limma powers differential expression analyses for RNA-sequencing and microarray studies, Nucl. Acids Res. 43 (7) (2015) e47.

[28] C. Ritz, S.M. Jensen, D. Gerhard, J.C. Streiberg, Dose-Response Analysis Using R, Chapman and Hall/ CRC, New York, 2019.

[29] J. Rooney, T. Hill, C. Qin, F.D. Sistare, J.C. Corton, Adverse outcome pathway-driven identification of rat liver tumorigens in short-term assays, Toxicol. Appl. Pharmacol. 356 (2018) 99–113.

[30] A. Serra, M. Fratello, G. del Giudice, L.A. Saarimäki, M. Paci, A. Federico, D. Greco, TinderMIX: Time-dose integrated modelling of toxicogenomics data, GigaScience 9 (5) (2020).

[31] O. Soufan, J. Ewald, C. Viau, D. Crump, M. Hecker, N. Basu, J. Xia, T1000: a reduced gene set prioritized for toxicogenomic studies, PeerJ 7 (2019) e7975.

[32] A.N. Spiess, N. Neumeyer, An evaluation of R2 as an inadequate measure for nonlinear models in pharmacological and biochemical research: a Monte Carlo approach, BMC Pharmacol. 10 (6) (2010).

[33] R. Su, H. Wu, B. Xu, X. Liu, L. Wei, Developing a Multi-Dose Computational Model for Drug-Induced Hepatotoxicity Prediction Based on Toxicogenomics Data, IEEE/ACM Trans. Comput. Biol. Bioinf. 16 (4) (2019) 1231–1239.

[34] H. Wickham, ggplot2: Elegant Graphics for Data Analysis, Springer-Verlag, New York, 2016.

[35] H. Wickham, R. François, L. Henry, K. Müller, dplyr: A Grammar of Data Manipulation, R package version 1 (2022) 10.

[36] M.E. Willy, B. Manda, D. Shatin, C.R. Drinkard, D.J. Graham, A study of compliance with FDA recommendations for Pemoline (Cylert), J. Am. Acad. Child Adolesc. Psychiatry 41 (7) (2002) 785–790.

[37] F. Yang, S. Hanon, P. Lam, P. Schweitzer, Quinidine revisited, Am. J. Med. 122 (4) (2009) 317–321.