Working Paper Series, N. 11, December 2013

Department of Statistical Sciences
University of Padua
Italy

# Toxicity of Tamoxifen on Daphnia pulex

**Paola Tellaroli**
Department of Statistical Sciences
University of Padua, Italy

**Myriam Borgatta**
Faculté des Géosciences et de l'environnement
Université de Lausanne, Switzerland

**Alessandra Brazzale**
Department of Statistical Sciences
University of Padua, Italy

**Nathalie Chèvre**
Faculté des Géosciences et de l'environnement
Université de Lausanne, Switzerland

**Céline Hernandez**
CIG - Protein Analysis Facility Genopode - University of Lausanne
SIB Swiss Institute of Bioinformatics, Vital-IT group, Lausanne, Switzerland

**Patrice Waridel**
CIG - Protein Analysis Facility Genopode
University of Lausanne, Switzerland

**Abstract:** *Daphnia pulex* is a water flea considered an environmental indicator species. In this experiment we exposed *Daphnia* to Tamoxifen in low or high concentrations, dissolved in dimethyl sulfoxide with water, and we measured the amount of proteins at day 2 and 7. With the R package `maSigPro` we selected proteins changing significantly over time among the four experimental groups and we developed a cluster analysis for the behavior of profiles over time, to understand which and how these specific proteins change according to the treatment received. The information obtained from this study represents an important first step towards characterizing patterns specific to environmental contaminants.

**Keywords:** Daphnia pulex; ecology; environment; hierarchical clustering; `maSigPro`; toxicology.

UNIVERSITÀ DEGLI STUDI DI PADOVA
DIPARTIMENTO DI SCIENZE STATISTICHE

# Contents



**Department of Statistical Sciences**
Via Cesare Battisti, 241
35121 Padova
Italy

tel: +39 049 8274168
fax: +39 049 8274170
`http://www.stat.unipd.it`

**Corresponding author:**
Paola Tellaroli
tel: +39 049 827 4174
`tellaroli@stat.unipd.it`

# Toxicity of Tamoxifen on Daphnia pulex

**Paola Tellaroli**
Department of Statistical Sciences
University of Padua, Italy

**Myriam Borgatta**
Faculté des Géosciences et de l'environnement
Université de Lausanne, Switzerland

**Alessandra Brazzale**
Department of Statistical Sciences
University of Padua, Italy

**Nathalie Chèvre**
Faculté des Géosciences et de l'environnement
Université de Lausanne, Switzerland

**Céline Hernandez**
CIG - Protein Analysis Facility Genopode - University of Lausanne
SIB Swiss Institute of Bioinformatics, Vital-IT group, Lausanne, Switzerland

**Patrice Waridel**
CIG - Protein Analysis Facility Genopode
University of Lausanne, Switzerland

**Abstract:** *Daphnia pulex* is a water flea considered an environmental indicator species. In this experiment we exposed *Daphnia* to Tamoxifen in low or high concentrations, dissolved in dimethyl sulfoxide with water, and we measured the amount of proteins at day 2 and 7. With the R package `maSigPro` we selected proteins changing significantly over time among the four experimental groups and we developed a cluster analysis for the behavior of profiles over time, to understand which and how these specific proteins change according to the treatment received. The information obtained from this study represents an important first step towards characterizing patterns specific to environmental contaminants.

**Keywords:** Daphnia pulex; ecology; environment; hierarchical clustering; `maSigPro`; toxicology.

## 1 Introduction

The subject of our ecotoxicological study is *Daphnia pulex*, a freshwater crustacean commonly used for environmental monitoring of pollutants around the world. It is the established model species for toxicological studies because of its role in aquatic food networks and its geographical distribution. In fact, daphnids serve as an important source of food for fish and other aquatic organisms. Furthermore, they are

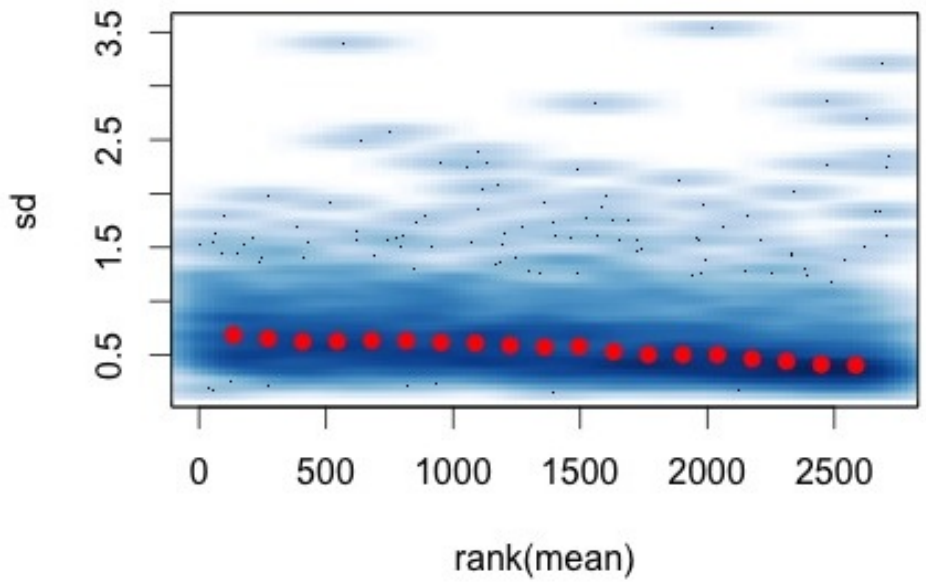


Figure 1: Standard deviation versus mean relationship. For each feature, the plot shows the empirical standard deviation $\sigma_i$ of the normalized and glog-transformed data on the $y$-axis versus the rank of the mean $m_i$ on the $x$-axis. The red dots, connected by lines, show the running median of the standard deviation.

strongly sensitive to changes in water chemistry, are simple and inexpensive to raise in an aquarium and are able to inhabit very different environments throughout the world, proving highly adaptive. They mature in just a few days, so it does not take long to grow a culture of test organisms. As a result, *Daphnia* is recognized as an indicator of environmental problems (Shaw et al., 2008).
Tamoxifen is a synthetic drug to treat breast cancer and infertility in women. It acts as an estrogen antagonist. We expose daphnis to Tamoxifen because it is an anti-cancer drug widely prescribed in the world, an endocrine disruptor, causes adverse effects in humans, has very powerful metabolites and has been studied very little. Here, we want to measure the production of certain specific proteins produced by an organism under stress.

# 2 Data

## 2.1 Experimental set-up

Data are divided into four experimental groups: water (M4), dimethyl sulfoxide and water (DMSO), Tamoxifen in low concentration dissolved in water and DMSO (C1), Tamoxifen in high concentration dissolved in water and DMSO (C2). Comparisons are made at two time instants, representing two stages of development of the daphnids: at day 2, when it is considered a baby, and at day 7 (actually between 6 and 8 days), when the daphnids are adult and laid once. Every experiment has been made in 2 replicates.

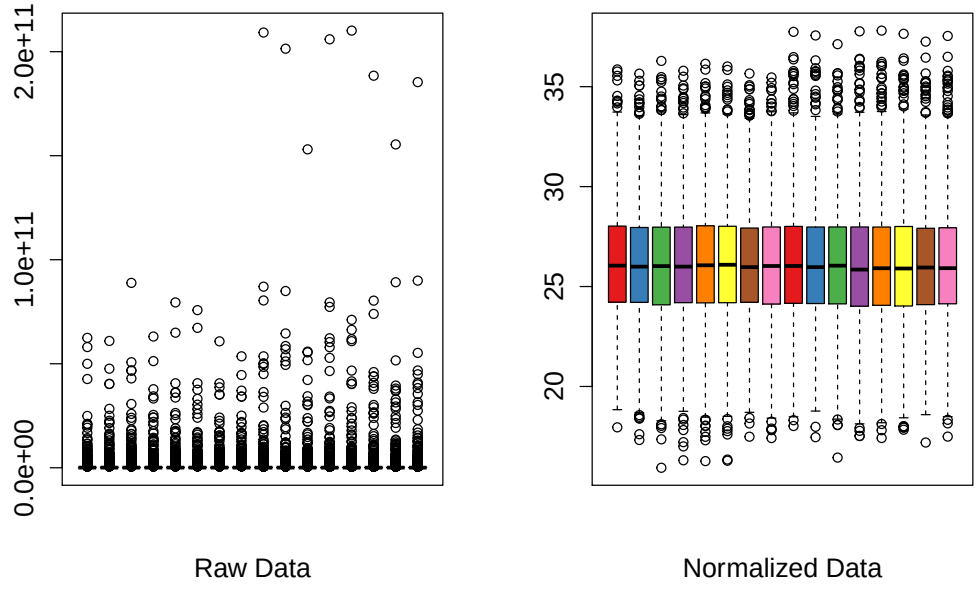


Figure 2: Boxplots of every experimental condition both for raw and normalized values.

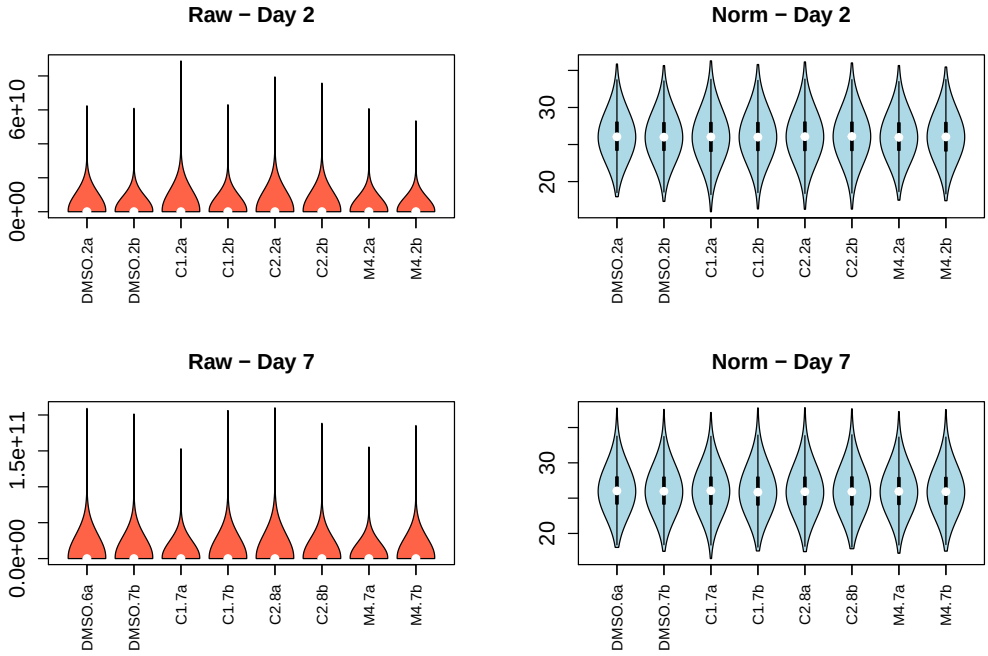


Figure 3: Violin plots of raw and normalized data. Dots correspond to the median values.

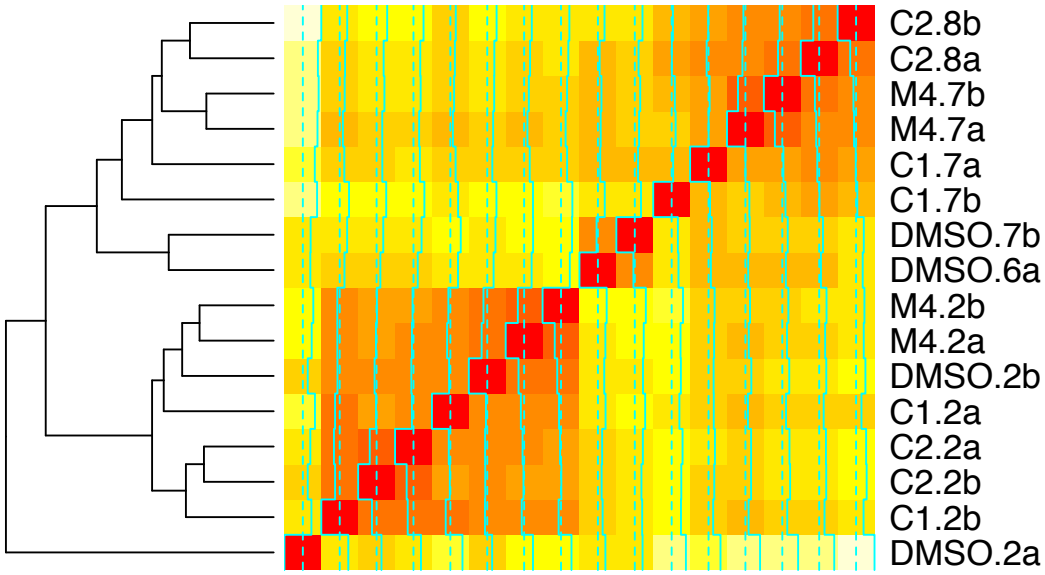


Figure 4: Heatmap of the replicates for quality control.

## 2.2 Pre-processing

We removed background noise and normalized our dataset composed by 2,720 proteins using the *variance stabilizing method* (`vsn`) proposed by Huber (2002). The advantage of this combined approach is that information across arrays can be shared to estimate the background correction parameters, which are otherwise estimated separately for each array. An important tool for assessing whether the `vsn` fit worked is the plot of the means, $m_i$, versus the empirical standard deviations $\sigma_i$. Fig. 1 indicates that the distribution of $\sigma_i$ is concentrated at small values and there is no significant trend of these values as a function of the means.

In Fig. 2 box plots of raw and normalized values are reported for every experimental conditions, while in Fig. 3 the output data of the normalization are reported in violin plots, which are combinations of box plots and kernel density plots.

## 2.3 Quality control

We checked the quality of the experiment performing a cluster analysis on the replicates. We computed the Euclidean distance and applied it in an hierarchical clustering. The heatmap is shown in Fig. 4; it seems to indicate that the two groups of replicates cluster better at day 7. The overall quality of the experiment seems good, except for the sample DMSO.2a, which clusters as an outlier; this is why we decided to leave it out from the analysis. This decision has been motivated also by the outlier analysis we performed using Bland Altman plots as reported in Fig. 5 and Fig. 6. Here we see how the DMSO samples at day 2 have a larger number of

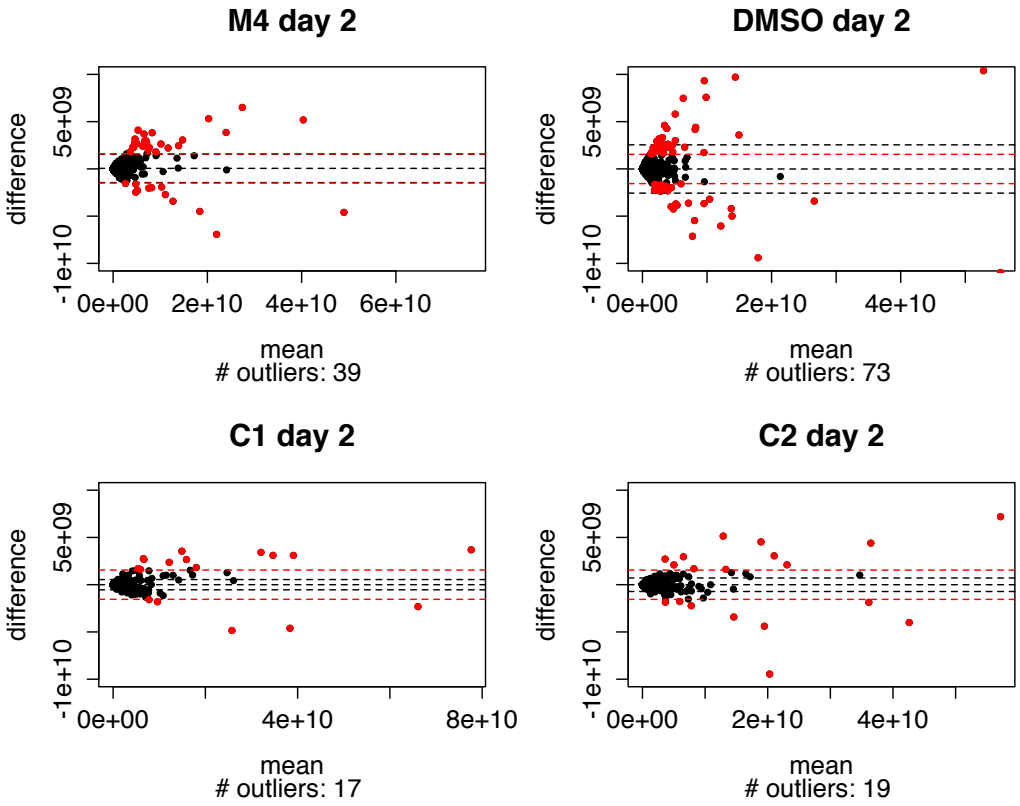


Figure 5: Bland Altman plots for day 2. Black boundaries are at ± two standard deviations of the difference between the two replicates from the same experiment; the red bandwidths are calculated with a pooled standard deviation, calculated on all differences at day 2.

outliers.

# 3 Model and methods

Our aim is to identify differentially expressed proteins and to group proteins that behave in a similar manner. This can be done using the R package `maSigPro` (Conesa et al., 2006) for the analysis of time-course experiments. This method follows a two steps regression strategy to find significant temporal expression changes and significant differences between experimental groups.

## 3.1 The `maSigPro` model

The `maSigPro` procedure makes use of a regression model. Let there be $I$ experimental groups, identified by a qualitative variable, and $J$ time points. Assume that protein expression is measured for $N$ proteins in $R_{ij}$ replications. Let $y_{ijr}$ denote the normalized and transformed expression value of each protein under condition $ijr$. We define $I-1$ dummy variables $D$ to distinguish between each group and a reference group. To explain the evolution of $y$ along time $t$ we consider the following

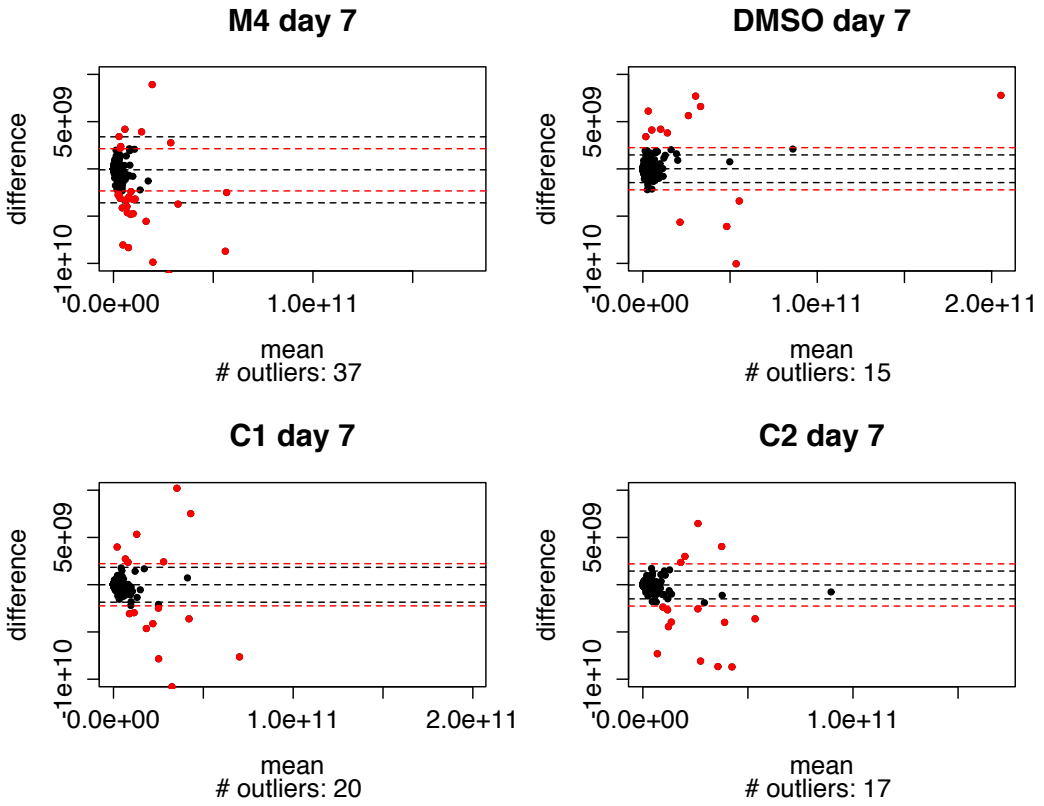


Figure 6: Bland Altman plots for day 7. Black boundaries are at ± two standard deviations of the difference between the two replicates from the same experiment; the red bandwidths are calculated with a pooled standard deviation, calculated on all differences at day 7.

polynomial model, which includes simple time effects and the interactions between the dummies and time:

$$\begin{aligned} y_{ijr} = \beta_0 + \beta_1 D_{1ijr} + \cdots + \beta_{(I-1)ijr} D_{(I-1)ijr} + \\ + \delta_0 t_{ijr} + \delta_1 t_{ijr} D_{1ijr} + \cdots + \delta_{(I-1)} t_{ijr} D_{(I-1)ijr} + \epsilon_{ijr}, \end{aligned}$$

where:

1. $\beta_0, \delta_0$ are the regression coefficients corresponding to the reference group;

2. $\beta_i, \delta_i$ are the regression coefficients that account for specific linear differences between the $(i+1)$-th group profile and the reference group profile;

3. $\epsilon_{ijr}$ is the random variation associated with each protein owing to all sources other than those that have already been incorporated into the model.

The control group of our analysis is the one exposed to water (M4).

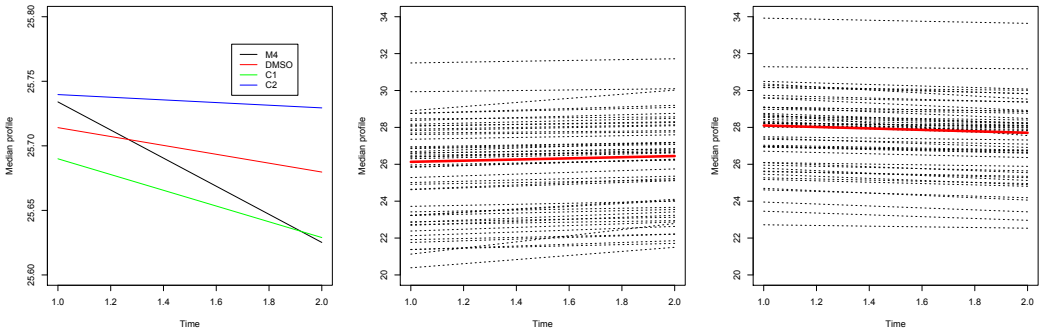


Figure 7: Left panel: median profiles of the 1,903 proteins showing no significant trend excluded at the first selection step of `maSigPro` for each experimental conditions. The absolute median change in time is less than 0.05 (IQR: $< 0.1$). Middle and right panels: profiles of the 108 proteins excluded at the second selection step of `maSigPro`. These proteins only change in time, with a median change (red solid line) of 0.32 (IQR: 0.24) and of $-0.33$ (IQR: 0.24) for, respectively, the upwards and downwards trends.

### 3.2 Two-step analysis

The first step is to estimate, for each protein, using least-squares the parameters of the described general regression model, and then to test:

$$H_0 : \beta_1 = \ldots = \beta_{I-1} = \delta_0 = \delta_1 = \ldots = \delta_{I-1}$$

against

$$H_1 : \exists i \mid \beta_i \neq 0 \vee \delta_i \neq 0, (i = 1, \ldots, I-1).$$

This generates an ANOVA table for each protein. FDR correction is applied in the protein selection. Proteins excluded from the first step behave as if there was no change both in time and among groups.

The second step aims at identifying statistically significant profile changes. This can be done by inspecting, through backward stepwise regression, the regression coefficients of the protein models not excluded at the first step. This additional selection step is based on the $R^2$ value, which we ask to be $\geq 0.60$. The identified proteins can then be clustered in order to identify different patterns. It is advisable to perform a sensitivity analysis, changing algorithm of clustering and distance measure. In our analysis we combined complete linkage associated with the correlation distance measure.

## 4 Results

### 4.1 At step 1

After the first selection step, only proteins showing statistically significant coefficients between the reference group (M4) and any other experimental group and/or time will be kept. Our analysis excluded 1,903 proteins at this step, which exhibit a flat behavior over time and among comparisons; see the left panel of Fig. 7.

| Protein | $\beta_0$ | $\beta_{DMSOvsM4}$ | $\beta_{C1vsM4}$ | $\beta_{C2vsM4}$ | $\beta_t$ | $\beta_{t\cdot DMSO}$ | $\beta_{t\cdot C1}$ | $\beta_{t\cdot C2}$ |
|---|---|---|---|---|---|---|---|---|
| A1C331 | 26.688 | 0 | 0 | 0 | 0.479 | 0 | 0 | 0 |
| E9FQR8 | 29.035 | 0.279 | 0.924 | 0 | 0 | 0 | −0.273 | −0.058 |
| E9FQS5 | 27.166 | 0 | 1.037 | 0 | 0.221 | 0 | −0.173 | −0.101 |
| E9FQT2 | 24.545 | −1.33 | 0 | −1.031 | 0 | 0.27 | 0.122 | 0.281 |
| E9FQT5 | 25.961 | 0 | 0 | 0 | −0.185 | 0.238 | −0.178 | −0.121 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... |

Table 1: Table reporting the coefficients of the first 5 proteins among the 817 selected after the first step of `maSigPro`.

## 4.2 At step 2

In the second step our aim is to select proteins which vary significantly over time and across experimental conditions. Table 1 reports the coefficients of the first 5 proteins out of the 817 proteins selected after the first step of `maSigPro`. For example, the protein "A1C33" passed the first selection because time has a significant effect on his variation, though the experimental conditions do not. The second protein "E9FQR8" shows a strong influence of DMSO and C1 and also a negative interaction between them and time.

*Special patterns* By analyzing the full table of coefficients we can understand a lot about the behavior of proteins. It results that 108 proteins present a change only due to time[1]; see the middle and right panels of Fig. 7. We, furthermore, found that 6 proteins passed to the second step of `maSigPro` only because of the variation due to the experimental conditions, while time is irrelevant.

[1]In fact, these are only 100, but `maSigPro` excludes further 8 proteins. Why is currently unclear to us.

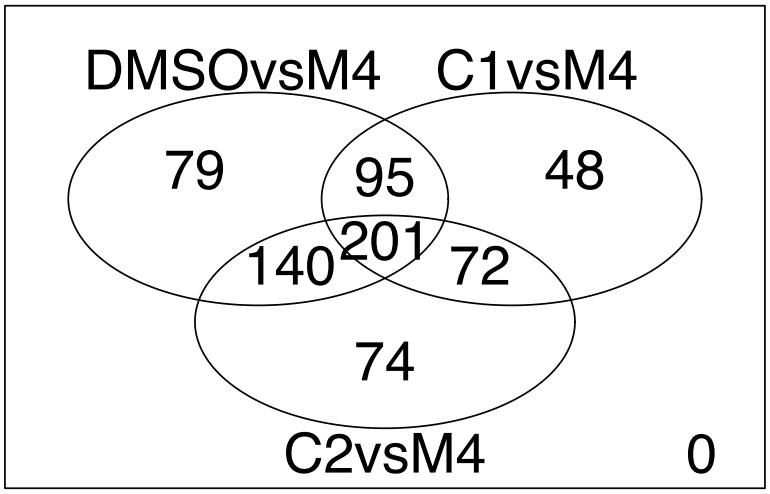


Figure 8: Venn diagram of the 709 significant proteins for each comparison.

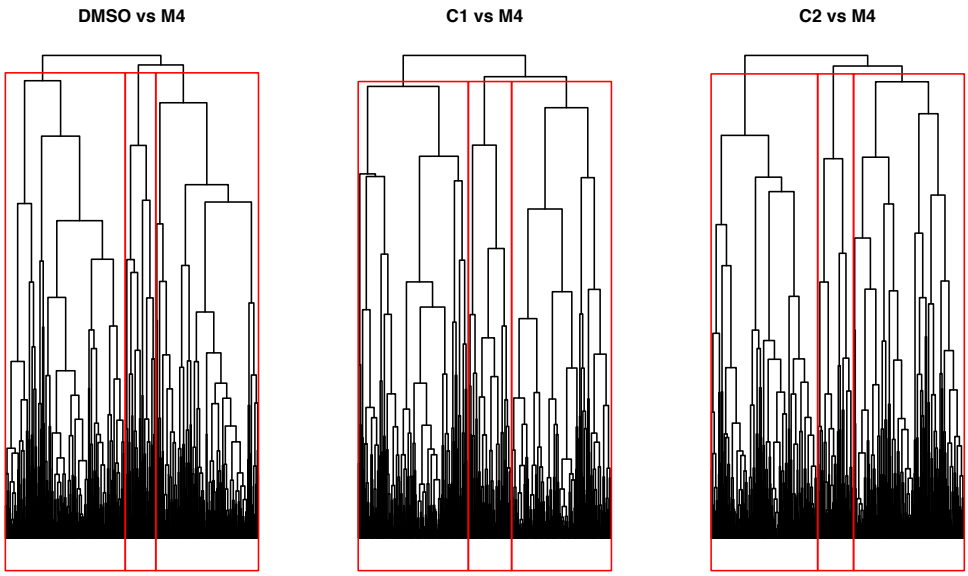


Figure 9: Dendrograms of the cluster analysis.

*Profile clustering* In Fig. 8 we show how the 709 proteins with a significant expression change can be grouped according to the experimental conditions: only 201 change significantly in all three comparisons, while 79 proteins show a change in time and also between treatment groups DMSO and M4, but not in other comparisons.

Dendrograms for the three clustering procedures are reported in Fig. 9, while Fig. 10 and Fig. 11 report how the proteins behave in the three groups and their median profiles for each group. From the median profiles we can always recognize an increasing and decreasing trend in time and also a third group with a less defined structure, though with strong differences between groups.

Fig. 12–14 show the heatmaps of the significant proteins; note that less contamination is present at day 7. Fig. 15 reports the heatmap of the proteins excluded from the analysis.

## 5 Conclusion

In our experiment Tamoxifen was given to daphnids, which was chosen as subject of the experiment because this water flea is easy to manipulate in laboratory, produces clones, reproduces quickly and is at the base of the food chain. Tamoxifen is a widely prescribed anticancer drug worldwide and an endocrine disruptor, it causes side effects in humans, has very powerful metabolites and is very little studied.

Here we wanted to study the response of proteins in organisms under stress, dividing the population in four experimental groups and measuring their production at day 2 and 7. We used a two step regression strategy to identify significantly differentially expressed proteins and to study their behavior, finding that only 709 proteins out of the 2,720 initially considered have to be retained, as they change

in time and among conditions. We grouped proteins on the basis of their profile in three clusters, finding that there is always an increasing, a decreasing and a less well defined pattern in time. These results are at the beginning of a deeper biological investigation about functions and meaning of these patterns.

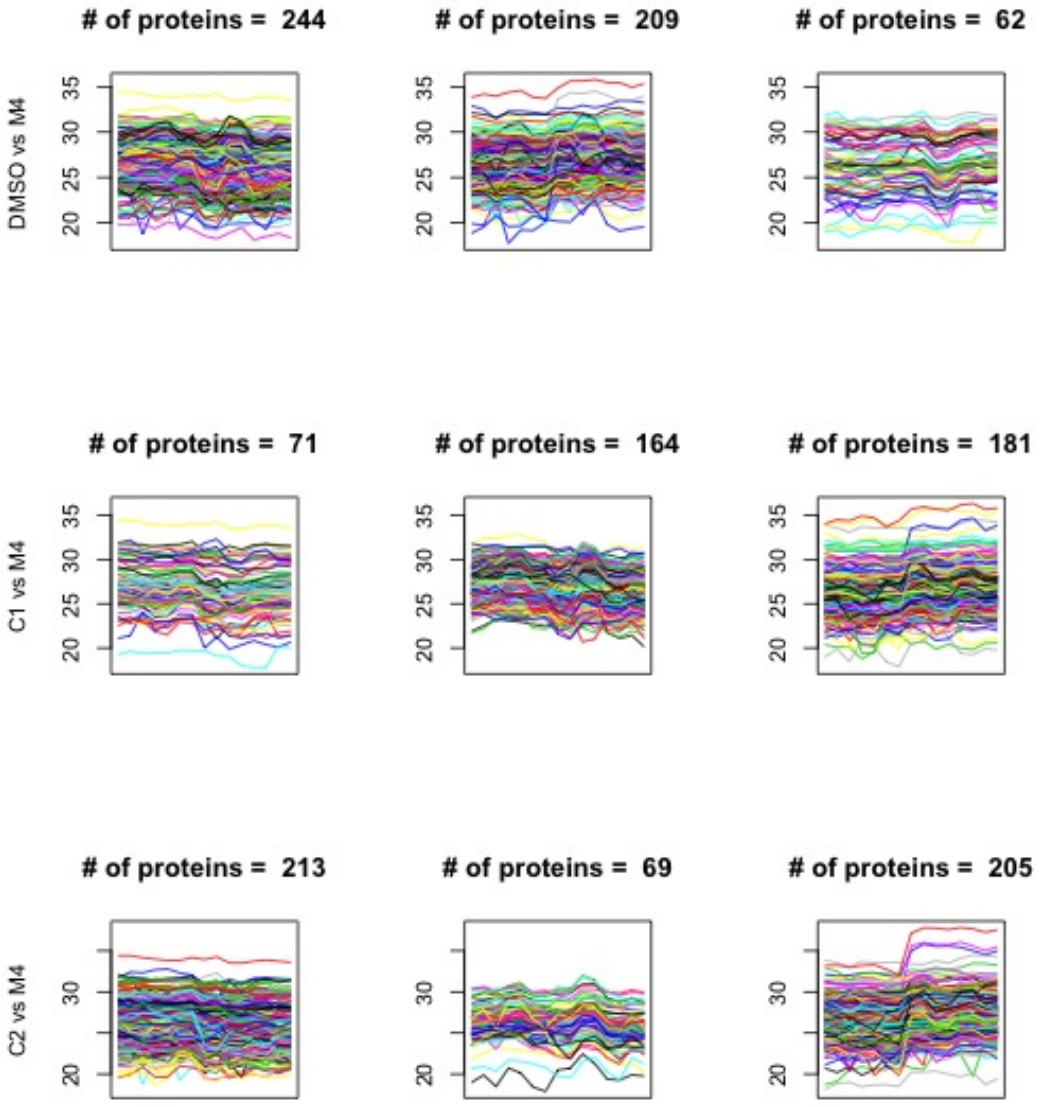


Figure 10: Cluster analysis of significant proteins.

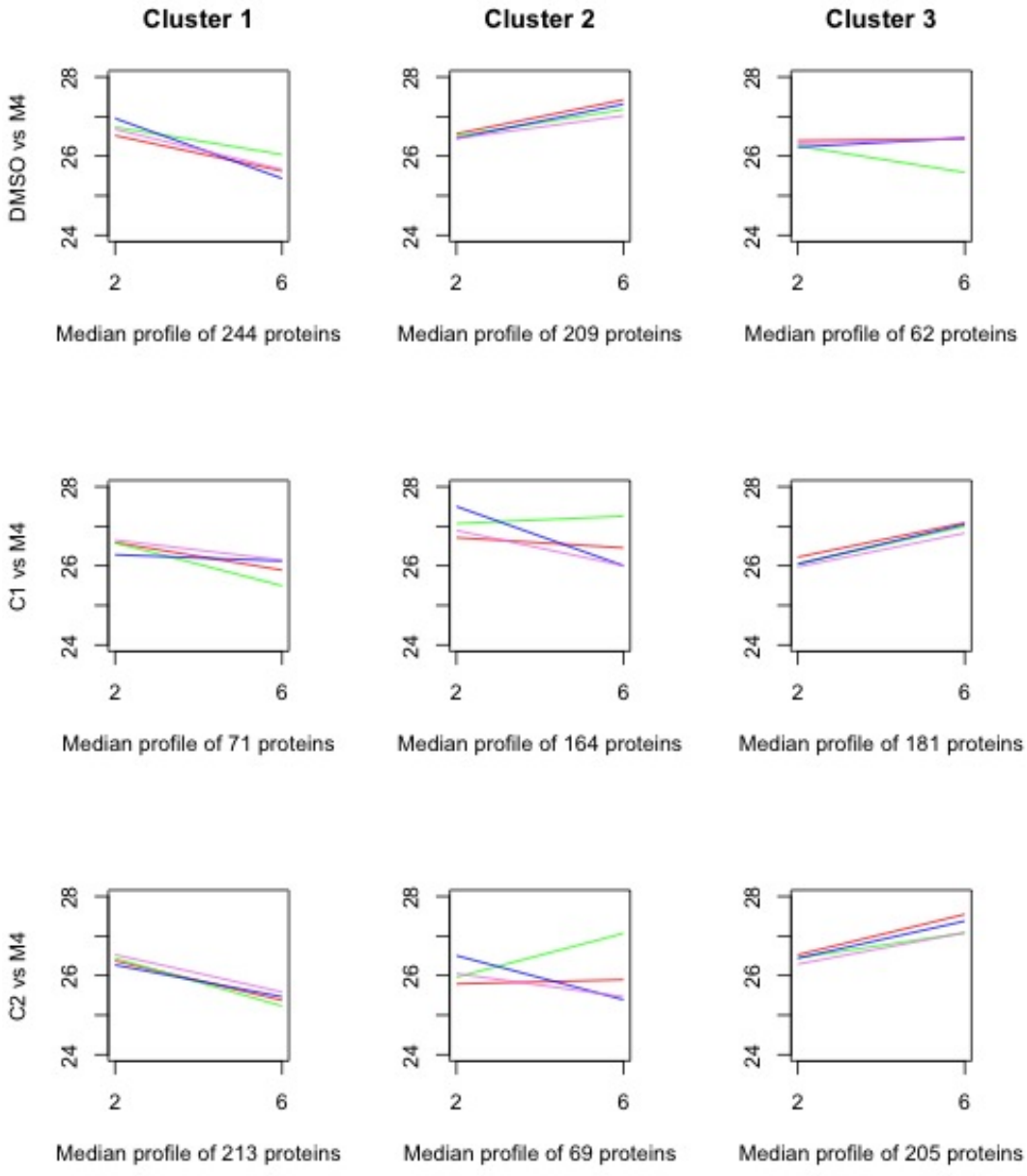


Figure 11: Median expression profiles of significant proteins. Red lines indicate DMSO profiles, green lines represent C1 profiles, blue lines C2 profiles, while the violet ones represent M4 profiles.

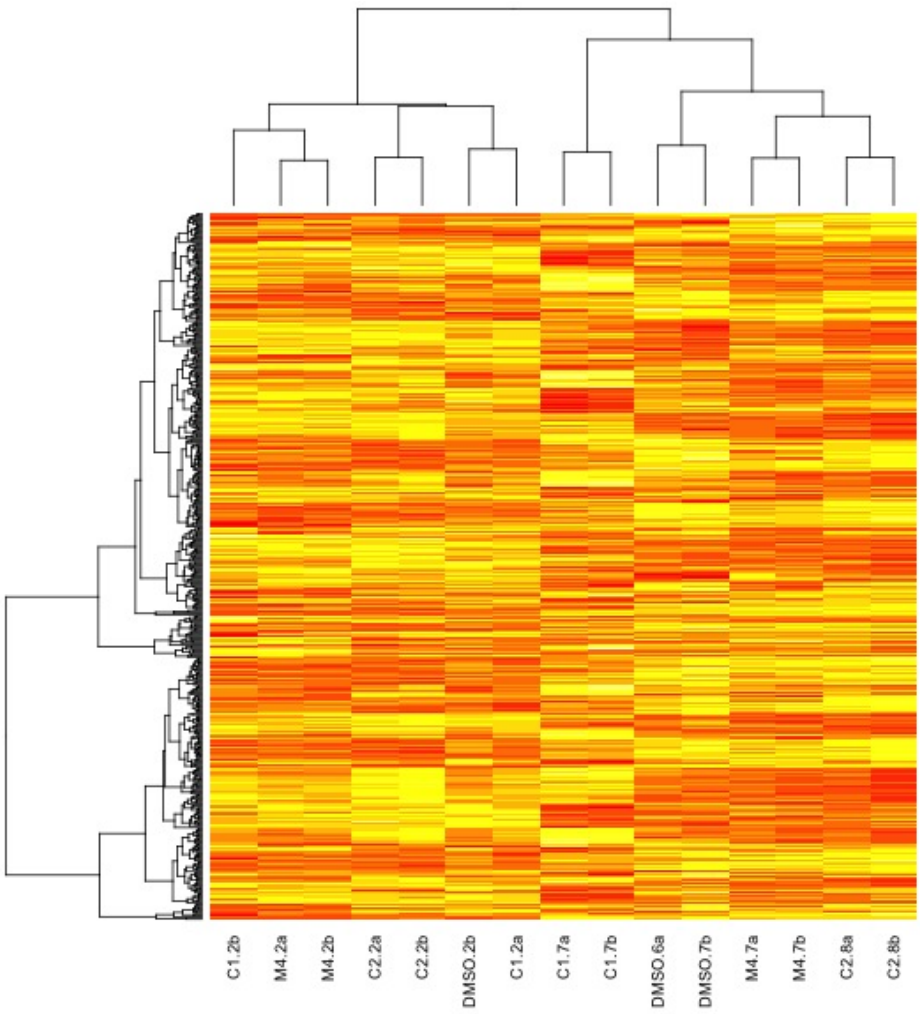


Figure 12: Heatmap of the 515 significant proteins in the comparison DMSO vs M4.

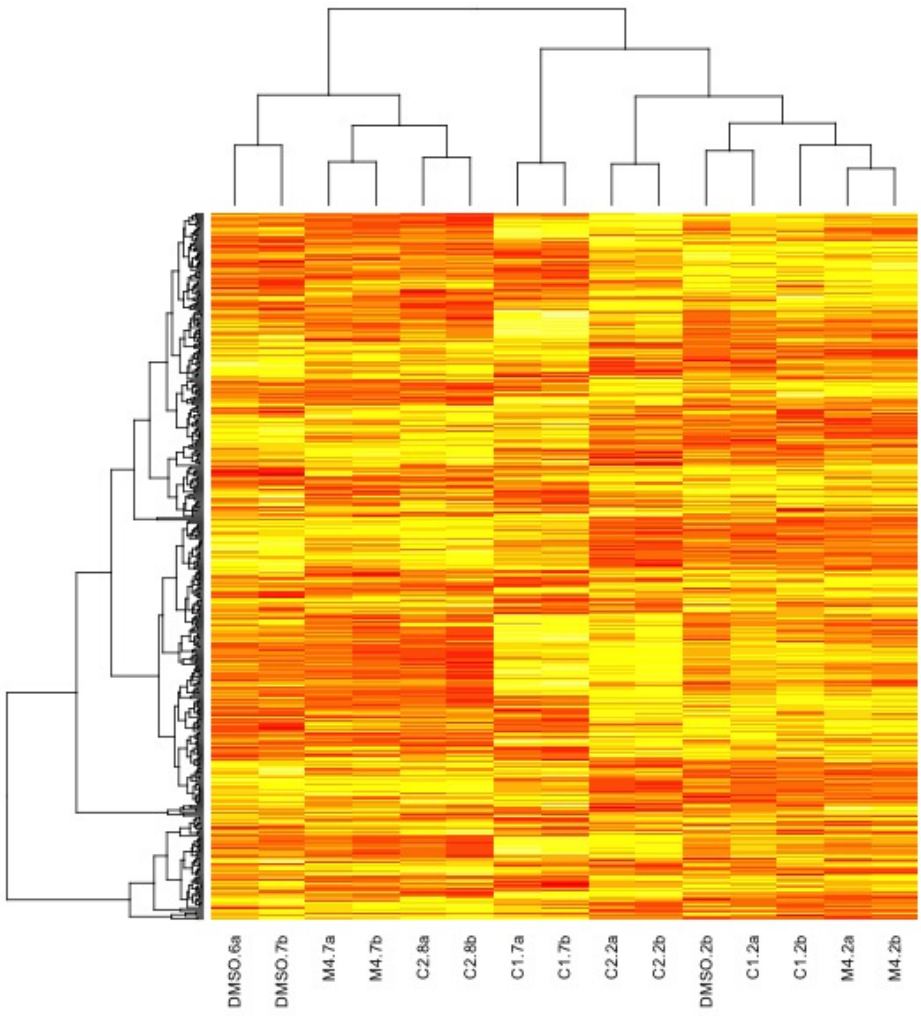


Figure 13: Heatmap of the 416 significant proteins in the comparison C1 vs M4.

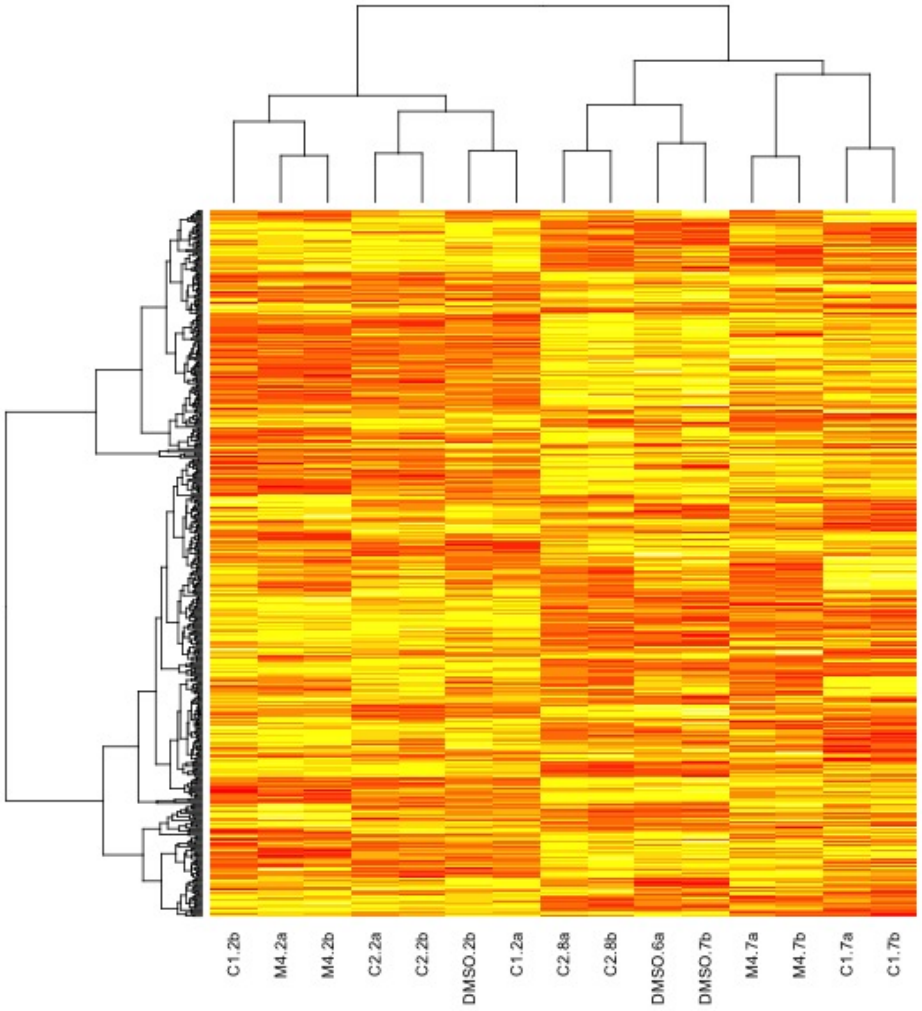


Figure 14: Heatmap of the 487 significant proteins in the comparison C2 vs M4.

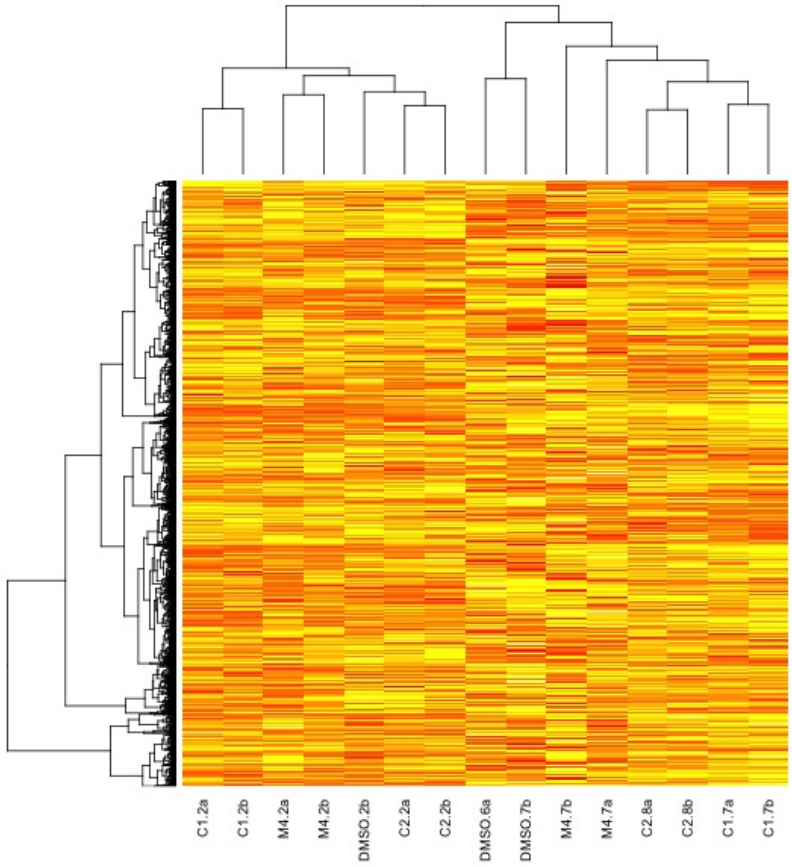


Figure 15: Heatmap of the 2,011 proteins excluded from the analysis.

# 6 References

Conesa, A., Nueda, M. J., Ferrer, A., Talón M. (2006) *maSigPro: a method to identify significantly differential expression profiles in time-course microarray experiments* Bioinformatics, 22 (6), 1096-1102.

Huber, W., von Heydebreck, A., Sültmann, H., Poustka, A. (2002). *Variance stabilization applied to microarray data calibration and to the quantification of differential expression* Bioinformatics, 18, suppl. 1, S96–S104.

Shaw, J.R., Pfrender, M. E., Eads, B. D., et al. (2008). *Daphnia as an emerging model for toxicological genomics* Advances in Experimental Biology, 2, 165–328.

# A Files

| file name | contents |
|---|---|
| `1-original_database.txt` | all 2,720 proteins |
| `2-excluded_at_step1.txt` | 1,903 proteins excluded at the first `maSigPro` step |
| `3-excluded_at_step2.txt` | 108 proteins excluded at the second `maSigPro` step |
| `4-retained_at_step2.txt` | 709 proteins retained after the second `maSigPro` step |
| `4a-venn_1intersect_DMSO.txt` | 79 proteins at the M4-DMSO intersection of the Venn diagram |
| `4b-venn_1intersect_C1.txt` | 48 proteins at the M4-C1 intersection of the Venn diagram |
| `4c-venn_1intersect_C2.txt` | 74 proteins at the M4-C2 intersection of the Venn diagram |
| `4d-venn_2intersect_DMSOC1.txt` | 95 proteins at the M4-DMSO-C1 intersection of the Venn diagram |
| `4e-venn_2intersect_DMSOC2.txt` | 140 proteins at the M4-DMSO-C2 intersection of the Venn diagram |
| `4f-venn_2intersect_C1C2.txt` | 72 proteins at the M4-C1-C2 intersection of the Venn diagram |
| `4g-venn_3intersect.txt` | 201 proteins at the M4-DMSO-C1-C2 intersection of the Venn diagram |
| `5-only_cond.txt` | 6 proteins which change only across experimental condition (but not in time) |
| `6-only_time.txt` | 100 proteins which change only in time (but not across experimental condition) |

## Working Paper Series
## Department of Statistical Sciences, University of Padua

You may order paper copies of the working papers by emailing wp@stat.unipd.it

Most of the working papers can also be found at the following url: http://wp.stat.unipd.it

UNIVERSITÀ DEGLI STUDI DI PADOVA
DIPARTIMENTO DI SCIENZE STATISTICHE

Department of Statistical Sciences
*University of Padua*
*Italy*