CLASSIFICATION OF CHEMICAL COMPOUNDS BASED ON THE CORRELATION BETWEEN IN VITRO GENE EXPRESSION PROFILES

Takeshita, Jun-ichi
Research Institute of Science for Safety and Sustainability, National Institute of Advanced Industrial Science and Technology (AIST)

Toyoda, Akinobu
School of Science and Engineering, University of Tsukuba

Tani, Hidenori
Environmental Management Research Institute, National Institute of Advanced Industrial Science and Technology (AIST)

Endo, Yasunori
Faculty of Systems and Information Engineering, University of Tsukuba

他

https://doi.org/10.5109/4476055

出版情報：Bulletin of informatics and cybernetics. 53 (4), pp.1-14, 2021. 統計科学研究会
バージョン：
権利関係：

CLASSIFICATION OF CHEMICAL COMPOUNDS BASED ON THE CORRELATION BETWEEN *IN VITRO* GENE EXPRESSION PROFILES

by

Jun-ichi Takeshita, **Akinobu Toyoda**, **Hidenori Tani**, **Yasunori Endo**
and
Sadaaki Miyamoto

Reprinted from the Bulletin of Informatics and Cybernetics
Research Association of Statistical Sciences, Vol.53, No. 4

FUKUOKA, JAPAN
2021

Bulletin of Informatics and Cybernetics, Vol. 53, No. 4, 2021

CLASSIFICATION OF CHEMICAL COMPOUNDS BASED ON THE CORRELATION BETWEEN *IN VITRO* GENE EXPRESSION PROFILES

By

Jun-ichi Takeshita[*] **Akinobu Toyoda**[†] **Hidenori Tani**[‡] **Yasunori Endo**[§]
and
Sadaaki Miyamoto[¶]

Abstract

Toxicity evaluation of chemical compounds has traditionally relied on animal experiments; however, the demand for non-animal-based prediction methods for toxicology of compounds is increasing worldwide. Our aim was to provide a classification method for compounds based on *in vitro* gene expression profiles. The *in vitro* gene expression data analyzed in the present study was obtained from our previous study. The data concerned nine compounds typically employed in chemical management. We used agglomerative hierarchical clustering to classify the compounds; however, there was a statistical difficulty to be overcome. We needed to properly extract RNAs for clustering from more than 30,000 RNAs. In order to overcome this difficulty, we introduced a combinatorial optimization problem with respect to both gene expression levels and the correlation between gene expression profiles. Then, the simulated annealing algorithm was used to obtain a good solution for the problem. As a result, the nine compounds were divided into two groups using 1,000 extracted RNAs. Our proposed methodology enables read-across, one of the frameworks for predicting toxicology, based on *in vitro* gene expression profiles.

Key Words and Phrases: Statistical classification, Multiobjective combinatorial optimization, Chemical toxicity, Alternatives to animal experiments, *In vitro* gene expression, Mathematical formulation

1. Introduction

Traditionally, toxicity evaluation of chemical compounds has relied on animal experiments [Eaton and Gilbert (2015)]. However, in terms of time, cost efficiency, and animal welfare concerns, there is an increasing demand for the development of non-animal-based

[*] Research Institute of Science for Safety and Sustainability, National Institute of Advanced Industrial Science and Technology (AIST), 16–1 Onogawa Tuskuba 305-8569 Japan. tel +81–50-3648-3533 jun-takeshita@aist.go.jp

[†] School of Science and Engineering, University of Tsukuba, 1–1–1 Tennodai, Tsukuba, 305-8573 Japan.

[‡] Environmental Management Research Institute, National Institute of Advanced Industrial Science and Technology (AIST), 16–1 Onogawa Tuskuba 305-8569 Japan.

[§] Faculty of Systems and Information Engineering, University of Tsukuba, 1–1–1 Tennodai, Tsukuba, 305-8573 Japan.

[¶] Faculty of Systems and Information Engineering, University of Tsukuba, 1–1–1 Tennodai, Tsukuba, 305-8573 Japan.

methodologies for predicting chemical toxicity. Recently, several elements have been proposed for use in toxicity prediction methods, such as (quantitative) structure-activity relationships ((Q)SAR), quantitative activity-activity relationships (QAAR), and read-across [OECDa (2014)].

Read-across is a method whereby the toxicity of a given compound is predicted without the use of animal test data, but instead based on the animal toxicity data for similar compounds. One study, for example, conducted read-across based on similarity in chemical structure and toxicology expert judgment [Mellor *et al.* (2017)]. For read-across, it is important to properly group compounds using data that does not rely on animal experiments. In order to facilitate the grouping of compounds, the Organisation for Economic Co-operation and Development (OECD) published the OECD QSAR Toolbox [Madden (2013), Section 4.3], [OECDb (2017)]. In addition, the Japanese government and participating academic institutes developed the Hazard Evaluation Support System Integrated Platform (HESS) [Sakuratani *et al.* (2013)], [Madden (2013), Section 4.4], [NITE (2020)]. These platforms group compounds based on *in silico* parameters, that is, chemical structures and essential physicochemical parameters; and by employing them, we can successfully group compounds based on these parameters.

In order to predict toxicity, it is useful to use not only *in silico* parameters but also *in vitro* parameters, because the latter can reflect certain biological characteristics of compounds. In fact, in the case of predicting hepatotoxicity, one of the most prevalent forms of toxicity, two studies have reported that using *in vitro* parameters increased the accuracy of discriminative models for predicting the presence or absence of hepatotoxicity, compared to using *in silico* parameters [Liu *et al.* (2015)], [Low *et al.* (2011)]. These studies suggest the hypothesis that grouping compounds based on *in vitro* parameters would increase the accuracy of read-across approaches for predicting chemical toxicity, compared to grouping based on *in silico* parameters.

The present study proposes a methodology for grouping compounds using *in vitro* gene expression data. In order to group compounds, agglomerative hierarchical clustering was applied; however, a statistical difficulty appeared. The gene expression data, from which it was necessary to properly extract a limited number of RNAs for clustering compounds, included more than 30,000 RNAs. In order to overcome this difficulty, we introduced a multiobjective combinatorial optimization problem with respect to both gene expression levels and the correlation between gene expression profiles. Then, we applied the simulated annealing algorithm, a metaheuristic algorithm, to obtain a good solution for the multiobjective combinatorial optimization problem.

2. Methods and materials

2.1. Gene expression data

The present study used the data reported in Tani *et al.* (2019). In that study, cell-based assays were conducted, in duplicate, using mouse embryonic stem cells, for nine compounds typically employed in chemical management: bis-phthalate, *p*-dicholorobenzene, phenol, trichloroethylene, benzene, chloroform, *p*-cresol, and tri-*n*-butyl-phosphate. In other words, each compound had two samples, and each group of nine samples had the same control condition. Then, the gene expression levels were quantified using the fragments per kilobase of exon per million mapped fragments (FPKMs), and reported for a total of 32,586 RNAs. We used these RNAs and their FPKMs in the present study's

analysis.

2.2. Gene expression ratio

For any compound-treated group, the following gene expression ratio was used for each gene:

$$f(x, y) = \log_2 \left(\frac{y}{x} \right),$$

where x and y denote the gene expression levels of the control and compound-treated groups, respectively. Note that if x or y was zero, then the next smallest value in the respective group was used.

2.3. Clustering

In order to group compounds, agglomerative hierarchical clustering (the average linkage between the merged groups) was used, because this method can be used for any dissimilarity measures. The following dissimilarity measure was used for the hierarchical clustering: for any two compounds, x and y, the dissimilarity measure, or distance, between x and y, say, $d(x, y)$ was defined by

$$d(x, y) = \frac{1 - \mathrm{corr}(x, y)}{2},$$

where $\mathrm{corr}(x, y)$ is the correlation coefficient between the respective FPKM vectors of Compound x and Compound y. The dissimilarity measure takes a value between 0 and 1.

2.4. Selection of RNAs

In order to extract RNAs that clearly revealed the difference between compounds, the present study introduced the following combinatorial optimization problem to extract n RNAs for a given natural number n:

$$\begin{aligned}
&\text{objective function} \\
&\quad U = (1 - \alpha)U_1 + \alpha U_2 \ (0 \le \alpha \le 1), \\
&\quad U_1 = \frac{1}{\mathrm{Count}(w)} \times \\
&\qquad \sum_{x,y \in E, x \neq y} w_{x,y} \frac{|(1/n)\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})|}{\{(1/n)\sum_{i=1}^{n}(x_i - \overline{x})^2\}^{1/2}\{(1/n)\sum_{i=1}^{n}(y_i - \overline{y})^2\}^{1/2}}, \\
&\quad U_2 = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{\|r_i\|}{\max_j \|r_j\|} \right), \\
&\text{subject to} \quad \{x_i\}, \{y_i\} \subset \Gamma \text{ and } \#\{x_i\} = \#\{y_i\} = n,
\end{aligned}$$

where E and Γ denote the set of compound-treated groups and the set of all the RNAs (32,586 RNAs), respectively; x_i and y_i denote the expression ratios of RNA_i for x and $y \in E$, respectively; $\overline{x}$ and $\overline{y}$ denote the means of x_i and y_i, respectively; $w_{x,y}$ denotes weights and takes a value in $\{-1, 0, 1\}$; and $\mathrm{Count}(w)$ denotes the number of $w_{x,y}$ taking

a number of 1. Note that, for some x and y, if $w_{x,y} = 1$, we can extract RNAs that increase the correlation between Compound x and Compound y; and if $w_{x,y} = -1$, we can extract RNAs that decrease the correlation between Compound x and Compound y; otherwise, we are not interested in the correlation between the two compounds. In addition, r_i designates the FPKM vector for RNA_i, and $\|\cdot\|$ represents the Euclidean norm.

The function U_1 takes a value in $[0, 1]$, since the sigma component is divided by $\mathrm{Count}(w)$. The function U_2 also takes a value in $[0, 1]$. Thus, the objective function U takes a value in $[0, 1]$, since α is a parameter taking a value in $[0, 1]$. Roughly speaking, the function U_1 describes the strength of the correlation between two compound-treated groups, x and y, based on extracted RNAs. The function U_2 describes the gene expression level of extracted RNAs. Then, the function U is a linear combination of U_1 and U_2.

The simulated annealing algorithm, which was originally introduced by Cerny (1985) and Kirkpatrick *et al.* (1983), was used to obtain a good solution for the combinatorial optimization problem. Let n be the number of RNAs we want to extract, T be the initial temperature, T_t $(0 < T_t < T)$ be the final temperature, and γ $(0 < \gamma < 1)$ be the cooling rate. Then, the following is the algorithm to extract RNAs.

Step 1 Choose n RNAs randomly as the initial state. Let R be the set of the chosen n RNAs, and then calculate the objective function U using the set R. In addition, set the initial temperature T.

Step 2 As a neighbor solution, generate a set R', in which random elements of R and $\overline{R}$ (the complement set of R) are exchanged. Then, calculate the objective function U' using the set R'.

Step 3 If $U < U'$, then $R := R'$; otherwise $R := R'$ with the following probability P:

$$P = \exp\left(-\frac{|U' - U|}{T}\right).$$

Step 4 $T := \gamma T$. If $T \geq T_t$, output the set R as the final state; otherwise, go to Step 2.

3. Results and discussion

3.1. Grouping compounds using all the RNAs

Figure 1 shows a dendrogram obtained by applying agglomerative hierarchical clustering (the average linkage between the merged groups) to the data set of gene expression ratios, and Figure 2 a dendogram obtained by applying similar clustering to the set of gene expression levels, for the nine compounds in duplicate and all the RNAs (32,586 RNAs). In both figures, the y-axes show the dissimilarity measures where two clusters were merged, while the x-axes show the distribution of compounds. Note that subscripts 1 and 2 refer to the sample numbers; that is, if any two compounds have the same subscript number, the control conditions for the two compounds are identical.

In Figure 1, we can see that there are two robust clusters, and each cluster consists of the nine compounds with the same subscript number. This implies that the two compound groups strongly depend on the gene expression levels in the control conditions. In Figure 2, the two control conditions are closely proximate, but there is no significant

difference between the 18 samples (the nine compounds in duplicate). These results indicate that there is no significant difference between the gene expression patterns of the two control conditions. Therefore, we may infer that the significant difference between the two control conditions in Figure 1 was the result of the incremental accumulation of difference among the RNAs, as more than 30,000 RNAs were in the data set. These results indicate that it is unreliable to use the information of all the RNAs for classifying compounds, but instead we must extract a limited number of RNAs in order to properly classify compounds using *in vitro* gene expression data.

3.2. Grouping compounds using extracted RNAs

We used the parameters $n = 3,000, 1,000, 100$, and $\alpha = 0.0, 0.1, 0.2, 0.3$, for the combinatorial optimization problem to extract RNAs; and $T = 0.01$, $T_t = 0.00001$, and $\gamma = 0.9999$ for the simulated annealing algorithm. Figures 3, 4, and 5 show dendrograms obtained by agglomerative hierarchical clustering (the average linkage between the merged groups) applied to the data set with the nine compounds in duplicate, and $3,000$, $1,000$, and 100 extracted RNAs, respectively. Each figure has four panels. The upper-left (a), upper-right (b), lower-left (c), and lower-right (d) panes correspond to the cases of $\alpha = 0.0, 0.1, 0.2$, and 0.3, respectively. In each panel in these figures, the y-axis shows the dissimilarity measures where two clusters were merged, while the x-axis shows the distribution of compounds. The meaning and function of the subscripts are as in Figures 1 and 2.

In Figures 3, 4, and 5, it is clear that the two samples are initially merged for all the compounds, and then different compounds are merged. However, when $3,000$ RNAs were extracted (Figure 3), there are roughly 0.05 dissimilarity measures between the two samples of each compound, although the maximum ranges are roughly 0.20. This result suggests that $3,000$ RNAs are not suitable for classifying compounds because the dissimilarity measures between the two samples of each compound are not small enough, compared to the dissimilarity measures between the compounds. On the other hand, when $1,000$ or 100 RNAs were extracted (Figures 4 and 5, respectively), and $\alpha = 0.1$ or 0.2, the two samples are merged with sufficiently small dissimilarity measures, and we can clearly see the difference between compounds. These results demonstrate the necessity of limiting the number of RNAs when classifying compounds to assess the effects of compounds on gene expression patterns.

Next, we compare the cases when $1,000$ and 100 RNAs were extracted (Figures 4 and 5, respectively). When $\alpha = 0.1$, neither case shows clear cluster structures; however, when $\alpha = 0.2$, with $1,000$ extracted RNAs, the nine compounds are divided into two groups, one consisting of three compounds (bis-phthalate, trichloroethylene, and tri-n-butyl-phosphate) and the other of the remaining six compounds. These results indicate that not only limiting the number of RNAs but also reflecting the sizes of RNAs is effective for revealing the difference between the compounds. Figures 6 ($\alpha = 0$) and 7 ($\alpha = 0.2$) show the scatter plots of all $32,586$ RNAs between Samples 1 and 2, in the case of bis-phthalate. The red plots indicate the $1,000$ extracted RNAs, and the blue plots the remaining RNAs. When $\alpha = 0.0$, it is clear that, for the most part, only RNAs with near-zero FPKMs are extracted; whereas, when $\alpha = 0.2$, there is an increased number of RNAs of significant size. Thus, using $n = 1,000$ and $\alpha = 0.2$ in the combinatorial optimization problem would be best for extracting RNAs in order to classify these compounds.

4. Concluding remarks

The present study applied agglomerative hierarchical clustering methods and a multiobjective combinatorial optimization problem, to classify chemical compounds based on *in vitro* gene expression data, an approach that enables read-across based on *in vitro* parameters, and the prediction of chemical toxicity. Further investigations could be considered, based on this research; and thus, to conclude the present study, this section will discuss some further challenges.

First, the present study applied agglomerative hierarchical clustering, a traditional clustering method, well-known in many fields; but several clustering methods present distinctive challenges for classifying chemical compounds [Raymond *et al.* (2003)]. In particular, methods based on graphs describing chemical structures are relatively new and, compared to traditional methods, notable for their ability to present different aspects of the given data. To apply graph-based methods, however, we must consider how to construct graph structures based on *in vitro* gene expression data. This presents an interesting challenge for future research.

Second, the present study used $U = (1 - \alpha)U_1 + \alpha U_2$ as the objective function to extract RNAs, because correlation coefficients and the Euclidean norms are well-known and common to several fields. However, there are several means, other than utilizing U_1 and U_2, for describing the strength of the correlation between two compound-treated groups and the gene expression level of extracted RNAs. Performing the related analysis using such different means, and comparing the respective results, is another interesting avenue for future research.

Third, the combinatorial optimization problem here included two functions (U_1 and U_2) to be optimized. In general, multiobjective optimization problems have a set of Pareto optimal solutions, because two objective functions have a trade-off relationship. In other words, an optimal solution is decided for each weight between the two objective functions (α in the present study), when a linear combination of the two objective functions is considered. This implies that we do not have a general method for determining the best α (see essential textbooks on the theory of multiobjective optimization problems; e.g., Sawaragi *et al.* (1985)). The present study, therefore, analyzed the cases of $\alpha = 0.0, 0.1, 0.2$, and 0.3, and then chose the best α among them. Statistically, the best α could be obtained using cross-validation methods, but the present study obtained gene expression data for only nine chemical compounds, and the number of samples was insufficient for utilizing cross-validation methods. It is difficult to obtain *in vitro* gene expression data for a large number of chemicals, due to the time and cost involved in conducting the necessary experiments; and scenarios in which the number of chemical compounds is small, but each compound has a host of characteristic parameters, are currently very common. Thus, several statistical methods have been proposed, in the recent years, to address such scenarios (Aoshima *et al.*, 2018). Investigation of chemical compound classification methods based on such new statistical methods presents a further challenge.

Finally, we should note that, even if our results can properly classify chemical compounds using *in vitro* parameters, the results may not show the same similarities among chemical compounds as in *in vivo* experiments. There may be gaps, in other words, between the respective similarities based on *in vitro* parameters and *in vivo* experiments. To reduce such gaps, a means must be developed for extracting RNAs such that the results will reflect the similarities determined by *in vivo* experiments, and

this requires collaboration with toxicology experts. This is a consideration for the future, but our approach here would aid in such analysis.

Acknowledgement

The authors would like to express appreciation to Mr. Ryosuke Abe, Dr. Hiroshi Aoki, Dr. Hiroaki Sato, Dr. Masaki Torimura, and Dr. Masashi Gamo for thier fruitful discussions for the first version of manuscript. This study is supported by JSPS KAKENSHI Grant Numbers JP20K12203 and JP20H05797.

References

Aoshima, M. (2018). A survey of high dimension low sample size asymptotics, *Aust. N. Z. J. Stat.*, 60, 4-19.

Cerny, V. (1985). Thermodynamical approach to the traveling salesman problem : An efficient simulation algorithm, *J. Optim. Theory Appl.*, 45, 41-51.

Eaton, D.L. and Gilbert, S.G. (2015). Principles of toxicology, In: Klaassen, C.D. and Watkins, III J.B. (eds.) *Casarett Doull's Essentials Toxicology*, Third Edition, 5-20, McGraw-Hill Education, NY, USA.

Kirkpatrick, S., Gelatt, C.D., and Vecchi, M.P. (1983). Optimization by simulated annealing, *Science*, 220, 671-680.

Kramer, A *et al.* (2014). Causal analysis approaches in Ingenuity Pathway Analysis, *Bioinformatics*, 30, 523–530.

Liu, J. *et al.* (2015). Predicting hepatotoxicity using ToxCast in vitro bioactivity and chemical structure, *Chem. Res. Toxicol.*, 28, 738-751.

Low, Y. *et al.* (2011). Predicting drug-induced hepatotoxicity using QSAR and toxicogenomics approaches, *Chem. Res. Toxicol.*, 24, 1251-1262.

Madden, J.C. (2013). Tools for grouping chemicals and forming categories, In: Cronin, M. *et al.* (eds) *Chemical Toxicity Prediction: Category Formation and Read-Across*, 72-97, Royal Society of Chemical Publishing, London.

Mellor, C.L. *et al.* (2017). Read-across for rat oral gavage repeated-dose toxicity for short-chain mono-alkylphenols: A case study, *Comput. Toxicol.*, 2, 1-11.

NITE: Hazard Evaluation Support System Integrated Platform (HESS) (2020). `http://www.nite.go.jp/en/chem/qsar/hess-e.html`, Accessed 5 January 2021.

OECD (2014). Guidance on grouping of chemicals, Second edition, Number 194, ENV/JM/MONO(2014)4, Paris, France.

OECD (2017). The OECD QSAR Toolbox. `http://www.oecd.org/env/ehs/risk-assessment/theoecdqsartoolbox.htm`, Accessed 5 January 2021

Raymond, J.W., Blankley, C.J., and Willett, P. (2003). Comparison of chemical clustering methods using graph- and fingerprint-based similarity measures, *J. Mol. Graph. Model.*, 21, 421-433.

Sakuratani, Y. *et al.* (2013). Hazard Evaluation Support System (HESS) for predicting repeated dose toxicity using toxicological categories, *SAR QSAR Environ. Res.*, 24, 351-63.

Sawaragi, Y., Nakayama, H., and Tanino, T. (1985). *Theory of Multiobjective Optimization*, Academic Press.

Tani, H. *et al.* (2019). Identification of RNA biomarkers for chemical safety screening in mouse embryonic stem cells using RNA-seq, *Biochem. Biophys. Res. Commun.*, 512, 641-646.

Received: March 3, 2021
Revised: May 13, 2021
Accept: May 21, 2021

Figures

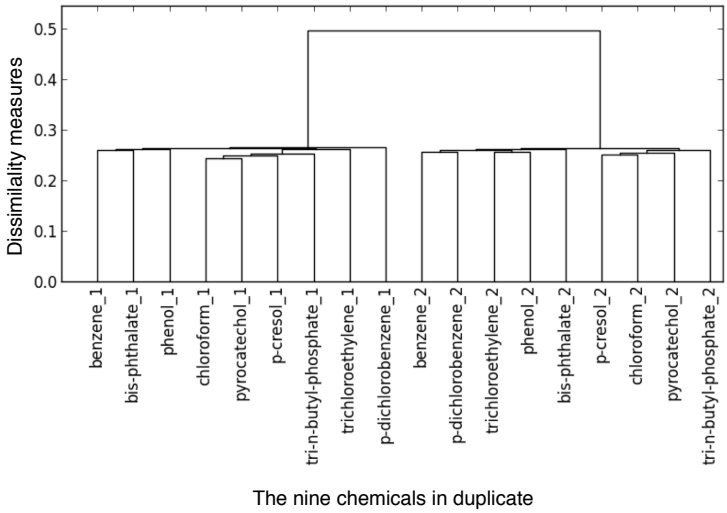

Figure 1: A dendrogram obtained by applying aggregative hierarchical clustering (the average linkage between the merged groups) to the data of the gene expression ratios for the nine compounds and 32,586 RNAs. The y-axis marks the dissimilarity measures at which the clusters merge, and the x-axis the distribution of the nine compounds in duplicate.

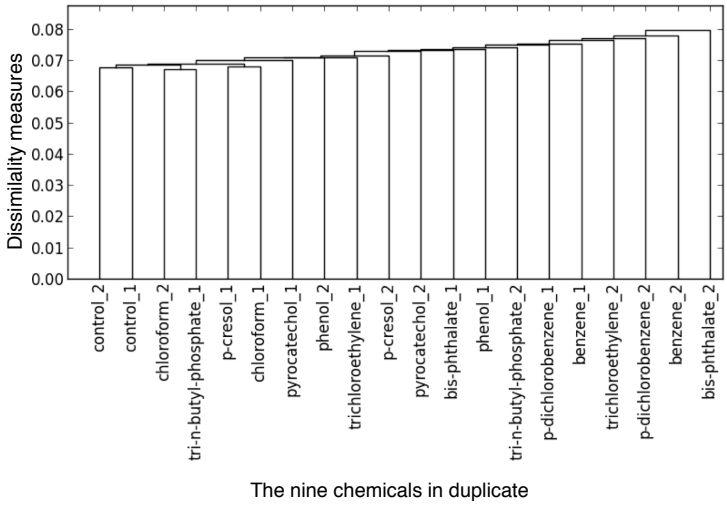

Figure 2: A dendrogram obtained by applying aggregative hierarchical clustering (the average linkage between the merged groups) to the data of the gene expression levels for the nine compounds and 32,586 RNAs. The y-axis marks the dissimilarity measures at which the clusters merge, and the x-axis the distribution of the nine compounds in duplicate.

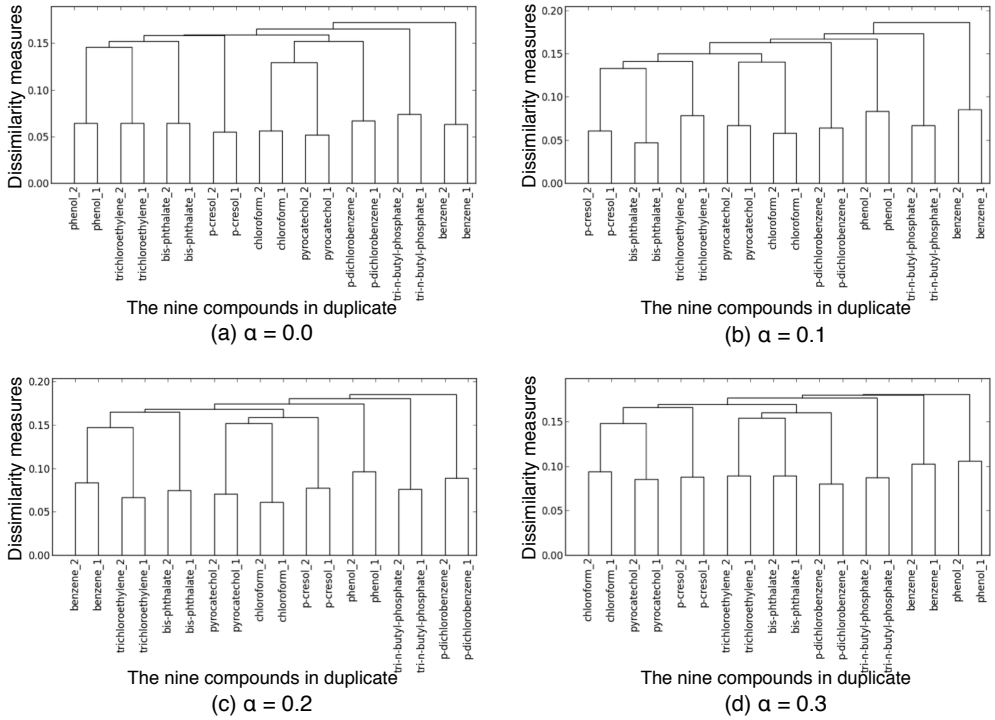

Figure 3: Four dendrograms obtained by applying aggregative hierarchical clustering (the average linkage between the merged groups) to the data of the gene expression ratios for the nine compounds and 3,000 extracted RNAs. The upper-left (a), upper-right (b), lower-left (c) and lower-right (d) panels are the cases of $\alpha = 0.0, 0.1, 0.2$, and 0.3, respectively. In each panel, the y-axis marks the dissimilarity measures at which the clusters merge, and the x-axis the distribution of the nine compounds in duplicate.

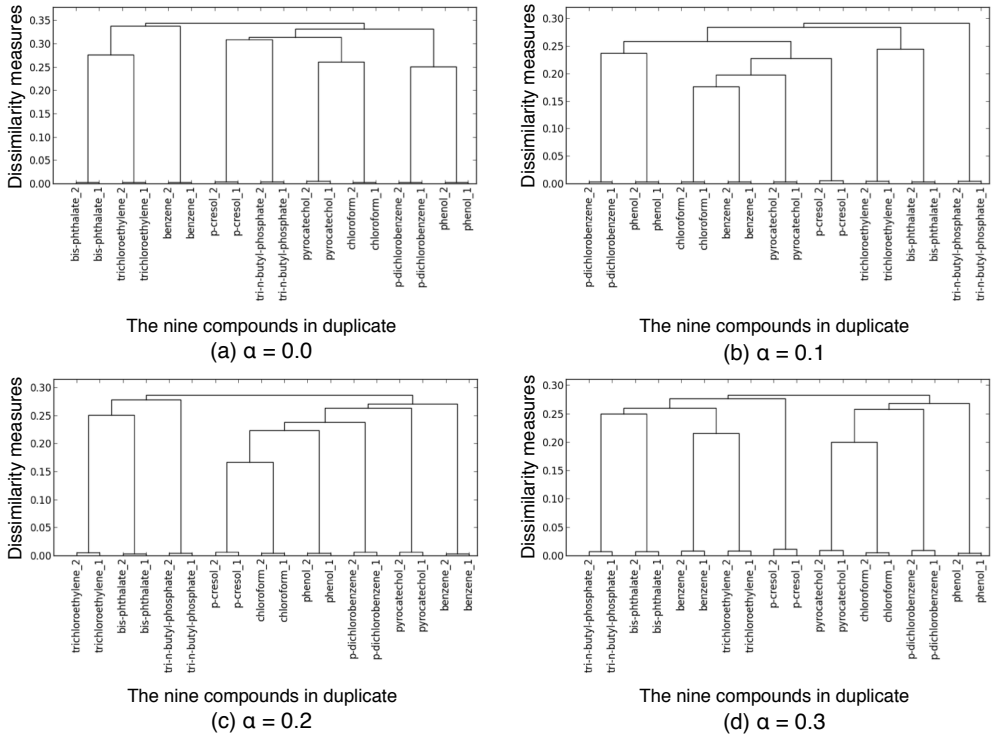

Figure 4: Four dendrograms obtained by applying aggregative hierarchical clustering (the average linkage between the merged groups) to the data of the gene expression ratios for the nine compounds and 1,000 extracted RNAs. The upper-left (a), upper-right (b), lower-left (c) and lower-right (d) panels are the cases of $\alpha = 0.0, 0.1, 0.2$, and 0.3, respectively. In each panel, the y-axis marks the dissimilarity measures at which the clusters merge, and the x-axis the distribution of the nine compounds in duplicate.

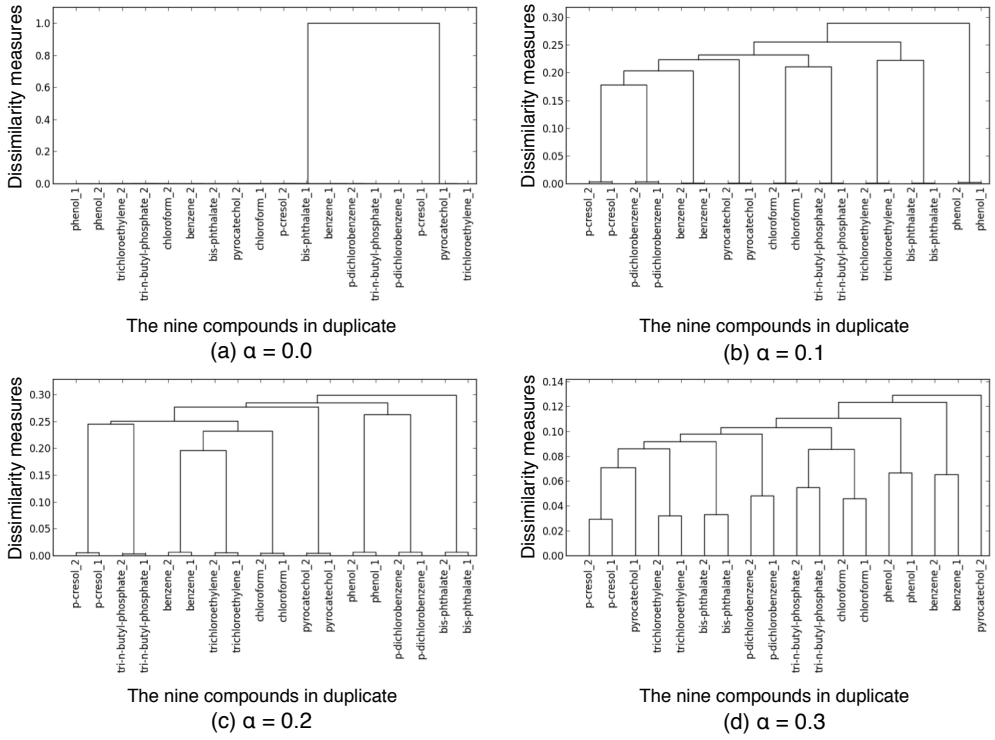

Figure 5: Four dendrograms obtained by applying aggregative hierarchical clustering (the average linkage between the merged groups) to the data of the gene expression ratios for the nine compounds and 100 extracted RNAs. The upper-left (a), upper-right (b), lower-left (c) and lower-right (d) panels are the cases of $\alpha = 0.0, 0.1, 0.2$, and 0.3, respectively. In each panel, the y-axis marks the dissimilarity measures at which the clusters merge, and the x-axis the distribution of the nine compounds in duplicate.

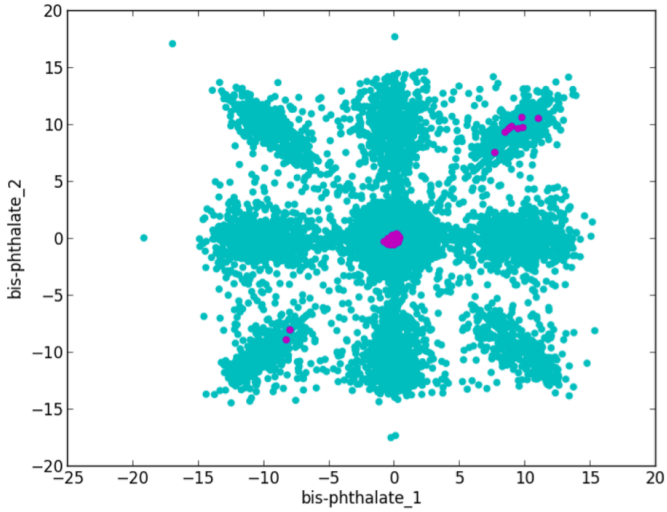

Figure 6: Scatter plot of all the RNAs (32, 586 RNAs) between the sample 1 and 2 in case of bis-phthalate. The red plots indicate the 1, 000 extracted RNAs in case of $\alpha = 0.0$, and the blue plots are the rest RNAs. The sizes of the extracted RNAs are almost zeros.

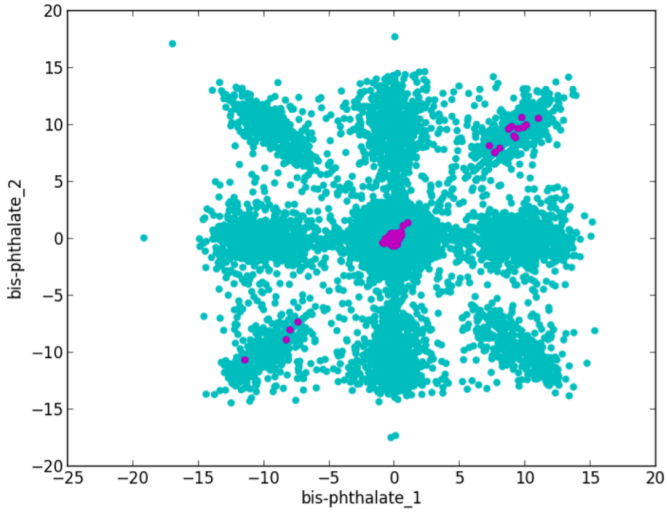

Figure 7: Scatter plot of all the RNAs (32, 586 RNAs) between the sample 1 and 2 in case of bis-phthalate. The red plots indicate the 1, 000 extracted RNAs in case of $\alpha = 0.2$, and the blue plots are the rest RNAs. The number of RNAs whose sizes are not zeros increase, compared to the case of $\alpha = 0.0$.