Grouping levels of exposure with same observable effects before class prediction in toxicogenomics.

Vincent Guillemot*†, Cathy Philippe*, Arthur Tenenhaus*, Jérôme Rollin*, and Vincent Frouin*
*CEA, Laboratoire d'Exploration Fonctionnelle des Génomes
2 rue Gaston Crémieux - 91000 Évry.
†Department of Signal and Electronic Systems, Supélec
Plateau de Moulon, 3 rue Joliot-Curie, Gif sur Yvette, 91192, France

***Abstract*—Gene expression profiling in toxicogenomics is often used to find molecular signature of toxicants. The range of doses chosen in toxicogenomics studies does not always represent all the possible effects on gene expression: several doses of toxicant can lead to the same observable effect on the transcriptome. This makes the problem of dose exposure prediction difficult to address. We propose a strategy allowing to gather the doses with similar effects prior to the computing of a molecular signature. The different gathering of doses are compared with criteria based on likelihood or Monte Carlo Cross Validation. The molecular signature is then determined via a voting algorithm. Experimental results point out that the obtained classifier has better prediction performances than the classifier computed according to the original labeling.**

***Index Terms*—Classification likelihood, Classification, Monte Carlo Cross Validation, Molecular signature, Microarrays**

I. Introduction

Microarray gene-expression profiling is recognized to bring valuable information as regards diagnosis or prognosis (e.g. oncology, new drugs testing, etc.). Many works now aim at applying this high-throughput tool to toxicological studies [5], [10], for which the ultimate purposes are to know whether an individual has been intoxicated and if so, to identify the toxicant and possibly to predict the exposure level. Because the clinical signs are the same for a wide range of toxicants, a molecular imprint yielded by gene expression, the so called *molecular signature*, of these toxicants would help the design of a fast and efficient diagnostic tool. A typical toxicogenomic study consists in administering a toxicant to a model organism at different doses within a range, and getting the corresponding gene expression data.

Some Machine Learning algorithms are dedicated to finding the molecular signature from gene expression. They have to be used cautiously to provide reliable results. The genes of the signature are determined in a cross-validation framework in order to limit the risk of over learning bias, as recommended in recent works [2], [12], [9].

In this work, we consider as mandatory to explore the possibility that different doses may have the same effect on the expression levels of genes. To perform this task, we propose two methods to apply prior to the determination of the molecular signature: one is a likelihood-based method and the other, introduced here, is based on the Monte Carlo Cross Validation (MCCV) algorithm. Then, once the doses of toxicant with similar effect are gathered, a machine learning algorithm is run, aiming at determining a molecular signature and its predictive power.

Our approach is applied to two toxicogenomic studies. The results obtained demonstrate the possibility to gather similar doses, and the interest of this grouping in order to estimate a corresponding classifier with better prediction performances than the one related to the original range of doses.

II. Methods

We present here the notations adopted throughout this paper.

- n is the number of observations
- g is the number of variables (genes), $g \gg n$
- N is the number of doses
- X is the $n \times g$ real matrix containing all the expression profiles
- x is a $g \times 1$ observation (or individual)
- $P_{orig} = (1, ..., N)$ is the vector of the N administered doses labels and is called the original partition.
- y is the $n \times 1$ vector of the administered dose of the toxicant for each observation. Each element of y takes its value in P_{orig}
- $P = (P^{(1)}, ..., P^{(N)})$ is a new labeling of the doses, consisting in a permutation with repetitions of the N elements of P_{orig} among its first $K < N$ elements: $(P^{(i)} \in \{1, ..., K\}$ with $i = 1, ..., N)$. $P^{(i)}$ is the new class label replacing the class label i in P_{orig}. Concretely, $P^{(1)}$ is the class label of the lowest dose

and $P^{(N)}$ the class label of the highest dose. P is called a new partition
- z is the new vector of the observable doses, deduced from a new partition P
- g_{filt} is the number of genes kept after each filtering step (see figure 1)
- g_s is the number of genes kept in the molecular signature

To clarify the notion of partition, let us consider an example of a five dose exposure experiment. The initial partition is $P_{orig} = (1,2,3,4,5)$ and $y = (1,1,2,2,3,3,4,4,5,5)$ is the vector describing the class of each observation (two observations per dose). If a partition P consists in aggregating the weakest doses $P^{(1)}_{orig}$ and $P^{(2)}_{orig}$, it is noted $P = (1,1,2,3,4)$; the new vector of classes is $z = (1,1,1,1,2,2,3,3,4,4)$. It is worth noticing that P could be indifferently noted $P = (1,1,2,4,3)$, or $P = (2,2,1,3,4)$. Finally, the partition $P_1 = (1,1,2,3,4)$ aggregates the weakest doses $P^{(1)}_{orig}$ and $P^{(2)}_{orig}$, whereas the partition $P_2 = (1,2,3,4,1)$ the weakest dose $P^{(1)}_{orig}$ with the strongest dose $P^{(5)}_{orig}$.

We present thereafter two different methods used to estimate a partition $\widehat{P}$ which describes the observable effects of the toxicant in the dataset X (see paragraphs II-A and II-B1). Then, given $\widehat{P}$, we classically determine a molecular signature and the test error rate (see paragraph II-B2).

A. Determination of the best partition

Let observation x be drawn from a multivariate mixture density:

$$f(x,\Theta) = \sum_{k=1}^{K} p_k f_k(x,\theta_k) \tag{1}$$

$\Theta = (p_k;\theta_k)_{k=1,...,K}$, p_k is the probability for an observation to be in the class k and θ_k is the parameter vector of f_k. The choice of a mixture model allows to derive a Classification Log-Likelihood (CLL), as already proposed by [4]. Let P be the current partition. This leads to a partition of the sample $X = [x_1;...;x_n]$ into K classes $C_k, k = 1,...,K$.

$$\mathcal{L}_c(X,P,\Theta) = \sum_{k=1}^{K} \sum_{x \in C_k} \log\left(p_k f_k(x|\theta_k)\right) \tag{2}$$

Equation (2) is commonly used within the Classification Expectation-Maximization algorithm (CEM) [6]. We propose to use a Bayesian Information Criterion (BIC) based on the CLL (equation (3)) which will characterize the quality of P while taking into account the complexity of the corresponding mixture model.

$$BIC = -2\mathcal{L}_c + \nu \log(n) \tag{3}$$

where $\nu = gK + 1$ is a parameter depending on the complexity of the model, assuming that the observations follow a Gaussian mixture model.

The results of the BIC approach allow the selection of a presumably optimal partition. Yet, two characteristics of this criterion are debatable when considering the final objective of molecular signature finding:
- the BIC value does not have any signification, especially if one wants to characterize the test error rate associated to the partition P
- it depends strongly on the Gaussian mixture assumption

In the next section, a prediction model and its cross-validated error rate are computed for each partition. We investigate whether the partitions proposed by the BIC approach provide the smallest test error rates.

B. Classification

Some papers dealing with discrimination from microarray data have been severely criticized in recent works [2], [9], [12]. For instance, Michiels *et al.* [12] emphasizes the fact that numerous papers use methodology resulting in an overoptimistic estimation of the error rate. The approach proposed in this paper was designed to meet the quality requirements suggested in [12] and advocates the use of validation by repeated random sampling, leading to an accurate methodology (cf. figure 1) to get both a molecular signature and the discrimination model associated to a test error rate.

1) Monte Carlo Cross-Validation (MCCV) testing of each partition : To obtain a robust estimation of the test error rate, the learning phase is embedded in a cross-validation framework, presented by figure 1.

For each partition P, the test error rate evaluation is a 3-steps MCCV algorithm consisting in repeating B times:
(a) a split step: split randomly the dataset in a training and a testing set, respecting a 2:1 ratio.
(b) a filtering step: select the g_{filt} relevant variables from the learning dataset via a K-sample F-test with a Bayesian regularization [13].
(c) a classification step: compute a prediction function. To deal with the $g \gg n$ setting, linear Support Vector Machines (SVM) [7] were used to build the classifier. When a multiclass situation is encountered, a *One versus One* strategy is applied [1]. The regularization parameter of the classifier is determined on the learning set by a Leave One Out Cross Validation technique (not shown on figure I).

g_{filt} is set to 200 for all P and MCCV iterations. Additional runs (not depicted in this paper) showed that g_{filt} has no influence on the ranking of P, it was chosen small enough with regard to g to significantly reduce the number of variables. B is set to 50, allowing the estimation of the mean test error rate. Each step of this MCCV

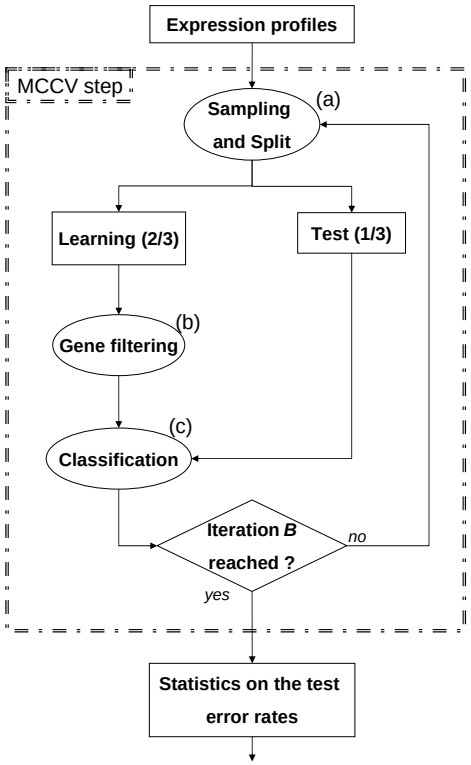


Fig. 1. Algorithm of the MCCV estimation of the classification error rate, P being fixed. This algorithm is run for each P.

algorithm respects the recommendations suggested in [3].

To achieve the grouping of doses with same observable effect, we choose the partition $\widehat{P}$ as a trade-off between minimizing the BIC and the test error rate. Knowing $\widehat{P}$, we can estimate a classifier to discriminate between doses with observable effect.

2) Signature and classifier : In this section, $P = \widehat{P}$ is known, and we compute a molecular signature of g_s genes. Between two MCCV iterations, the g_{filt} genes selected are not likely to be the same: it strongly depends on the split step 1. Thus, the B lists of variables provided by the MCCV are very heterogeneous and to obtain a consensus list, a voting method is required.

We consider two voting techniques:

- "Unanimity": the g_s genes which are selected unanimously by all the B iterations of the MCCV procedure
- "Quorum": genes are sorted according to the number of occurrences in the B iterations. The first g_s genes of this sorting are selected.

Once the signature is determined, we finally build the classifier from the whole dataset and test it on an unseen set of observations, leading to an estimation of the generalization error associated to the g_s signature genes.

TABLE I
DESCRIPTION OF THE DATASETS

	Toxicant	
	Ricin (Tox_1)	Mustard Gas (Tox_2)
Organism	*Mus musculus*	*Rattus norvegicus*
Biological tissue	total blood	lung
# of doses	5	4
# of samples per doses	10, 7, 7, 9, 7	20, 10, 10, 11
# of variables	24111	15923
# of partitions	43	14

III. RESULTS

A. Description of the datasets

We applied our approach to two toxicogenomic datasets Tox_1 and Tox_2, described in table I. Tox_1 corresponds to an in-house experiment (not yet published data), and Tox_2 includes data described in [8] available on the GEO repository (GSE1888). The animals have been sacrificed and messenger RNA (mRNA) has been extracted from the appropriate biological tissue according to usual protocols. The obtained samples have been hybridized on microarrays. For both cases the control sample consists in mRNA from animals injected with the vehicle of the toxicant.

Following classical experimental plans, the two selected experiments are designed as follows: among the doses of toxicants injected to the animals, at least one has known effects, for instance, the dose for which 50% of the exposed animals die (Lethal Dose 50, LD50). Exposure to all these doses can have very different effects on the tissue under study, with possibly no visible phenotype. As proposed earlier, the issue is then to cluster the doses which have the same effects on gene expression.

Figure 2 and 3 depict the projection of the n observations onto the 2-dimensional space spanned by the two first components of the Partial Least Squares regression [14] of y on X. Symbols represent the class membership.

For Tox_1, we point out a clear linear discrimination between low and high doses. Samples corresponding to doses 2 and 3 seems to be quite similar in a transcriptomic point of view and can be considered as belonging to the same class (say the low level exposure class). Samples associated with doses 1, 4 and 5 constitute the null, medium and high level exposure classes respectively. The corresponding partition is $P = (1, 2, 2, 3, 4)$.

For Tox_2, we conjecture that samples corresponding to doses 2 and 3 are quite similar and can be considered as belonging to the same class (say the medium level

exposure class) whereas samples associated with doses 1 and 4 constitute two other distinct classes, say the null level exposure class and the high level exposure class. The corresponding partition is $P = (1, 2, 2, 3)$.

Those remarks illustrate the need to formalize a way to characterize the "optimal" class structure by clustering the doses with the same observable effect.

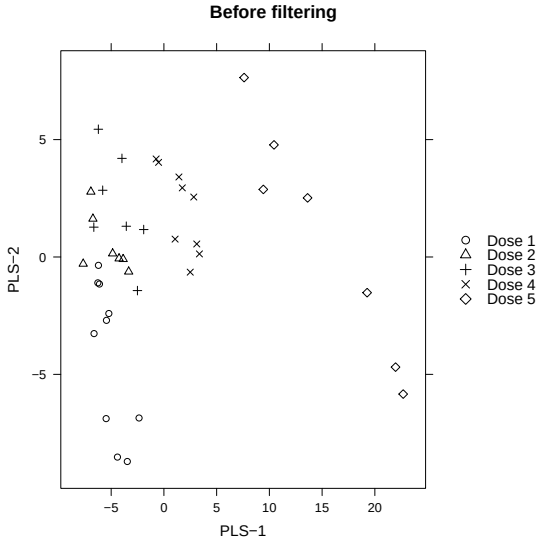


Fig. 2. Projection of the observations on the two first components of a Partial Least Squares Regression of y on X for Tox_1.

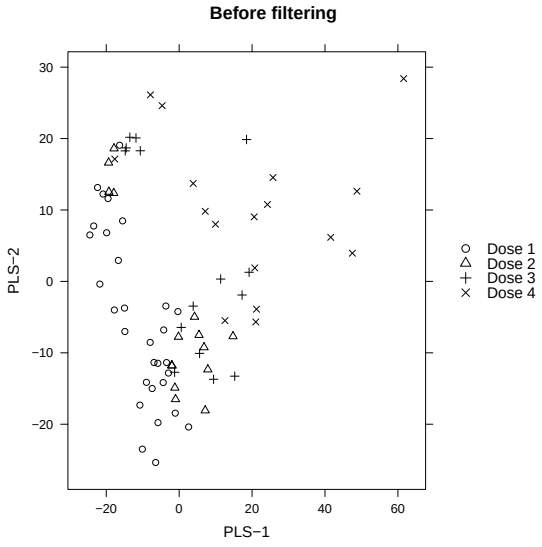


Fig. 3. Projection of the observations on the two first components of a Partial Least Squares Regression of y on X for Tox_2.

B. Choice of $\widehat{P}$

For each P, the BIC is estimated. Figures 4(a) and 5(a) depict BIC as a function of P for each dataset. BIC declared as optimal respectively the partitions $(1, 1, 1, 1, 2)$ and $(1, 1, 2, 2)$. The partitions suggested by figures 2 and 3 are ranked respectively 26$^{\text{th}}$ out of 43 and 9$^{\text{th}}$ out of 14 by BIC. We then compared the BIC and classification results in figures 4(b) and 5(b). As expected, the test error rates associated with the partitions top-ranked by BIC are the smallest. The partitions corresponding to the lowest test error rates are $(1, 1, 1, 2, 3)$ for Tox_1 and $(1, 1, 1, 2)$ for Tox_2. Moreover these partitions are biologically interesting: from a range of 5 doses, we are able to deduce a range of 3 observable effects as regards gene expression for Tox_1 and from a range of 4 doses, to deduce a range of 2 observable effects for Tox_2. In the Tox_1 case, this new partition is all the more interesting because it keeps apart samples associated with the LD50 from the non lethal doses.

TABLE II
GENERALIZATION ERROR RATES, AS A FUNCTION OF g_{filt} AND g_s AND THE VOTING STRATEGY FOR Tox_1.

(a) Unanimity

	Number of filtered genes g_{filt}				
	25	50	100	150	200
Error rate	31%	23%	31%	31%	31%
g_s	12	17	39	66	80

(b) Quorum

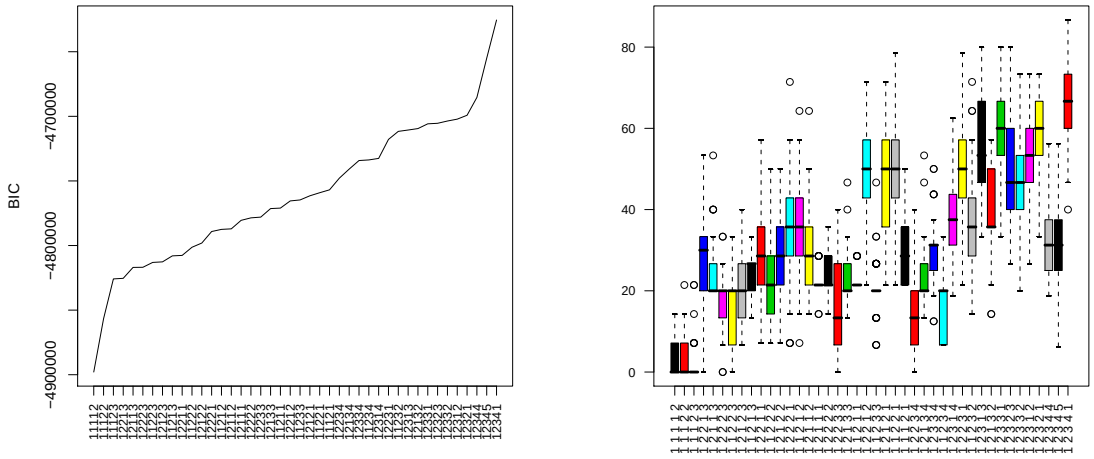
	Number of filtered genes g_{filt}				
g_s	25	50	100	150	200
5	54%	31%	23%	31%	31%
10	38%	23%	31%	38%	31%
15	15%	23%	31%	31%	31%
20	46%	62%	31%	31%	31%
25	23%	23%	23%	31%	31%
30	23%	23%	38%	38%	31%
40	23%	31%	23%	23%	38%

C. Classification knowing $\widehat{P}$

Finally, we give the generalization error rates, as a function of g_{filt} and g_s, for $\widehat{P}$ in tables II(b) and III, obtained from 13 unseen observations for Tox_1 (resp. 50 for Tox_2). We set g_{filt} and g_s to sufficiently small values to significantly reduce the number of variables after the filtering step and to allow the biological validation of the signature. For both datasets, the original partitions $(1, 2, 3, 4, 5)$ and $(1, 2, 3, 4)$ show poor performances either according to their BIC ranking, to their ranking with the MCCV procedure (see figures III-B III-B) and to the generalization error rates associated to their respective signatures (minimum of gnagna and 38% respectively, data not shown).

It is worth pointing out that the number of genes kept by the filtering step has a non negligible effect on the determination of a consensus list of genes, although it has no effect on the ranking of the partitions.

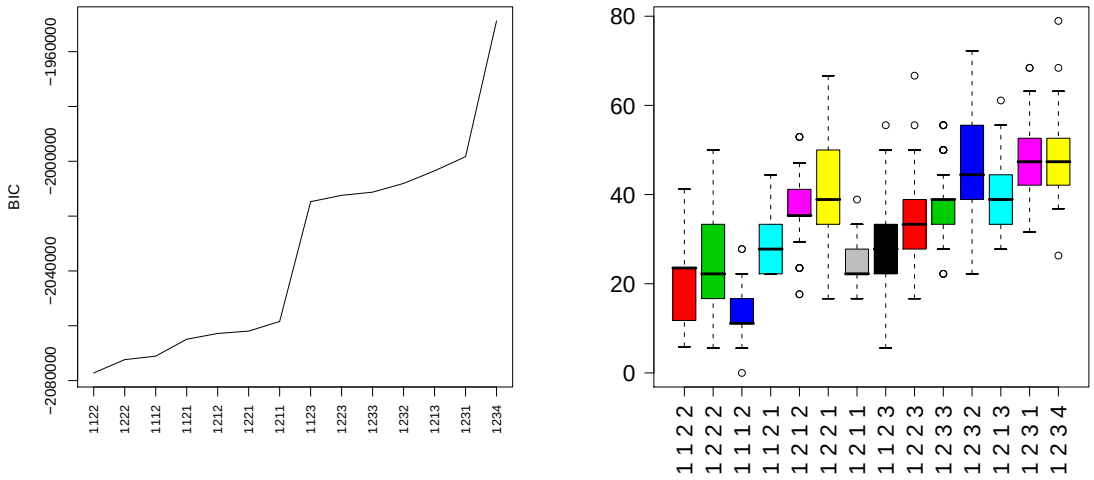
There is an optimal value (15%) when $g_s = 15$ for Tox_1. These genes are present in each signature of length greater than 15. The lowest generalization error rate for Tox_2 is 8% and is associated to a $g_s = 40$ signature.



(a) BIC as a function of P for Tox_1.

(b) Boxplots of the B test error rates as a function of the partition P for Tox_1.

Fig. 4. Bayesian Information Criterion and test error rates for each partition for Tox_1. On the x axis, partitions are ranked by increasing values of BIC.



(a) BIC as a function of P for Tox_2.

(b) Boxplots of the B test error rates as a function of the partition P, for Tox_2

Fig. 5. Bayesian Information Criterion and test error rates for each partition for Tox_2. On the x axis, partitions are ranked by increasing values of BIC.

Our method provided both a classifier able to predict the dose exposure of a new observation and the best subset of genes in terms of prediction.

IV. Conclusion

In the framework of toxicogenomics, studies aim at determining the molecular signature of a given toxicant from a tissue sample. We propose a two-fold methodology to be applied to usual dose-range gene expression experiments, consisting in: first, the discovery of sets of doses with the same observable expression effect and second, the determination of the molecular signature using a MCCV approach. The results presented on two datasets show the impact of the preliminary step on the generalization error. The results presented on two datasets show that gathering similar doses yields a classifier with better prediction performances than the one related to the original range of doses.

Future work will focus on alternative methods to filter

TABLE III
GENERALIZATION ERROR RATES, AS A FUNCTION OF g_{filt} AND g_s AND THE VOTING STRATEGY FOR Tox_2.

(a) Unanimity

	Number of filtered genes g_{filt}				
	25	50	100	150	200
Error rate			19%	15%	19%
g_s	0	0	1	4	2

(b) Quorum

	Number of filtered genes g_{filt}				
g_s	25	50	100	150	200
5	23%	23%	15%	12%	23%
10	19%	19%	23%	23%	27%
15	23%	23%	19%	23%	15%
20	35%	23%	27%	23%	12%
25	15%	27%	31%	23%	15%
30	15%	19%	15%	23%	19%
40	23%	15%	8%	8%	19%

variables [11] and on the automatic selection of the best partitions. A special care will be given to the use of contingency tables rather than test and generalization error rate in order to better account for multiclass discrimination.

V. ACKNOWLEDGMENT

We thank Laurent Le Brusquet and Etienne Thévenot for their constructive remarks.

REFERENCES

[1] E. L. Allwein, R. E. Schapire, and Y. Singer, *Reducing multiclass to binary: A unifying approach for margin classifiers*, Journal of Machine Learning Research **1** (2000), 113–141.

[2] C. Ambroise and G. J. McLachlan, *Selection bias in gene extraction on the basis of microarray gene-expression data*, Proc Natl Acad Sci U S A **99** (2002), no. 10, 6562–6566.

[3] A.-L. Boulesteix, C. Strobl, T. Augustin, and M. Daumer, *Evaluating microarray-based classifiers: an overview*, Technical Report, Department of Statistics, University of Munich **5** (2007).

[4] P. G. Bryant, *Large-sample results for optimization-based clustering methods*, Journal of Classification **8** (1991), no. 1, 31–44.

[5] P. R. Bushel, A. N. Heinloth, J. Li, L. Huang, J. W. Chou, G. A. Boorman, D. E. Malarkey, C. D. Houle, S. M. Ward, R. E. Wilson, R. D. Fannin, M. W. Russo, P. B. Watkins, R. W. Tennant, and R. S. Paules, *Blood gene expression signatures predict exposure levels*, Proc Natl Acad Sci U S A **104** (2007), no. 46, 18211–18216.

[6] G. Celeux and G. Govaert, *A classification em algorithm for clustering and two stochastic versions*, Computational statistics and data analysis **14** (1992), 315–332.

[7] C. Cortes and V. Vapnik, *Support-vector networks*, Machine Learning **20** (1995), 273–297.

[8] J. F. Dillman, C. S. Phillips, L. M. Dorsch, M. D. Croxton, A. I. Hege, A. J. Sylvester, T. S. Moran, and A. M. Sciuto, *Genomic analysis of rodent pulmonary tissue following bis-(2-chloroethyl) sulfide exposure*, Chem Res Toxicol **18** (2005), no. 1, 28–34.

[9] A. Dupuy and R. M Simon, *Critical review of published microarray studies for cancer outcome and guidelines on statistical analysis and reporting*, J Natl Cancer Inst **99** (2007), no. 2, 147–157.

[10] R. D. Fannin, J. T. Auman, M. E. Bruno, S. O. Sieber, S. M. Ward, C. J. Tucker, B. A. Merrick, and R. S. Paules, *Differential gene expression profiling in whole blood during acute systemic inflammation in lipopolysaccharide-treated rats*, Physiol Genomics **21** (2005), no. 1, 92–104.

[11] B. Krishnapuram, L. Carin, and A. Hartemink, *Gene expression analysis: Joint feature selection and classifier design*, ch. 14, pp. 299–318, MIT press, 2004.

[12] S. Michiels, S. Koscielny, and C. Hill, *Prediction of cancer outcome with microarrays: a multiple random validation strategy*, Lancet **365** (2005), no. 9458, 488–492.

[13] G. K. Smyth, *Linear models and empirical bayes methods for assessing differential expression in microarray experiments*, Stat Appl Genet Mol Biol **3** (2004), Article3.

[14] S. Wold, L. Martens, and H. Wold, *The multivariate calibration problem in chemistry solved by the PLS method*, Proceedings Conf. Matrix Pencils, Ruhe A. & Kåstrøm B, Lecture Notes in Mathematics, Springer Verlag, 1983, pp. 286–293.