Using an attribute estimation technique for the analysis of microarray data

J. Mary[1], G. Mercier[2], J-P. Comet[3], A. Cornuéjols[1], Ch. Froidevaux[1] & M. Dutreix[2]

[1] LRI - CNRS UMR 8623 – Bât 490 – Université Paris Sud – F- 91 405 Orsay
{mary, antoine, chris}@lri.fr
[2] Institut Curie – CNRS UMR 2027 – Bât 110 – Centre Universitaire – F- 91405 Orsay
{geraldine.mercier, marie.dutreix}@curie.u-psud.fr
[3] LaMI – Université d'Evry – Tour Evry 2 – 523 Place des terrasses de l'agora – F - 91000 Evry
comet@lami.univ-evry.fr

Abstract

We present an original method for the analysis of microarray data based on an attribute estimation technique. We show, in the context of radioactive environmental stress, that this method can be used to detect a small number of genes that, highly probably, are involved in the response to this stress in yeast. We also studied the different processes in which these informative genes are involved. Moreover, we ranked the genes according to their ability to distinguish between two experimental conditions (irradiated versus non irradiated). Using this method, we highlighted the transcriptional response of yeast cells to low doses of radiation.

Résumé

Nous présentons une méthode originale pour l'analyse de données de puces à ADN, basée sur une technique d'estimation d'attributs. Nous montrons que dans le contexte d'un stress dû à la radioactivité, cette méthode peut être utilisée pour identifier un petit nombre de gènes de levure, qui sont très probablement impliqués dans la réponse à ce stress. Nous avons aussi étudié les différents processus dans lesquels interviennent ces gènes informatifs. Par ailleurs, nous avons ordonné les gènes selon leur capacité à distinguer les deux conditions expérimentales, irradiée et non irradiée. Ainsi, grâce à cette méthode, nous avons pu mettre en évidence la réponse transcriptionnelle des cellules de levure à de faibles doses de radiation.

1 Introduction

Yeast microarrays provide information about the transcription of all genes in a cell population. Genome-wide monitoring of transcript changes in yeast could provide information about previously unrecognized cellular responses to environmental stress and reveal specific genes that allow yeasts to survive in such conditions. As microarrays generate massive amounts of data, data analysis methods are required to determine whether changes in gene expression are indeed significant.

In our approach, we first looked for a method to analyze the transcriptome of these yeasts with the aim of identifying genes involved in the response to environmental changes. We then tried to identify the functional groups to which these newly highlighted genes belong.

In the first stage, we were particularly interested in answering several questions:

(1) the number of genes involved in the transcriptional response;
(2) the identity of these genes (to determine whether genes induced by low-level irradiation are also induced by acute irradiation);
(3) the possibility to use transcriptional data to predict whether a given sample belongs to one of the considered (environmental) classes (i.e. had the yeast been exposed to radiation or not).

This standard problem is named "supervised classification" [1]: some examples from both classes (treated yeasts and non-treated yeasts) are available for training purposes and the aim is to distinguish them by means of the expression levels of a subset of pertinent genes to be determined. This task is however difficult for several reasons.

(1) Available data are noisy due to:
   - imprecise measurements: classical noise assumed to be Gaussian
   - the fact that some data are aberrant due to spotting or hybridization problems.
(2) The number of attributes is huge: expression level of 6135 genes.
(3) The number of training examples is usually very low: 18 in our study.
(4) Classes are not well balanced (they do not contain the same number of slides): 12 non-treated cultures and 6 treated cultures in our study.
(5) The expression of genes is correlated (common regulatory pathways), which goes against the assumption of probabilistic conditional independence.

In an ideal learning process, the measurement of gene expression levels alone should make it possible to identify all the genes involved in a given environmental response. Unfortunately, with the learning methods available it is difficult to identify such sets of genes given the large number of attributes (the expression levels of genes) and the low number of training examples (the available microarray slides). Thus, we had to choose between two alternatives:

- The detection of almost all the genes possibly involved in the transcriptional response (small number of false negatives) even if there is a risk that genes not involved in the response will be considered to be pertinent (many false positives).
- The detection of only genes that are almost certainly involved (small number of true positives), with the possibility that a lot of involved genes will be overlooked (numerous false negatives).

As small subsets of genes make it easier to identify involved biological functions and as we wanted to limit the number of false positives, we chose the second approach. Using the RELIEF attribute estimation method (see section 3), we showed that it is possible to detect a small number of yeast genes that are almost certainly involved in the response to radioactive stress.

2. Biological design and data collection

Our environment is altered by human and industrial activities that release different kinds of chemicals. These products, which are potentially highly toxic for public health, are difficult to detect because of their very low concentrations and because they are mixed with other compounds. We started this study from the assumption that exposure to genotoxic agents alters the transcription of some genes and that these changes can be estimated by comparing the messenger RNA (mRNA) levels of two populations growing in different environments. We used DNA microarrays to measure all the mRNAs of the *Saccharomyces cerevisiae* yeast in different growth conditions. The aim of this study was to identify the specific response induced by a given chemical. We limited our study to the quantification of the biological effects of low doses of ionizing radiation, particularly when they are delivered at low dose rates, as is the case in most environmental conditions.

To analyze the effect of continuous exposure to low doses of radiation, we quantified the expression level of most of the yeast genes in cells grown with (I) and without (NI) low doses of irradiation. For this purpose, cells were exponentially grown in rich medium for 20 hours (corresponding to 12 division cycles) in the presence or absence of β radiation (1.71 Mev). The number and the distribution of cells at each stage of the cell cycle were then determined by microscopy: the frequencies of single cells (G1), budding cells (S) and doublets (G2/M) in the NI and I populations should be similar if growth is not affected by the irradiation. Doses were considered to be "low" when no differences in cell growth or mutation (or recombination) frequencies were observed between NI and I populations [7]. In the first set of experiments, the dose rate was between 10 and 30 mGy/h. Although we observed no physiological (growth) or genetic (recombination and mutagenesis) effects in the irradiated populations, the expression levels of numerous genes were significantly changed. The relative expression level of each gene was estimated by use of glass slide microarrays (produced by Corning) spotted with 6135 denatured DNA sequences corresponding to all of the open reading frames (ORFs) of *S. cerevisiae.* We labeled the control cDNAs with Cy3-dye and the NI or I cDNAs of interest with Cy5-dye. The same control cDNA was used for all the experiments. It was prepared from a pool of independent cultures grown without treatment.

Hybridized microarrays were scanned using a Genepix 4000B machine (Axon Instrument). Separate images were acquired for each type of fluorescence, at a resolution of 10 µm per pixel. Images were analyzed with Genepix pro 3.0 (Axon), after manually rectifying the outline of each spot. The median values for both types of fluorescence were used for each spot. A quality control standard (QCS) was estimated by calculating the difference between the median pixel values within the spot (F) and in the background (B) and correcting with the sum of the standard deviation (SD).

$$\frac{F\ median - B\ median}{\sqrt{\left(\frac{F(SD)\times\sqrt{\pi}}{\sqrt{2\times F\ pixels}}\right)^2+\left(\frac{B(SD)\times\sqrt{\pi}}{\sqrt{2\times B\ pixels}}\right)^2}} > Quantile\ normal\left(\frac{0.5}{2\times number\ of\ spots}\right)$$

Data with "non significant" QCS values for both types of fluorescence were considered to be missing.

The data generated by cDNA microarrays are affected by many experimental parameters other than differential expression and this systematically leads to variations in the measured intensities. These variations cannot easily be quantified by standard methods. Measured intensities must be normalized before we can compare measurements from different microarrays experiments. The main assumption underlying normalization is that there is a functional coherence between a true biological difference and the corresponding measured values. Looking for crude ratios of signal intensity in intensity-dependent dye normalization methods seems preferable to relying on global methods such as mean or median normalization. We used the location and scale normalization procedures originally developed by Yang [15]. These methods correct for intensity and spatial dye biases, by use of a robust local regression. They use the Splus LOWESS function (Insightful) to perform robust local regression and were applied to obtain a scaled within print group normalization to account for spatial dependence in dye biases, with scale adjustment between the blocks. To make this method more robust, we did not consider saturating points (with saturating fluorescence intensities) when estimating LOWESS fits. However, as these points contain relevant information, all measured intensities were normalized with the estimated LOWESS curves. The normalized data for each spot were defined as the estimated relative expression levels for experiments with irradiated populations (I values) and non irradiated populations (NI values).

3. Attribute estimation with the RELIEF algorithm

Given the characteristics of the available data (large number of attributes versus few examples), it is not worth trying to detect complex correlations, as most of the numerous correlations found will be spurious. Instead, we used approaches that examine the possibility of a direct correlation between each gene and class (I or NI). This is an attribute estimation problem that can be addressed by numerous techniques (see for instance [5, 8, 10, 12, 14]). For reasons that will become clear shortly, we used an adaptation of the RELIEF method [5, 6], which is an attribute estimation method that looks for the attributes that seem to be the most significantly correlated to the class to be predicted. The overall principle is to calculate a normalized weight for each gene. These weights are comprised between -1 and +1, with positive weights indicating that there is a positive correlation between the relative expression level of a gene in a given example and its class. The weight of a gene is a function of the variation of its relative expression level within each class compared to the variation between classes. Indeed, the correlation between class and gene expression seems to be stronger if the intra-class variation is small compared to the extra-class variation.

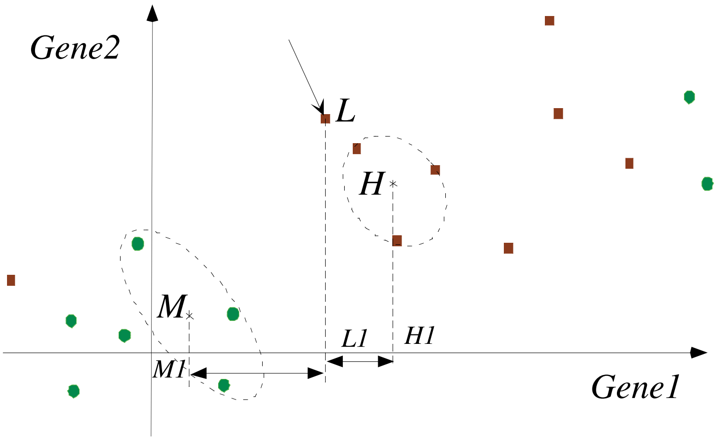

Figure 1 – This figure illustrates the RELIEF principle when only two genes are considered. Examples are in this case represented by points in the plan determined by the relative expression levels of the two genes. Here, there are 17 examples, 9 of which belong to the "square" class, and 8 to the "circle" one. In this case, the example *L* positively contributes to the overall estimated relevance of Gene1.

The weight associated with a gene (an attribute) is calculated as follows. One example *L* can be considered as a point in the attribute space (a 6135-dimensional space). For each of the *m* examples, labeled I or NI, we determine its *k* nearest neighbors in the same class and calculate the corresponding barycenter *H* (for nearest Hit), and then determine its *k* nearest neighbors in the other class and calculate the corresponding barycenter *M* (for nearest Miss). For each gene *g* in turn, we calculate the projection of the points *L*, *H* and *M* on the associated axis. For instance, the projection *L1* of point *L* on the gene *g* corresponds to the expression level of the gene *g* for the example *L*, whereas the projection of point *H* (respectively *M*) corresponds to the mean expression level of gene *g* for the *k* nearest neighbors of the same class (respectively of the other class). We then calculate, on one hand, the distance between the projection of *L* and the projection of *M*, and, on the other hand, the distance between the projection of *L* and the projection of *H* (see figure 1). The difference distance(*L1, M1*) - distance(*L1, H1)* between these

two values provides the contribution of this example to the estimated relevance of gene *g*. This is repeated for all examples in turn. The sum of all the contributions thus obtained, divided by the number of examples, defines the weight associated with each gene.
This can be summarized by the following formula:

$$\mathrm{w}eight(gene) = \frac{1}{m}\sum_{L=1}^{m}\left\{\left[\mathrm{expr}_{gene}(L) - \mathrm{expr}_{gene}(M)\right] - \left[\mathrm{expr}_{gene}(L) - \mathrm{expr}_{gene}(H)\right]\right\}$$

where $\mathrm{expr}_{gene}(x)$ is the projection of point x on the axis associated with the gene, and m is the total number of examples.
The weight that is calculated in this way for each gene is an approximation of the difference of two probabilities (see [6]) as follows:
Weight(gene) = P(gene has a different value / *k* nearest neighbors in a different class)
- P (gene has a different value / *k* nearest neighbors in the same class).
In our work, we chose to use the Manhattan distance to determine the neighbors in the 6135-dimensional space instead of the more classical Euclidian one, because the latter tends to overvalue the genes (attributes) that exhibit very noisy measures or aberrant values.
As well as its low computational cost (after we had optimized the code), one important advantage of RELIEF is that, unlike many other statistically oriented techniques, it does not assume that genes are independent and does not need information on the data distribution (e.g. Gaussian). Moreover, the parameter *k* makes it possible to control the tradeoff between sensitivity and robustness to noise. In our case, *k*=3 was determined empirically as the best choice. It avoids an unreasonable sensitivity to aberrant values while being more precise than methods that take into account only mean values.

4. Determining informative genes

It is difficult to decide how many genes are required to determine whether a sample has been irradiated or not by merely examining the weights obtained. Indeed, it is difficult to choose a minimal weight (threshold) beyond which a gene can be considered to be involved in the response to radiation. Moreover, the small number of samples, together with the intrinsic noise in the data, makes the task even more complex, meaning that some genes may appear to be strongly correlated by pure chance. To estimate the probability that irradiation does not lead to any variations in relative gene expression (null hypothesis), we compared the correlation between the relative expression level of each gene and the class in the experiment with the same correlation for random permutations of class labels. We labeled samples such that one class contained 12 and the other one 6 experiments, to respect our original distribution, as RELIEF is sensitive to proportions. For each permutation, we then re-applied RELIEF to obtain new weights for each gene. The process was repeated until the curve of average number of genes having a given weight stopped varying (see figure 2). The process was repeated 2000 times and the curve started stabilizing after 500 runs.
If there is indeed a correlation between the relative gene expression levels and the class, more genes should appear to be correlated in the experimental condition than in the null hypothesis condition for a given weight. Figure 2 shows the curves of the number of apparently correlated genes for a given weight corresponding to the two conditions (experimental and randomized). A difference is clearly visible, thus confirming that exposure to low radiation doses affects the expression of some genes.
More precisely, this figure allows us to select a set of genes that are highly likely to be involved in the adaptive response to low doses of radiation. One possibility would be to consider the point where no genes appear to be correlated with the class in the null hypothesis condition (beyond the

0.58 weight). At this threshold, 13 genes were correlated with class in the experiments and the probability was very low that any of them could be attributed to chance. It must be emphasized that other genes are certainly involved. To obtain a larger set of genes for the purpose of biological interpretation, we are willing to accept that 10% of the selected genes are false positives. This ratio is obtained for a threshold of 0.3: if we consider the genes with a weight greater than 0.3, 171 genes are selected in the experimental condition compared to about 17 in the randomized condition (see figure 2). The 171 genes with a weight greater than 0.3 are considered as "informative".

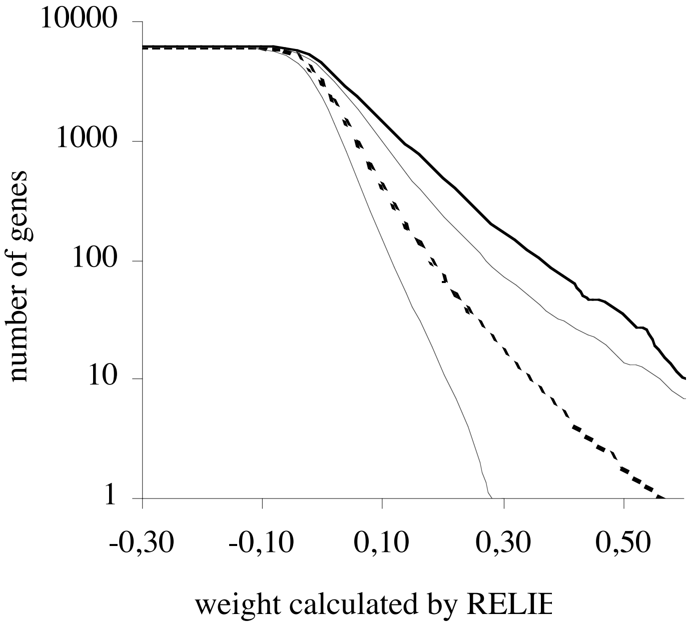

Figure 2 - Experimental (continuous bold) and randomized (dashed) curves showing the number of genes, the weight of which is greater than or equal to the threshold given on the x axis. The thin lines indicate the 95% confidence interval of the random distribution.

5. Interpreting biological results

In addition to its ability to separate the samples depending on whether they were exposed to radiation or not, the method allowed us to identify 171 informative genes responding to low doses of radiation. The analysis of the cellular functions involving the proteins coded by these genes could provide new information about the cellular target of the radiation and the functions induced in the response.

We analyzed the first 171 genes ranked by RELIEF. These genes were grouped according to the cellular process in which they participate. The processes implicated in radiation response were identified by comparing the proportion of genes involved in a given process within the first 171 ones with the proportion of genes involved in the same process within the 6135 measured genes. More precisely, the second column of the table indicates the number of ORFs induced (top part of the table) or repressed (bottom part) in response to radiation that take part in the cellular process indicated in column 1. The third column gives the proportion of the genes responding to each cellular process among the responding genes (91 induced genes in the top part and 80 genes in the bottom part). The fourth column indicates the proportion of the number of genes responding to the cellular process, represented on the microarray. The last column indicates the level of over-representation (ratio) of each cellular process among the 171 responding genes.

Only the processes involving more than four genes are considered. Three processes (highlighted in the table) are clearly induced by exposure to low levels of radiation. They all participate in the oxidative phosphorylation cascade. The genes participating in these processes are 14 to 30 times more frequent in the first 171 genes ranked by RELIEF than in the whole of the 6135 genes on the microarray (Table 1).

Without explaining the biological interpretation of these functional pathways in detail (ongoing work), our confidence in the method used for gene selection is strengthened by the fact that the functions revealed by the selection of the informative genes are known to be involved in the elimination of some cellular products caused by ionizing radiation (e.g. free radicals).

function of 91 induced genes/171	number of genes	% in this list	% total genes (6135)	ratio
unknown	38	41.8	50.4	0.8
oxidative stress response	4	4.4	0.3	14.3
oxidative phosphorylation	9	9.9	0.3	30.5
transport	4	4.4	2.2	2.0
gluconeogenesis	1	1.1	0.1	16.9
protein processing & synthesis	3	3.3	2.0	1.6
ATP synthesis	7	7.7	0.4	20.6
glucose repression	1	1.1	0.2	4.8
respiration	2	2.2	0.1	22.0

function of 80 repressed genes/171	number of genes	% in this list	% total genes (6135)	ratio
unknown	45	56.3	50.4	1.1
stress response (putative)	1	1.3	0.2	7.0
glycerol metabolism	2	2.5	0.1	30.8
protein processing & synthesis	3	3.8	2.0	1.9
secretion	2	2.5	2.0	1.3
transport	4	5.0	2.2	2.3
glycolysis	2	2.5	1.0	2.5

Table 1

6. Comparison with other methods

The use of RELIEF to select up- and down-regulated genes is original. We compared it to the very common ANOVA (ANalysis Of VAriance) method [2, 10], which is a generalization of the multi-class case of the Student's t-test, a statistical method that is at the root of the SAM method (Significance Analysis of Microarrays [12]). Analysis of variance has already been used for gene expression data [4] and led to the proposal that microarray data should be normalized, which in turn made it possible to analyze some aspects of the data (including the search for informative genes). In our study, we used analysis of variance to look for informative genes after applying a LOWESS normalization (see section 2).

ANOVA tests the equality of several means of a variable (here relative gene expressions) according to independent variables (here the class of microarray slides: I or NI). It is assumed that the data samples are chosen randomly and are independent, and that the populations have the same variance (this is a strong hypothesis not always true of microarray data). The principle of the method is to estimate the variance, first taking into account the classes and secondly without taking into account them. If the class is not correlated to the measured variable, these two estimated variances should be similar. We applied ANOVA to each of the 6135 genes, thus giving a number measuring the statistical correlation between relative gene expressions and classes for each gene. To see if the Fisher statistics given by ANOVA agree with the null hypothesis that there is no correlation between relative gene expressions and classes, we used the same approach as for RELIEF: shuffling classes and calculating a large number of new Fisher statistics (see section 3). The informative genes are those for which the Fisher statistic (for the real classes I and NI) is very high as compared to the statistics obtained for the random cases.
We also compared the ANOVA results with those obtained with the SAM software [12], which identifies the most significant genes by ordering them according to a statistic based on the variations in gene expression levels rescaled with the standard deviation. Thus, the selected genes are those for which the rank is far away from the mean rank calculated with the shuffling process.
We compared the three methods by analyzing the intersections of the 500 top-ranked genes. For instance, the ranking obtained by SAM is relatively similar to that obtained by ANOVA, with 82% of the top 500 genes being identical. We then compared the results obtained by the RELIEF and ANOVA methods; 257/500 genes were found to be common to both rankings. The probability of obtaining an overlap of this size between two groups of 500 genes randomly selected among 6135 is very low (about 10^{-160}) according to the hypergeometric law. This probability is in the same range for different measured overlaps: if we take the first q ranked genes obtained by both methods (for q between 100 and 2000), the overlap size is greater than the half of the chosen number of genes ($q/2$). Thus, we can conclude that although ANOVA and RELIEF are based on different principles, they give partially the same information from the same data. Given the absence of independence between gene expression in RELIEF and the fact that it can tolerate noise, the biological interpretation that we gave concerns the genes obtained by RELIEF (see Table 1).

7. Conclusion

The use of microarrays to distinguish between organisms living in various environmental conditions and to reveal the biological processes involved in survival in certain conditions is a great challenge. One of the main reasons for this stems from the fact that very few measurements are usually available compared to the large number of genes. Another reason is that the usual statistical assumptions regarding both the distribution of gene expression levels and the independence of genes do not appear to hold true (groups of genes are known to be correlated). We therefore propose an original attribute estimation method that remedies these drawbacks and is robust to noise. The analysis carried out with RELIEF made it possible to bring to the fore the transcriptional response induced by low doses of radiation. We ranked the genes according to their ability to distinguish between samples exposed to radioactivity or not. This revealed that some of the induced genes are involved in specific processes: three of these processes are clearly induced by exposure to low-level radiation and participate in the same oxidative phosphorylation cascade. This suggests that it could be possible to use dedicated microarray data (based on selected genes) for diagnostic purposes. We also plan to exploit the same approach for other biological studies such as, for instance, the classification of tumors in humans [13].

Acknowledgements

The authors gratefully acknowledge the "Consensus" research group for stimulating discussions, and the Genopole and the DRIDF for supporting group meetings financially. This work was supported by the Institut National de Recherche et de Sécurité (convention n°5011888), the CNRS, the Curie Institute, the Association pour la Recherche sur le Cancer (#5659) and Biogen n°74 (BioIngénierie 2001).

References

[1] Cornuéjols A. and Miclet L. : Apprentissage artificiel. Concepts et algorithmes. Eyrolles, 2002.
[2] Glantz S.A. and Slinker B.K., Primer of Applied Regression & Analysis of Variance, McGraw-Hill/Appleton & Lange, 2nd edition, 2000.
[3] Grant G., Manduchi E. and Stoeckert C., Using non-parametric methods in the context of multiple testing to determine differentially expressed genes, in Methods of Microarray Data Analysis: Papers from CAMDA'00, eds Lin SM. and Johnson KF., Kluwer Academics, pp. 37-55, 2000.
[4] Kerr M.K., Martin M. and Churchill G.A., Analysis of variance for gene expression in microarray data, Journal of Computational Biology, 2000, 7(6), 818-837.
[5] Kira K. and Rendell, L., A practical approach to feature selection. In Proc. International Conf. on Machine Learning, Aberdeen, D. Sleeman and P. Edwards (Eds.), Morgan Kaufmann, 1992, pp. 249-256.
[6] Kononenko I., Estimating Attributes : Analysis and Extensions of RELIEF, Proc. of the European Conference on Machine Learning, ECML-94, 171-182, 1994.
[7] Mercier G., Denis Y., Marc P., Picard L., Dutreix M., Transcriptional induction of repair genes during slowing of replication in irradiated *Saccharomyces cerevisiae*, Mutation Research 487, 2001, 157-172.
[8] Ng A. and Jordan M., Convergence rates of the voting Gibbs classifier, with application to Bayesian feature selection, citeseer.nj.nec.com/445984.html.
[9] Park P., Pagano M., Bonetti M., A non parametric scoring algorithm for identifying informative genes from microarray data, Pacific Symposium on Biology: 52-63,2001.
[10] Pavlidis P. and Noble W., Analysis of strain and regional variation in gene expression in mouse brain, Genome Biology, 2001, 2(10).
[11] Troyanskaya O., Garber M., Brown P., Botstein D. and Altman R., Nonparametric methods for identifying differentially expressed genes in microarray data, Bioinformatics, Vol.18, no.11, pp. 1454-1461, 2002
[12] Tusher V., Tibshirami and Chu GG., Significance analysis of microarrays applied to the ionizing radiation response, PNAS, April, 2001, Vol 98, n°9, 5116-5121.
[13] Van't Veer L., Dai H., van de Vijver M., He Y., Hart A., Mao M., Peterse H., van der Kooy K., Marton M., Witteveen A., Schreiber G., Kerkhoven R., Roberts C., Linsley P., Bernards R. and Friend S., Gene expression profiling predicts clinical outcome of breast cancer, *Nature*, 415, January, 2002.
[14] Xing E., Jordan M. and Karp R., Feature selection for high-dimensional genomic microarray, Proc. of the Int. Conf. on Machine Learning, ICML-2001, 601-608, 2001.
[15] Yang YH, Dudoit S, Luu P, Lin DM, Peng V, Ngai J, Speed TP. Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. Nucleic Acids Res. 2002 Feb 15; 30(4):27-28.