Data Preprocessing and Reducing for Microarray Data Exploration and Analysis

Fadoua Rafii
LIST Laboratory
UAE Tangier, Morocco

M'hamed Aït Kbir
LIST Laboratory
UAE Tangier, Morocco

Badr Dine Rossi
LABIPHABE Laboratory
UAE Tangier, Morocco

ABSTRACT
The advances known by Microarray technology have provided birth to enormous ameliorations and investigations in different domains, such as medicine, the pharmaceutical, biotechnology, agrochemical and food industries. The exploitation of Microarray data is still complex for many researchers, due to its huge quantity generated by different experiments. The produced Microarray data must be treated in order to get more valuable information, compare data by improving its clear visualization, make further analysis and respond to crucial hypotheses. Many researchers have found out the biological significance of Microarray data as the greatest challenge. This task couldn't be achieved without preprocessing and taking into consideration biases caused by the presence of variation sources in the Microarray experiment steps. This article will highlight the importance of implementing the preprocessing and the Data mining techniques on Microarray data. It will demonstrate the usefulness of results obtained after these techniques application, and the efficiency of PCA technique for analyzing Microarray data.

General Terms
Microarray technology

Keywords
Microarray data, Preprocessing techniques, Analysis, PCA technique

1. INTRODUCTION
The appearance of high-throughput technology has produced Microarray which is a promising tool that allows biologists to measure hundreds of thousands of gene expressions concurrently [1]. This advanced technology has given to biomedical researchers the opportunity to control expression for a whole genome, where previously they could only look at a handful of genes at the same time [2]. Despite the fact that DNA Microarrays have been widely used to understand the critical events underlying growth, development, homeostasis, behavior and the beginning of disease, the management of the resulting data has received little attention [3]. For this reason, it has been focused on exploring the worthwhile data obtained by the Microarray technology. To achieve this object, this paper represents a continuation of two accepted works, the first one was at the international conference VSST'2015 [4] and aims to shed light on the efficiency of integrating Microarray data from different sources, and the second one was at the international conference BDCA'15 [5] and its object is to implement the preprocessing techniques for Microarray data and to compare the results obtained from each method in order to select the best one for further analysis. On the present paper, it has been applied preprocessing techniques to filter the significant Microarray data and to get rid of noisy data. After that, the results obtained by the application of three methods (CBM, LV and LR) were compared to choose the most adequate one for implementing PCA technique. Since the goal of most Microarray applications is to detect differences in transcript levels calculated from fluorescence ratios, it is necessary to apply the normalization techniques for fluorescence signals to compensate for systematic variations [3]. Thus, before proceeding with a formal Microarray data, many aspects should be taken into account. These aspects are representing the key for choosing the effective techniques to obtain significant Microarray data.

2. MICROARRAY TECHNOLOGY

2.1 Presentation of the Microarray Technology
In the past several years, a new technology recognized as Microarray appeared in the biological field and recently evolved in molecular diagnostics, attracting interest among biologists and biomedical researchers and significantly influencing the approach to disease discovery and genome understanding [6]. The utilization of miniaturized Microarrays for gene expression profiling was first described in 1995 [7] and the first Microarray with the complete eukaryotic genome of the yeast (Saccharomyces cerevisiae) was produced in 1997 [8].

Microarrays are miniaturized arrays of hundreds to thousands of DNA fragments or synthetic oligonucleotides that have been joined to a solid substrate using automated printing equipment such that each spot in a fixed position on the array corresponds to a unique DNA [9]. Microarray technology has produced large data sets that can provide information on gene expression when cells are subjected to various treatments [10].

2.2 Microarray Experiment Steps
Gene expression Microarray experiments are made for one of two goals: the evaluation of differential gene expression between groups which referred to as class comparison; or for the classification studies, that referred to as class discovery and class prediction [11]. The Microarray assays are composed of five experimental steps including the biological question, the preparation of samples, the biochemical reaction, the detection, visualization and modeling of the data [7].

To measure the gene expressions on a sample, there are four steps:

- Sample preparation and labeling: this step consists of the extraction of the RNA from an identified tissue, and the labeling depends on the chosen technology.
- Hybridization: it is when the DNA probes on the glass and the labeled DNA or RNA target from heteroduplexes via Watson-Crick are base-paired. The detection of hybridization can be accomplished optically [12], electrochemically [13], or using devices sensitive to the mass [14].

- Washing: this step comprises the elimination of hybridization solution excess from the array.
- Image acquisition: it is the final step that consists of producing an image of the hybridized array surface.

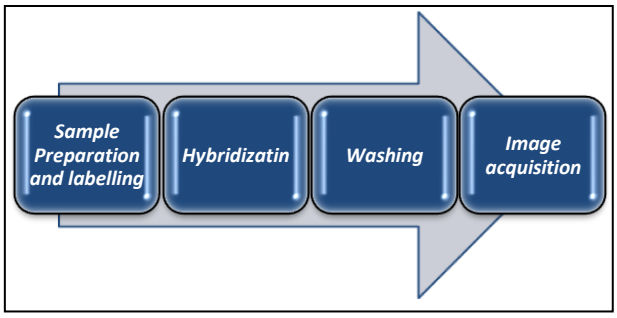

Fig 1: The Microarray experiment steps

3. MICROARRAY DATA

3.1 Microarray Data Resources

The continuous use of Microarray technology over the years has resulted large amount of data. These data are deposited in many different databases. Microarray databases are used to store the obtained data of finished experiments and to make the data available to other users and applications, either directly or by downloading [15]. The existing data represents reliable information that can be extracted, depending on the issues of researchers. On the following table, some of public Microarray databases were mentioned.

Table 1. Some Microarray Databases

Microarray Database	Description
ArrayExpress	ArrayExpress is a public database for Microarray data that is serving as an archive providing access to Microarray data supporting publications and building a knowledge base of gene expression profiles [16].
SMD	The Stanford Microarray Database is a research tool and archive that allows hundreds of researchers worldwide to store, annotate, analyze and share data generated by Microarray technology [17].
caArray	caArray is an open-source, web and programmatically accessible Microarray data management system supporting the annotation of Microarray data using MAGE-TAB and web-based forms [18].
BASE	BASE stores all Microarray experiment related information, biomaterial data, and annotations regardless if analysis tools for specific techniques or data formats are readily available [19].

3.2 Microarray Data Format

The data obtained from Microarray technology is stored in the format of large matrices of genes expression levels. The genes are represented by the rows, and they have been under different experimental conditions or samples represented by the columns.

$$\mathrm{GE} = (x_{ij}) = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1M} \\ x_{21} & x_{22} & \dots & x_{2M} \\ . & . & & . \\ . & . & & . \\ . & . & & . \\ x_{N1} & x_{N2} & \dots & x_{NM} \end{bmatrix}$$

GE represents Microarray matrix of gene expression which is described by *N* x *M*.

- *N* is the genes dimension of the Microarray experiment.
- *M* is the dimension of samples or conditions implicated on the specified Microarray experiment.

x_{ij} is the expression level of the gene i and the sample or the condition j:

$$\begin{cases} i = 1, \dots M \\ j = 1, \dots N \end{cases}$$

The matrix of Microarray data contains the values of gene expression within multiple experimental conditions or samples.

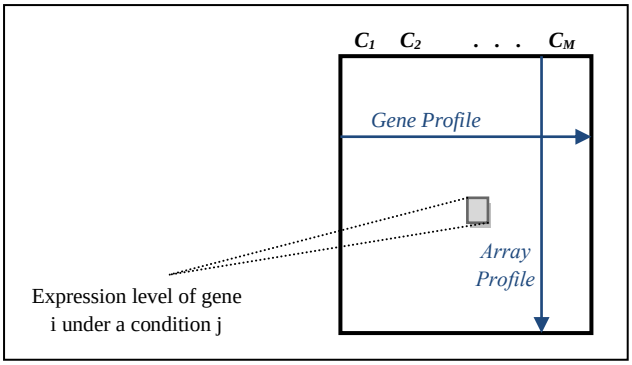

Fig 2: The structure of Microarray matrices of gene expression

In the structure of the Microarray matrix, two types of profiles are distinguishable:

- Gene Profile is the description of expression values for one gene in many samples or conditions.
- Array Profile is the description of expression values for many genes in one sample or condition.

4. PROBLEMATIC

The use of Microarrays by multiple kinds of users including investigators, biologists, laboratories, and across disciplines has necessitated the reporting of data that can be rapidly compared and analyzed. The mass of numbers produced by a single Microarray experiment could be tens of thousands of data points for thousands of genes [3]. This huge amount of the produced data represents the key information for responding to crucial biological questions and hypothesis. In the object of getting the reliable information from Microarray data, it is necessary to appeal the preprocessing techniques to extract the accurate data. The main issues of Microarray data preprocessing is to filter out the existing noise on experimental data, correct the systematic errors and eliminate the data that are not representing significant information for analysis. The existent outliers on Microarray data are the result of the variation sources presence on the experiment process. The sources of systematic errors comprise array

surface chemistry, Microarray printing, labeling methods, hybridization parameters, image analysis and RNA isolation. For interpreting the Microarray data, the data need to be analyzed by implementing Data mining techniques in order to distinguish different kind of data that will be involved in further analysis.

5. PREPROCESSING MICROARRAY DATA ISSUES

The powerful of Microarray technology is apparent when the data produced are affecting on the questionable fields. Because Microarray experiments produce large amounts of data, systematic methods are required to extract the meaningful expression relations [20]. In order to compare the expression levels of genes in a specific Microarray experiment, it is necessary to carry out transformations of data to adjust the resulted matrix values and to eliminate the measurements that are non-significant and with low quality.

5.1 Logarithm transformation

Even though ratios provide an intuitive measure of expression changes, it has the disadvantage of treating regulated genes differently [21]. There is a general convention that the better transformation procedure is to take the logarithm base 2 value of the expression ratio. The logarithm transformation of most Microarray data provides a good approximation of normal distribution with minor exceptions.

5.2 Missing values

Microarray data are often distinguished by a remarkable proportion of missing values [22]. Missing values are caused by different factors, such as the corruption of image, insufficient resolution, dust or scratches on the slide and the robotic methods used to create the arrays.

5.3 Filtering data

The reliability of measurements could be increased by keeping only the array elements that are significant above the background [23]. The filtering concept consists of getting a small data matrix from the origin one. The large matrix contains hybridizations arranged on the columns and substances on the rows and it may have lot of different observations types of each hybridization arranged on columns. The filtering implements three general concepts:

- Selection
- Average
- Estimation

6. MICROARRAY DATA NORMALIZATION METHODS

6.1 Correlation Based Method (CBM)

When the components changes of vector attribute are not of the same order of magnitude, the distance between forms may not be sensitive to changes in certain attributes, which is due to the difference of magnitude order between them. Consequently, a pretreatment step based on normalizing vectors is essential to remedy this problem.

When there is a correlation between the attributes, another form which is more adequate was introduced by Fukunaga [24]:

$$\widetilde{X}^n = \Lambda^{-1/2} U^t (X^n - \overline{X})$$

- Λ is a diagonal matrix containing Eigen values of the covariance matrix of all the existent observations.
- *U* is the matrix composed of columns containing the Eigen vectors of the covariance matrix.

It's evident that the new attributes of vectors have null means and an identity covariance matrix.

6.2 Linear scaling to unit variance (LV)

In order to transform an expression value X_{ij} with zero mean and unit variance, it has been calculated:

$$* X_{ij} = (X_{ij} - \text{mean}(A_j)) / sd(A_j)$$

- A_j represents the j^{th} array profile of the Microarray matrix
- X_{ij} is the origin expression value
- $* X_{ij}$ is the new expression value
- $\text{mean}(A_j)$ is the mean of the j^{th} array profile
- $\text{sd}(A_j)$ is the standard deviation

6.3 Linear scaling to unit range (LR)

Given a lower bound and an upper bound for an expression value X_{ij}, it has been calculated:

$$* X_{ij} = (X_{ij} - \min(A_j)) / (\max(A_j) - \min(A_j))$$

- A_j represents the j^{th} array profile of the Microarray matrix
- X_{ij} is the origin expression value
- $* X_{ij}$ is the new expression value
- $\max(A_j)$ is the maximum of the j^{th} array profile
- $\min(A_j)$ is the minimum of the j^{th} array profile

7. PCA TECHNIQUE

In this paper, the principal component analysis PCA was implemented. PCA technique reduces the dimensionality from m to d, where $d < m$. PCA allows transforming the original data to a new set of coordinates or variables representing a linear combination of the original variables.

In general, the important goals of PCA are to summarize patterns of correlation, to reduce the number of observed variables, and to provide the basis for predictive models [25]. A principal component could be seen as a rotation of the discerned data points [26]. Principal Components Analysis is a multivariate statistical technique that permits exploration and simplifying complex data sets.

8. MICROARRAY DATA PREPROCESSING

8.1 Microarray Data filtering

Microarray data should be filtered in order to select the valuable data from those representing the biases on data. The treatments were implemented to get rid of these biases. The types of gene expression profiles depicting the biases on data are:

- Gene expression profiles where the values are marked as NaN (Not a Number)
- Gene expression profiles where the variance is less than the 10^{th} percentile of the variance

- Gene expression profiles with low entropy expression values
- Gene expression profiles with small profile ranges where the ranges are less than the 10^{th} percentile.

8.2 Microarray Data processing steps

The following steps represent the Microarray data processing steps that were used in order to get the desired results:

- Beginning by selecting from GEO database a specific experiment which is responding to biological questions and hypotheses
- Getting the Microarray matrix of the selected experiment
- Transforming the initial gene expression data by taking logarithm base 2 of data
- Applying techniques for filtering the genes representing valuable information
- Implementing the predefined adopted methods on Microarray data
- Comparing the results obtained from each method (CBM, LV and CR) by exploring them on the PCA technique application
- Getting the significant Microarray data for further analysis

9. PREPROCESSING LUNG CANCER DATA

9.1 Microarray Dataset

In the present study, the Microarray Dataset was selected from the Gene Expression Omnibus (GEO) database. This Microarray experiment identified by GSE7670 [27] is described by the following elements:

- The submission date to GEO: April 30, 2007
- The last update date: May 14, 2015
- The title of this experiment: Expression data from Lung cancer

Table 2. The imported Microarray experiment features

GEO Accession	Features			
	Rows	Columns	Organism	Experiment type
GSE7670	22283 Gene profiles	66 Samples	Homo sapiens	Expression profiling by array

9.2 Experimental Results

To depict the difference between the original values and those transformed to Logarithm base 2, two graphs of genes were traced, eight genes were randomly chosen from the data set, to compare and visualize clearly the characteristics. By comparing the two graphs of the figure 3, it has been observed that the not log-transformed gene expressions are not distinguishable, the graph shows few genes on the interval over 0 and others are all skewed on the same level which makes analysis and comparison very difficult. In the second graph, where the gene expressions are log-transformed, genes are comparable very easily. The ratio data, in the first graph are skewed, where a large proportion of the values are confined to the lower end of the observed scale. Gene expression values, not log-transformed and represented under expression, are all crammed into the interval between 0 and 1. The reason for performing this operation is to pick up the levels characterized by skewness of the ratio data. When the values are normally spread around zero, it simplifies the data analysis and comparisons.

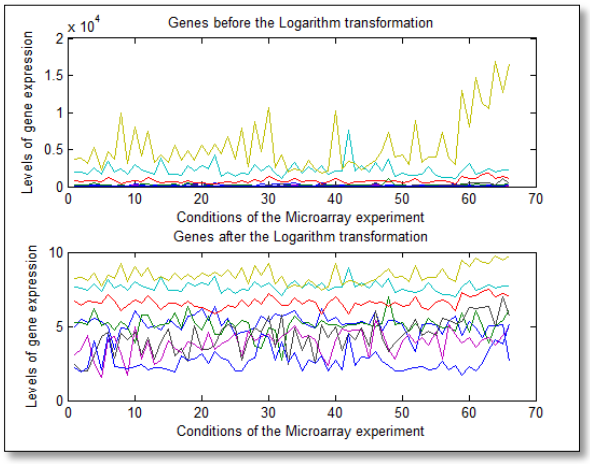

Fig 3: Comparison between the Microarray experiment genes before and after Logarithm transformation

It has been focused on the indispensable task of data preprocessing to make meaningful comparisons of expression levels and to select the useful genes for further analysis and Data mining techniques application. To illustrate the difference between the selected preprocessing methods, the methods were implemented on Microarray data and the results were represented by box plots. Each box is characterized by the central mark which is the median, the edges of box are representing the 25^{th} and 75^{th} percentiles, the whiskers that are extending to the most extreme data points are not considered as outliers by the algorithm, and the outliers are plotted separately.

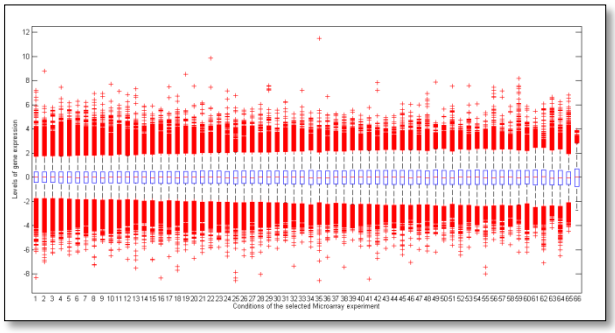

Fig 4: Box plot of Microarray data after CBM transformation

The box plots on the figures 4, 5 and 6 show a difference on the interval of gene expressions levels, the median and the outliers detected on each transformation. The method based on correlation (CBM) has produced different results compared to the other methods. The outliers are not only related to high extreme data points, but low extreme data points are also concerned. This normalization method spread out the condition values as much as possible, which improves the discrimination capabilities of the similarity measures that can be used for Data mining algorithms.

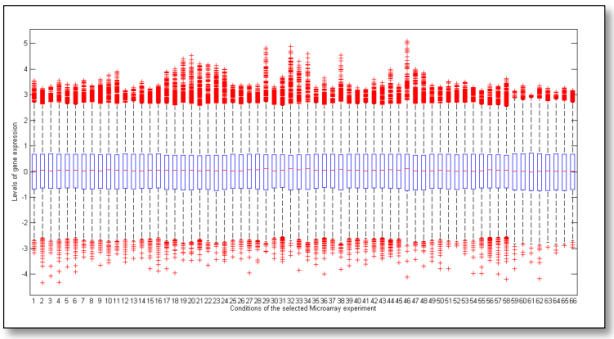

Fig 5: Box plot of Microarray data after LV transformation

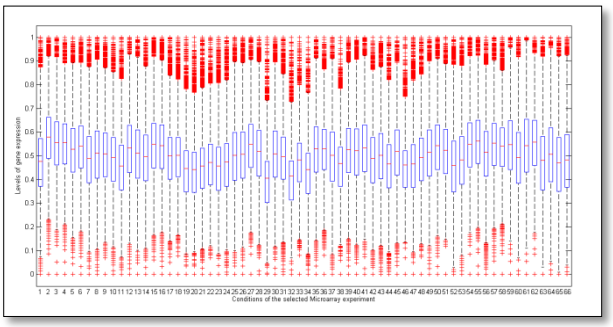

Fig 6: Box plot of Microarray data after LR transformation

For reducing the selected Microarray experiment dimension, PCA technique was implemented. The Eigen vectors and its values have been obtained. In the object of choosing the value of k which represents the number of the new matrix dimensions, the cumulative percentage of each Eigen vector index was calculated on the log-transformed Microarray data before the filtering techniques application (table 3). And the PCA technique was also applied on the data obtained from the implementation of each method (CBM, LV and LR).

Table 3. Cumulative percentage values of the explained variance resulted from PCA application on log-transformed data (before filtering techniques application)

Index of Eigen Values	Cumulative percentage of the variance
1	97,47
2	0,51
3	0,27
4	0,14
5	0,10
6	0,08

By comparing the results depicted on the two tables (table 3 and table 4), it shows the clear difference between the cumulative percentage values obtained. The first PC (Principal Component) from the table 3 explained more than 97% of the total variations, where the Microarray data are logarithm transformed. On the table 4, the first PC is explained by 76.65%. Thus, the differences depicted on the two tables illustrate the importance of the preprocessing methods implementation. On the table 4, the first PC is explained by 76.65%. Thus, the differences depicted on the two tables illustrate the importance of the preprocessing methods implementation. On the other hand, PCA based on correlation based method (CBM) is performed by 16243 genes, where the first thirty six PCs from the table 4 explained more than 95%.

Table 4. Cumulative percentage values of the explained variance resulted from PCA application on CBM transformed data

Index of Eigen Values	Cumulative percentage of the variance
1	76,65
2	4,93
3	1,74
4	1,17
5	0,97
6	0,75
7	0,50
8	0,43
9	0,42
10	0,38
11	0,38
12	0,36
13	0,35
14	0,34
15	0,33
16	0,32
17	0,31
18	0,31
19	0,30
20	0,29
21	0,29
22	0,28
23	0,27
24	0,27
25	0,26
26	0,26
27	0,25
28	0,25
29	0,25
30	0,24
31	0,24
32	0,23
33	0,22
34	0,22

35	0,21
36	0,21

For CBM transformation, the number of the principle components is 36. On the other side, for LV and LR transformations, the number of PCs is 35.

Table 5. Number of the first PCs explaining more than 95% for each adopted method

Index of Eigen Values	Cumulative percentage of the variance
CBM	36
LV	35
LR	35

By summarizing the obtained results, the Microarray data preprocessing represents an indispensable task for further analysis. It shows that the results provided by PCA technique can be used not only for representing data in reduced dimension spaces or for checking the validity of hypothesis implemented by the Microarray data based experiments, but also as input for more sophisticated Data Mining techniques.

10. CONCLUSION

The Microarray technology contributes to the improvement of decision making in many fields such as pharmacology, diagnostic of diseases and organism studies. In order to facilitate the use of this powerful technology, it has been chosen to work on preprocessing the data obtained from Microarray experiments. Due to the huge amount of Microarray data, many researchers and especially biologists find problems for treating Microarray data. The work presented in this paper has shed light on the existing differences between Microarray data before and after applying the proposed preprocessing techniques. It looks forward to develop a complete solution to minimize the obstacles facing investigators for responding to all hypothesis of the Microarray data based experiments.

11. REFERENCES

[1] C.-R. Chen, W.-Y. Shu, M.-L. Tsai, W.-C. Cheng, and I. C. Hsu, "THEME: A web tool for loop-design microarray data analysis", Computers in Biology and Medicine, vol. 42, no. 2, pp. 228–234, Feb. 2012.

[2] N. E. Olson, "The microarray data analysis process: from raw data to biological significance", NeuroRx, vol. 3, no. 3, pp. 373–383, 2006.

[3] J. Craig and N. C. Wong, Eds., "Epigenetics: a reference manual", Norwich, Norfolk, UK: Caister Academic Press, pp. 143-159, 2011.

[4] Fadoua Rafii, M. Aït Kbir and B. D. Rossi Hassani, "Microarray Data Integration to Explore the Wealth of Sources Generated by Modern Molecular Biology", Veille Stratégique Scientifique et Technologique, Granada, Spain, 11 - 13 may 2015.

[5] Fadoua Rafii, M. Aït Kbir and B. D. Rossi Hassani, "Microarray Data Preprocessing To Improve Exploration on Biological Databases", International Conference on Big Data, Cloud and Applications, Tetuan, Morocco, 25 - 26 may 2015.

[6] G. Ventimiglia and S. Petralia, "Recent Advances in DNA Microarray Technology: an Overview on Production Strategies and Detection Methods", BioNanoScience, vol. 3, no. 4, pp. 428–450, Dec. 2013.

[7] Schena, M., Shalon, D.; Davis, R. W.; Brown, P. O., "Quantitative monitoring of gene expression patterns with a complementary DNA microarray", Science 270, pp. 467-470, 1995.

[8] Lashkari D A, DeRisi J L, McCusker JH, Namath A F, Gentile C, Hwang SY, Brown PO, Davis RW, "Yeast microarrays for genome wide parallel genetic and gene expression analysis", Proc. Natl. Acad. Sci. U.S.A. 94, pp. 13057–13062, 1997.

[9] Schena, M., Heller, R., Theriault, T., Konrad, K., Lachenmeier, E., and Davis, R. W., "Microarrays: biotechnology's disovery platform for functional genomics", Trends Biotech. 16, pp. 301–306, 1998.

[10] K. Kafadar and T. Phang, "Transformations, background estimation, and process effects in the statistical analysis of microarrays", Computational Statistics & Data Analysis, vol. 44, no. 1–2, pp. 313–338, Oct. 2003.

[11] Miller LD, Long PM, Wong L, Mukherjee S, McShane LM and Liu ET, "Optimal gene expression analysis by microarrays", Cancer Cell, pp. 353–361, 2002.

[12] Pavlov, V., Xiao, Y., Gill, R., Dishon, A., Kotler, M., Willner, I., "Amplified chemiluminescence surface detection of DNA and telomerase activity using catalytic nucleic acid labels", Analytical Chemistry 76, pp. 2152-2156, 2004.

Gooding, J. J., "Electrochemical DNA hybridization biosensors", Electroanalysis 14, pp. 1149 -1156, 2002.

[14] Fawcett N. C., Evans J. A., Chien L. C., Flowers N., "A quartz crystal detector for DNA", Analytical Letters 21, pp. 1099-1110, 1998.

[15] Sánchez-Pla, A., "DNA Microarrays Technology: Overview and Current Status", Comprehensive Analytical Chemistry, vol. 63, Elsevier, pp. 1–23, 2014.

[16] A. Brazma, M. Kapushesky, H. Parkinson, U. Sarkans, and M. Shojatalab, "[20] Data Storage and Analysis in ArrayExpress", Methods in Enzymology, vol. 411, Elsevier, pp. 370–386, 2006.

[17] J. Demeter, C. Beauheim, J. Gollub, T. Hernandez-Boussard, H. Jin, D. Maier, J. C. Matese, M. Nitzberg, F. Wymore, Z. K. Zachariah, P. O. Brown, G. Sherlock, and C. A. Ball, "The Stanford Microarray Database: implementation of new analysis tools and open source release of software", Nucleic Acids Research, vol. 35, no. Database, pp. D766–D770, Jan. 2007.

[18] https://array.nci.nih.gov/caarray/home.action

[19] Gimbrone, Jr, et al., "Argus-A New Database System for Web-Based Analysis of Multiple Microarray Data Sets", pp. 1603-1610, 2001.

[20] http://biochem218.stanford.edu/Projects%202012/Yu.pdf

[21] J. Quackenbush, "Microarray data normalization and transformation", Nature Genetics, vol. 32, no. Supp, pp. 496–501, Dec. 2002.

[22] Xiaofeng Zhou, Hiroshi Egusa, Steven W. Cole, Ichiro Nishimura, and David T.W. Wong, "Methodology of

Microarray Data Analysis", Volume 3: Molecular Genetics, Liver Carcinoma, and Pancreatic Carcinoma pp. 17-29, 2005.

[23] T. K. Karakach, R. M. Flight, S. E. Douglas, and P. D. Wentzell, "An introduction to DNA microarrays for gene expression analysis", Chemometrics and Intelligent Laboratory Systems, vol. 104, no. 1, pp. 28–52, Nov. 2010.

[24] K. Fukunaga, "Introduction to pattern recognition (Second ed.)", Academic Press (San Diego), 1990.

[25] Tabachnick B.G., Fidel L.S., "Using Multivariate Statistics 3rd Edition", Harper Collins College Publisher, pp. 635-708, 1996.

[26] S. Raychaudhuri, J. M. Stuart, R. B. Altman, and others, "Principal components analysis to summarize microarray experiments: application to sporulation time series", in Pac Symp Biocomput, vol. 5, pp. 455–466, 2000.

[27] http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE7670