PhUSE 2014

Paper HE02

From Gene Expression to Expression Cartography
Correspondence Analysis Application in Class Comparison Studies

Edyta Winciorek, PAREXEL, Warsaw, Poland

ABSTRACT
DNA microarrays and gene expression profile analysis give new possibilities in clinical trials analysis. An important goal of microarray studies is the reliable detection of genes that shows significant changes in observed expressions of gens between treatment and control group of patients in clinical trials. Taking into account that DNA microarrays can simultaneously measure the expression level of thousands of genes within a particular mRNA sample a specialized statistical techniques are required for the analysis of high-dimensional microarray data. In this article the overview of the applications of Correspondence Analysis (CA) to microarray data will be presented. CA methods will be used to detect the differences and associations in gene expression among treatments and controls. The results will be discussed using measures of the "*representativness*" of the trends, as well as measures of their "*regularity*". The concept of "*overrepresentation*" maps will be used for the visualization of trends.

INTRODUCTION
Thanks to the microarrays DNA technology it is possible to examine the behavior of all the genes of an organism under different conditions during the single experiment. Gene expression microarrays are a standard tool for large-scale measurement of gene expression. This methodology is widely used to detect genes that are differentially expressed across different groups. The results of such experiments can change the methods employed in the diagnosis and prognosis of disease in obstetrics and gynecology. However, the analysis studying the patterns of gene expression profiles in large microarray dataset is challenging statistical task. The main issue is the complexity of the problem, the data generated by these experiments may consist of thousands to millions of gene profiles with only several number of samples available in microarray matrix. There are several methods that can be applied to microarray data analysis - from classical statistical methods like ANOVA, t-test, to the use of test statistics developed specifically for the microarray context. In this paper, we introduce the Correspondence Analysis (CA) for the analysis of microarray gene expression data. This is quite novel approach which scales he relationship between genes and tissues in multidimensional microarray gene expression data into 2-dimensional graphs in such a way, that genes which primarily distinguish certain types of tissue are spatially close to those tissues.

MICROARRAY GENE EXPRESSION PROFILE – PRELIMINARIES
All human being organisms consist of trillions of cells and each cell contains a complete copy of the genome which is encoded in DNA. A gene is a segment of DNA that specifies how to make a protein. **Gene Expression** is the process by which the information encoded in a gene is converted into an observable phenotype. At the end of this process we get the degree to which a gene is active in a certain tissue of the body, measured by the amount of mRNA in the tissue (Alshamlan, 2013).

DNA Microarray is a technology which enables the researchers to analyze the expression of the large number of genes within a particular mRNA sample. A typical microarray experiment involves the hybridization of an mRNA molecule to the DNA template from which it is originated. More precisely, mRNA is extracted from the subject tissues or cells. Next, some of its molecules are labeled with a fluorescent dye. The resulting labelled transcripts are called targets. Such prepared sample is deposited over the array and left onside a hybridization chamber for some hours where the labelled targets bind by hybridization to the probes on the array. When the hybridization process is finished the array is washed out in order to eliminate not hybridized targets. For hybridized targets the quantity of labelled target should be proportional to the level of expression of the gene represented by that probe. It means that the amount of fluorescence measured at each sequence specific location is directly proportional to the amount of mRNA with complementary sequence present in the sample analyzed. In order to estimate the amount of sample hybridized the microarray is illuminated by a laser light. Thanks to that, the labelled molecules emit fluorescence proportionally to their quantity. The next step is to generate an image using laser-induced fluorescent imaging. Such created image is finally transformed into numbers and will be the basis of the statistical analysis (Sánchez and Villa, 2008). The process how the images are turned into numbers (quantify) will not be discussed here, however the principle behind the quantification of expression levels is to provide the data that can be used to compare the expression level among conditions and genes (e.g., health vs disease) (Gibson, 2002.).

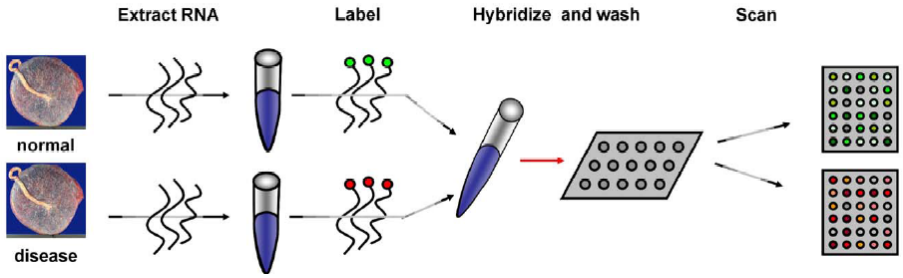

Figure 1 Schematic representation of the steps involved in microarrays (Tarca, 2006).

All the steps involved in microarrays are illustrated on Figure 1. Here the experiment is designed to compare the mRNA expression profile of tissues between healthy subject and patients with a disease. The normal and disease mRNA are labeled with two different dyes (usually green and red), mixed and then hybridized on the same array. After washing, the array is scanned at two different wavelengths to yield two images: one for normal patient and one for patient with disease.

Once the experiments are done and images are acquired, the measurement of relative expression of thousands of genes are captured in the **Gene Expression Matrix** – the matrix, where rows represent genes and columns represent measurements from different experimental conditions measured on individual arrays (for example ill or normal tissue). More precisely, let *i* be integer identifying a known gene G_i, and *j* be integer identifying a particular hybridizations H_j (experiment trial). *I* genes and *J* hybridizations are collected into the *IxJ* matrix *N* with elements n_{ij} - the gene expression level for each gene G_i in hybridizations H_j. Figure 2, illustrates an example of the hybridization with the gene expression matrix *N* for *n* genes assayed by *m* microarray experiments. Typically, microarray matrixes contain thousands of rows and dozens of columns.

	H1	H2	H3	H4	...	Hm-1	Hm
G1	0.6	4.4	1.3	1.0	...	3.1	2.2
G2	1.5	2.6	5.2	0.8	...	2.8	2.9
G3	0.7	3.7	2.4	1.9	...	1.5	1.6
G4	0.3	0.7	0.2	1.3	...	4.9	3.0
G5	3.1	3.0	2.1	1.4	...	4.2	0.9
...	...	...	...	...	...	...	...
Gn-1	1.8	2.5	1.8	0.7	...	2.7	3.1
Gn	0.5	3.4	3.0	0.5	...	1.8	2.5

Figure 2 Example of microarray gene expression matrix (Simon, 2009)

Prior to the analysis and interpretation of the results the gene expression matrix needs to be preprocessed, for example the logarithm of the raw intensity values is taken or normalization of data is performed in order to compensate for the different dye efficiency in two channel microarray experiments using green and red color. Preprocessing is a step that extracts or enhances meaningful data characteristics and prepares the gene expression matrix for the application of statistical analysis methods. The preprocessing steps will not be disused here, but which can be considered to be relatively reliable and stable.

TYPES OF DNA MICROARRAY EXPERIMENTS

The are several types of studies that can be conducted with DNA microarrays. Depending on the objectives of the study we can distinguish the following main categories of microarray analysis: class discovery, class prediction, class comparison and pathway analysis (Simon, 2003).

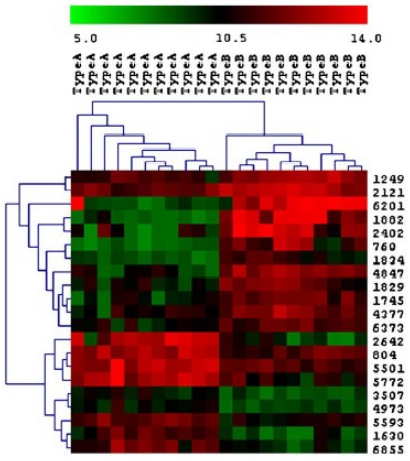

Figure 3 An example of simultaneous clustering of arrays (discovery of 2 types of patients A and B) and genes (discovery of co-regulated groups of genes) (Tarca and all, 2006).

Class Discovery (*Clustering*) is very often the first step of gene expression matrix analysis. Taking into account that that co-regulated and functionally related genes are very often expressed (go up or down) simultaneously we can identify them, group together similarly expressed genes and finally try to correlate results with investigate biological process. Moreover, clustering analysis can be used to identify spatial or temporal expression patterns and by this we can identify new biological classes (Tarca, 2006). For example it has been shown that certain types of leukemia present some subclasses that are very hard to distinguish morphologically but which can be classified using gene expression. What is more, clustering can reduces the dimensionality of the gene expression matrix and by this, the management of huge gene expression matrix is easiest.

Class discovery analysis is usually based on unsupervised machine learning method such as hierarchical clustering, k-means clustering or self-organizing maps (Young, 2009). The result of the clustering is often illustrated graphically combines so called "*dendrogram*" and ''*heat map"* - see Figure 3. On dendrogram the clusters creates a hierarchical, tree-like structure, while heat map refers to any display in which intensities of gene expression are mapped on a color scale (color red represents high expression, while green represents low expression). The rows represent genes identified by the numbers on the right of the figure, while the columns represents the individual patient samples.

Class Prediction Experiments involves classification techniques in order to identify the sample's class membership basing on its gene expression profile. An example of class prediction experiments would be to predict if subject will develop (or not) pre-eclampsia based on her blood expression profile. For this purpose we need to construct a classifier i.e. a) determine mathematical model well describing the classification rule used to distinguishing the pre-defined classes, b) estimate the parameters of the mathematical function used in this model, and c) estimate the accuracy of the predictor. One of the most popular method used to determine mathematical function distinguishing the pre-defined classes is discriminant analysis (Dudoit, 2002). This method allows to classify binary or multiple outputs using a discriminant function of expression values of selected genes obtained by maximum likelihood function. Once the model is established then the training process is activated. The purpose of this training to validate the mathematical model, i.e.: calculate its specificity and sensitivity and predicted values. Figure 4 illustrate the principles of class prediction experiments (Golub, 1999). On left side the process of predictor building and validation are illustrate, while on right side is show how this predictor can be used to predict the subjects status basing on his gene expression profile.

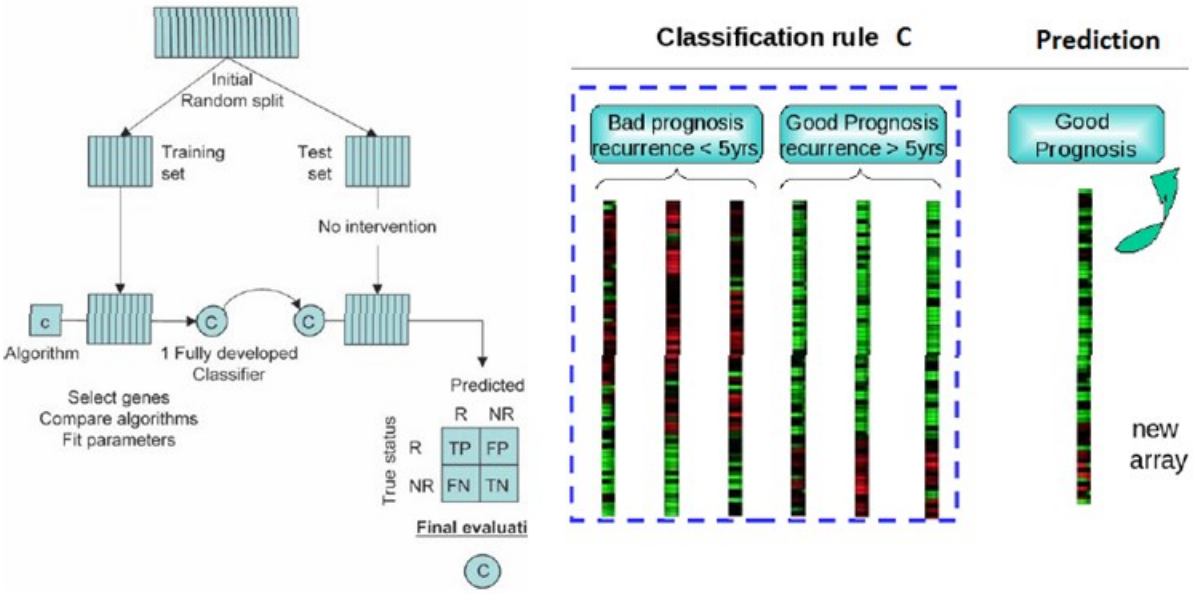

Figure 4 Class Prediction example: assignment of type to a new sample of gene expression matrix (Sánchez and Villa, 2008).

Class Comparison Experiments called also **Differential Analysis** are used to find the differences in expression levels of gens between two or more groups of patients. This kind of experiments can address the question about the impact of treatments on genes expression or can be used for comparison the gene transcription of healthy vs diseased subjects.

The statistical context of such analysis is the comparison of two populations according to the level of its gene expression verifying the *null hypothesis* standing that a given gene on the array is not differentially expressed between the two conditions under study against the *alternative hypothesis* that the expression level of that gene is different. There have been developed many models and methods for such kind of analysis - some are based on parametric models (like t-test or ANOVA), others are based on nonparametric approaches (like SAM method) allowing the hypothesis testing without distributional assumptions. In ANOVA procedure a single model is for testing hypothesis for all the genes simultaneously. On the other hand, the global tests like SAM, analyze each gene separately, using a common model. However, using this methodology one has to face up the large multiplicity problem where thousands of hypotheses are tested simultaneously. This increased the chance of false positives signals in hypothesis testing problem - i.e., a type I error which occurs when the null hypothesis is incorrectly rejected.

The next step after receiving the list of genes (very often long) which behave differently is a **Pathway Analysis**. This methodology is used to assign the biological interpretation to given list of genes, i.e. the interpretation that reflects their participation in common biochemical processes. Usually the biological interpretation process is basing on reprocessing the list genes with one or more functional annotation databases such as the Gene Ontology (GO), Kyoto Encyclopedia of Genes and Genomes (KEGG) or others (Draghici, 2005).

In this article we will focus our attention to Correspondence Analysis (CA) and in the next section we present how CA can be applied to detect the differences and associations in gene expression among treatments and controls group. This is a methodology where no parameterization is needed. Moreover CA reveals and visualizes interdependencies (correspondence) between hybridization – genes at the same time.

CORRESPONDENCE ANALYSIS OF DNA MICROARRAYS

The **Correspondence Analysis** is a quite novel approach used for visualization the relationship between genes and tissues as 2-dimensional graphs (Waddel, 2000). This projection is oriented so that the distances between genes and distances between tissues are preserved as well as possible. Thus, the genes which primarily distinguish certain types of tissue are spatially close to those tissues.

Correspondence analysis was originally developed for contingency tables in order to show associations between particular rows and columns – in the sense of deviations from homogeneity, measured by the χ^2-statistic (see Table 1). The main purpose of CA is to investigate whether the differences among the rows (columns) of the table are large enough to reject the hypothesis that the columns (rows) are homogenous. In terms of microarray experiments we may investigate if the discrepancies between the observed gens' profile and an average gene profile - expected in case of homogeneity – are statistically significant. The analytical process of correspondence analysis bases on row (map 1) and column (map 2) profiles.

Table 1 Contingency table for microarray experiment

Genes	Experiments		Total
$gene1$	a_1	b_1	T_1
$gene2$	a_2	b_2	T_2
$gene3$	a_3	b_3	T_3
.....	....	....	....
Totals	a_T	b_T	T

For contingency Table 1 the **Row Profile for** Gene G_i is calculated as follows: $a_{ri} = \frac{a_i}{T_i}$, $b_{ri} = \frac{b_i}{T_i}$ while the **Column Profile for Gene G_i** is: $a_{ci} = \frac{a_i}{a_T}$, $b_{ci} = \frac{b_i}{b_T}$. The **Average Gene Profile** (expected profile) is the profile of the marginal distribution calculated as follows: $n_a = \frac{a_T}{T}$; $n_b = \frac{b_T}{T}$. The **Gene Masses** (or the Marginal Profile) consists of the relative frequency distributions of the sums of the rows, i.e.: $m_1 = \frac{T_1}{T}$; $m_2 = \frac{T_2}{T}$.

For row profile (map 1) – the average of the row profiles is the treated as **centroid** and it is placed at the origin on the principal axes. During the analysis each gene profile is next compared with this centroid using the distance corresponding to the chi squared separation:

$$d(i,i') = \sqrt{\sum_j \frac{a_{ij} - a_{i'j}}{a_{.j}}} \qquad (1)$$

where $d(i,i')$ is the "chi square" distance between the points i and i', a_{ij} are the elements in the row profile, while $a_{.j}$ are elements in the average row profile. In the same way, the column profiles are performed. The difference between observed gene profile and centroid (expected value) is squared and subsequently divided by the expected value. These distances, calculated for all elements of the table, sum up to the **χ2-statistic**. Using this value we may verify the hypothesis testing problem about the homogeneity of gene expression matrix, namely the higher value of the χ2-statistic the better association between rows and columns in contingency tables. Moreover, in CA procedure the chi square distance is used for scaling the multidimensional gene expression matrix to 2-dimensional space, thus we can easily present the data and find the potential patterns.

CORRESPONDENCE ANALYSIS APPLICATION IN CLASS COMPARISON STUDIES – EXAMPLE

In this section we discuss an example of correspondence analysis applied to the data available on the web at http://genomics-pubs.princeton.edu/oncology/ (Alon, 1999). This is the example of class comparison studies, where data are a series of 62 Affimetrix GeneChip experiments upon normal (N) and cancerous (T) colon tissue. In order to perform CA the **PROC CORRESP** from SAS 9.1® has been used. The standard SAS Output of this procedure run for the data is shown below:

Correspondence Analysis - plangoa

The CORRESP Procedure

Inertia and Chi-Square Decomposition

Singular Value	Principal Inertia	Chi-Square	Percent	Cumulative Percent	11 22 33 44 55
0.42252	0.17852	2668.35	53.27	53.27	**********
0.39571	0.15658	2340.44	46.73	100.00	**********
Total	0.33510	5008.79	100.00		

Degrees of Freedom = 6342

Column Coordinates

	Dim1	Dim2
bp	-0.1598	-1.2932
cc	1.7256	0.5600
mf	-0.8651	0.8802

Summary Statistics for the Column Points

	Quality	Mass	Inertia
bp	1.0000	0.3706	0.2947
cc	1.0000	0.2330	0.4038
mf	1.0000	0.3963	0.3015

In example, the total χ2-statistic, which is a measure of the association between the rows and columns is 5008.79 and is explained equally for both the dimensions – i.e. about 53.27% explain Dimension 1 and 46.73% explain Dimension 2. This indicates that the association between the row and column categories is essentially two dimensional. Using the column coordinates (see Dim1, Dim2 in SAS output presented above) these genes were located on a 2-dimensional map (*biplot*) shown in Figure 5. If a column determines an outstanding entry of a row (and vice versa), then the corresponding row and column points tend to lie on a common line through the centroid. For a positive association two points will lie on the same side of the centroid (the larger the distance to centroid the stronger the association). A negative association will cause the column-point and the row-point to lie on opposite sides of the centroid (Fellenberg, 2002).

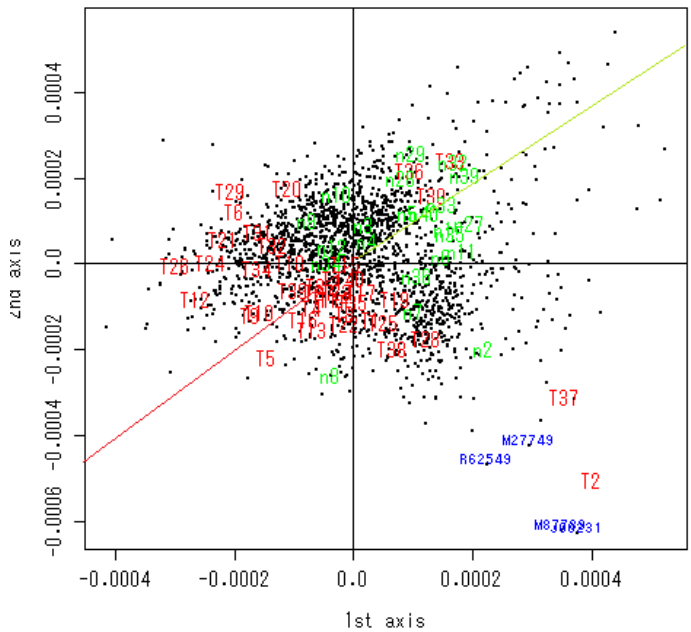

Figure 5 Overlaid scatter plot of genes and cells upon normal (green) and cancerous (red) colon tissue.

It is easy to notice that the normal cells are mostly distributed in the upper-right area of biplot, while the tumor cells are distributed in the lower-left region. By analyzing the graph we can easily separate the case and control group exceptions are T30, T33, T36 that appear in the upper-right region and n8 that appears in the lower-left region. Moreover, tumor tissues T2 and T37 are located in the peripheral area of the scatter plot, which implies that changes in the expression pattern of a small number of genes typifies them. Reflecting the fact that T37 is an outlier, a number of genes with typically low expression levels, are very high in this tissue and distinguish it and T2 from other tissues. Indeed, there are four genes, J00231, M27749, M87789, R62549, are located near T37 and T2. They are all immunoglobulin related genes.

CONCLUSION

Microarrays DNA experiments is quite new technology that enables the researchers to monitor the expression levels of thousands of genes simultaneously. The result of these experiments can be used in medicine for comparing clinically relevant groups (e.g., healthy vs diseased), Moreover, gene expression matrix can be used to detect the new subclasses of diseases, protect clinically important outcomes, such as the response to therapy and survival. Thus there is still growing application of DNA microarray in clinical trials. In this article we present the application of correspondence analysis to class comparison studies. This methodology can be used to measure the homogeneity in gene expression matrix. Moreover, the correspondence diagram seems to be giving a reasonable picture of highly expressed genes associated with particular tissue types.

REFERENCES

Alon, U. B., Barkai D., Notterman D. and all (1999). Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc. Natl. Acad. Sci. USA, 96*, 6745–6750.

Diong, C. (2004). *A method for linking microarray data to database information (*http://mysite.science.uottawa.ca/ardsg/microarrayfiles/Diong.pdf*)*

Draghici., P. K. (2005). Ontological analysis of gene expression data. *Bioinformatics*(18), 3587-3595.

Dubitzky W, G. M. (2002). Data mining and machine: Methods of microarray data analysis. *Cambridge (MA): Kluwer Academic*, 5-22.

Dudoit, S, Fridlyand, J. and Speed, T. (2002). Comparison of discrimination methods for the classification of tumors using gene expression data. *Journal of the American Statistical Association, 97*, 457.

Fellenberg, K. (2002). *Storage and analysis of microarray data.* Heidelberg.

Gibson G, Maue S. (2002.). A primer of genome science. *Sunderland, MA: Sinauer Associates, Inc.;*.

Golub, D. K., Slonim, D., Tamayo, P., Huard K. and all (1999). Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science, 286*.

Hala M., Alshamlan, G. H. (2013). A Study of Cancer Microarray Gene Expression. *Proceedings of the World Congress on Engineering, Vol II*.

Richard M. Simon, E. L. (2003). Design and Analysis of DNA Microarray Investigations. *Springer-Verlag*.

Schena, M., Shalon, D., Davis, R. and Brown P. (1995). Quantitative monitoring. *Science* 270:467-70.

Simon, R. (2009). Analysis of dna microarray expression data. *Best practice and research Clinical haematology, vol. 22, no. 2*, 271–282.

Tarca, A. L., Romero, R., Draghici, S. (2006). Analysis of microarray experiments of gene expression profiling. *American journal of obstetrics, 195, no. 2*, 373–388.

Waddel, P.J. Kishino, H. (2000). Correspondence Analysis of Genes and Tissue Types and Finding Genetic Links from Microarray Data. *Genome Informatics, 11*, 83-95.

Sánchez, A. and Ruíz de Villa, M. (2008). *A Tutorial Review of Microarray Data Analysis (*http://www.ub.edu/stat/docencia/bioinformatica/microarrays/ADM/slides/A_Tutorial_Review_of_Microarray_data_Analysis_17-06-08.pdf*)*

Y. T. Young. (2009). Efficient multi-class cancer diagnosis algorithm,using a global similarity pattern. *Comput. Stat. Data Anal., 53*(3), 756–765.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Author Name: **Edyta Winciorek**
Company: **PAREXEL**
Address: **Zwirki i Wigury 18a,**
City / Postcode: **02-092 Warsaw, Poland**
Work Phone**: +48.22.4521.305**
Email: **Edyta.Winciorek@PAREXEL.com**