PRE-PROCESSING AND CLUSTERING ALGORITHMS FOR MICROARRAY DATA IN CANCER RESEARCH

Ioana ILEA, Monica BORDA
Technical University of Cluj-Napoca
26-28 George Baritiu St., 400027 Cluj-Napoca, Romania, Tel: +40-264-401226, Fax: +40-264-597083,
{Ioana.Ilea, Monica.Borda}@com.utcluj.ro

Abstract: The development of microarray technologies makes it possible to understand the molecular basis of human diseases, like cancer. Microarray experiments may be used to simultaneously asses gene expression values for thousands of genes. Along with bioinformatics and data mining techniques, they are important in finding relationships between gene expression values and clinical outcomes. One of these techniques is represented by the clustering analysis. This paper describes a typical workflow for microarray data and different hierarchical clustering algorithms for DNA microarray data, underlining their advantages and disadvantages.

Key words: *microarray, gene expression, pre-processing, normalization, hierarchical clustering, cancer.*

I. INTRODUCTION

In recent years, there has been a grate development in the area of biotechnologies. New high-throughput microarray technologies make it possible to gain comprehensive insight into the molecular basis of human diseases. An important area of research in this field is concerned with various aspects of cancer disease and gene expression values.

Microarray experiments usually start with a biological question and follow several steps in order to find an appropriate answer to that initial problem. After defining the biological question, a microarray experiment is designed and carried out. As a result, images are produced, and quantified to obtain intensity values. The next step in the workflow for microarray data analysis is preprocessing, or the low-level analysis, which produce expression values and quality assessment. Expression values may than be used in higher level analysis to answer the biological question. One of these methods might be the class discovery analysis, or clustering. The goal of clustering methods is to group the objects in a data set in subsets, such that objects within a subset are more similar to one another than objects in different subsets [4]. The partition should be made without prior knowledge of class labels.

This paper illustrates the workflow for microarray data analysis. In the next part, some elements of genetics and microarray technology are introduced. Then, in the third part, pre-processing techniques will be presented, for obtaining values that will be further on used in high level analysis. Clustering algorithms represent the technology that will be described in the fourth part. The last part, "Experimental results", will highlight how different parameters for hierarchical clustering may influence the final results. The clustering technique will be applied for a data set correlated to bladder cancer patients, in order to classify them according to the clinical outcome. In the end, some conclusions on this topic will be mentioned.

II. ELEMENTS OF MICROARRAY TECHNOLOGY AND BIOINFORMATICS

1. Basic biology of cell expression

The cell is considered the basic unit of life. Most of the important functions performed in cells involve proteins. A protein can be represented as a linear sequence of amino acids joined by peptide bonds. Typical proteins contain between 100 and 1000 amino acids.

Proteins do not self-assemble. They are assembled based on information contained in DNA. In proteins formation, messenger RNA (mRNA) acts as an intermediate: mRNA is synthesized using DNA as a template and then it is used for translation into protein. All RNA molecules consist of a sequence of nucleotides, or bases. There are four types of nucleotides in RNA: adenine (A), cytosine (C), guanine (G), and uracil (U), while in DNA molecules thymine (T) replaces uracil (U).

The synthesized RNA molecule is complementary to the DNA sequence in the gene, the complementarity of bases being a crucial feature of DNA and the basis of both cell reproduction and gene expression.

Cells differ from each other in function. These functional differences are determined by differences in the abundance of various types of proteins and can be measured using gene expression.

Cancer researchers are interested in finding biological differences between cancerous tissue and normal tissue, in order to find a cure for this disease. Cancer cells often have patterns of gene expression that differ from their normal cell counterparts. Some genes are expressed at higher levels (*over-expressed genes*), and others are expressed at lower levels, or not at all (*under-expressed genes*). Genes that are over-expressed in cancerous tissue are of particular interest because over-expression is an expected characteristic of a

Manuscript received January 29, 2014; revised February 28, 2014

gene that is involved in cancer growth. Microarray gene expression analysis could be used to facilitate the identification of molecular prognostic markers that correlate with bladder cancer outcomes.

2. DNA microarray technology

DNA microarrays are assays for quantifying the types and amounts of mRNA transcripts present in a collection of cells [4]. The number of mRNA molecules derived from transcription of a given gene is an approximate estimate of the level of expression of that gene.

Microarray consists of a solid surface on which strands of polynucleotide have been attached in specified positions, called *probes*. Probes are substances or molecules, for instance nucleic acids, affixed to an array or a chip. These molecules interact in a selective and predetermined way with target substances to measure a specific property or quantity. The obtained numbers should represent an estimation of the level of gene expression [4].

Chip technologies differ in type, number and distribution of probe sequences across the chip, as well as in the number of samples that are hybridized to a chip. One of these microarray platforms is Illumina BeadArray.

Illumina's BeadArray Technology, manufactured by Illumina Inc., San Diego, CA, is based on 3-micron silica beads that self-assemble in microwells on either of two substrates: fiber optic bundles or planar silica slides. On these two substrates, beads are randomly assembled and have a uniform spacing of ~ 5.7 microns. Each bead is covered with hundreds of thousands of copies of a specific oligonucleotide that act as the capture sequences in one of Illumina's assays [7]. The data set in the experimental part is obtained with this technology.

III. PRE-PROCESSING MICROARRAY DATA

Every microarray experiment produces images. Image analysis software reduces these images to raw intensity data. To be useful for a data analyst in order to apply data mining techniques [3] the raw intensity data needs to be converted into gene expression measures. *Pre-processing* is used to describe these procedures.

The most important purpose of the low-level analysis of a microarray data experiment is to provide better expression measures which can be used in higher-level analysis. Ideally, expression values should be both precise (low variance) and accurate (low bias). Another important aspect of low-level analysis is to be able to assess the quality of the microarray data.

The pre-processing analysis contains of some others stages: quality control of the data set, background correction, data normalization, logarithm transformation and data filtering.

1. Quality control

Quality assessment is an important part of a low-level analysis involving both data inspection and decision making.

Microarray data quality assessment can be carried out at several levels. It is recommended to be performed as the first step in the low-level analysis, because it is important to carefully examine distributional properties of the data and to assess their quality prior to any further analysis. This step should also be performed after normalizing the dataset, in order to highlight the results and the efficiency of normalization techniques. The goal of quality control methods is to examine whether the data are appropriate for any subsequent analysis.

Usually basic summary statistics and graphical assessments of the data distribution offer important information about data's quality. Parameters like: median, mean, first quartile, third quartile, variance and standard deviation or minimum and maximum values, offer information on data's distribution. Graphical assessment is often used for statistical analysis. Some of the most useful methods include: box plots, histograms, density plots, scatter plots, M-A plots. These techniques offer a visual representation and interpretation of the distribution of the data.

2. Background correction

Background correction is an important step required for microarray data because typically there is some level of binding producing detectable signal even when specific biological material is not in the original sample, and its purpose is to remove the non-specific signal from total signal.

Sometimes this step is already performed by the software used for processing Illumina BeadArray data, like *BeadStudio*. As this is not always true, several methods for background correction are available and the choice should take into consideration the analysis that will be carried out later. For instance, if a logarithmic transformation will be applied, the background correction method should force all values to be positive. This can be accomplished by adding an offset, so that all negative values are converted into positive ones.

Some other methods are background subtraction that produces negative values or the robust multi-array analysis (RMA) background correction [6].

3. Data normalization

Normalization is the process of removing unwanted non-biological variation that might exist within and between arrays in a microarray experiment. In other words, the normalization process makes the measurements from different arrays inter-comparable.

DNA microarray technologies are characterized by specific methods for normalization. One of the methods used for normalizing Illumina data is quantile normalization.

Quantile normalization is a method used to make the distribution of probe intensities the same for every sample (or array). The method is based on the rationale behind the quantile-quantile plots (QQ-plots). Quantiles, are the sorted measurements or values, of a data set that are plotted against the quantiles of another data set. A QQ-plots can be used for analyzing the distributions of two datasets. For example, two data vectors have the same distribution if the plot is a straight diagonal line. If this requirement is not accomplished, and the plot is other than a diagonal line the distribution of the vectors is different. The same concept can be extended to an n dimensional space. If n vectors have the same distribution, then by plotting the quantiles in an n dimensional space a straight line along the line given by the unit vector $(1/\sqrt{n},...,1/\sqrt{n})$ will be obtained. This suggests that a set of data have the same distribution if the points of the n dimensional quantile plot are projected onto the diagonal.

Let $q_k = (q_{k1},...,q_{kn})$ for $k=1,\ldots, p$ be the vector of the

k^{th} quantiles for all n arrays and $d = (1/\sqrt{n},...,1/\sqrt{n})$ be the unit diagonal. To transform the quantiles so that they all lie along the diagonal, the projection of q onto d should be considered:

$$proj_d q_k = \frac{q_k \cdot d}{d \cdot d} d = \frac{1}{\sqrt{n}} \sum_{j=1}^{n} q_{kj} d = (\frac{1}{n} \sum_{j=1}^{n} q_{kj},...,\frac{1}{n} \sum_{j=1}^{n} q_{kj}) \quad (1)$$

4. Logarithmic transformation

DNA microarray datasets are usually transformed from raw intensities into binary logarithms (base 2) intensities before proceeding with analysis. One of the goals of this data transformation is to make the transformed distribution close to a Gaussian statistical distribution. Furthermore, features should be reasonably even spread across the intensity range. In addition, the variability should be constant at all intensity levels and the distribution of experimental errors should be approximately normal.

5. Data filtering

Unspecific filtering refers to methods of excluding a certain part of the data without any knowledge of the grouping of the samples. An unspecific filtering method is filtering by standard deviation. In this case, all the elements that have a standard deviation smaller than a threshold are eliminated from the data set.

IV. ALGORITHMS FOR HIERARCHICAL CLUSTERING

According to the structure of the final results, data can be clustered in a hierarchical or nonhierarchical manner. Hierarchical clustering allows detecting higher-order relationships between clusters of profiles whereas most of the nonhierarchical classification techniques allocate profiles into a predefined number of clusters, without any assumption on the inter-cluster relationships [1].

Hierarchical clustering is used for grouping objects into a tree of clusters. This technique offers the possibility of analyzing the relationships between clusters at different levels. According to the method used for computing distance metrics, different algorithms can be implemented.

Usually, microarray data are represented as two dimensional matrices, where the lines represent genes and columns represent patients. In this case, there are two methods used to determine the distance between two data objects: Euclidean distance and Pearson's correlation coefficients, as mentioned in [2].

Euclidean distance:

$$d(o_i, o_j) = (\sum_{k=1}^{p} | o_{ik} - o_{jk} |^2)^{1/2} \quad , \quad (2)$$

where: o_i and o_j are data objects and p is the number of dimensions for multidimensional objects.

Pearson's correlation coefficients:

$$Pearson(o_i, o_j) = \frac{\sum_{k=1}^{p} (o_{ik} - \mu_{o_i})(o_{jk} - \mu_{o_j})}{\sqrt{\sum_{k=1}^{p} (o_{ik} - \mu_{o_i})^2} \sqrt{\sum_{k=1}^{p} (o_{jk} - \mu_{o_j})^2}}, \quad (3)$$

where: μ_{o_i} and μ_{o_j} the means for the objects o_i and o_j and p is the number of dimensions for multidimensional objects.

Hierarchical clustering algorithms can be further divided into two categories: agglomerative and divisive clustering.

Divisive clustering is a top-down method. At the beginning, all objects are part of the same cluster and at each step the best split is found. The process of dividing the data set stops when each object forms an individual cluster, or when a termination condition is reached. This algorithm is computationally demanding and it is not used in data analysis.

Agglomerative clustering starts by considering each object as a distinct cluster. At each step, the distance between two clusters is computed, and the closest pairs are merged. This is an iteratively process and it continues until all objects are in the same cluster, or a termination condition is reached. Based on the method used to determine the proximity between two clusters, agglomerative clustering techniques can be differentiated into single linkage, complete linkage and average linkage [2].

In *single linkage* clustering method, the distance between two clusters is computed as the distance between the two closest elements in different clusters (shortest inter-cluster distance). This technique has some disadvantages. First of all, it can produce long "string-like" clusters (a phenomenon known as chaining, as mentioned in [4]) and it is also influenced by the noise in the data sets.

In *complete linkage* method, the distance between two clusters is computed as the maximum distance between two objects in different clusters (furthest inter-clustering distance). Complete linkage algorithms can have, as a result, clusters with approximately the same number of objects.

Average linkage is a technique situated between simple linkage, which offers a clear separation between clusters and complete linkage, characterized by compact clusters. In this case, the distance between two clusters is the average of all distances between pairs of objects. In other words, it is the mean distance between elements of each cluster.

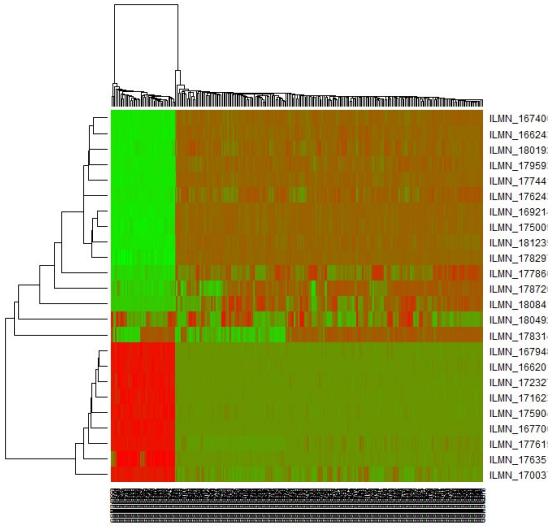
Due to the fact that it is a sequential algorithm, the hierarchical clustering approach has some disadvantages. First of all, even if a bad decision regarding the splitting criterion or the merging one has been made, the algorithm must continue until the end. Usually, bad decisions can be taken when there is noise in the datasets. This means that the method is sensitive to noise. Furthermore, it is characterized by high time and space complexity.

V. EXPERIMENTAL RESULTS

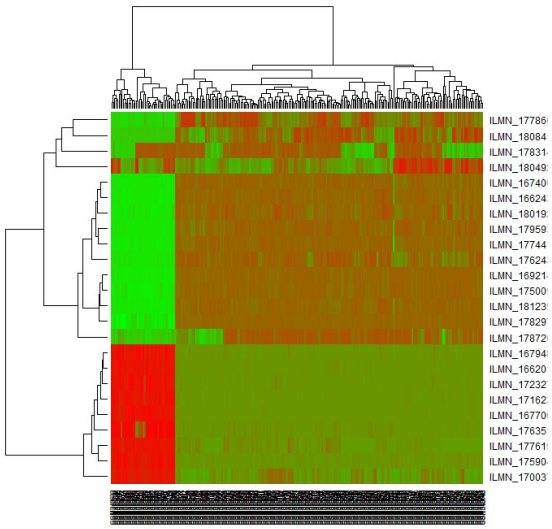
This section presents how different parameters for hierarchical clustering influence the final results. A data set of expression values from 272 primary bladder cancer specimens, obtained with Illumina BeadArray microarray technology, has been used in order to identify the genes and molecules that correlate with clinical outcomes in bladder cancer.

For analyzing these data, R statistical software and its expansion packages from Bioconductor projects were used. R is a language and environment for statistical computing and graphics, while Bioconductor provides tools for the analysis and comprehension of high-throughput genomic data [5].

For background correction an algorithm called "forcePositive" was applied, to force all expression values to be positive by adding an offset. To reduce the variation among microarrays, the intensity values for each array were normalized using quantile normalization. The values were log-2 transformed and filtered by standard deviation. Only genes for which the standard deviation was at least 2.5 were retained. Further on, the high level analysis was carried out. Different types of agglomerative hierarchical clustering methods were applied.



(a)



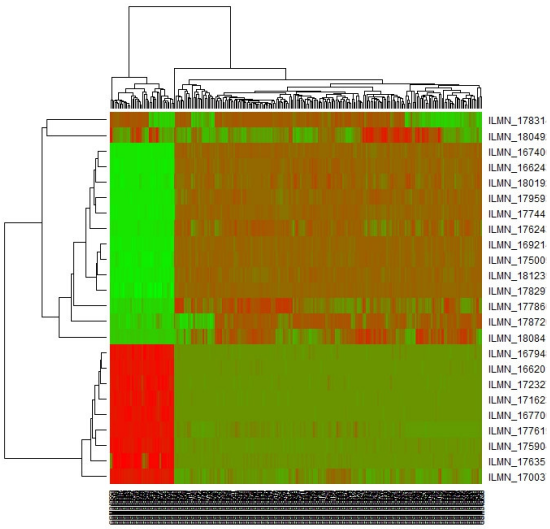
(b)

Figure 1. Heatmap for hierarchical clustering algorithm with Euclidian distance and single linkage (a) and complete linkage (b).

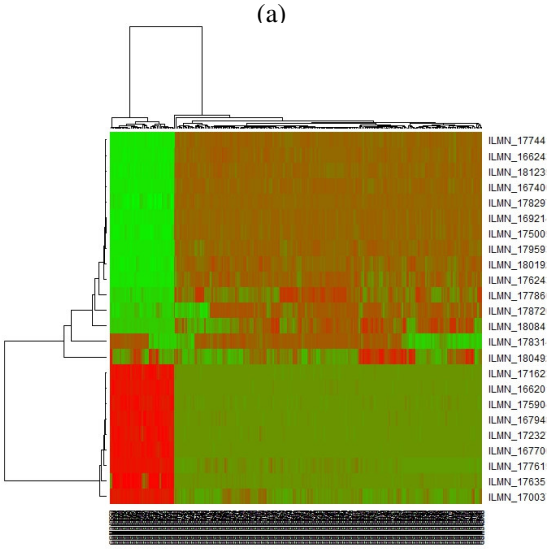
The output of these algorithms is represented as a heatmap. This is a graphical representation of the results, consisting of a matrix formed by blocks of different colors. The blocks' color represents the level of the gene expression value. For instance, over expressed genes are pictured with red, while under expressed genes are green. In this matrix, each column corresponds to a specimen and each row to a gene. The order of columns and rows is determined by two dendrograms, one for each dimension of the matrix, and it is in close relation with the linkage method.

First, for computing distance metrics, the Euclidian distance with single linkage was used (Figure. 1 (a)). As a result, long "string-like" clusters were obtained. At each step only a single gene was added to the existent clusters. In contrast, by using complete linkage (Figure 1 (b)) compact, well defined clusters were obtained.

Figure 2 shows the differences between the results for Euclidian distance (a) and Pearson's correlation coefficients (b), both with average linkage. One can observe that genes are arranged differently because of the algorithm used for computing distance metrics.



(a)



(b)

Figure 2. Heatmap for hierarchical clustering algorithm with Euclidian distance (a) and Pearson's correlation coefficients (b) with average linkage.

In all graphics, there can be observed that patients are grouped in two classes, but the best result was obtained by using Pearson's correlation coefficients and average linkage. Patients' separation in two groups is important, because it offers information on clinical outcomes. One of the most important clinical outcomes is progression, and by using this method, patients are grouped in two classes characterized by progression, or by no progression for their disease. The first one contains the patients for whom a number of genes are over-expressed, while the second one contains the patients that do not possess this type of genes. Genes are also grouped into two categories: over- expressed and under-expressed genes. In this case, the results are important because they help specialists to understand molecular mechanisms involved in bladder cancer progression.

By analyzing Figure 2, it can be seen that for the Pearson's correlation coefficients, genes are grouped in a more consistent manner. More precisely, all over-expressed genes and under-expressed genes are grouped in two big clusters. This observation is also true for the Euclidian distance, but in addition, for the Pearson's correlation coefficients case, all the genes that have not a precise behavior are also grouped together.

All these results can be further on used in the prediction of disease progression even before starting a treatment, enabling optimal personalized treatment strategies.

VI. CONCLUSIONS

Clustering of biological data is performed in order to obtain certain biological knowledge either about samples and systems or biological molecules, and the most optimal method is the one that provides accurate results and allows extraction of this useful information. In this case, hierarchical clustering with Pearson's coefficients and average linkage was the best option, because it provides a good separation between patients with disease progression and those with no progression, and also identifies related genes.

ACKNOWLEDGEMENTS

I would like to thank Alexadru Floares, Head of Artificial Intelligence Department, Cancer Institute of Cluj-Napoca, for his help and useful advice.

REFERENCES

[1] Dopazo, J., "Clustering- Class discovery in the post-genomic era", *Fundamentals of Data Mining in Geonomics and Proteomics*, W. Dubitzky, M. Granzow, and D. P. Berrar, (Eds.), New York: Springer Science + Business Media, pp. 123–148, 2006;

[2] Lee, J. K., *Statistical Bioinformatics- A guide for Life and Biomedical Science Researchers*, John Wiley & Sons, Inc., Hoboken, New Jersey, pp.89-127, 2010;

[3] Nisbet, R., Elder, J., Miner, G, *Handbook of Statistical Analysis and Data Mining- Application*, Elvier Inc., San Diego California, 2009;

[4] Simon, R. M., Korn, E. L., McShane, L. M., Radmacher, M. D., Wright, G. W., Zhao, Y., *Design and Analysis of DNA Microarray Investigations*, New York: Springer, pp. 121-157, 2003;

[5] Tuimala, J., *DNA Microarray Data Analysis Using Bioconductor*, CSC- IT Center for Science Ltd., 2008;

[6] Xie, Y., Wang X., Story M., "Statistical methods of background correction for Illumina BeadArray data", Bioinformatics, Vol. 25, No. 6, pp. 751-757, Oxford, England, 2009.

[7] Illumina Inc., information avalable: http://www.illumina.com/technology/beadarray_technology.ilmn.