Microarray Data Analysis and Management in Colorectal Cancer

Oscar García-Hernández[1], Guillermo López-Campos[1], Juan Pedro Sánchez[1], Rosa Blanco[1], Alejandro Romera-Lopez[2], Beatriz Perez-Villamil[2], and Fernando Martín-Sánchez[1]

[1] Medical Bioinformatics Department
Institute of Health 'Carlos III'
Ctra. Majadahonda-Pozuelo, km 2. 28220 Majadahonda, Madrid
{o.garcia,glopez,jpsanchez,rblanco,fmartin}@isciii.es
[2] Departamento de Oncología Médica
Hospital Clínico San Carlos
Martín Lagos s/n, 28040, Madrid
bperezvillamil.hcsc@salud.madrid.org

Abstract. The availability of microarray technologies has enabled biomedical researchers to explore expression levels of a complete genome simultaneously. The analysis of gene expression patterns can explain the biological basis of several pathological processes. Deepening in the understanding of the molecular processes underlying colorectal cancer might become of interest for the advance of its clinical management. This work presents the analysis of microarrays data using colon cancer samples in order to determine the differentially expressed genes underlying this disease process. The comparison of gene expression levels using a complete genome approach of tumor samples versus healthy controls allows the definition of a set of genes involved in the differentiation of both tissues. The analysis of these differentially expressed genes using Gene Ontology analysis permits the location of most prevalent processes that are altered during under this disease.

1 Introduction

The availability of new genomic based technologies for the massive screening and analysis of data has brought new scopes for the research and analysis of both clinical and biological data. Microarray technologies [1, 2] are one of the best examples of how these new technologies have changed the research. Microarray technology based experiments involve several steps and processes which can generate lots of information. From the initial step of manufacture with the annotation of probes until the final numerical data analysis several intermediate steps are done (sample processing, hybridization, scanning, etc ...). In each of those stages a great amount of data is generated and needs to be managed.

The most common application of microarray technologies since their origin has been the analysis of the gene expression under different conditions. Microar-

J.L. Oliveira et al. (Eds.): ISBMDA 2005, LNBI 3745, pp. 391–400, 2005.

rays offer the researchers the possibility of identifying and measuring simultaneously the complete set of genes expressed in a particular moment [3]. As results of this massive approach a new way of thinking the experiments has risen. In this new experimental approaches the objective is the analysis of complex systems as transcriptomes to elucidate the genes and the magnitude of the changes in their expression responsible for the adaptation of cells to the different conditions or even for complex diseases.

The evolution of the laboratory techniques for massive approaches that generates huge amounts of data has needed the evolution of bioinformatics to support them. In the case of microarray technology, bioinformatics is extremely interrelated with it. Almost all the processes in microarray experiments are supported by bioinformatics due to the huge amounts of data generated and managed. Therefore, microarray bioinformatics has been a very hot topic in recent years, specially the numerical data analysis related aspects.

The availability of public datasets as well as the increasing number of microarray publications has provided a substrate for the research in different methods to analyse gene expression data [4]. The bioinformatic analysis of gene expression studies is often related with the identification and explanation of the genes that are differentially expressed among the situations studied. There is an increasing number of techniques and algorithms and possible approaches that can be followed to achieve the final goal of understanding the underlying biology of the studied processes. These analysis usually involves several quality control steps like feature selection or outlier detection. Once this quality steps are done the data set is ready for undergoing a deeper analysis using both supervised and unsupervised techniques [5].

Cancer is a common disease nowadays in the most evolved societies. Due to the social, epidemiological and complexity cancer has become a paradigm in the application in research of new technologies for its study. Early in their development microarrays were used to study cancer processes [6]. Colon cancer is among the different types of cancer one of the most prevalent, affecting almost equally men and women. Therefore the study of the molecular basis of this type of cancer is very interesting from a clinical point of view as well as it is also an interesting problem for biology. The feasibility of analyzing with microarray technologies the gene expression profile of this disease to identify genes involved in the tumoral process is a way to deepen the available knowledge about the disease.

In this work we have used microarrays to study the gene expression profiles in colon cancer. The data generated was analysed to detect the genes differentially expressed among healthy tissue and tumour samples.

The paper is structured as follows. In section 2 we present the methodology used for microarray data preprocessing and filtering besides the information system developed for data storage and management. Section 3 presents the experimental results got from this study. Paper finishes with the conclusions got from the analysis and the future work that is going to be developed.

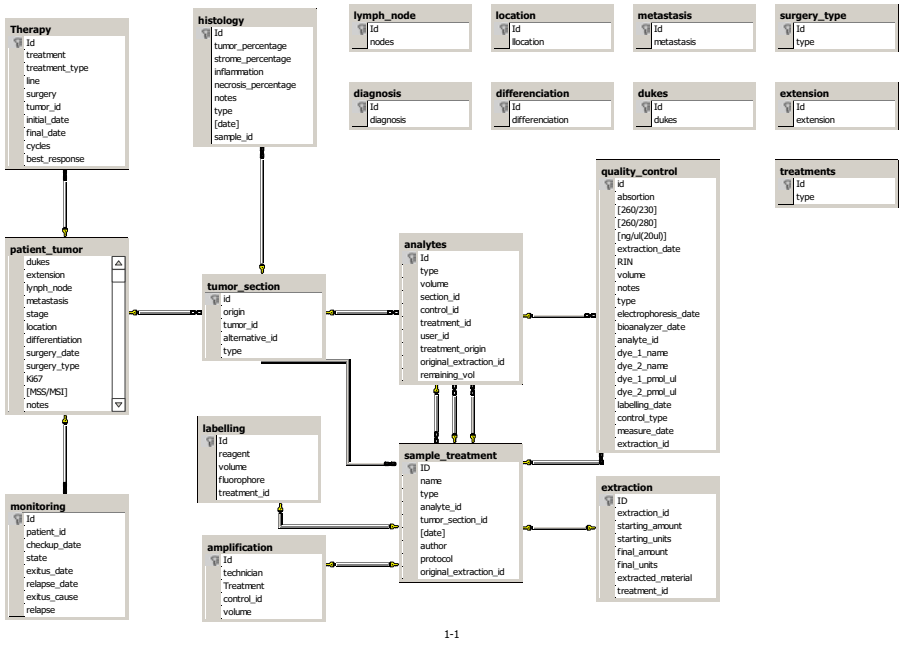


Fig. 1. Scheme of the patients, samples and experiments database

2 Microarray Data Analysis for Colorectal Cancer

Although, microarray data analysis is a wide and growing field, this study aims for detection of differentially expressed genes in colorectal cancer microarrays.

An information system is designed *ad-hoc* to correctly and easily store and manage the colorectal cancer microarray data. Furthermore, the database is extended for clinical annotation of the experiment.

2.1 Information System

An information system is designed and developed using SQL Server 2000 and MS Access 2003 technologies. This system is intended to manage information generated in the microarray expression analysis including patients and samples' information to be examined in the study.

The information system is composed by:

1. SQL Server databases
   (a) Managing information about the patients, samples and experiments –see Figures 1 and 2–.
   (b) Making easier the filter, transformation and normalization of microarray expression data.

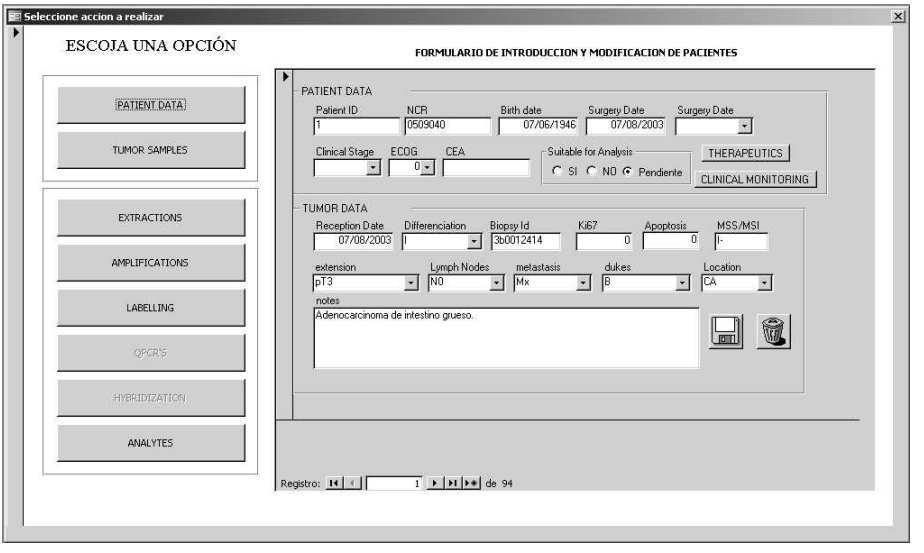


Fig. 2. Graphical user interface of the patients, samples and experiments database

2. Several Access clients which access to the SQL Server and allow the users to populate the databases and to browse, edit and search the data by means of the definition of different forms, views, queries and Visual Basic procedure defined in clients

Microarray expression data can be traced using the information system, which store information related to:

- Patient features. Under this topic is stored information related to the diagnosis and clinical aspects of the patients. These characteristics were selected in collaboration with clinical oncologists
- Therapies followed by the patients, including the chemotherapy drugs and surgeries suffered by the patients
- Biological samples. Biological description and annotation of the biopsies used as samples for the microarray experiments. This information contains aspects such as percentage of tumoral cells in the sample and some other parameters.
- Experimental processes applied to the samples. A series of tables are included to manage the information gathered during all the possible procedures done in the laboratory to the samples. As example of these procedures and reaction we may talk about histologies, PCR's, extractions, labellings, etc.
- Quality controls of both samples and processes done during the experiments
- Quantified expression data derived from the images generated by microarray experiments

Due to its user friendly environment and its facility to be learned and used, the MS Access 2003 technology is chosen to build the client applications. The

clinical environment where this tool was going to be used, required the development of an application with an easy and friendly interface. Both physicians and biological scientists involved in gathering the information and storing it in the database were already familiar with this tool.

As MS Access 2003 has a storage limit of 2 Gigabytes, and cause of the amount of data being produced within the microarray expression experiments (several Gigabytes), it is used in combination with SQL Server. With this configuration, we define all forms, queries and procedures in the Access clients while store the data in the SQL Server database.

2.2 Description of Microarray Data

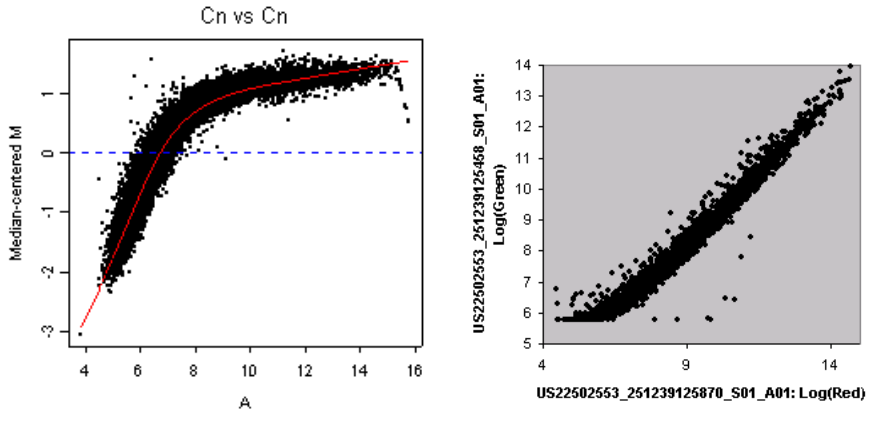
Microarray experiments are performed using Agilent's Whole Human Genome Oligo Microarray Kit. These oligonucleotide based arrays consist in a set of 43392 spots containing different probes for human genes as well as some internal positive and negative controls. RNA from 27 tumor samples is collected, amplified and labelled with a fluorescent label. 68 healthy colon RNA samples are collected and pooled together to act as healthy controls for the comparisons being amplified and labelled with a different fluorescent label than the one used for tumor samples. Both tumor and control samples are hybridized in pairs on the microarrays following the instructions of the microarray provider. Once the hybridization reaction is done the microarrays are washed and scanned with an Agilent scanner. The images generated by the scanner are then quantified using Agilent's Feature Extraction software using 'cookie cutter' segmentation algorithm.

2.3 Data Preprocessing

Microarray experiment data contain thousands of variables, several of which require a careful management due to their capacity to introduce noise into the data. Some examples of these variables are replicated probes or spot quantification processes. In some cases, these parameters are kept in mind by the quantification software. Therefore, a reliable and constant control on the setting of this software is necessary. Owing to these variability fonts, 3 preprocessing steps are followed in this study in order to polish and determine the data quality.

1. Controlling the experimental technique and reproducibility by checking labelling process for each experiment
2. Identifying required variables to perform the study and check the homogeneity in the variables and heterogeneity among the variables
3. Analyzing positive and negative controls

In this analysis, a labelling control is carried out every each 10 experiments and self-self control hybridizations are done. Subsequently, the numerical results of these hybridizations are plotted to determine whether the experimental technique is robust and reliable or not. The correlation analysis of these self-self



(a) (b)

Fig. 3. Representation of self-self control graphical analysis. (a) M-A plot of a control versus control hybridization before normalization. The graph shows an intensity dependant effect of the report that is solved with normalization. (b) Scatter plot of 2 non-normalized control samples from different days and dyes. It's clearly seen the correlation of both samples showing the good behaviour and overall reproducibility of the experiments

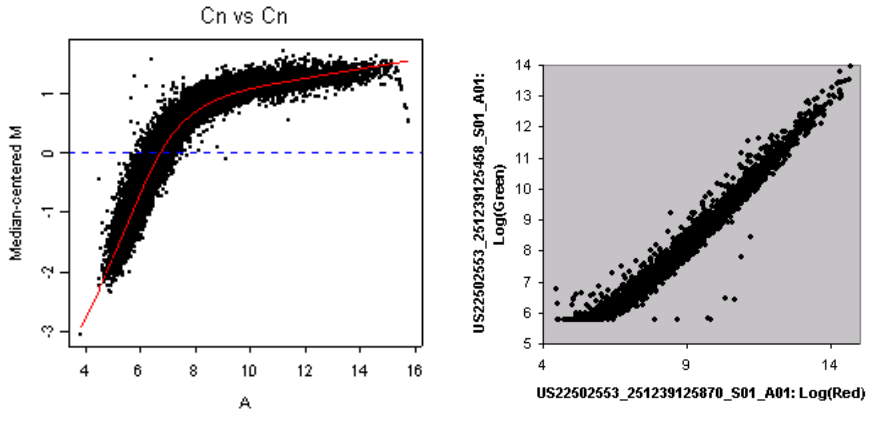
hybridizations is high, giving an overall correlation of 0.98 for the 3 microarray experiments.

At this step of the microarray data preprocessing, we remove the labelling background to obtain sharper images following literature recommendations [7]. Next, the ratio between the signal intensity in red and green channels are calculated to better management.

Data filtering. Data filtering is an important task in microarray analysis to obtain a reliable subset of genes. The flags provided by Agilent Feature Extraction Software are applied to gene filtering. So that, outliers and local background are removed. Then, genes with low intensity are filtered. Therefore, intensity of negative controls is studied. Spots whose signal intensity in red channel is below 16 and in green channel is below 54 are excluded. In this way, 31134 genes are attained.

Data transformation and normalization. In order to obtain smoother data, the $\log_2$ transformation is applied. Due to an observed effect of the signal intensity shown in self-self control M-A plots –see Figure 4–, microarray data are normalized by the *lowess smoother* function to remove this intensity effect. Further, a perfectly normalized data are reached to perform complex analysis.

Gene filtering. After all the spot filtering processes, a gene filtering step is carried out. Genes under the following set of restrictions are excluded from further analyses:



(a) (b)

Fig. 4. Representation of self-self control graphical analysis. (a) M-A plot of a control versus control hybridization before normalization. The graph shows an intensity dependant effect of the report that is solved with normalization. (b) Scatter plot of 2 non-normalized control samples from different days and dyes. It's clearly seen the correlation of both samples showing the good behaviour and overall reproducibility of the experiments

- genes with less than 20% of its values differs at least ±1.5 fold from median
- genes with more than 30% missing value

Then, the number of genes that passed filtering criteria was 3676.

3 Experimental Results

The analysis of microarray data coming from hybridizations of colorectal cancer patients is done using BRB-Array Tools from NCI [8]. The comparison between tumor and healthy samples is done using red and green channel comparison tool, instead of the simple class comparison. In this case, resulting log ratios are analyzed by the paired t-test. The null hypothesis means that the log ratio distribution equals zero. The significance level of the univariate t-test analysis is fixed to $p = 0.001$. The confidence level of false discovery rate is fixed to 95%. Other false discovery limits used in the analysis are the number of false positives and ratio of allowed false positives. In this study, the number of total false discoveries is set to 5 and the ratio of false positives to 0.01.

From these analysis we find that there are 2412 genes which discriminate between both studied classes (tumor versus healthy tissues) with $p = 0.001$. From the false discovery assessment we conclude that 2135 genes contain no more than 5 false positive with a probability of 95%. Additionally, in 2448 genes the false positive ratio is lower than 1%. It must be noted that only the 3676 genes selected after the data preprocessing are taken to account to this class comparison.

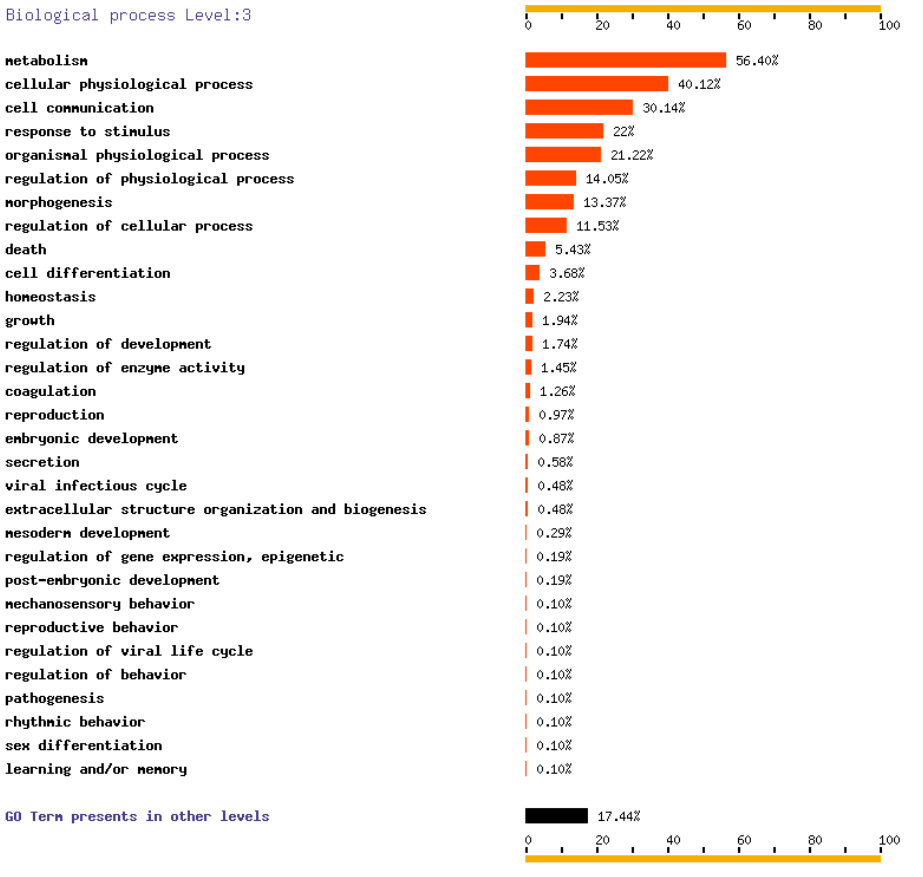


Fig. 5. Relation between genes and GO terms

The obtained genes are related to their biological categories and processes using the Gene Ontology (GO) annotation of genes. In order to discover the biological significance of the attained genes, FatiGO [9] is applied. The depth of the GO terms genes annotation is limited to the third level and the significance level is fixed to 0.001. A pitfall of this analysis is related to the lack of annotation for some of the probes used in the microarrays experiment. Therefore, almost half of the obtained genes are not included in the analysis.

Figure 5 shows the biological processes represented in GO. As it can be appreciated in the Figure, more than the 50% of the differentially expressed genes have a 'metabolism' annotation in GO. The 'cellular physiological process' annotation is related to more than the 40% of genes.

4 Conclusions and Future Work

The microarray technology allow researchers the exploration of new ways of study and analysis of diseases and therapies. Cancer has become in a common disease for the development and application of new methods and approaches.

The development of the information system described in this work allows clinicians and biologist an easy management of data related with clinical record and laboratory procedures. The use of such information system allowed the clinical annotation of the samples. This facilitates the incorporation of the clinical outcomes to the final data analysis and conclusions. Moreover, this information system facilitates the data filtering and processing. In the future the information system is going to be extended with new modules that allows the use of medical and bioinformatics standards such as MIAME [10] (Minimum Information About Microarray Experiments) and ICD (international Classification of Diseases)

In this work, a study of colorectal cancer microarrays is performed seeking for the differentially expressed set of genes. For this purpose, a group of filters are applied using Agilent's Feature Extraction and BRB-ArrayTools. In this way, the original 43392 spots in the microarrays are reduce to 2412 differentially expressed genes.

After the dimensionality reduction, the attained genes are related to GO terms in order to determinate their biological meaning. Resulting that more than the 50% of the genes have 'metabolism' GO term. This fact seems to be associated with a variation in the metabolic rate of the proliferating cells of the tumor.

As a future work line, a data filtering and gene selection based on function measure from machine learning field is developed. Moreover, a new classification of colorectal tumors type by their microarray expression profile is studied. For this purposes, techniques as unsupervised classification (clustering) and supervised classification are applied.

Acknowledgments

Analyses were performed using BRB-ArrayTools version 3.2 developed by Dr. Richard Simon and Amy Peng Lam.

References

1. Fodor, S., Read, J., Pirrung, M., Stryer, L., Lu, A., Solas, D.: Light-directed, spatially addressable parallel chemical synthesis. Science **251** (1991) 767–773
2. Pease, A., Solas, D., Sullivan, E., Cronin, M., Holmes, C., Fodor, S.: Light-generated oligonucleotide arrays for rapid DNA sequence analysis. Proceedings of the National Academy of Sciences **91** (1994) 5022–5026
3. Schena, M., Shalon, D., Davis, R., Brown, P.: Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science **20** (1995) 1008–1017
4. Lin, S., Johnson, K.: Methods of Microarray Data Analysis. Kluwer Academic Publishers (2000)
5. Brazma, A., Vilo, J.: Gene expression data analysis. Microbes and Infection **3** (2001) 823–829

6. DeRisi, J., Penland, L., Brown, P., Bittner, M., Meltzer, P., Ray, M., Chen, Y., Su, Y., Trent, J.: Use of a cDNA microarray to analyse gene expression patterns in human cancer. Nature Genetics **14** (1996) 457–460
7. Causton, H., Quackenbush, J., Brazma, A.: A Beginner's Guide. Microarray Gene Expression and Data Analysis. Blackwell Publishing (2003)
8. Simon, R., Lam, A.P., Ngan, M., Gibiansky, L., Shrabstein, P.: The BRB-ArrayTools development team (2005) http://linus.nci.nih.gov/BRB-ArrayTools.html.
9. Al-Shahrour, F., Díaz-Uriarte, R., Dopazo, J.: FatiGO: a web tool for finding significant associations of Gene Ontology terms to groups of genes. Bioinformatics **20** (2004) 578–580
10. Brazma A, Hingamp P, Quackenbush J, Sherlock G, Spellman P, Stoeckert C, Aach J, Ansorge W, Ball CA, Causton HC, Gaasterland T, Glenisson P, Holstege FC, Kim IF, Markowitz V, Matese JC, Parkinson H, Robinson A, Sarkans U, Schulze-Kremer S, Stewart J, Taylor R, Vilo J, Vingron M.: Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. Nature Genetics **29** (2001) 365–371