Circulation in Computer Science
Vol.2, No.1, pp: (5-8), Jan 2017
www.ccsarchive.org

Analysis of Microarray Data

Omar Moussati
Computer Science Department,
USTOMB, Oran, Algeria

Mohamed Benyettou
Computer Science Department,
USTOMB, Oran, Algeria

ABSTRACT
The computerized interpretation of biological information has taken a great interest in the scientific community, since it opens up very rich perspectives for the understanding of biological phenomena. These phenomena require collaboration between biologists, doctors, computer scientists, mathematicians and physicists. In this article we studied one of the most important subjects of bioinformatics, it is the biochip.We presented the various steps involved in the analysis of microarray data, Then we applied the KPPV method to the biochip data.

Keywords
Biochips , DNA chip, Microarray data .

1. INTRODUCTION
The biochip is a modern technique of analysis, it is widely used in several fields: in medicine, in pharmacology, agriculture and many other fields. With the rapid development of DNA chip technology in recent decades, it is now possible to simultaneously study the expression of thousands of genes. The expression data from this technology are observed and analyzed under different experimental conditions. These data are usually analyzed for various purposes. We will present in this work a history on the techniques and the methods which contributed to the appearance of this technique, in particular The PCR method that revolutionized research in several fields such as biology, Then we present the different steps of a DNA chip analysis, in order to apply a treatment on the data obtained from the biochips using the KNN algorithm.

2. THE PCR METHOD
The development of the polymerase chain reaction (PCR) technique by K. Mullis and colleagues in 1985 revolutionized molecular biology and molecular medicine [1].The polymerase chain reaction is a technique used to amplify with enzymes a specific region of DNA that is between two regions of known DNA sequence. Whereas in the past only very small amounts of a specific gene could be obtained, The PCR now allows to amplify even one copy of annoyance to a million copies in a few hours.

PCR techniques have become essential for many common procedures, Such as the cloning of specific DNA fragments, detection and identification of genes for diagnostic and forensic purposes, and research on gene expression patterns. More recently, PCR has allowed the exploration of new areas, such as control of the authenticity of foodstuffs, the presence of genetically modified DNA and microbiological contamination.

DNA contains complete genetic information that defines the structure and function of an organism. Three different processes are responsible for the transmission of genetic information:

- Replication.
- Transcription.
- Translation.

During replication, a double-stranded nucleic acid is duplicated to give identical copies. This process perpetuates genetic information. During transcription, a segment of DNA constituting a gene is read and transcribed into a single-stranded sequence of RNA. RNA moves from the nucleus to the cytoplasm.

Finally, during translation, the RNA sequence is translated into the amino acid sequence in the formation of the protein [2].DNA replication is the process on which the PCR is based, and is described below.

2.1. Principle of PCR
PCR is based on the mechanism of DNA replication: double-stranded DNA is unrolled in single-stranded DNA, then duplicated and "rewound". This technique includes the following repetitive cycles:

- Denaturation of DNA by high temperature fusion to convert double-stranded DNA to single-stranded DNA;
- Hybridization to the target DNA of two oligonucleotides used as primers;
- Extension of the DNA chain by addition of nucleotides from the primers using DNA polymerase as a catalyst in the presence of Mg2 + ions.

The oligonucleotides generally consist of relatively short sequences that are different from each other and complementary to the recognition sites flanking the target DNA segment to be amplified. The steps of denaturation of the matrix, primer hybridization and primer extension constitute a "cycle" in the polymerization chain reaction method.

In the final step of the PCR, a copy identical to that of the first is obtained.

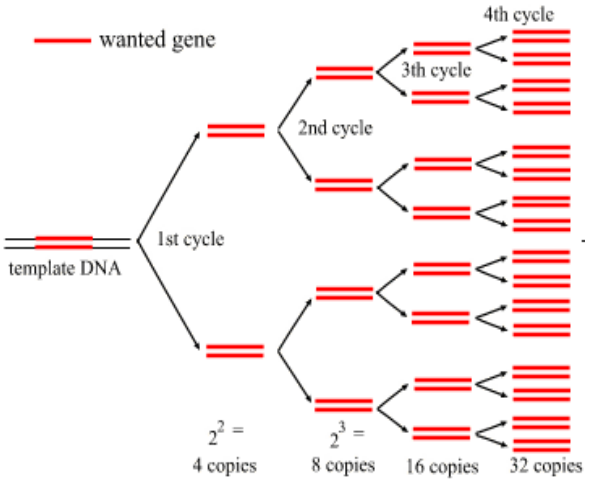


Figure 1. DNA amplification in PCR.

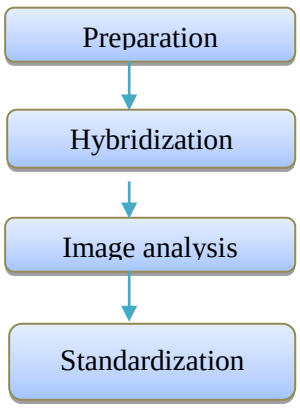
3. PRINCIPLE OF DNA CHIPS

The technology of DNA chips or microarrays, is currently experiencing an exceptional growth and attracts a great interest in the scientific community. This technology was developed in the early 1990s and allows the simultaneous measurement of the levels of expression of several thousand genes, or even an entire genome, in dozens of different conditions, physiological or pathological. The usefulness of this information is scientifically indisputable because the knowledge of the level of expression of a gene in these different situations constitutes an advance towards its function, But also to the screening of new molecules and the identification of new drugs and diagnostic tools. [5]

The functioning of the DNA chips rests on the principle of complementarity of the strands of the double helix of DNA and the property of hybridization between two complementary sequences of nucleic acids.

A DNA or RNA sequence can thus serve as a probe for capturing its complement (target) in a mixture of nucleic acids. A DNA microarray consists of DNA fragments immobilized on a solid support in an ordered manner. Each sequence location is carefully marked: the position (xi, yi) corresponds to the gene (i). A location is often referred to as a spot or probe. The hybridization of the chip with a biological sample which has been labeled with a radioactive or fluorescent substance makes it possible to quantify the set of targets it contains; the intensity of the emitted signal is proportional to the quantity of target genes it contains. [3] [4].

The different phases of a DNA chip analysis are shown in Figure 2.



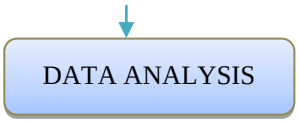


Figure 2. Steps of a DNA chip analysis.

3.1. Target Preparation and Hybridization

To compare expression levels in two biological samples or two conditions (reference and pathological), the first step consists in the preparation of the genome expressed in these two samples. It is a question of extracting the mRNAs from a biological sample to be analyzed, and the quality of the extraction is of course paramount for the success of the hybridization which will follow. Poor purification can lead to increased background noise on the blade. The second step is to label the two samples and then hybridize them using an oven and clean them using a washing station. The samples are labeled with fluorescent substances (Cy3 and Cy5), ie one culture is labeled with a green fluorochrome, while the second is labeled with a red fluorochrome. The hybridization is then carried out on a single chip (simple marking) or on two chips (double marking: one sample on each chip). The labeled DNAs are mixed (target) and placed on the DNA chip (probe). This process of hybridization is carried out in a fluidic station (furnace) to promote the links between complementary sequences [4]. The duration oscillates between 10 and 17 hours in liquid medium at 60 degrees, in fact at this temperature a single stranded DNA fragment or messenger RNA recognizes its complementary strand (cDNA) among thousands of others to form a DNA of Double strand (duplex or double helix). The step of cleaning or washing the chips is intended to remove unhybridized targets from the chip. The chip is washed several times so that only the perfectly matched strands remain on the blade.

3.2. Image Acquisition and Analysis

Following hybridization, a reading step of the chip makes it possible to identify the probes reacted with the sample tested. This reading is a key step [4]. Indeed, its quality significantly affects the accuracy of the data and therefore the relevance of the interpretations. The images are obtained by reading the chips on high-precision scanners, adapted to the markers used. The detection method combines two lasers to excite fluorochromes Cy3 and Cy5.

Two images are then obtained, the gray level of which represents the intensity of the fluorescence read. If gray levels are replaced by green levels for the first image and red levels for the second image, a false color image composed of spots ranging from green to red when one of the fluorophores dominates is obtained by superimposing them. Passing through the yellow (same intensity for the two fluorophores). Black symbolizes the absence of a signal. The intensity of the fluorescence signal for each pair (gene, spot) is proportional to the intensity of hybridization and therefore to the expression of the targeted gene (FIG. 2.2). The images are processed by analysis software which allows to measure the fluorescence of each spot on the slide (estimating the levels of expression for each of the genes present on the chip), but also to connect each probe to the corresponding annotation (Name of gene, number of cDNA used, sequence of oligonucleotide, etc.). Thus, for each spot, the intensity of each marker is calculated and then compared to the background noise.

3.3. Data Transformation

The ratios of the intensities of fluorescence in red and green are generally used to measure a variation in the expression of a gene between two conditions (reference and pathological, for example). Intensity data are rarely manipulated without transformation and the most commonly used transformation is that using the two-based logarithm. There are several reasons for this transformation. On the one hand, the variation of the logarithm of the intensities is less dependent on the magnitude of the intensities, and on the other hand, this transformation makes it possible to approach a symmetrical distribution and to obtain a better dispersion with fewer extreme values. The normalization consists in adjusting the overall intensity of the images acquired on each of the two red and green channels, so as to correct the systematic differences between the samples on the same slide, which do not represent biological variations between the samples and which tend to Unbalance the signal of one of the channels with respect to the other. This normalization procedure is defined by the reference genes. The reference genes on average must not change expression between two conditions. Normalization is carried out from all the probes present on the support to eliminate the differences between the different chips related to variations in starting quantity, marking or hybridization biases and variations in background noise [7].

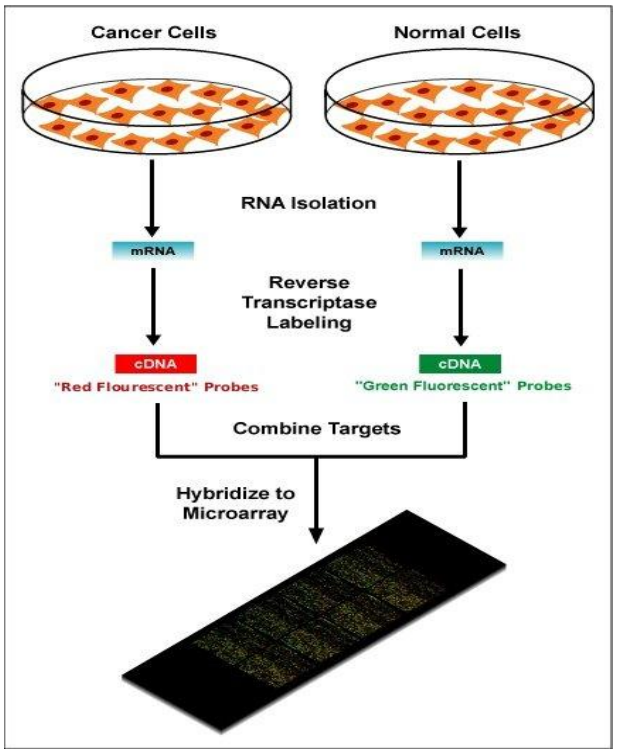


Figure 3. Image acquisition process.

Presentation of the DNA chip data: after the transformations described above, the data collected for the study of a given problem are grouped into a matrix with one line per pair (gene, probe) and one column per sample (table 1) Each value of mij is the measurement of the level of expression of the i-th gene in the j-th sample, where i = 1,. . . , M and j = 1,. . . , N [6].

Table 1. Gene expression matrix.

$Gene_{id}$	$Sample_1$	$Sample_2$	……	$Sample_N$
$Gene_1$	……	……	……	……
$Gene_2$	…..	……	……	……
…….	……	……	……	……
…….	……	……	……	……
$Gene_M$	m_{M1}	m_{M2}	……	m_{MN}

3.4. Standardization

Standardization is required to ensure that the observed differences in intensities are due to actual differences in expression and not to experimental artifacts. In the manufacture of DNA chips, there are many sources of variability. We can mention the amplification of the probes by the PCR technique and their positioning on the chip, probe / target hybridization, cleaning and drying of chips etc. The objective of standardization is to correct the systematic differences between measurements on the same chip that do not represent true biological variations. It allows the comparison of several replicas of the same experiment and focuses on the systematic errors, which help to over or under evaluate the measured values rather than on the stochastic errors. Before applying a logarithmic transformation, most of the measured intensities are small , the logarithmic transformation allows to refocus the distribution and make it symmetrical, which facilitates the use of statistics. Note that the logarithmic transformation based on 2 is the most used.

4. EXPERIMENTS AND RESULTS

We used the prostate cancer dataset, for a complete description of this dataset see: http://www.broad.mit.edu/cgibin/cancer/datasets.cgi

Under the heading «Gene Expression Correlates of Clinical Prostate Cancer Behavior".

Prostate cancer data contains nearly 12,600 genes and to better analyze this important number of genes, we use two statistical filtering methods: the Fisher test and the BW test.

After the reduction of the number of genes to The table below shows the results we found: 75 genes, we apply the KPPV method.

Table 2. Results obtained by the KPPV method.

KPPV	Prostate cancer		
	Classification rate	Number of genes	K
Fisher Test	80 %	11	2 ; 4
BW Test	91.42 %	09	4

In this study, we applied the Kppv method to prostate cancer data. The table shows that with a very small number of gene, we can identify whether a tissue is cancerous or not. But if we read the results well, we notice that the Kppv method is insufficient to make a final decision, so there are still improvements to be made.

5. CONCLUSION

In this study, we presented the PCR method which allowed developing the research in several fields including biology. Then we presented the different steps of a DNA chip analysis, such as target preparation and hybridization, image acquisition and analysis, and data transformation. We applied the Kppv method to analyze the biochip data.

We envisage in the future work to apply classification methods to better analyze the biochip data.

6. REFERENCES

[1] Saiki etal. (1985).Enzymatic amplification of ß-globin genomic sequences and restriction site analysis for diagnosis of sickle cell anemia. Science 230, 1350.

[2] Alberts et al. (1983). Molecular biology of the cell. Garland Publishing, Inc., New York.

[3] Hardin,J,et al robust measure of Correlation between two gens on amicroarray.BMC Bioinformatics 2007.

[4] E. M. Southern. DNA Arrays methods and protocols, chapter DNA Microarrays, pages 1–15. Humana Press, 2001.

[5] Golub et al.,. Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. Science, 286:531–537, 1999.

[6] Y.H. Yang, et al Speed.Normalization for cdna microarray data: a robust composite method addressing single and multiple slide systematic variation. Nucleic Acids Res, 30:1–12, 2002.

[7] Genome Resource Facility GRF, Microarray section, London School Of Hygiene and Tropical, Article technique, Medecine. 2006

[8] David M. Rocke et Blythe Durbin. A Model for Measurement Error for Gene Expression Arrays. Journal of Computational Biology, 8, 559–567, 2001. (55, 56)