Pol. J. Appl. Sci., **2015**, *1*, 23-26

# DNA PROCESSING BASED ON MICROARRAYS AND SEQUENCING

Igor Mäsiar, Michal Zábovský

*Institute of Information and Communication Technologies*
*University of Žilina, Žilina, Slovakia*

E-mail: michal.zabovsky@uniza.sk

**Abstract:** Since the early 1990s the evolution of the technologies has brought considerable change in the area of DNA processing. Although we can read and process DNA, the analysis of the collected data still remains a vast research area. This paper presents a brief overview of human DNA processing from an informatics (bioinformatics) point of view. The paper explains the basic structure of DNA and outlines the importance of DNA processing. The main part of the article describes the procedure of transformation and DNA processing. We will also discuss the historical development of the available data processing methods. We will look at the major differences in data sequencing and processing using the DNA chip. The article is concluded with an assumption of data processing.

**Key words:** DNA, Microarrays, DNA sequencing, Affymetrix

## DNA Characteristics

The human body is a complex system consisting of about 200 different types of more than $10^{13}$ cells, and we can identify over 20 different structure types in the context of those cells [1] [2].

DNA is shortcut for deoxyribonucleic acid. DNA and ribonucleic acid (RNA) are the nucleic acid contained in the cell nucleus. The role of RNA is the transcription of genetic information from DNA and the physical transfer to the place of translation into the final protein (an important feature for us provides only class mRNA).

DNA can very loosely be described as a kind of „architecture" or plan of cells formed from proteins. The gene is just a part of DNA, which must be transcribed into RNA form – which codes the certain information. The information bearer is part of a gene called exon.

DNA in cells is very similar for several individuals. In a particular cell a specific group of genes is always apparent – this fact is called gene expression. Gene expression is a complex process of expressing the genetic information contained in DNA, depending on several particular circumstances. Gene expression is characterized in the Central Dogma of Molecular Biology [3]. By virtue of the central dogma, genes are decoded to perform different functions, the best known of which is to synthesize proteins. This is done in a process that follows three stages:

1. Transcription,
2. Splicing,
3. Translation [4].

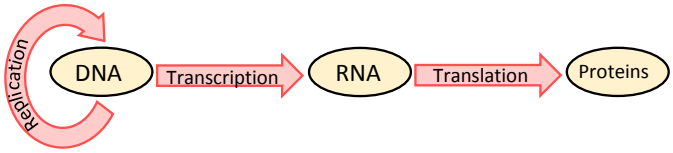


Fig. 1: The Central Dogma of Molecular Biology

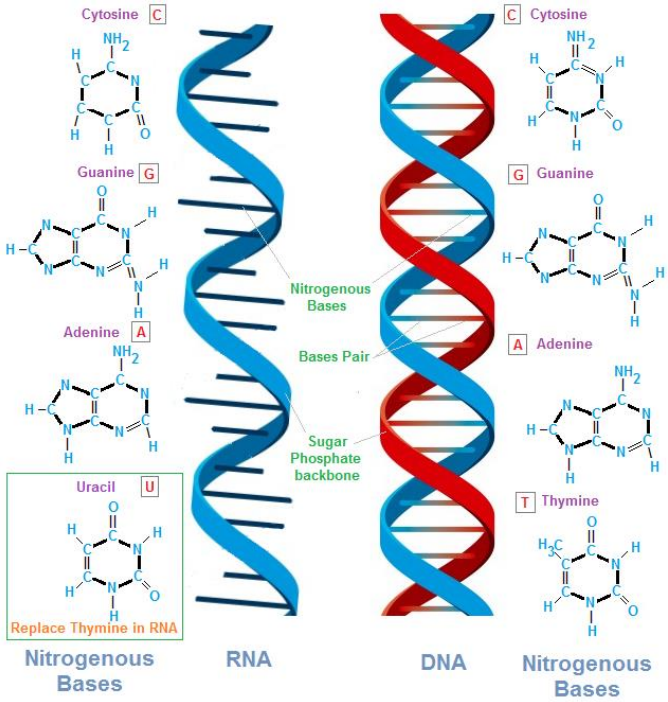


Fig. 2: DNA and RNA Comparison

There are many differences between DNA and RNA, but the basics are as follows [4]:

- DNA carries „code", or genetic information;
- DNA has a form of double helix(double-stranded), while RNA are usually single stranded;
- DNA consists also from nucleotides Thymidine (T) in addition to RNA, while RNA is formed also from molecules of Uracil (U), which are not included in DNA.

## Readable Information from Processed DNA

The ability to analyze and the subsequent explanations of information stored in DNA are very helpful for professionals. Doctors diagnose serious disease(s) by errors in DNA, or determine the likelihood of occurrence in the future. And we are talking only about the field of medicine – the use for forensic scientists is just one of the areas where we can use the information that carries DNA. What then can be detected from processed DNA? Theoretically, anything on a given subject. In practice, however, we encounter several problems:

- The complexity of the preparation and processing of DNA samples, and the possibility of errors during DNA preparation;
- Errors which arise in the processing of data and their repair;
- Identification of data – a huge data set.

There is an idea to create a system designed to process DNA sequences obtained by the patient and their subsequent identification. It could be accomplished thanks to the close cooperation between the University Science Park at University of Žilina and Jessenius Faculty of Medicine in Martin – Department of Clinical Biochemistry. In the case of a successful implementation it may be a useful tool for decision support for a doctor in determining the treatment of a patient's cancer.

## The Evolution of DNA Processing

A significant milestone in the understanding of DNA was Rosalind Franklin's (expert at x-ray diffraction) discovery in 1952. Her research led to the first images which revealed the exact structure of DNA [5]. The next important step was the development of DNA sequencing (called Sanger method) method by two-time NobelPrize winner Frederick Sanger, also known as the Sanger method [6]. Two years later another approach called Maxam-Gilbert DNA sequencing method was developed and published by Allan Maxam and Walter Gilbert. This method is based on the principles of nucleobase-specific partial chemical modification of DNA and the subsequent cleavage of the DNA backbone at sites adjacent to the modified nucleotides. [7].

The first signs of the automation of DNA sequencing can be found only in the early 1980s. Before the arrival of automation, DNA was sequenced by hand. It was a complicated and lengthy process, during which errors or omissions occurred very often. The first semi-automated commercialized DNA sequencer was introduced by GATC Biotech – German specialist in DNA and RNA sequencing. The principle was published in **DNA Sequencing with Direct Blotting Electrophoresis** [8].

The second generation of DNA sequencing was started in 1996 by Pål Nyrén and Mostafa Ronaghi at the Royal Institute of Technology in Stockholm. They introduce pyrosequencing in **Real-time DNA Sequencing Using Detection of Pyrophosphate Release** [9]. This generation of DNA sequencing was called next-generation sequencing (NGS). The evolution of DNA sequencing is heading to one molecule of DNA sequencing in real time, but today it is available in a more theoretical than practical form. A subtly different approach (or technology) associated with the processing of DNA is DNA microarrays. In the beginning of the 1990s technological progress allows „genomic revolution", this revolution enabled scientists to complete the sequences of a variety of organisms, including the culmination of the full draft sequence of the human genome [10]. In the late 1980s a team of scientists led by Stephen P.A. Fodor Affymetrix ®GeneChip technology (oligonucleotide microarray). In 1994 the company, Affymetrix, commenced the commercial sale of the ®GeneChip system for research use.

## DNA Chip (Microarrays) vs. DNA Sequencing

### DNA Microarrays

Hybridization has been used to identify genes in cellular DNA for more than 30 years now [5]. Microarrays are based on the same principle but differ in quantity. Whilst traditional hybridization techniques, such as Southern blot, can detect one gene at a time, microarrays are intended to do the same with thousands of genes in a single experiment.

The flowchart of a ®GeneChip System microarray experiment is shown at Fig. 3. Once the nucleic acid sample has been obtained, target amplification and labeling result in a labeled sample. The labeled sample is then injected into the probe array and allowed to hybridize overnight in the hybridization oven. The washing and staining of the probe array occur on the fluidics station, which can handle four probe arrays simultaneously. The probe array is then ready to be scanned in the Affymetrix GeneChip scanner, where the fluorescence intensity of each feature is read. Data output

includes an intensity measurement for each transcript or the detailed sequence or genotyping (SNP) information [11].

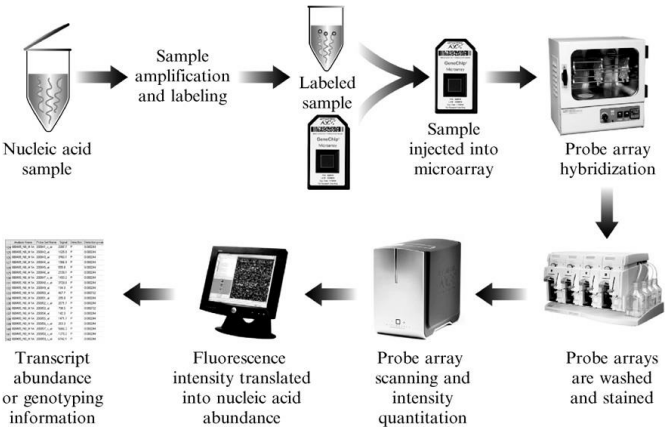


Fig. 3: Flowchart of a GeneChip System Microarray Experiment

When we have data, we can start the very complex process of their preparation and subsequent analysis. The „Output" of loaded chip with all the data is in the .CEL file. The treatment process should consist of the following steps:

1. **Verification of data quality**
   (a) RNA degradation plot, Relative Log Expression (RLE), ...
2. **Data normalization**
   (a) RMA normalization method, GCRMA, ...
   (b) Chip identification
3. **Phenodata – add/edit**
   (a) Define to which group of samples chip belongs
4. **Preprocessing**
   (a) Filter non-changing genes
5. **Analysis & Visualization**
   (a) Statistical testing
   (b) Clustering
   (c) Pathway analysis
   (d) Promoter analysis

An easy way to work with microarray data is to use the Chipster open-source platform for data analysis (*http:// chipster.csc.fi/*)

## DNA Sequencing

In electrophoresis, DNA to be sequenced is placed at one end of a gel-a slab of a gelatin-like substance. Electrodes are placed at either end of the gel and an electrical current is applied, causing the DNA molecules to move through the gel. Smaller molecules move through the gel more rapidly, so the DNA molecules become separated into different bands according to their size. The catch is that electrophoresis can only separate about 500 bases into clear bands-hence the need for chopping DNA up into small pieces in order to sequence it.

An automatic sequencing machine spits out what genome scientists call „raw" sequence. In a raw sequence, the reads or short DNA sequences are all jumbled together, like the pieces of a jigsaw puzzle in a just-opened box. Inevitably, a raw sequence also contains a few gaps, mistakes, and ambiguities. The process of polishing that raw sequence-transforming the fragmented rough draft into a long, continuous final product without breaks or errors-is called finishing. Finishing involves both assembly, in which individual reads are hooked together in the proper order, and a laborious process of double-checking and refining the sequence to eliminate mistakes and close gaps. Finishing often takes longer than the sequencing itself [12].

## Possibilities of Processing Acquired Data

A great deal of work in regard to the processing of DNA is spent on the conditioning and subsequent data analysis. Regardless of the method of data processing, one of the biggest problems is the identification of information hidden in the data. Several „genetic databases" are available which are a rich information source for identifying individual genes (Genome Trax). The system which would know based on the selected parameters browse the database and analyzed data to compare them with the data obtained from the DNA sample is one of the options for further work.

The crawling process given the implemented comparison algorithm and considering the fact that a huge set of data will require significant computing power, makes the task a suitable candidate for the use of parallelization.

Basic statistical data processing in conjunction with other information sources reveals the possibility of using a complex exploratory analysis, which will help to uncover hidden links and connections. Semi-automation of this analysis would require the creation of a comprehensive environment for a domain expert. The key role will by the implementation of appropriate knowledge base.

## Literature

[1] B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts, P. Walter. *Molecular Biology of the Cell*. Garland Science, 2007.
[2] H. Ibelgaufts. Cope - cytokines & cells online pathfinder encyclopaedia.
[3] F. Crick,. Central dogma of molecular biology. *Nature*, page 227, 1970.

[4] A. Sánchez and M. C. R. d. Villa. *A Tutorial Review of Microarray Data Analysis*. Universitat de Barcelona, 2008.

[5] P. Heidi Chial, C. Drovdlic, P. Maggie Koopman, P. D. Sarah Catherine Nelson, A. Spivey, P. D. Robin Smith and W. Communications. *Essentials of Genetics*. 2014.

[6] F. Sanger, A. Coulson. A rapid method for determining sequences in dna by primed synthesis with dna polymerase. *Journal of Molecular Biology*, 94(3):441–446, 1974.

[7] A. Maxam, W. Gilbert. A new method for sequencing dna. *Proceedings of the National Academy of Sciences of the United States of America*, 74(2):560–564, 1977.

[8] S. Beck, F. Pohl. Dna sequencing with direct blotting electrophoresis. *The EMBO Journal*, 3(12):2905–2909, 1984.

[9] M. Ronaghi, S. Karamohamed, B. Pettersson, M. Uhlén, P. Nyrén. Real-time dna sequencing using detection of pyrophosphate release. *Analytical Biochemistry*, 242:84–89, 1996.

[10] E. S. Lander, e. al. Initial sequencing and analysis of the human genome. *Nature*, 409:860–921, 2001.

[11] D.-W. D.D., W. J., T. E.Y. and M. C.G. The affymetrix genechip platform: An overview. *Methods in Enzymology - METH ENZYMOLOGY*, 410:3–28, 2006.

[12] J. Craig. *Genome News Network*. Venter Institute.

[13] J. Zvárová, e. al. *Metody molekulární biologie a bioinformatiky*. Nakladatelství Karolinum, Praha, 2012.