Deep sequencing – a new method and new requirements of gene expression analysis

Jan Czerniecki[1,2], **Sławomir Wołczyński**[3]

[1] Department of Biology and Pathology of Human Reproduction, Institute of Animal Reproduction and Food Research of Polish Academy of Sciences in Olsztyn

[2] Department of Cytobiochemistry, Institute of Biology, University of Bialystok

[3] Department of Reproduction and Gynecological Endocrinology, Medical University of Bialystok

Abstract. The determination of gene expression is a very common scientific method used in modern laboratory for a variety of applications. One of the most popular is the real time PCR, a quantitative modification of the classic PCR method where the increase of the amplify nucleic acid is examined cycle by cycle after every amplification step. The analysis of the PCR product during the amplification process allows to compare the initial amount of cDNA synthesized from isolated RNA and calculate the number of particular RNA copies present in examined material. In spite of obvious advantages of real time PCR there are also some inconveniences of this method. First of all, there is no possibility of analyzing more than one gene in a single reaction mixture. It is limited by the necessity of design and usage of different pairs of primers for each analyzed gene. Therefore, it is necessary to predict the cell, tissue or organism response for applied treatment, examined condition, etc. The development of microarray methods enables to overcome these problems and parallel analyze all known genes in the single sample at the same time. There is no need to predict which gene expression might be changed under studied conditions because the microarray data is a comprehensive pattern of the expression of all known genes, which probes are implemented on the microchip surface. Although the microarray data is an excellent method for gene expression comparison, the estimation of the extent of change fold is not very precise and usually is confirmed and determined by real time PCR with respect to selected genes. The method which combines the quantitative precision of real time PCR and the possibility to analyze broad spectrum of the genes is a deep sequencing method also called next generation sequencing. It is a new method developed for the analysis of the whole RNA isolated from a sample without the need to design primers and thus any knowledge of expressed genes sequence. The advantages of this method include the possibility of finding unexpected expression of completely unknown DNA fragments, alternative splicing variants of the genes and differences in DNA sequence. The deep sequencing provides an extremely large amount of information, much more than microarray data, and to analyse it new bioinformatics methods and tools especially designed for this purpose are required.

Introduction

Every cell of the same organism has a complete set of genes written in DNA. But morphology, structure and function of cells of the body of

ISBN 978–83–7431–296–7 ISSN 0860-150X

multicellular organism are very diverse. For instance human body has four main types of tissue and numerous types of cells [1]. This diversity is created during the differentiation process occurring in fetal development from one zygote formed from two gametes. All these changes are controlled by activation and silencing of particular genes expression according to the local requirements. Despite having a complete genetic information the fully differentiated cells use just part of it and transcribe into ribonucleic acids. Moreover, part of genetic information even in mature cells is activated under specific circumstances as a response to external signals of chemical or physical nature [2]. The pattern of active genes is similar in the same type of intact cells under the same condition. However gene expression of tumors is changed in comparison to their origin cells [3]. The process of carcinogenesis is characterized by specific expression profiles during different stages leading to the transformation of the intact cell into a cancer cell [4]. It concerns especially genes responsible for cell cycle regulation. Determination the sequence of transformation process events and its key points is very important in cancer research.

The eukaryotic genome organization and the process of gene expression

There is less than thirty thousands genes in the human genome [5] but the number of known proteins is much higher [6]. It is obvious that the rule "one gene one protein" is false. The eukaryotic genes are divided into coding parts called exons and non-coding parts – introns. Before the protein synthesis introns are removed from mRNA during RNA splicing which is the RNA maturation process. The pattern of mRNA maturation is not the same in every cell and tissue [7]. Often this process has alternative character and in this way one gene can encode different proteins with different structure and properties. The final protein can be modified in the process of posttranscriptional modification what is the source of additional increase in the number of one gene product diversity.

The knowledge of the genome sequences and gene expression of many species, including alternative splicing processes, has increased considerably in the last decade. Therefore the modern microarrays contain probes allowing for analysis of different variants of the mRNA [8]. However it is always possible that under specific circumstances in the case of particular organism or tissue the expression profiles of the gene (even well known) can be changed and a new type of the protein can appear due to unknown splicing manner.

Part of the expressed genetic information is not translated into pro-

teins but is important as a regulatory factors responsible for the genetic information expression. There are many types of miRNAs or siRNAs which can effect target transcripts of messenger RNA and activate or inhibit their translation. Some investigators suggest that miRNA represents 1% of human genome and regulates about 10% of translated proteins [9]. Probably we still do not know the entire active (transcribed into RNA) genetic information written in DNA even if the whole sequence is already known like in the case of human species.

Moreover, it is necessary to be aware that the concept of one species genome is just a theoretical entity and it does not exist in the reality, even though the internet databases contains thousands sequences of genes, both as a genomic DNA or cDNA library for many species. Every specimen of the species (excluding twins or clones) has a specific set of its own genes responsible for its individual distinctiveness. Children are not exactly the same like parents, grandparents or any relatives. It means that the genome of each individual is unique. Of course the similarity of the genome is really high. For example, human DNA sequence in the whole population is the same in 99.9% but the remaining 0.1% of diversity is enough to create all observed differences between people. There are a lot of single nucleotides polymorphisms (SNPs) and mutations in every population of the organisms belonging to the same species which makes that transcriptome of each specimen different. These individual properties of the expressed genetic material can be sometimes responsible for the subtle differences in physiology and lead to an individual response to the same treatment [10]. In the last few years the idea of the personally dedicated drugs according to patient's genetic specificity have become more and more popular. It indicates that in spite of significant similarity within the same species the occurring genome diversity can not be ignored. Moreover, eukaryotic DNA contains a lot of highly repeated fragments which are not translated into proteins. Previously it was thought that this part of genetic information does not play any role. It was even called "junk DNA". But some results suggest that at least part of the "junk DNA" is expressed into RNA and plays regulatory function of the processes transcription and translation [11]. It means that the transcriptome is far more complex than previously thought.

The limitation of the microarray and real time PCR method

The methods of gene expression analysis used so far in the laboratory required at least the basic knowledge of the analyzed material structure. The order of nucleotides at the beginning and the end of the amplified

sequence is necessary to design primers for real-time PCR. There is no possibility of synthesizing oligoprobes for microarray analysis of completely unknown genes without their prior sequencing. These problems are overcome by the use in gene expression studies new method developed recently – the next generation sequencing. Innovative approach to the sequencing allows to establish the sequence of all transcripts in the sample in a single run: determine SNPs, find new transcript isoforms, identify regular RNAs, characterize intron-exon junction and determine quantitative relation between the transcripts [12].

Classical sequencing versus next generation sequencing

In classical Sanger sequencing the method developed in the middle of 70's the determination is made using 2'-3'-dideoxynucleotides triphosphates (ddNTPs), molecules that differ from deoxynucleotides by the having a hydrogen atom attached to the 3' carbon rather than an OH group. Incorporation of such kind of molecule into newly synthesized complementary strand of DNA leads to termination of this process because without the OH group there is no possibility of forming a phosphodiester bound with the next nucleotide. The presence of a low concentration of four types of dideoxynucleotides in the reaction mixture altogether with higher concentration of deoxynucleotides results in the synthesis of the mix of randomly terminated double stranded DNA which differs in the length of one nucleotide. This mix is separated by gel or capillary electrophoresis according to the length. While four separated reaction mixtures contain one type of ddNTP (separation by gel electrophoresis) or all of them are differentially fluorescently labeled (capillary electrophoresis with fluorescent detector) it is possible to identify the terminated nucleotide and thereby the sequence of analyzed fragment. For successful sequencing with the Sanger method it is necessary to separate analysed fragments of nucleic acids because even with differentially fluorescent labeling of ddNTPs it is possible to determine only one sequence of limited length in one reaction mixture [13].

The situation is different for the next generation sequencing. Despite of the existence few commercially available platforms for the next generation sequencing of total transcriptome which are differ in details there are some similarities between them. In short, the total RNA is purified and randomly fragmented into pieces of required length and special short fragments called adapters are ligated on both ends of every analyzed RNA fragment. The presence of adapters with a known sequence allows for the synthesis

of cDNA and quantitative amplification of all of the fragments of RNA isolated during the sample preparation. Finally, the amplified products are immobilized on the surface of a special chip as an aggregation of identical clones. Contrary to the classical Sanger method the immobilized fragments are sequenced during elongation of second complementary strand of DNA. Moreover, the analyzer can determine the number of copies of the same fragments what is the approximate equivalent of the quantitative analysis of the real time PCR method. The final result of this analysis is a complete sequence of the whole expressed genetic material combined with the information concerning quantitative relation between the number of the copies of particular transcripts [14].

The new requirements of molecular research

The next generation sequencing is a very powerful tool in molecular biology and it has a great potential. But its successful use required development new techniques of data processing and analysis [15]. First of all the deep sequencing generates a massive amount of data, much more than any other laboratory method used so far. The microarray analysis has already forced the development of new statistical method because of the non-typical structure of the data, small number of considered attributes and hundreds (or even thousands) of analyzed cases: genes and spots. The amount of data derived using next generation sequencing is bigger by several orders of magnitude in comparison to microarray analysis. Besides the quantitative comparison of the analyzed genes (similar to microarray analysis) it is necessary to compare the determined sequences in order to identify the expressed material: the type of the gene, presence of mutations, polymorphisms and alternative splicing. It needs a fast progress in creation of complex sequence databases of as much as possible species. Without that any analysis is impossible. It is a crucial thing to know the typical (the most common) sequences of the genes if we want to find any abnormalities.

On the other hand, new data analysis techniques are needed to select among thousands of genes in the genome which are important in physiological responses to particular treatment or specific condition. Several programs dedicated to the analysis of biochemical pathways and ontologies already exist and are used for microarray analysis [16]. But the use of such programs and databases is becoming more and more complicated. In modern molecular biology experiments very often the most complicated work starts after the end of laboratory work and requires knowledge of advanced stati-

stical methods. Many laboratory scientists are facing serious problems with the use of increasingly complex mathematical models and statistical tools. Therefore, they increasingly require the assistance of bioinformatic specialists. Some specialized centers, usually associated with equipment suppliers, were established recently and they offer support in professional analysis of molecular data. The increase in demand for specialists fully conversant with both molecular biology and advanced mathematics will be observed in the near future as a response to the needs of modern science.

Conclusions

The progress in molecular biology research required not only the development of new laboratory methods but also the development of new bioinformatics tools and statistical methods for handling of the growing number collected data. Very often, the end of laboratory work and data gathering is just the beginning of a time consuming process of their selection and statistical analysis. This would not be possible without computer programs which are able to process massive number of data. The existence of internet databases gives an opportunity of access to the constantly growing number of data sets. All these databases will probably grow rapidly in the near future as a response to the increasing number of carried out next generation sequencing experiments. But this constant progress is not be possible without the growth of computing power and the development of new scientific computer programs.

REFERENCES

[1] Stevens A., Histologia człowieka, Wydawnictwo Lekarskie PZWL, Warszawa, 2000.

[2] Moran J. L., Li Y., Hill A. A., et al., Gene expression changes during mouse skeletal myoblast differentiation revealed by transcriptional profiling, Physiological Genomics, 10, pp. 103–111, 2002.

[3] Bignotti E., Tassi R. A., Calza S., et al., Differential gene expression profiles between tumor biopsies and short-term primary cultures of ovarian serous carcinomas: Identification of novel molecular biomarkers for early diagnosis and therapy Gynecologic Oncology, 103 (2), pp. 405–416, 2006.

[4] Ho A., Dowdy S. F., Regulation of G1 cell-cycle progression by oncogenes and tumor suppressor genes Current Opinion in Genetics & Development, 12 (1), pp. 47–52, 2002.

[5] Harrison P. M., Kumar A., Lang N., et al., A question of size: the eukaryotic proteome and the problems in defining it, Nucleic Acids Research, 30 (5), pp. 1083–1090, 2002.

[6] Lamesch P., et al., hORFeome v3.1: A resource of human open reading frames representing over 10,000 human genes, Genomics, 89 (3), pp. 307–315, 2007.
[7] Kalsotra A., Cooper T. A., Functional consequences of developmentally regulated alternative splicing, Nature reviews, Genetics, 12 (10), pp. 715–729, 2011.
[8] Relógio A., Ben-Dov C., Baum M., et al., Alternative splicing microarrays reveal functional expression of neuron-specific regulators in Hodgkin lymphoma cells, The Journal of Biological Chemistry, 280 (6), pp. 4779–4784, 2005.
[9] John B., Enright A. J., Aravin A., et al., Human MicroRNA targets, PLoS Biology, 2 (11), pp. 363, 2004.
[10] den Hoed M., Smeets A. J., Veldhorst M. A., et al., SNP analyses of postprandial responses in (an)orexigenic hormones and feelings of hunger reveal long-term physiological adaptations to facilitate homeostasis, International Journal of Obesity, 32 (12), pp. 1790–1798, 2008.
[11] Hultén M. A., Stacey M., Armstrong S. J., Does junk DNA regulate gene expression in humans?, Clinical Molecular Pathology, 48 (3), pp. 118–123, 1995.
[12] Voelkerding K. V., Dames S. A., Durtschi J. D., Next-Generation Sequencing: From Basic Research to Diagnostics, Clinical Chemistry, 55 (4), pp. 641–658, 2009.
[13] Sanger F., Nicklen S., Coulson A. R., DNA sequencing with chain-terminating inhibitors, Proceedings National Academy of Sciences USA, 74, pp. 5463–5467, 1977.
[14] Carninci P., Constructing the landscape of the mammalian transcriptome, The Journal of Experimental Biology, 210, pp. 1497–1506, 2007.
[15] Quackenbush J., Extracting biology from high-dimensional biological data, The Journal of Experimental Biology, 210, pp. 1507–1517, 2007.
[16] Stępniak P., Handschuh L., Figlerowicz M., DNA microarray data analysis, Biotechnologia, 4 (83), pp. 68–87, 2008.