4. GLOBAL APPROACH TO BIOMEDICINE: FUNCTIONAL GENOMICS AND PROTEOMICS

Krešimir Pavelić, Marijeta Kralj, Sandra Kraljević, Mirela Sedić

Division of Molecular Medicine, "Rudjer Boskovic" Institute, Zagreb, Croatia

4.1 Abstract

Functional genomics (transcriptomics and proteomics) is a global, systematic and comprehensive approach to identification and description of the processes and pathways involved in the normal and abnormal physiological states. The most applied methods of functional genomics today are DNA microarrays and proteomics methods, primarily two-dimensional gel electrophoresis coupled with mass spectrometry. Up to date interesting research have been carried out, representing the milestones for future implementation of functional genomics/proteomics in biomedicine. Still, further systematic examination of differentially regulated genes and proteins in tissues and fluids in healthy vs. diseased subjects will be required. However, high-throughput technologies reflect biological fluctuations and methodological errors. Large amount of such different data challenges the performance and capacity of statistical tools and softwares available at the moment. Yet, further major developments in this field are pending and the intellectual investment will certainly result in clinical advances.

4.2 Introduction to functional genomics

As an emerging discipline, functional genomics has many different definitions, which depend on the research area. Functional genomics (transcriptomics and proteomics) is a global, systematic and comprehensive approach to identification and description of the processes and pathways involved in the normal and abnormal states. Why is it such an important experimental approach nowadays? It is estimated that approximately 30% of the open reading frames in a fully sequenced organism have unknown function at the biochemical level and are unrelated to any known gene. This is why recently the interest of researchers has shifted from genome mapping and sequencing to determination of genome function by using the functional genomics approach (Figure 4.1.).

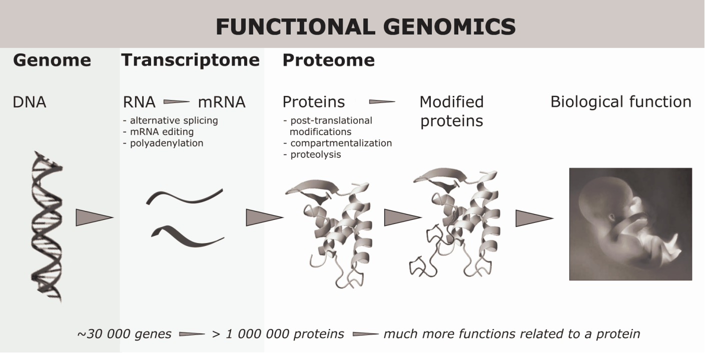

Figure 4.1. *A single gene can give rise to multiple gene products. RNA can be alternatively spliced or edited to form mature mRNA. Besides, proteins are regulated by additional mechanisms such as posttranslational modifications, compartmentalization and proteolysis. Finally, biological function is determined by the complexity of these processes.*

Techniques of functional genomics include methods for gene expression profiling at the transcript level (differential display, expressed sequence tags, serial analysis of gene expression and DNA microarrays) as well as methods for proteome analysis [Celis *et al*, 2000]. Due to recent technological advances in fabrication of very precise high-throughput instruments, a complex functional genomics approach has become possible. Processing large quantity of experimental data requires powerful information systems – bioinformatics, which encompasses development of new computational methods and application of those methods to solve biological problems. Bioinformatics has also a large service component in which computational resources such as databases are operated for the benefit of the research community. It also finds its implications in other aspects of biomedical research involved in functional genomics, such as laboratory information management systems, medical record systems as well as documentation of clinical trial results for regulatory agencies [Bogusky and McIntosh, 2003].

All those high-throughput experimental procedures enabled and accelerated new and important basic discoveries, especially in the field of molecular medicine. The ultimate aim is, however, to bring these discoveries closer to clinics and therefore substantially improve daily clinical practice.

The benefits of functional (integrative) genomics approach and applications to biomedicine are likely to be: a) the understanding of physiological or pathophysiological processes, b) identification of genes/proteins which can be used for screening, diagnosis or monitoring disease severity and c) discovery of novel genes/proteins suitable for therapeutic manipulation.

The most applied methods of functional genomics today are DNA microarrays and proteomics methods, primarily two-dimensional electrophoresis (2-DE) coupled with mass spectrometry (MS).

4.3 Introduction to DNA arrays

DNA microarrays may be defined as miniaturized, systematic immobilization of nucleic acid fragments derived from individual genes on a solid support, which by specific hybridization, enables the simultaneous analysis of thousands of genes in parallel [Dudda-Subramanya *et al*, 2003; Lockhart and Winzeler, 2000]. Microarray technique is based on new technology in which DNA probes are robotically spotted on miniature slides. The length of probes varies from kilobases, for standard cDNA arrays used for RNA expression analysis, to tens of basis in oligonucleotide arrays for both RNA expression and DNA sequence analysis [Dudda-Subramanya *et al*, 2003].

Nylon DNA arrays, often termed macroarrays are nylon-membrane based cDNA arrays for broad-scale expression profiling. They continue to be used because of their accessibility and flexibility. This is not only due to somewhat expensive use of microarrays but also to the fact that nylon DNA arrays offer a promising alternative in a number of situations, particularly when small amounts of sample are available or when specific pathways are being studied, which is itself significant for a number of research and clinical applications.

There are several different implementations of the DNA microarray principle for expression measurement. Microarrays are the most promising in terms of throughput and should eventually allow simultaneous measurement of expression of the whole set of genes. DNA microarrays have already had various applications: messenger RNA expression profiling for improved disease classification; genotyping of polymorphisms affecting disease susceptibility;

identification of genetic lesions within malignancies; design and discovery of new therapeutic drugs and sequencing of DNA [Celis *et al*, 2000].

4.4 DNA arrays in transcriptomics studies

DNA microarrays are the preferred and wide-accepted method for transcriptomic research as most disease processes are accompanied not only by characteristic macroscopic or histological changes, but also by systematic changes in gene expression patterns. For some pathological processes such as cancer, inappropriate gene expression is a fundamental aspect of pathogenesis. For other pathological processes, the gene expression programs, both in cells directly affected by a disease and in healthy cells responding to the local and systemic effects of a disease, can provide a detailed molecular picture of the pathogenic process [Diehn *et al*, 2000].

When used for documentation of gene expression at a genome-wide scale, microarray-based transcriptional profiling allows the identification of a set of genes that defines differential biological states (Figure 4.2.). This is a crucial step in the development of novel approaches for complete diagnosis of a disease. Molecular classification of a disease combined with the ability to place a certain patient into a specific genetic subtype, holds a promise of a better comprehension of a disease and thus development of individualized health care delivery. Microarray analysis can help in prediction of disease outcomes and prognosis.

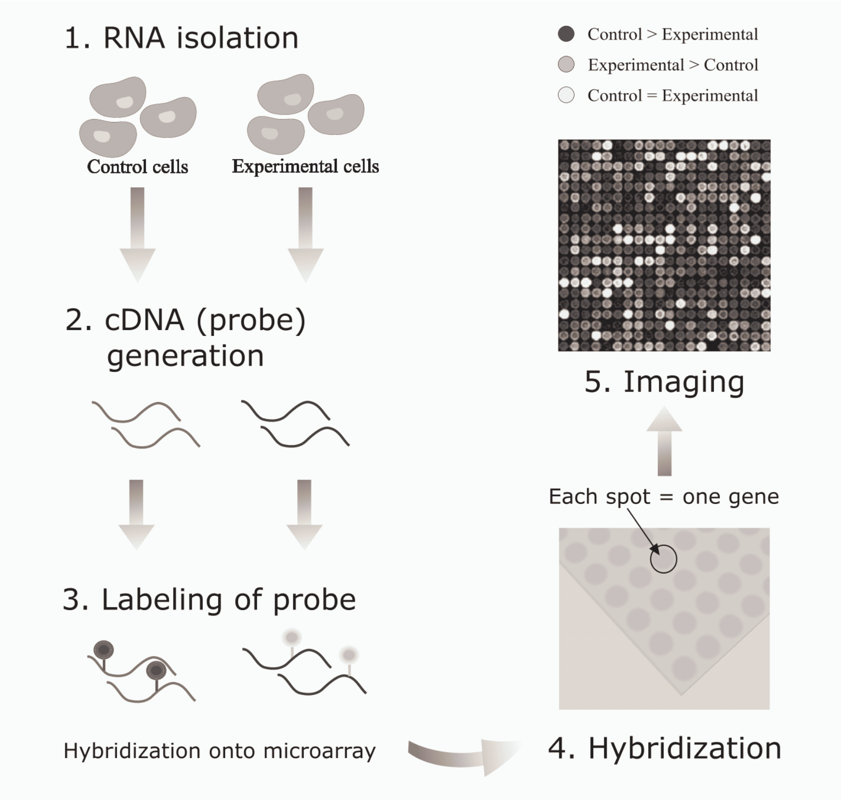

Figure 4.2. *Comparative gene expression analysis by using DNA microarray.*

Microarrays are also being applied increasingly to mutation analysis (polymorphisms analysis) by minisequencing. Single nucleotide polymorphisms (SNPs) are the commonest source of mutation in man and can be used as markers in whole genome linkage analysis of families or in association studies of individuals in a population. Single base differences between human genomes, or polymorphisms underlie differences in susceptibility to or protection from a host of diseases. Scientists believe SNP maps will help them identify the multiple genes associated with complex diseases such as cancer, diabetes, vascular disease, and some forms of mental illness. These relationships are difficult to establish with conventional gene-hunting methods because a single altered gene may make only a small contribution to the disease.

For instance, when applied to pregnancy research, the DNA array technique is extremely powerful. In that regard, interesting papers were published on newly discovered cDNAs/proteins which play a role in various stages of placentation mostly using mouse models and purified cell culture systems.

4.5 Proteomics

The term proteome was first introduced to describe the entire PROTEin complement of a given genOME, i.e. the complete set of proteins, which are expressed by the entire genome. Since the proteome is quite dynamic and it changes along with the development of an organism and with any alterations in the environment, it can be referred to as the array of proteins expressed in a biological compartment, such as cell, tissue or organ, under particular environmental circumstances at a particular time. Proteomics is a powerful tool for examining differential protein expression comparing hundreds of proteins simultaneously.

Majority of our cells contain the same genome regardless of the cell type, stage of the development or environmental conditions, unlike the proteome, that varies significantly under these diverse circumstances due to different patterns of gene expression and protein modification. There are more proteins in the proteome in comparison with the genes in a genome and it has been estimated that the human proteome is at least an order of magnitude more complex than the human genome, since it is assessed that there might be as many as a million human proteins. For example, in spite of the same genome, the same growth factor may have diverse signaling pathways in different cell types or cell states, implying that the study of the genome exclusively could not clarify the molecular mechanisms of diseases, aging and cell response to external stimuli [Godovac-Zimmermann and Brown, 2001].

It has been postulated that the average number of protein forms per gene is one or two in bacteria, three in yeast and three or more in humans. Proteomics deals with the characterization, identification and quantitative analysis of proteins in cells, tissues and body fluids. It is particularly suitable for the analysis of some body fluids, such as serum and urine that are deprived of mRNA, and thus cannot be studied by mRNA analysis.

Proteomic studies also include the analyses of the variations in the protein expression levels in cells or tissues under two different conditions and identification of the main proteins likely to have important functional roles in the cells' response to those conditions, for example, healthy versus diseased tissue or cells. A protein found only in a diseased sample may represent a useful drug target or diagnostic marker. These proteomic comparative approaches are analogous to microarray experiments, which examine whether genes are turned on or off under diverse conditions.

Furthermore, proteomics studies the numerous possible interactions among proteins as well as the molecular composition of the particular cellular structures (organelles). Knowing exactly which proteins interact with one another could help determining, for example, components of a particular enzymatic pathway.

Protein modifications not obvious from the DNA sequence, such as isoforms and post-translational modifications (e.g. phosphorylation and glycosylation) can be determined solely by the proteomic studies. It is assessed that approximately 200 diverse types of post-translational protein modifications occur.

One aspect of proteomics studies is directed towards determining the subcellular localization of proteins in order to construct an overall three-dimensional protein map of the cell, which could give an insight into the regulation of protein function.

Taken together, the scope of proteomics is quite broad and its practical application goes far beyond merely analyzing large numbers of proteins in complex mixtures [Graves and Haystead, 2002].

4.6 Short overview of proteomics methods

The method of polymerase chain reaction (PCR) amplification was probably the largest advance in genome studies, which enabled many applications for genome study to date. However, in protein-based studies it may not be possible to exploit such a tool. For that reason and because protein chemistry and enzymology differ substantially from those of the DNA, proteomics is expected to be a more complex task than genome sequencing. To meet this challenge, many new and different technologies for the whole proteome characterization must be applied synergistically.

A detailed overview of proteomics techniques could be found elsewhere [Garfin, 2003; Gevaert K, J Vandekerckhove, 2000]. In general, proteome studies include the following experimental stages: isolation and separation of proteins from a cell line, tissue or organism; characterization and identification of the particular protein species; and the information storage in databases (Figure 4.3.). Two-dimensional polyacrylamide gel electrophoresis is the most widely used method for the separation of proteins in proteomics, which allows simultaneous analysis of hundreds to thousands of gene products. The first dimension, isoelectric focusing (IEF) is an electrophoretic method that separates proteins according to their isoelectric points (pI), which are defined as the specific pH at which the protein net charge is zero. The second, perpendicular dimension, SDS-PAGE (sodium dodecyl sulfate-polyacrylamide gel electrophoresis) resolves polypeptides according to their molecular weights. The ultimate result of two-dimensional gel electrophoresis is a gel with spots corresponding to individual proteins. The most critical step in 2-DE is the sample preparation, which depends on whether the aim of the study is to examine the complete protein profile of the cell or only proteins present in the particular cellular compartments (such as membranes, subnuclear structures, small organelles and vesicles).

Subsequent to 2-DE, the protein spots to be identified are cut out of the gel and the proteins digested into shorter peptides by a protease, most often trypsin. The peptide fragments are then analyzed by mass spectrometer for the purpose of identification. The resulting peptide masses make so called fingerprint, which is characteristic for the particular protein. This fingerprint is juxtaposed to theoretically expected peptide masses for each protein entry in the database. Mass spectrometer measures the mass of unknown molecules by ionizing, separating and detecting ions according to their mass-to-charge ratios and consists of three main components: an ion source, a mass-selective analyzer and an ion detector. High-throughput protein identification from 2-DE gels is dominated by the use of matrix assisted laser desorption/ionization method of sample ionization and time-of-flight mass analyzer (MALDI-TOF).

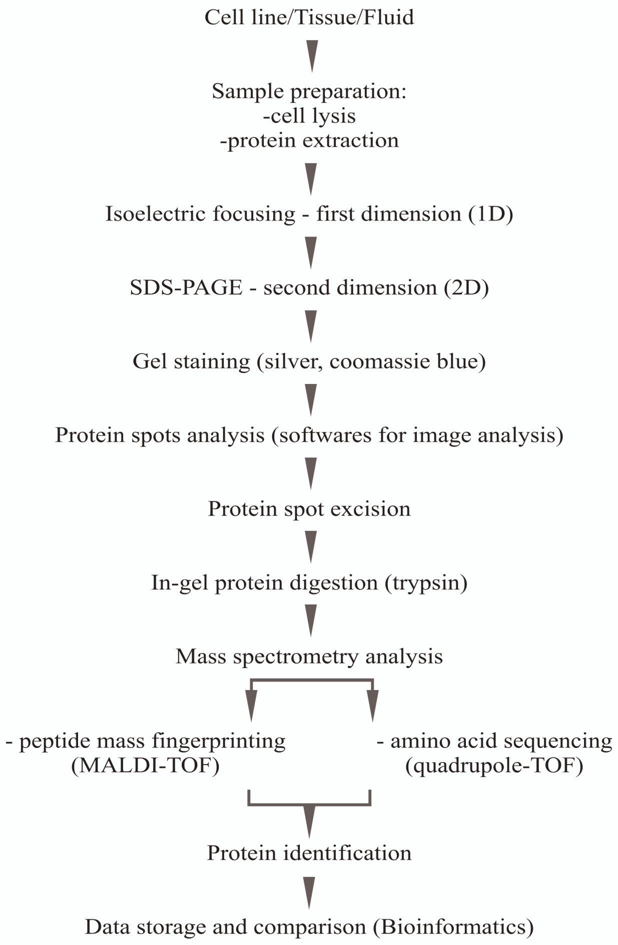
Cell line/Tissue/Fluid
Sample preparation:
-cell lysis
-protein extraction
Isoelectric focusing - first dimension (1D)
SDS-PAGE - second dimension (2D)
Gel staining (silver, coomassie blue)
Protein spots analysis (softwares for image analysis)
Protein spot excision
In-gel protein digestion (trypsin)
Mass spectrometry analysis
- peptide mass fingerprinting
(MALDI-TOF)
- amino acid sequencing
(quadrupole-TOF)
Protein identification
Data storage and comparison (Bioinformatics)

Figure4.3. *Practical approach to proteome research*

If the peptide mass fingerprint fails to identify the protein, it is possible to determine a small piece of amino-acid sequence from one or more of the peptides by fragmentation of the peptide ions in the mass spectrometer that uses quadrupole orthogonal time-of-flight mass analyzer (Q-TOF). The analysis of these new masses provides some partial amino acid sequence from the peptides, which is normally sufficient to clearly identify a protein.

Together with all broadly accepted standard proteomics methods, a high-throughput, analytical tool for rapid identification of differences between large number of samples has currently being developed – protein arrays. It would be extremely powerful to analyze hundreds or thousands of protein samples using a single protein array and thus acquire large amounts of data. However, making protein arrays is far more difficult than making DNA arrays. This is mainly due to the fact that protein function depends on its precise three-dimensional structure. Out of narrow range of environmental conditions, proteins denature. For protein arrays, it is thus critical to choose the right capture agent on the basis of its specificity and affinity for the target protein. The capture agent is immobilized on the array surface and, when exposed to samples containing complex mixture of proteins, it binds the target protein. The proteins that remain bound to the capture agents on the surface are usually detected by fluorescence and identified by mass spectrometry. Protein arrays will possibly be used as sensitive, throughput and robust technology only if they will further maturate and be constantly applied to authentic biological samples. This technology is very useful in linking gene-array expression data with protein discovery and, unlike its gene-based counterpart, can be used to examine post-translational modification of proteins.

4.7 Functional genomics in practice: is it all that perfect?

In spite of constant technological improvements that are being made in the field of functional genomics, there are still many technical and statistical obstacles hampering their usage in routine clinical practice.

When it comes to microarray technology, data analysis has been criticized for inter-, and sometimes intra- laboratory variability. Microarray experiments are quite complex and the data have proven to be noisy, susceptible to systematic errors, dependent upon sample heterogeneity (e.g. tissue) and easily affected by technical problems (e.g. sampling error, DNA/RNA isolation method, variation in protocols and handling) [Kraljevic *et al*, 2004]. When planning experiments with microarrays, some critical points should be taken into account starting from standardization of all methods and protocols, through the experimental design and ending with careful analysis and validation of the data. Before starting a microarray analysis, the reported measurements should be normalized or modified to make them comparable. The goal of normalization is to adjust for effects that are due to variations in the technology rather than the biology. Methods of interpretation of large sets of microarray biologic data are still being developed. Questions arise as to which method is considered the "right one" because of the variety of the possible outcomes. However, patterns of gene expression revealed by data analysis are just the beginning. In many cases, greater biological understanding can be attained by using expression data in conjunction with sequence data, pathway, and biomedical text sources. The limiting steps performing microarray experiments are hence not only sample handling or the analysis itself but also determination of what the obtained results actually mean which depends, as previously mentioned, on the smart use of bioinformatics tools that allow integrated analysis of multiple data types (mRNA levels coupled with proteomic analysis) resulting in the improvement of the identification of the clinical endpoints biomarkers.

However, in comparison with microarray technology, proteomics encounters some other specific problems. In particular, rather small fraction out of the total number of proteins expressed by eukaryotic cell can be routinely separated on 2-DE gel. The reason for this might be that some proteins simply fail to get into the gel due to poor solubilization (hydrophobic membrane proteins, nuclear proteins and proteins that tend to aggregate, e.g. tubulin and keratins) or molecular weight size (very large and very small proteins), others are not resolved by the pH gradient (basic and acidic proteins), or even limitations in the sensitivity of the gel staining method (disability to detect low-abundance proteins, in particular those playing an important role in the cell cycle, signal transduction and receptors).

Sample preparation is considered as the bottleneck of 2-DE in terms of quality and protein distribution, and its success lies in the efficient extraction and solubilization of proteins of interest. Unfortunately, there is no universal protocol adequate for all proteins, so that each sample (cell culture, tissue or body fluid) represents new challenge with respect to sample preparation. An important issue in this respect is the removal of non-proteinaceous particles that might interfere with 2-DE (such as salts, lipids, nucleic acids, polysaccharides) as well as of abundant proteins (e.g. albumin in human serum or fibrinogen in plasma) that might hide some low-copy proteins of biological significance.

Another matter to be considered is the artifactual modifications of proteins that might occur during 2-DE, i.e. proteolysis that occurs upon liberation of proteases subsequent to cell disruption and carbamylation, caused by the degradation of urea used in sample solutions to cyanate, which in turn reacts with the amine groups on proteins and changes their net charge thus affecting IEF separation [Garfin, 2003].

Owing to the variability from gel to gel resulting in the discrepancy in the observed amounts and number of the individual protein spots, some spots observed in one gel might not be displayed in another gel of the same sample run in parallel. Consequently, it is often difficult to come to a firm conclusion on whether the particular spot on one 2-DE gel actually matches the same protein spot on a different gel. Also, biological variations amongst samples render it difficult to establish normal protein expression profiles that can be compared with the diseased states.

Finally, 2-DE is a quite manual technique and does not seem to be easily adapted to automation for the purpose of high-throughput analyses. Nevertheless, it proves to be a method of choice for protein separation due to its excellent resolving power as well as the ability to separate different protein isoforms, and we believe it will continue to play an important role in future proteomics research.

4.8 Conclusion

The sequencing of human genome has provided the first look at all genes. The next steps however require powerful transcriptomics tools in order to verify and refine the predicted, incomplete gene models as well as define new models. Combined with complete coding sequences, protein sequences can be examined for motifs, domains, and biochemical characteristics. Proteomics is complementary to transcriptomics and it can be used to confirm the existence of an individual gene thus serving to figure out the total number of genes in a particular genome, so called “functional annotation” of a genome. Future implementation of functional genomics/proteomics in biomedicine will require a systematic examination of differentially regulated genes and proteins in tissues and fluids in healthy vs. diseased subjects. However, high-throughput technologies reflect biological fluctuations and methodological errors. Large amount of such different data challenges the performance and capacity of statistical tools and softwares available at moment. Further major developments in this field are pending and the intellectual investment will certainly result in clinical advances.

Recommended literature:

1. Bogusky MS, McIntosh MW. Biomedical informatics for proteomics. Nature 2003;422:233.
2. Celis JE, Kruhoffer M, Gromova I, Frederiksen C, Ostergaard M, Thykjaer T, *et al*. Gene expression profiling: monitoring transcription and translation products using DNA microarrays and proteomics. FEBS Lett 2000;480:2-16.
3. Diehn M, Alizadeh AA, Brown PO. Examining the living genome in health and disease with DNA microarrays. JAMA 2000;283:2298-9.
4. Dudda-Subramanya R, Lucchese G, Kanduc D, Sinha AA. Clinical applications of DNA microarrays analysis. J Exp Ther Oncol 2003;3:297-304.
5. Garfin DE. Two-dimensional gel electrophoresis: an overview. Trends Anal Chem 2003;22:263-72.
6. Gevaert K, Vandekerckhove J. Protein identification methods in proteomics. Electrophoresis 2000;21:1145-54.
7. Godovac-Zimmermann J, Brown LR. Perspectives for mass spectrometry and functional proteomics. Mass Spectrom Rev 2001;20:1-57.
8. Graves PR, Haystead TAJ. Molecular biologist's guide to proteomics. Microbiol Mol Biol Rev 2002;66:39-63.
9. Kraljevic S, Stambrook PJ, Pavelic K. Accelerating drug discovery. EMBO Reports 2004;5:837-42.
10. Lockhart DJ, Winzeler EA. Genomics, gene expression and DNA arrays. Nature 2000;405:827-36.