High-Throughput Methods for Postgenomic Research

P.V. Sergiev
Belozersky Institute of Physico-Chemical Biology, Lomonosov Moscow State University
Department of Chemistry, Lomonosov Moscow State University
E-mail: petya@genebee.msu.ru

The rapid improvement in high-throughput genome sequencing has resulted in an avalanche-like accumulation of data on nucleotide sequences now stored in databases. Yet, our understanding of the function of genes, specifically the mechanisms underlying their expression and mutual influence, remains sketchy. High-throughput study of the expression, interactions and functional role of genes could be considred as the primary goal of postgenomic research. A methodological basis for postgenomic biology is being quickly developed in order to attain this goal.

Information on the development and functioning of living organisms, as well as data concerning the supposed response of these organisms to external stimuli, is encoded in their genome. Today, genome sequencing is an important stage in the study of any species. The genome size can vary from several hundred thousand nucleotides in some bacteria to several hundred billions in some eukaryotes. The number of genes increases along with the genome size, but only up to a certain level. Several high-throughput methods for genome sequencing have thus far been elaborated. Among these sequencing platforms are Roche/GS-FLX Titanium (500 million nucleotides per day), Illumina/HiSeq 2000 (55 billion nucleotides per day), and the ABI/SOLiD 5500xl (up to 30 billion nucleotides per day). Nowadays, the amount of nucleotide sequences continues to increase at a rapid rate. By early 2011, the major GenBank database contained as many as 126, 551, 501, 141 nucleotides. The sequencing procedure is getting cheaper, and in the foreseeable future the metagenome of the entire biosphere could well be determined.

With the advance in high-throughput methods for sequencing, new data on nucleotide sequences was obtained much faster than our undersanding of individual genes functions. In 2000, Peer Bork articulated the following problem: the function of approximately 30% of genes in each new genome as yet remains unknown. Furthermore, the validity of predicting the function of the other 70% of genes is also approximately 70%. In other words, today we are at the same point as geographers were in the epoch of great geographical discoveries. The general outline has already been made clear, but much remains to be done in order to understand the integral worldview. When studying a living organism, we are to determine the type and extent of mutual influence of all the genes whose products have an effect on each other.

To study a gene means to answer several questions (*Fig. 1*). First, it is necessary to know what happens to a cell or an organism if a gene is inactivated or temporarily "switched off." This also relates to the question of whether the mutations in the genome regions that are distant from the given gene reduce or increase the effects of the inactivation of this gene. All genes function under different conditions. Some genes are required all the time, while others are required only under certain circumstances. Studying the conditions in which a gene functions, i.e., when demand for the functioning of this gene emerges, provides a significant amount of information that helps to understand the question of the gene's role. Since gene functioning comprises several stages, our task is to study them all. Secondly, it is necessary to determine when a gene is transcribed; thirdly, we must find out when it is translated. Fourthly, one needs to know what molecules (proteins, RNA, DNA, small molecules) it interacts with. If a gene encodes an enzyme, one needs to know the reactions it catalyzes. If it is not an enzyme encoded by a gene, one needs to know the processes in which it participates. Efficient elaboration of post-genomic technologies is impossible without a procedural basis enabling the study of the functional interactions of a number of genes and their products. Moreover, it is necessary to elaborate methods that would enable the study of both

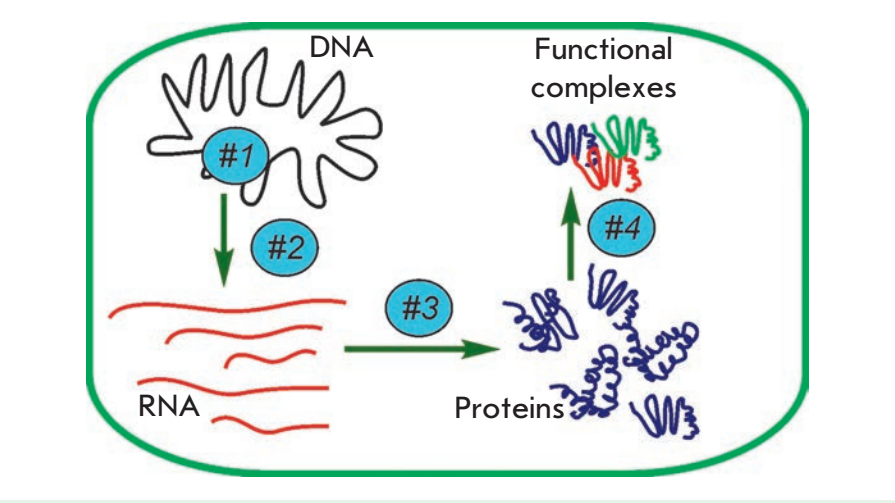

Fig. 1. Gene function scheme and questions needed to be addressed. DNA, RNA, proteins, and functional complexes are schematically shown and labelled. The gene expression pathway is shown by arrows. The questions to be answered are in blue circles: (#1) what is the phenotype of gene inactivation? (#2) How is a gene transcribed? (#3) How does mRNA correspond to a particular gene translated? (#4) What are the interaction partners of the gene expression product?

the expression of all genes under certain conditions, as well as those based on genetic manipulations with each gene from the full gene pool of an organism.

METHODS FOR STUDYING GENE EXPRESSION AT THE TRANSCRIPTION STAGE

As follows from the central dogma of molecular biology proposed by Francis Crick, gene expression involves two stages. Firstly, a gene is transcribed by RNA polymerase, yielding the proper RNA. Sometimes, it is RNA that is the functional product of a gene; in this case, gene expression is identical to RNA transcription and maturation. More frequently, the RNA copy of a gene is an mRNA and is translated by ribosomes, yielding a protein product. In order to "measure" gene expression, methods that allow to determine the amount of RNA or a protein are used. When performing post-genomic studies, it is necessary to measure the quantitative characteristics of the expression of a set of genes in the organism (ideally, all genes).

Microchip hybridization is currently the standard method for measuring the amount of all types of RNA in cells (*Fig. 2A*). The total RNA is extracted from the cells and cDNA is constructed by reverse transcription, in order to determine the gene expression level using microchips. This cDNA is modified by fluorescent dyes. Two cDNA samples stained with Cy3 and Cy5 dyes are typically compared. The mixture of cDNAs stained with different dyes is hybridized on a microchip with immobilized oligonucleotides that are complementary to the individual types of cDNA. The ratio between the fluorescence intensities of Cy3 and Cy5 in a certain spot on the microchip corresponding to a certain gene is a measure of the relative expression of this gene in the samples. Specific oligonucleotide samples, numbering between several tens of thousands to over a million, are used in modern microchips; they include the known genes of almost any model organism with a multifold excess.

Microchip technology is being gradually replaced by high-throughput sequencing technology (*Fig. 2B*). The same technologies as those used for high-throughput determination of the nucleotide sequence of individual genomes can be used to determine the whole range of RNAs present in a cell. This experiment is used to determine a large number (up to several billions) of short sequences contained in the total RNA (transcriptome). A computer analysis can be applied to align these short sequences with the genome sequence and, thereby, determine which regions of the genome are transcribed. The number of short RNA fragments referring to this gene which are detected by high-throughput sequencing can serve as a measure of gene expression.

High-throughput RNA sequencing has considerable advantages over the microchip technology, since no requirements to the preliminary annotation of this genome region as a gene are set. Thus, many earlier unknown transcripts are detected. Moreover, high-throughput RNA sequencing also permits the unbiased (i.e., not based on earlier known hypotheses) determination of the beginning and end of the transcript, as well as the variants formed by alternative processing. The results of studies of the transcriptome obtained using both microchips and high-throughput RNA sequencing are typically verified using quantitative PCR of cDNA (RT qPCR) in the case of particularly interesting genes (*Fig. 2B*). This method is based on the amplification of one cDNA fragment (amplicon) involving the quantitative determination of the resulting product, depending on the PCR cycle. With this purpose in mind, the fluorescence intensity of either a DNA-intercalating dye or a dye bound to a specially selected DNA probe is measured. In terms of its reliability, the RT qPCR method is superior to the microchip hybridization method and, to a certain ex-

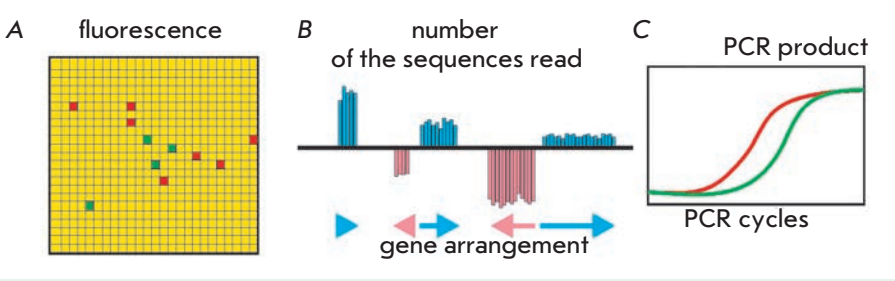

Fig. 2. High-throughput methods for studying gene transcription. *A* – Scheme of evaluation of the transcription levels by microchip hybridization. Reverse transcription is used to introduce the fluorescent label. Cy3- and Cy5-labelled cDNA derived from the samples to be compared fluoresce green and red. The rectangular grid is a microchip with immobilized oligonucleotide probes. Yellow areas correspond to the regions with equal gene expression in the compared samples, while green and red areas are those where cDNA labelled with Cy5 or Cy3, respectively, is predominant. *B* – Scheme of high-throughput transcriptome sequencing. The result of such an analysis is a plot of RNA reads of the distribution over the genome sequence. The frequency of the reads that belong to a particular transcript shown as blue and pink bars is used as a quantitative measure of the expression. *C* – Scheme of gene expression study by the quantitative polymerase chain reaction of cDNA (RT qPCR). Accumulation of two amplicons corresponding to two genes with passage of the PCR cycles is schematically shown in the plot. The earlier emergence of the PCR product attests to abundant mRNA in the transcriptome.

tent, even to high-throughput RNA sequencing. The major drawback of RT qPCR is that only one transcript can be measured in an experiment. There are several ways to overcome this disadvantage. Firstly, the use of 96- and 384-well PCR plates and proper instruments allows for the simultaneous detection of many transcripts. With modern instruments, such as the 7900HT (Applied Biosystems) or the CFX384 (Bio-Rad), the level of fluorescence intensity is measured simultaneously at four or five wavelengths. Moreover, the capability of automatic plate loading from a stack increases the maximum performance of the instrument to 20, 000 samples per run. This performance is sufficient to study the expression of all genes of bacteria and even primitive eukaryotes. Automated stations, such as the station based on Janus Extended (Perkin Elmer) mounted at the Centre for Collective Use (Moscow State University), are used to pipette this amount of PCR.

METHODS FOR STUDYING GENE EXPRESSION AT THE TRANSLATION STAGE

The expression of the genes encoding proteins involves two stages, according to the central dogma of molecular biology. mRNA formed as a result of gene transcription has to be read (translated) by a ribosome. Protein is synthesized by the ribosome according to the information encoded in mRNA. Although gene expression is regulated mostly at the transcription stage, mRNA translation can also be regulated. A number of fascinating mechanisms which regulate the translation of individual mRNAs remains to be understood. For a systemic comprehension of the mechanisms of regulation of gene expression, one should determine the relative protein amount in a cell, in addition to measuring the levels of RNA. Proteomics deals with measuring the amount of protein. This field of science is undoubtedly part of post-genomic technologies and has its own tools.

Two-dimensional protein gel electrophoresis remains the standard method for studying the combination of proteins in a cell (*Fig. 3A*). Protean (Bio-Rad) is the most commonly used system of two-dimensional gel electrophoresis. Similar to the study of the transcriptome using microchips, the comparison of two samples modified with different fluorescent dyes is most informative. Cy3 and Cy5 are typically used; they are bound to the proteins by the reaction between the hydroxysuccinimide esters of the dyes and the lysine residues in protein molecules. Protein samples modified with Cy3 and Cy5 are mixed and separated according to the isoelectric point value. The proteins which had become neutral at various pH values could be separated in this manner. The proteins are then separated according to their molecular weight using gel electrophoresis in the presence of an anionic detergent, sodium dodecyl sulfate. After protein separation, the gel is scanned on a fluorescence scanner. The relative amount of a certain protein in the initial samples can be assessed from the Cy3/Cy5 fluorescence ratio. The resolving power of each method of separation is believed to be approximately 100 protein bands. Thus, the number of protein types that can be distinguished in theory is 10, 000. Unfortunately, such a resolving power cannot be achieved in practice, because of several reasons. Firstly, the distribution of proteins in a cell over the isoelectric point and weight is not ideal. The properties of most proteins are quite similar. Secondly, the abundance of protein species in a cell varies by several orders of magnitude. Abundant proteins can be detected easily using two-dimensional gel electrophoresis, whereas it is almost impossible to detect rare proteins using this procedure. Thus, two-dimensional gel electrophoresis can be considered reliable only for the determination

of the amount of several hundreds of the most common proteins.

Protein identification from the fluorescing spots of a two-dimensional gel is performed via mass spectroscopy. Accordingly, a protein is cleaved into fragments by a specific protease, such as trypsin. Then, MALDI mass spectrometry is used to analyze fragment weights. Such instruments as Ultraflex (Brucker) and AB SCIEX 5800 (AB Sciex) are the most commonly used in modern proteomic laboratories.

Liquid chromatography, coupled with electrospray ionization mass spectrometry, is an alternative and supplementary method to two-dimensional gel electrophoresis in separating the whole cell proteome (*Fig. 3B*). The use of such systems makes the analysis of the whole proteome possible; however, the main difficulty consists in the extreme variety of the proteolytic fragments resulting from the hydrolysis of the entire pool of cell proteins. In the absence of protein modification with fluorophores, which is used in the study of proteome using two-dimensional gel electrophoresis, specific methods that would allow to compare the protein amounts in two samples using mass spectrometry only are required. Isotope labelling with iTRAQ is such a method. When using this method, protein samples are modified with chemically identical appendages consisting of two fragments. They can be split easily into special mass spectrometers that possess a fragmentation mode. Prior to fragmentation, the total weights of the appendages used to label two samples are equal. Thus, two identical peptides with chemically identical appendages that have equal weights originating from two protein samples are simultaneously analyzed in the mass spectrometer. The differences in weight appear only after the appendages are split into two fragments. Fragment weights differ, since they have a different isotopic composition. The ratio between the number of fragments with different weights will be the same as the protein ratio in the initial mixture.

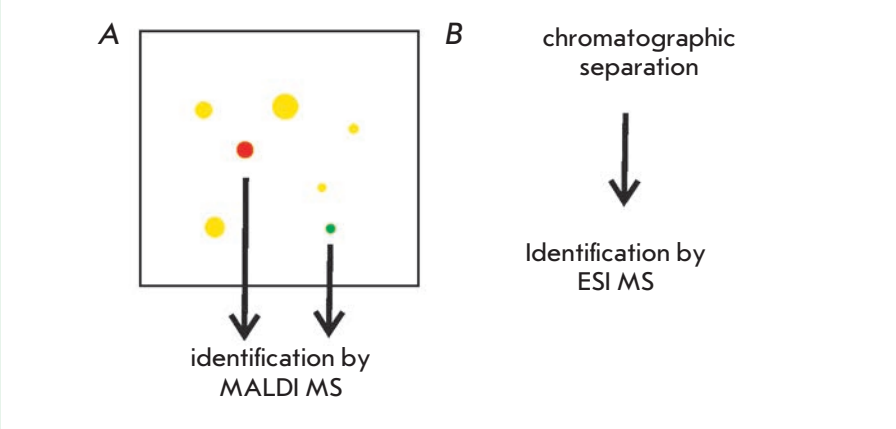

Fig. 3. Methods for studying proteome. *A* – Scheme of a quantitative comparison of proteomes in two samples by 2D gel electrophoresis. The protein samples to be compared are separately labelled with succinimide esters of the fluorescent dyes Cy3 and Cy5. After labelling, the protein samples are mixed and subjected to separation on a 2D gel. Yellow spots correspond to proteins with equal abundance in both samples, while green and red spots correspond to the proteins with different amounts. Protein identification in excised spots is accomplished via protease hydrolysis followed by peptide fingerprinting by mass spectrometry. *B* – Scheme of quantitative proteome analysis by chromatography. The total protein sample is treated with protease; the resulting peptides are separated by liquid chromatography. Peptide identification is usually performed by mass spectrometry with electrospray sample ionization.

METHODS FOR STUDYING GENE EXPRESSION USING REPORTER CONSTRUCTS

Applying the modern methods of proteomic analysis permits one to compare the amount of most cell proteins in different samples. Nevertheless, the stage at which the expression changed (transcription or translation) cannot be determined by proteomic analysis. Reporter constructs are used to study each individual expression stage, as well as the gene elements that are important for a certain mechanism of expression regulation. When using this method, the gene under study is replaced by the gene encoding the protein, whose amount in the cell can be measured easily. The genes of b-galactosidase, luciferases of different origins, and fluorescent proteins are widely used as reporter constructs. The individual elements of the gene subjected to a study that are responsible for transcription (promoter) and translation (usually, the 5'-untranslated region) can be used for the creation of a reporter construct. The amount of the reporter protein in a cell is assessed by measuring the amount of the products of the model enzyme reaction or fluorescence intensity in the case of fluorescent proteins. A similar reporter gene, whose expression is independent of the regulatory elements of the gene under study, is used as an internal control. The major problem in the use of reporter constructs is the complexity of creating them and detecting expression for the set of genes studied. The reporter construct method is very informative for one or several genes, but it is quite labour-consuming when studying a set of genes (ideally, all genes in the organism). This drawback can be overcome by using

the automated methods for cloning reporter constructs and determining their expression products. The facilities of the automated station based on Janus Extended (Perkin Elmer) mounted at the Centre for Collective Use (Moscow State University) is used in our laboratory. The automated station allows one to perform cloning, bacterial transformation, and to detect the expression of the reporter genes in automatic high-throughput mode. The genes of the red fluorescent protein from *Entacmaea quadricolor* and a cyan-modified variant of the protein from *Aequorea macrodactyla* are used as reporter genes (*Fig. 4*). As opposed to b-galactosidase and luciferases from *Photinus pyralis* and *Renilla reniformis*, the amounts of fluorescent proteins can be measured without damaging the cell or using enzymatic reactions. These advantages make the analysis of the expression of a set of reporter constructs in a single experiment considerably simpler and less expensive.

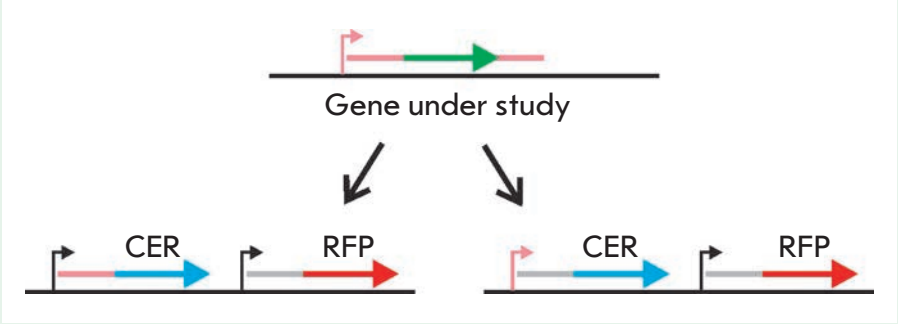

Fig. 4. Scheme for studying gene expression at the transcription and translation stages using reporter constructs. The gene under study is shown in green. The promoter and untranslated regions are shown in pink. The red fluorescent protein gene is shown in red, while the cyan fluorescent protein gene is shown in blue. The promoter (right) or 5′-untranslated region (left) under study is inserted in front of the cyan fluorescent protein gene. Expression of the red fluorescent protein is used as an internal standard.

METHODS FOR STUDYING THE GENE FUNCTION

Studying the gene function does not encompass their expression. For comprehension of the functional role of gene products, one has to ascertain what the gene product interacts with and what the consequences of the absence of a gene for a cell are. The cell components that interact with the product of a gene under study are typically detected by affinity co-purification with the gene product (protein or RNA). The cell components that are extracted, together with the protein under study or RNA, are fractionated and identified. Either the antibody to the protein under study or standard affinity tags attached to a protein by gene modification for this protein is used for affinity extraction. This procedure is easy to implement in the case of one or several proteins, but it becomes extremely labour-consuming when studying a set of proteins. At the time of writing, it is impossible to produce antibodies to each cell protein, as opposed to high-throughput-changing genes in the genome. Technologies of the same level are to be used to study the phenotype of cells lacking one of the proteins. If gene inactivation is not lethal, a strain or cell line lacking this gene is to be produced (i.e., gene knockout is to be performed). If gene inactivation results in non-viability, an artificially regulated promoter is to be incorporated in front of the gene (i.e., gene knockdown is to be performed). In order to increase the scale of these studies, methods of genome manipulations must be automated for the entire set of genes. Currently, such opportunities exist only for bacteria and yeast. Vast collections of strains of bacteria *Escherichia coli* and yeast *Saccharomyces cerevisiae* with inactivation of one of the genes have already been created. There are also partial collections of strains of these organisms in which one of the genes contains a fragment encoding the affinity tag. These collections have already made it possible to carry out a partial experimental analysis of the phenotypes of gene inactivation and study the partial cellular interactome; i.e., to perform cataloguing of the contacts between cell components.

Today, high-throughput analysis of the functional role of genes is in its early days of development. Even for the model bacterium *E. coli*, the phenotypes of the entire set of knockout strains have been analyzed only according to the rate of colony formation under various growth conditions. No overall collection of strains in which the genes important for viability are controlled by the regulated promoters has been created as of yet. Studies on the introduction of the tetracycline-activated promoter in front of *E. coli* genes that are essential for viability have been launched at the automated station based on Janus Extended (Perkin Elmer) mounted at the Centre for Collective Use (Lomonosov Moscow State University).

Combination of high-throughput methods for genome manipulations, such as creating strains containing gene “knockouts” and “knockdowns,” with methods of high-throughput analysis of gene expression could hold great promise. Such combination could allow one to uncover the effect that the activity of a gene has on another gene. Filling such a “matrix of mutual effects” will allow one to completely ascer-

tain all the regulatory pathways of a cell. The major difficulty in solving this problem is rooted in the fact that it is necessary to perform an extremely large number of experimental studies, which increases in proportion to the square number of genes. Hence, determination of the "matrix of mutual effects" for all *E. coli* genes will require 17 million experiments. It is clear that this is the maximum number, which can be reduced by restricting the screening at the expense of experiments. Systemic investigation of the influence the genes have on one another will require that the experiments be completely automated and considerably cheap to perform. Thus, the use of reporter constructs based on fluorescent proteins only, a method which was started in our laboratory, seems feasible in analyzing gene expression. This method requires significant investments to be made at the stage of creating reporter constructs; however, in future it will make possible the performance of measurements based only on fluorescence, without the use of any enzymatic reactions.

The systemic analysis of all of the stages of gene expression, the mutual effect of gene expression, functional role of gene products, and interactions between gene products is the most significant challenge in post-genomic biology. The constructed "matrix of mutual effects" will allow one to understand the functioning of the entire network of regulatory interactions inside a cell, and it has the potential of allowing control of any intracellular process. ●

The author is grateful to O.A. Dontsova for valuable comments and contribution to the manuscript.

REFERENCES

1. Watson J.D. // Genome Res. 2001. V. 11. № 11. P. 1803–1804.
2. Lewin B. Genes VIII. Upper Saddle River, N.J.: Pearson Prentice Hall. 2004. 1027 p.
3. Benson D.A., Karsch-Mizrachi I., Lipman D., Ostell J., Sayers E.W. // Nucl. Acids Res. 2011. V. 39 (Database issue). P. D32–37.
4. Kennedy J., Flemer B., Jackson S.A., Lejon D.P., Morrissey J. P., O'Gara F., Dobson A.D. // Mar. Drugs. 2010. V. 8. № 3. P. 608–628.
5. Bork P. // Genome Res. 2000. V. 10. № 4. P. 398-400.
6. Crick F.H. // Symp. Soc. Exp. Biol. 1958. V. 12. P. 138–163.
7. Ramsay G. // Nat. Biotechnol. 1998. V. 16. № 1. P. 40–44.
8. Hoen P.A., Ariyurek Y., Thygesen H.H., Vreugdenhil E., Vossen R.H., de Menezes R.X., Boer J.M., van Ommen G.J., den Dunnen J.T. // Nucl. Acids Res. 2008. V. 36. № 21. P. e141.
9. VanGuilder H.D., Vrana K.E., Freeman W.M. // Biotechniques. 2008. V. 44. № 5. P. 619–626.
10. Mathews M., Sonenberg N., Hershey J.W.B. Translational control in biology and medicine. 3rd ed. Cold Spring Harbor monograph ser. Cold Spring Harbor, N.Y.; Cold Spring Harbor Laboratory Press, 2007. 934 p.
11. Sonenberg N., Hershey J.W.B., Mathews M. Translational control of gene expression. 2nd ed. Cold Spring Harbor monograph ser. Cold Spring Harbor, N.Y.; Cold Spring Harbor Laboratory Press, 2000. 1020 p.
12. Shevchenko A., Wilm M., Vorm O., Mann M. // Anal. Chem. 1996. V. 68. № 5. P. 850–858.
13. Chakravarti B., Gallagher S.R., Chakravarti D.N. // Curr. Protoc. Mol. Biol. 2005. Chapter 10. P. Unit 10 23.
14. Link A.J. Methods in molecular biology. Totowa, N.J.: Humana Press, 1999. V. xvii, 601 p.
15. Brewis I.A., Brennan P. // Adv. Protein Chem. Struct. Biol. 2010. V. 80. P. 1–44.
16. Chalkley R. // Methods Mol. Biol. 2010. V. 658. P. 47–60.
17. Treumann A., Thiede B. // Expert Rev. Proteomics. 2010. V. 7. № 5. P. 647–653.
18. Ghim C.M., Lee S.K., Takayama S., Mitchell R.J. // BMB Rep. 2010. V. 43. № 7. P. 451–460.
19. Merzlyak E.M., Goedhart J., Shcherbo D., Bulina M.E., Shcheglov A.S., Fradkov A.F., Gaintzeva A., Lukyanov K.A., Lukyanov S., Gadella T.W., Chudakov D.M. // Nat. Methods. 2007. V. 4. № 7. P. 555–557.
20. Rizzo M.A., Springer G.H., Granada B., Piston D.W. // Nat. Biotechnol. 2004. V. 22. № 4. P. 445–449.
21. Casadaban M.J., Chou J., Cohen S.N. // J. Bacteriol. 1980. V. 143. № 2. P. 971–980.
22. Nordeen S.K. // Biotechniques. 1988. V. 6. № 5. P. 454–458.
23. Baba T., Ara T., Hasegawa M., Takai Y., Okumura Y., Baba M., Datsenko K.A., Tomita M., Wanner B.L., Mori H. // Mol. Syst. Biol. 2006. V. 2. P. 2006–2008.
24. Chu A.M., Davis R.W. // Methods Mol. Biol. 2008. V. 416. P. 205–220.
25. Butland G., Peregrin-Alvarez J. M., Li J., Yang W., Yang X., Canadien V., Starostine A., Richards D., Beattie B., Krogan N., et al. // Nature. 2005. V. 433. № 7025. P. 531–537.
26. Hu P., Janga S.C., Babu M., Diaz-Mejia J.J., Butland G., Yang W., Pogoutse O., Guo X., Phanse S., Wong P., et al. // PLoS Biol. 2009. V. 7. № 4. P. e96.
27. Howson R., Huh W.K., Ghaemmaghami S., Falvo J.V., Bower K., Belle A., Dephoure N., Wykoff D.D., Weissman J.S., O'Shea E.K. // Comp. Funct. Genomics. 2005. V. 6. № 1–2. P. 2–16.
28. Nichols R.J., Sen S., Choo Y.J., Beltrao P., Zietek M., Chaba R., Lee S., Kazmierczak K.M., Lee K.J., Wong A., et al. // Cell. 2011. V. 144. № 1. P. 143–156.

EDITORIAL NOTE

The subject matter raised is of absolute urgency. Indeed, some of the most modern equipment, including the unique kind, has recently appeared in many research institutions. This makes it possible to considerably enhance research capabilities not only at these institutions, but at other centers as well, provided that access to information is adequately ensured. The editorial board is bound on honor to spread this information, and this topic will be discussed in the next issues of the journal *Acta Naturae*.