REVIEW —Current Perspective—

Recent Progress in Human Molecular Biology and Expression Profiling of Active Genes in the Body

Kenichi Matsubara and Kousaku Okubo

Institute for Molecular and Cellular Biology, Osaka University, Osaka 565, Japan

Received July 24, 1995

ABSTRACT—Recent progress in molecular biology based upon rapidly developing DNA technology is reviewed. Emphasis is placed on the developing human genome project that includes structural as well as functional analyses of the genome. Expression profiling of active genes in the body helps construct a database for the functional aspects of the human genome.

Keywords: Human genome project, DNA technology, Expression profiling of genes, Body map, cDNA analysis

Molecular biology has made the transition from phase 1, dealing with prokaryotic genes and DNA, to phase 2, dealing with genes of eukaryotic multicellular organisms. This transition came about because of DNA technology that has evolved since the early 1970s. Now we are at the beginning of the transition to phase 3, through the emerging human genome project.

In conjunction with the human genome efforts, the large-scale collection of partial cDNA sequences is becoming more important. Similarity or motif searches in DNA databases using these partial cDNA sequences have facilitated the discovery of new genes of interest. By collecting and registering large numbers of partial sequences with a well-designed non-biased cDNA library, an expression profile of active genes in a particular tissue can be obtained. Tissue-specific or stage-specific genes can be discovered by comparing the profiles from different tissues or from a tissue at different stages of development, respectively. The compilation of such expression profiles enables genes to be mapped to the tissue(s) where they are actively transcribed. The large-scale collation of gene sequences actively expressed in the body into databases, called "body mapping", complements efforts directed towards the structural analysis of the genome, with the ultimate aim of decoding all the genetic information carried in the human genome.

Revolution of life sciences by molecular biology

Toward the end of the 20th century, we have been witnessing a revolution of the life sciences because of molecular biology. For example, reports on the structure and function of genes in such complex processes as regulation of cell proliferation, differentiation, cell-cell communication and morphogenesis are appearing at much greater speed than had been expected only 20 years ago.

The molecular biology that evolved shortly after World War 2 succeeded, among other things, in elucidating the basic mechanisms of heredity. However, the target for this research was limited to prokaryotes. In the mid-70s, it entered into phase 2, the period of DNA research, and many biological phenomena unique to multicellular organisms have become amenable to study by isolating and studying the relevant genes. Now, through the development of the human genome project, it is entering phase 3, in which we are beginning to have a global understanding of genomes, rather than just the individual genes. There has never been such an ambitious project in the life sciences.

DNA research and the medical sciences

DNA research is also having an impact on medical science through the development of human molecular biology. Up to the mid-70's we had no way to obtain and study a new gene out of the total set of human genes. But now, only 21 years after the introduction of DNA technology, we have more than 3000 human genes that have been isolated in their pure form and studied in detail. A union between basic biology and the medical sciences has opened up new areas in 1) pathogenetics, 2) diagnosis, 3) pharmaceutical science and 4) treatments that include gene therapy.

As an example, let us take a look at the emerging new pathogenetics. The newly developed positional cloning technique combined with the genetic mapping of the disease gene along chromosomes have made this new research possible. Isolation of the genes for Duchenne muscular dystrophy, cystic fibrosis and Huntington's chorea are but a few of the achievements. This new technology will prove to be invaluable also in studies with genes for cancer (its development and suppression), cardiovascular diseases and aging. More subtle problems such as mental disorders as represented by manic depression and alcoholism may also be amenable to this technology. More will be added to the list, including the genes determining susceptibility to disease.

Genes to genome

Our understanding of disease(s) has been greatly improved by gathering information on many genes whose functions are interrelated. An example can be seen with problems in growth regulation. Collection of information on growth factors, cytokines and their receptors, and signal transduction is escalating. Knowledge of these areas is necessary to design new plans for research on cell growth and its control. Another remarkable example can be seen with cancer research, where a global view of more than 40 oncogenes and no less than ten antioncogenes is required for advanced research.

Under these circumstances, Dulbecco (Science, 1986) discussed two options: either to try to discover the genes important in malignancy by a piecemeal approach as we have up to now or to sequence the whole genome of a selected animal species. He then proposed that sequencing the whole human genome should be initiated as soon as possible.

Similar proposals were made independently by several researchers, and an exciting discussion as to whether the life science community should initiate the human genome project took place. By the end of 1988, most had agreed to initiate the human genome project, recognizing that this will mark a turning point in the life sciences, without sacrificing ongoing research activities.

The human genome project

By the spring of 1989, many developed countries like the United States, Britain, France, Japan and Italy had launched national human genome projects, and since then, many more have joined in this effort.

Human genome analysis aims at:

- registering all the genes in the human genome
- elucidating the structure of these genes and other DNA
- understanding the principle that underlies arrangement of genes and other DNA in the genome
- analyzing regulatory mechanisms that act in gene expression

Because the human genome DNA consists of some 3 billion nucleotides and contains roughly 100 thousand genes, it is by no means easy to complete its sequencing and to identify all the genes with the currently available technologies. Therefore, the human genome project has organized concerted international efforts for genetic and physical mapping and to prepare ordered arrays of sequencing. It also aims at developing new, highly efficient and automated technologies for DNA analysis. At the same time, it also aims at developing new informatics with which one can manage and analyze large amounts of structural and genetic human genome information. The resulting database from the human genome efforts will prove invaluable for human health and welfare. At the same time, the new DNA technologies and systems for analyses will greatly influence the future development of the life sciences.

Current state of human genome efforts

As noted earlier, the currently ongoing human genome analyses consist of two major activities: analyses of structure of the human genome DNA and analyses of the function of the human genome. In parallel, structural analyses of several model organisms that carry smaller genomes are being carried out. The basic strategy for the structural analyses is to construct maps of genes and genetic traits of disease along chromosomes and simultaneously to construct physical maps of all or part of the chromosomes. For future analyses at the nucleotide level when the relevant technology matures, collection of DNA fragments and arraying them along the chromosome are also being actively worked out. Recent advances in genetic and physical mapping, in conjunction with positional cloning, are also greatly facilitating isolation and characterization of genes for important diseases.

With smaller genomes, active sequencing work is already underway. Thus the genomes of *Escherichia coli*, *Bacillus subtilis*, yeast, the little worm nematode, the fly and the small plant *Arabidopsis* have been actively sequenced all around the world; and in a few years from now, we will have nucleotide sequences with all, or at least significant, fractions of these genomes. There is an increasing demand that massive efforts must be organized to identify the biological functions of the novel genes discovered through these sequencing studies.

The cDNA project

The analyses of genome function are carried out by analyzing mRNA, the transcript of the gene. For easy analysis, mRNA is converted to cDNA, and hence the effort is often called the cDNA project. Through the cDNA analysis, we can learn the minimum structure of each gene, that can be extended up to an expected 100,000 genes. With the help of currently available technology, cataloguing of the total of genes in the human genome is expected to be nearly complete in only a few years. The main activities of the currently ongoing cDNA strategy is to collect a large number of partial cDNA sequences and to use these to discover interesting genes by carrying out similarity searches. These efforts are stimulated by human genome mapping activities, and therefore, most of the data are of human genes, although this strategy is rapidly expanding to include other organisms. Collecting partial cDNA sequences is equivalent to collecting signatures of genes. Each partial sequence may cover only a fraction of the full-size gene transcript, but it irrevocably identifies a gene, unless it is a member of a very closely related gene family. More than 60% of the randomly selected clones in a cDNA library obtained from a human cell or tissue identify novel genes. This value is lower than the percentage of uncharacterized genes present in the genome, because any given cDNA library contains multiple recurring clones.

Types of cDNA library and data collection

Several different types of cDNA library can be constructed, and their design can greatly influence the information available for analysis. First, a cDNA library can be prepared from a selected cell or organ (e.g., brain or liver) so that the library constituents reflect the physiology of that cell or organ. With appropriate measures, one can construct a library that faithfully represents the abundance of gene transcripts in the original mRNA population. Alternatively, by eliminating or reducing redundant components, one can construct a 'single book' or 'normalized' library, which is suited for surveying large numbers of gene transcripts. To maximize the complexity of the constituents, several cDNA sub-libraries from various tissues may be combined. In earlier attempts, cDNA were hybridized in liquid with unfractionated mRNA or cDNA before cloning to eliminate the readily reassociating abundant species. However, because the resulting library tends to consist of fragmented cDNA, 'clone first and then select' has recently become more popular. In this approach, unselected cDNA clones are immobilized on a filter, hybridized to a total cDNA probe, and appropriate clones (whose abundance is reflected in the intensity of the hybridization signal) are selected for subsequent tests.

Second, the library may be designed to comprise full-length inserts or partial cDNAs. The latter type of insert can be made either by random priming (1) or by directed synthesis to cover a defined region of each mRNA such as the 3′ or 5′ ends. Among these, 3′-directed libraries (made using appropriate measures) can faithfully represent the composition of the mRNA population (2). Thus, this library may be used for qualitative as well as quantitative analyses of mRNAs, although a significant fraction of the library constituents may lack an amino acid coding sequence. On the other hand, in currently available 5′-directed libraries, some (perhaps ~40%) of the clones may code for leader sequences and amino-terminal amino acid sequences, but others represent internal portions of the mRNA. With full-size or randomly primed libraries, full or partial amino acid sequences may be obtained, but such libraries are not suited for quantitative analyses.

Expression sequence tags for discovering interesting genes by similarity searches

Using randomly primed cDNA libraries from human brain, Adams et al. (1, 3) have collected 3272 partial cDNA sequences (average size 370 nucleotides), terming each sequence an expression sequence tag (EST). After subtracting redundant sequences and sequences that do not represent unique genes, 2723 different ESTs were obtained. Homology searches in the DNA databank revealed that 2092 of these ESTs represented novel genes. These workers have subsequently expanded the list to more than 10,000. The EST project has permitted thousands of expressed genes in a given tissue to be scanned for similarity. Genes of particular interest (e.g., homologs of gene encoding receptors, signal transducers, and proteins with a particular amino acid sequence motif) can be sorted using this approach. It can also be used to pinpoint genes of evolutionary interest (with or without conserved sequences) and genes known to be associated with genetic disorders.

The power of the partial sequencing strategy is its applicability to a comprehensive search of new and interesting genes, including those that may have commercial significance. This strategy has been applied to many cell- and tissue-types other than the brain, including liver, lymphocytes, testes, skeletal and heart muscle, kidney and thymus, and will soon have covered most of the important cell types in our body. It has also been extended to other organisms, such as *Caenorbabditis elegans, Arabidopsis thaliana* and rice. To maximize the available amino acid coding information, efforts are being made to rapidly collect long (hopefully, full-size) sequences, but several hurdles must be cleared in order to realize effective technologies for cDNA library construction and sequencing. Construction of 5′-directed cDNA libraries using enriched capped mRNAs is becoming more realistic. Generally speaking, however, the use of the EST collections that contain only primary nucleotide sequences found in the source tissue is limited solely for homology/similarity searches. For discovering interesting genes, biological information added to a particular sequence can greatly improve the quality of the library. In this connection, the library of nematode partial cDNA sequences, in which both sequence data and mapping information are combined, is a model system.

Another problem inherent to the EST approach is the lack of quantitative evaluation of the data. An average cell carries 150,000–300,000 mRNA molecules. Suppose that one database entry represents an mRNA whose abundance is more than 1000 molecules per cell, and another entry represents an mRNA whose abundance is less than 0.1 molecule per cell. Without information concerning their abundance, how can one compare the biological significance of these two genes? For the same reason, even after intensive analyses of ESTs, it is not clear what fraction of the total mRNA is represented in a collection. Thus, the best use of the EST collection and homology search efforts is to find new genes in a tissue of interest, such as the brain. An EST taken from a randomly primed cDNA library does not necessarily have a 1 : 1 correspondence to a gene. For example, two ESTs could be from two genes or from two different regions of the same gene giving overestimates of new 'hits'. For efficient searches of new genes, it is obvious that two ESTs should represent different genes. This can be achieved only with directed cDNA libraries.

Expression profiles: qualitative and quantitative analysis of gene transcripts in a tissue

Okubo et al. (2) have constructed a 3′-directed cDNA library from a liver carcinoma cell line, using techniques such that the library composition faithfully represents the mRNA population. By randomly selecting clones and collecting 3′-sequences, which they termed gene signatures, they compiled 982 gene signatures to create an expression profile. The data, in which the frequency of the appearance of a clone is proportional to the abundance of mRNA, demonstrate that three highly abundant genes are represented in this cell line, serum albumin, elongation factor 1α and translationally controlled tumor protein, each representing more than 1% of the total mRNAs. These authors obtained 170 different genes of medium abundance (35 were known and 135 were novel) and 468 low-abundance genes, among which 415 were novel. As a rough estimate based on the abundance of identified genes in the public DNA databank, about one third of the expressed genes in this cell are related to protein synthesis, another third are likely to be tissue-specific and the rest encode other functions. Thus, by analyzing about 1000 clones, one can obtain an idea of the class or type of abundance of mRNA in which major active genes in the cell are represented. To obtain a larger collection of gene signatures, however, elimination of high-abundance class genes is necessary. The list describing the genes expressed in a given tissue along with their relative activities of transcription, termed an expression profile, is not only of practical relevance to finding new genes, but also has biological significance. This is because the quantitative approach allows one to know the pattern of gene expression in the tissue in addition to sequence data.

Comparison of expression profiles: discovering tissue- or stage-specific genes

Collecting expression profiles from as many tissues as are available is advantageous, as has been suggested by Okubo et al. (2). By comparing two (or multiple) profiles, one can readily observe global differences in gene activities, and thus have a rational means for discovering cell-specific or stage-specific genes. This method may be termed 'subtraction in the database'. Common 'housekeeping' genes can be similarly identified. In contrast to liquid hybridization for subtraction, an expression profile, once made, can be compared with any other expression profile, so long as they are made in the same way. It can also be used to compare multiple samples for 'subtraction'. Comparison of normal tissues with affected (or malignant) tissues to identify up-regulated or down-regulated genes is another important application of this strategy. This comparison is not possible with sequence collections of normalized or partially selected cDNA libraries.

Body mapping: compiled expression profiles

By collecting and compiling expression profiles from as many different types of cells as possible, a body map of expressed genes man be made in which individual genes are assigned to the site(s) in the body where they are active. Such efforts may seem formidable, especially as the human body consists of some 60 trillion cells, but the basic types of tissues are estimated to number around 200. This approach may be extended to the examination of gene activities in subregions of a complex organ such as the brain and for describing gene activities during the course of development. As the amount of data collected expands, the reliability of this map will be improved.

Perspectives

The cDNA strategy has become an important way to discover interesting genes and is attracting an increasing number of researchers. This trend has placed a high demand on technological improvements in comprehensive data collections (e.g., refinement of automation, long-range sequencers and the handling of smaller samples). Of course, full-size sequencing will become an important issue as the sequencing capacity of equipment is improved.

Since the emergence of the controversy over the patenting of ESTs by the National Institutes of Health, much has been discussed concerning the rules of ownership of genes and unified standards for claiming property rights on the sequenced genes. The prevailing arguments are for free access to the collected partial cDNA sequence of human genes that are being stored in private databanks. Such collections of partial sequences may be surveyed over and over against sequence motifs, consensus sequences or gene sequences of interest. Once an interesting gene(s) is discovered, it may then be subjected to independent, intensive investigation. Even so, as discussed above, little information (other than the name of the tissue(s) in which a gene is expressed) can be retrieved by a researcher who queries his newly identified gene against a databank that collects only partial cDNA sequences and the records of similarity searches conducted without other relevant biological information.

We propose the following approach towards cDNA database construction in the future. Ideally, in addition to the nucleotide sequences and resulting amino acid sequences, a cDNA database should contain the following information: first, the tissue distribution of the expression of individual genes; second, an expression profile of genes in a particular tissue; third, the stage-specific, physiology-specific or disease-specific mode of expression of individual genes; fourth, *in situ* hybridization cytology; fifth, mapping of the gene on the genome; sixth, the effect of mutating the gene; and seventh, any corresponding disease. As seen above, the first three characteristics (which are relevant to gene expression control) can be obtained by the cDNA strategy for expression profiles and body mapping. At least some of the biological information should be available when a novel cDNA is checked with the database.

In the framework of current human genome efforts, the successful decoding of all the genetic information carried in the human genome will require the combination of structural analyses that has been discussed in the early part of this review and functional analyses. A cDNA database that combines sequences of genes, together with relevant biological information, will serve this purpose.

Lastly, I would like to add few words to this discussion on genome analyses. These efforts will never be restricted within the framework for understanding the mechanisms acting in the human body. Similar databases with many other organisms will emerge, keeping pace with the development of the human genome database. Then we will begin to understand the relationship of humans to other beings. In addition to the tremendous usefulness in the application of such information to our welfare, such progress, of necessity, will be able to deepen our understanding of the perpetuation and ever-changing forms of life. This will lead us to better insights into life.

REFERENCES

1 Adams MD, Kelley JM, Gocayne JD, Dubnick M, Polymeropoulos MH, Xiao H, Merril CR, Wu A, Olde B, Moreno RF, Kerlavage AR, McCombie WR and Venter JC: Complementary DNA sequencing expressed sequence tags and human genome project. Science **252**, 1651–1656 (1991)

2 Okubo K, Hori N, Matoba R, Niiyama T, Fukushima A, Kojima Y and Matsubara K: Large scale cDNA sequencing for analysis of quantitative and qualitative aspects of gene expression. Nature Genet **2**, 173–179 (1992)

3 Adams MD, Dubnick M, Kerlavage AR, Moreno R, Kelley JM, Utterback TR, Nagle JW, Fields C and Venter JC: Sequence identification of 2375 human brain genes. Nature **335**, 632–634 (1992)