FUNCTIONAL GENOMICS

Huang Caihong and Yang Qian*

Department of Life Science and Engineering,
Harbin Institute of Technology, Harbin 150001, P R China

ABSTRACT

Molecular biology has advanced a lot in recent years. So far, the complete genome sequences of several eukaryotes have been published. The first draft of the human genome has been completed ahead of schedule. Thus, much of the sequencing work is done. Now, the focus in genomic research is shifting from "structural genomics" towards "functional genomics" Functional genomics is a relatively new field of molecular biology that studies how genomic information defines the functions of proteins in living organism. It combines high-throughput experimental methodologies with statistical and computational analysis of the results to study genes or proteins in a systematic and systemic fashion. Study on bioinformatics is of crucial importance for our ability to exploit the wealth of information contained in the human genome for applied research. In this paper, we present information about genomic function and show several methods used in functional research.

Keywords: genomic function analysis methods

1. INTRODUCTION

The term "genome", first used by H. Winkler in 1920, was created by elision of the words GENes and chromosomes, and that is what the term signifies: the complete set of chromosomes and their genes. The term "genomics" was proposed by Thomas H. Roderick in 1986 for the discipline and for the journal named GENOMICS that was then being planned. Genomics can be divided into two aspects: structural genomics aimed at genome-scale sequencing and functional genomics aimed at functional identification of genome. Structural genomics stands for the early step of genome analysis while functional genomics is the new step, the latter can use the information supplied by the former and is characterized by high-throughput, big-scale experimental methods with statistical and computational analysis (Bond

et al., 2001). In 1990, while the human genome project was in progress, sequencing analysis of several species of prokaryotes has been completed. The research emphasis has shifted from disclosing all the genetic information to the study of gene function on a molecular scale.

Genome research is, in itself, multidisciplinary. We can anticipate commensurate growth in knowledge with greater understanding of complex genomes. Progress in laser technology, oligonucleotide synthesis, pattern recognition or cloning strategies will lead to cost curtailment (Bloker, 1995) Advances in computational and biological methods during the last decade have remarkably changed the scale of genome research. Sequencing machines and assembly algorithms enable sequencing of complete genomes within months and even weeks. (Hagit, 2002) This has resulted in a flood of sequenced and structural data which herald the advent of the post-genome era that is marked by functional genomics. Also the human genome project was catalyst for developing several high-throughput technologies, making it possible to map and sequence complex genomes. (Steffen, 2002) Nevertheless, knowing an organism's genome sequence is only an initial step in the quest to understand life's essential process. The research paradigm is shifting from genome sequencing and mapping toward the description of genomic and proteomic functions, that is to say, genomics is moving from a focus on structure to a focus on function.

Functional genomics surveys the dynamic biological function of genome; it changes the way of traditional biology and becomes the heated point of research in the field. Its research not only will be the most important part of whole cell biology in 21 century, but also can provide medicine, agriculture and industry with new ideas. In this paper we mainly discuss several methods used in functional genome research.

2. EXPERIMENTAL METHODS FOR IDENTIFICATION OF GENOMIC FUNCTION

Genome research project was resulted in new experimental approaches. It would be a challenge for bioinformatics to turn data into knowledge, i.e., integrate the ever-growing data to ascribe functions of proteins, cells, and ultimately organism. There is substantial research effort in assigning function to all identified open reading frames and orphan genes. (Bailey, 1999) Presently about 80 microbial genomes have been completely sequenced and a large number of microbial sequencing projects are in progress. Furthermore, besides the human genome several genomes of other higher eukaryotes have been fully sequenced and many genomic sequencing programs are in progress. Of all the genomes sequenced typically only about 50-70% of the sequenced genes have a known function and therefore much research is being done in the field of functional genomics. This has led to the development of many powerful analytical methods for analysis of cellular functions.

Experimental approaches to the analysis of gene function on a genomic scale are presented as follows:

2.1 Gene Expression Patterns

A great deal of a cell's functions can be determined from its pattern of gene expression. One can study the functional and toxological properties of a compound based on its effects on the gene expression of a particular cell or gene. The survey of gene expression is to compare the difference between different tissues, different growth stages, normal status and disease status. Many traditional methods such as RT-PCR, RNase protection and RNA hybrid can analyze one or several genes, however, new high-throughput technologies can process more.

Several methods have been developed for determining gene expression levels of different genes. Some of these methods for detecting and quantifying gene expression levels include Differential Display-CR (DD-PCR), Serial Analysis of Gene Expression (SAGE), S1 nuclease protection, DNA chips and Complementary DNA (cDNA) microarray. The expression of all genes in the genome may be measured through analysis of the mRNA pool. Different techniques have been applied for genome-wide transcription analysis, but the basic principles are the same-namely that the mRNA levels are quantified by nucleotide-nucleotide hybridization.

2.1.1 DNA Microarrays

cDNA microarrays were established by Brown's group at Stanford university in the United States. It is one of the major breakthroughs in genomic research and is a key element in today's functional genomics toolbox (Aharoni and Vorst, 2002). The power of the method lies in miniaturization, automation and parallelism permitting large-scale and genome-wide acquisition of quantitative biological information from multiple samples. cDNA microarrays are currently fabricated and assayed by two main approaches involving either in situ synthesis of oligonucleotides ('oligonucleotide microarrays') or deposition of pre-synthesized DNA fragments ('cDNA microarrays') on solid surfaces. To date, the main applications of microarrays are in comprehensive, simultaneous gene expression monitoring and in DNA variation analyses for the identification and genotyping of mutations and polymorphisms (Aharoni, 2002). cDNA microarray techniques make use of the information available in the public databases which contain more than one million expressed sequence tags (ESTs) representing 50-90% of all human genes. The microarray method for gene expression involves the arraying of cDNA probes (cDNAs of known sequences, representing genes or parts of genes) onto a glass slide and hybridization with fluorescently or radioactively labeled cDNA target sequences. Expression arrays containing up to 8000 genes are printed onto a 2×4 cm membrane or a microscope glass slide with a probe diameter of 75-100 μm and a 150μm distance between probes.(Whitchurch, 2002)

A microarray is a device that measures gene expressions across the entire genome and provides insight into the genomic behavior (Li, 2003). While microarrays are an extremely valuable tool, analysis of the data produced is a complex task. Depending on whether the processing data samples are labeled with classes, microarray data analysis can be broadly distinguished into unsupervised learning (Eisen *et al.*, 1998 and Tamayo *et al.*, 1999) and

supervised learning (Brown *et al.*, 2000). In the latter category, the problem of constructing a classifier based on gene expression profiles has gained increasing interest, following the success in demonstrating that gene expression on the genomic scale differentiated between two types of leukemia (Golub *et al.*, 1999). To build an optimal classifier from a small number of samples of known class with each sample described by thousands of gene expressions is a difficult problem because of the tendency to find a degenerate solution. Recent results have suggested the feasibility of constructing a classifier with reasonable predictive accuracy under this circumstance (Guyon *et al.*, 2002).

2.1.2 Serial Analysis of Gene Expression (SAGE)

SAGE developed in 1995 by Velculescu, Vogelstein and Kinzler at Johns Hopkins University (Baltimore, MD, USA), is a sequence-based genomics tool that features comprehensive gene discovery and quantitative gene expression capabilities. As an 'open' system, SAGE can reveal which genes are expressed and their level of expression rather than merely quantifying the expression level of a predetermined, and presently incomplete, set of genes as carried out by 'closed' system gene expression profiling platforms such as microarrays. These distinguishing attributes enable SAGE to be used as a primary discovery engine that can characterize human disease at the molecular level while illuminating potential targets and markers for therapeutic and diagnostic development, respectively. (Madden *et al.*, 2000) This method allows the quantitative and simultaneous analysis of a large number of transcripts. In this strategy, short diagnostic sequence tags were isolated, concatenated, and cloned.

The principle is that gene expression profile for a given cell is the list of all expressed gene, together with each gene expression level defined as the number of cytoplasmic mRNA transcripts in the cell. It is also a method to collect gene expression data in which complete segments of mRNA are sequenced instead of just a 10 base tag. However, with this method only a small amount of data can be gathered at very high cost. The SAGE method was introduced in 1995. The same group proposed that the SAGE method could be used to study the differences between cancerous and normal cells. (Zhang *et al.*, 1997)

2.2 Proteome Analysis (or Proteomics)

Most of the life activity of cell occurs at protein level rather than at RNA level, the integrated function of gene is due to the performance of proteins coded by these gene. Proteome analysis involves two steps: protein isolation and protein identification. All the proteins within the cell are analyzed using 2D electrophoresis. Through analysis of the individual spots on the gels the proteins can be analyzed in further detail using mass spectrometry (typically MALDI-TOF MS), e.g. with respect to the degree of phosphorylation or other post-translational modifications. Proteome analysis can be a valuable complement of cDNA microarrays and EST analysis.

2.3 Protein-protein Interactions

The interaction between different proteins is analyzed using the yeast two hybrid system, techniques based on fluorescence resonance energy transfer (FRET), etc. (Schwikowski, 2000; Uetz *et al.*, 2000; Ito, *et al.*, 2001; Gavin *et al.*, 2002) This has even been scaled-up by spotting 6000 yeast transformants on an array, and thereby the possible interaction between 192 *Saccharomyces cerevisiae* proteins with all the other possible *S. cerevisiae* proteins could be investigated (the interactome) (Uetz *et al.*, 2000). This approach is particularly important for unraveling signal transduction pathways that are believed to play a role in connection with overall regulation of cell function. In addition, one-hybrid system, three-hybrid system and reverse hybrid system can be used in protein-protein analysis. (Zhao *et al.*, 2000)

2.4 Protein-DNA interactions

The interaction of proteins with genomic DNA is characterized using DNA arrays. (MacBeath and Schreiber, 2000) Using a recently developed methodology for the analysis of genome-wide protein-DNA interactions 200 new putative transcription factors were identified in *S. cerevisiae* (Iyer *et al.*, 2001).

2.5 Metabolome Analysis

A large number of intra-cellular metabolite concentrations can be quantified, typically using GC-MS or LC-MS. Using GC-MS it is possible to measure 150 intracellular metabolites (Rosessner *et al.*, 2000), and through the use of this kind of data it is possible to reveal the phenotype of otherwise silent mutations (Raamsdonk *et al.*, 2001).

2.6 Metabolic Flux Analysis

The flows of carbon through the different metabolic routes within the cell are quantified. The fluxes can be quantified using simple metabolite balancing, but a more robust estimation of the fluxes is obtained through the use of 13C-labeled substrates followed by analysis of the labeling patterns of intracellular metabolites (Raamsdonk *et al.*, 2001; Gombert, 2001; Maaheimo, 2001; Wiechert, 2001).

2.7 Genetics Analysis

2.7.1 Gene Disruption through Transposon Insertion or Deletion

Mutantion is one of the most effective tools of functional analysis. Through insertion or deletion of transposon we can create a new individual. The combination between reported gene and mutant can be used to identify the expression of gene in the inserted position, so that we can select special cell and the individual expressed by gene under special environment, then we can isolate the special gene.

2.7.2 Reverse Genetics Technology

In recent years, RNA interference technology and synthesis of antisense nucleic acid have attract scientists' attention. The introduction of ds RNA into cells can interfere with the expression of an ensogenous gene homologous to the ds RNA, which results in the loss-of-function phenotype. This response is called RNA interference. The machanism of RNAi has been brought to light by the identification of RISC comlex and an enzyme named Dicer. As a reverse genetic tool, RNAi can be used to isolate loss-of-function or reduction-of-function mutants and to identify the biological function of genes. Therefore, RNAi can be applied for functional genomic analysis on a large scale (Liu, 2002).

As inhibitor, antisense nucleic acid can inhibit gene expression by combination with the nucleic acid in cell and form cross molecule, so that it can adjust the gene expression at transcription and translation level (Wang, 2002).

2.8 Comparative Genomics Analysis

By comparing gene sequences we can collect information concerned with gene function. Expressed Sequence Tag (EST) and Analysis method of combination with motif and COG (Chen, *et al.*, 2002) can do better nowadays.

Currently existing collections of expressed sequence tags (ESTs) are growing in size and are proving to be useful for various computational studies. ESTs offer a rapid route to gene identification (Adams, 1991 and 1992; Boguski, 1994; Okubo, 1991) analysis of expression and regulation data (Vasmatzis, 1998), and reconstruction of untranslated regions (Gautheret, 1998), and highlight multigene family diversity and gene alternative splicing (Brett, 2000; Kan, 2001; Mironov, 1999; Wolfberg and Landsman, 1997). EST matches may identify more than half of the known human genes (Hillier *et al.*, 1996). By aligning EST compared with BLAST we can find the similar structure and the similar function. As for the major part gene fragments unknown, cloning its full-length cDNA and analysis later. There are many ways to amplify genes. Rapid Amplification of cDNA Ends (RACE) is fast and simple (Zhu *et al.*, 2001).

2.9 Informatics Analysis

A common feature of the tools described above is that they give information about cellular processes at the global level, i.e. the complete mRNA pool is measured or a large fraction of the total protein pool is measured. This information supplies information of all the molecular interactions in the cell, but obviously it is difficult to reconstruct these interactions from the raw data. Advanced computational techniques therefore play an important role in connection with this type of analysis, and bioinformatics is consequently positioned as the focal point of functional genomics (Nielsen *et al.*, 2002). At the interface between biology, medicine, methematics, and computer science (Dalevi, 2001), bioinformatics (the information

technology infrastructure, databases, and software for the life science) (Gatto , 2003), provide great value for the research of the post-genomics, help to access and analyse the growing databases of experimental results and speed the identification and validation of gene targets and possibly shorten the time to identify their function.

3. APPLICABLE RANGE OF METHODS

Each of the methods described aim at defining gene function from different angles; therefore, each has its own strengths and weaknesses. Thus, we should choose different methods to solve different problems. cDNA microarrays measure messenger RNA (mRNA) expression on a genomic scale under various conditions, and thus indirectly indicate each gene's involvement in certain biological processes, and which genes may have related cellular function. Yeast two-hybrid assays explore protein–protein interactions in a pairwise fashion, whereas TAP tagging is useful for detecting protein complexes of two or more proteins. Proteome chip can measure both the biochemical activity of proteins and the interaction of proteins with other molecules, such as other proteins, metabolites, or drugs. All these approaches aim to elucidate gene function in terms of molecular interactions, the caveat being that the experimental systems do not exactly mimic physiological conditions; therefore, the results obtained may not agree with individual in vivo assays. Gene disruption measures the resulting phenotype following disablement of each gene and thereby explicates gene function in terms of physiological activity of the organism. However, this is applicable to only a subset of genes whose interruption causes discernible phenotypic changes. Computational approaches can be used to organize the genome-scale data into clusters of functionally related genes or to indicate the involvement of genes in certain biological processes, whereas the precise function of a gene needs to be determined through either individual experimental assays or homology-based prediction. (Lan *et al.*, 2002)

REFERENCES

Adams, M.D., J. M. Kelley *et al.*, Complementary DNA sequencing: Expressed sequence tags and human genome project, *Science*, *252*, **1991**, 1651-1656.

Adams, M.D., M. Dubnick, A. R. Kerlavage, *et al.*, Sequence identification of 2,375 human brain genes, *Nature*, *355*, **1992**, 632–634.

Aharoni, A., O.Vorst, DNA microarrays for functional plant genomics. *Plant Mol. Biol.*, *48*(1-2), **2002**, 99-118.

Bailey, J. E., Lessons from metabolic engineering for functionl genomics and drug discovery, *Nat. Biotechnol.*, *17*, 1999, 616-618.

Blocker, H., Genome research and molecular biotechnology, *Journal of Biotechnology*, *41*, **1995**.

Boguski, M.S., C. M. Tolstoshev, and D. E. Bassett Jr., Gene discovery in dbEST, *Science*, *265*, **1994**.

Bond, U., S.G.Campbell *et al.*, A model organism for genomic and postgenomic studies, *Engineering in medicine and biology*, *7-8*, **2001**, 22-32.

Brett, D., J. Hanke, *et al.*, EST comparison indicates 38% of human mRNA's contain possible alternative splice forms, *FEBS Lett.*, *474*, **2000**, 83–86.

Brown, M.P., W. N. Grundy, D. Lin *et al.*, Knowledge-based analysis of microarray gene expression data by using support vector machines, *Proc. Nat. Acad. Sci. USA*, *97*, **2001**, 262–267.

Chen, T., X. R. Liao, J. F. Du, *et al.*, The anvancement of functional genomics research, *Biotechnology Bulletin*, *2*, 2000, 1-6.

Dalevi, D. and S.G.E. Andersson, Discovering the dynamics of microbial genomes, Engineering in Medicine and Biology, July/August, **2001**, 55-60.

Eisen, M.B., P. T. Spellman, P. O. Brown, and D. Botstein, Cluster analysis and display of genome-wide expression patterns, *Proc. Nat. Acad. Sci. USA*, *95*, **1998**, 14863–14868.

Gatto, J.G., The changing face of bioinformatics, *Drug Discov Today*, *8*(9), **2003**, 375-376.

Gautheret, D., O. Poirot, F. Lopez, S Audic, and J. M. Claverie, Alternate polyadenylation in human mRNAs: A large-scale analysis by EST clustering, *Genome Res.*, *8*(5), **1998**, 524–530.

Gavin, A. C. *et al.*, Functional organization of the yeast proteome by systematic analysis of protein complexes, *Nature*, *415*, **2002**, 141–147.

Golub, T.R., D. K. Slonim, P. Tamayo *et al.*, Molecular classification of cancer: class discovery and class prediction by gene expression monitoring, *Science*, *286*, **1999**, 531-537.

Gombert, A.K., M.M. dos Santos, B. Christensen, and J. Nielsen, J., Network identification and flux quantification in the central metabolism of Saccharomyces cerevisiae at different conditions of glucose repression. *J. Bacteriol. 183*,**2001**, 1441-1451.

Guyon, I., J. Weston, S. Barnhill, and V. Vapnik, Gene selection for cancer classification using support vector machines, *Machine Learning*, *46*, 2002, 389-422.

Hillier, L., N. Clark, T. Dubuque, *et al.*, Generation and analysis of 280000 human expressed sequence tags, *Genome Res.*, *6*, **1996**, 807-828.

Ito, T., T. Chiba, R. Ozawa, M. Yoshida, M. Hattori, and Y. Sakaki, A

comprehensive two-hybrid analysis to explore the yeast protein interactome, in *Proc. Nat. Acad. Sci. USA*, *98*, **2001**, 4569–4574.

Iyer, V.R., C.R. Horak, *et al.*, Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF. *Nature*, *409*, **2001**, 533-538.

Kan, Z., E. C. Rouchka, W. R. Gish, and D. J. States, Gene structure prediction and alternative splicing analysis using genomically aligned EST's, *Genome Res.*, *11*, **2001**, 889–900.

Lan, N., J. Ronald *et al.*, Toward a systematic definition of protein function that scales to the genome level: defining function in terms of interactions, *Progressings of the IEEE*, *90*(12), **2002**, 1848-1857.

Li, M.F., and S.Y. Eun, Improving Reliability of Gene Selection From Microarray Functional Genomics Data, *Transactions on Information Technology in Biomedicine*, *7*(3), **2003**, 191-196.

Liu, M., Y. Cao, Y. Jiang, Large scale functional genetics analysis with RNAi, *Chinese Bulletin of Botany*, **2002**, 19(4), 491-495.

Maaheimo, H., J. Fiaux, Z.P. Cakar *et al.*, Central carbon metabolism of Saccharomyces cerevisiae explored by biosynthetic fractional ^{13}C labeling of common amino acids. Eur. J. Biochem. *268*, **2001**, 2464-2479.

MacBeath, G., and S. L. Schreiber, Printing proteins as microarrays for high throughput function determination, *Science*, *289*, **2000**, 1760–1762.

Madden, S.L., C.J. Wang, and G. Landes, Serial analysis of gene expression: from gene discovery to target identification, *Drug Discovery Today*, 5(9), **2000**, 415-425.

Mironov, A. A., J. W. Fickett and M. S. Gelfand, Frequent alternative splicing of human genes, *Genome Res.*, *9*, **1999**, 1288–1293.

Nielsen, J. *et al.*, An expanded role for microbial physiology in metabolic engineering and functional genomics : moving towards systems biology, *FEMS Yeast Research*, *2*(2), **2002**, 175-181.

Okubo, K., H. Hori, *et al.*, A novel system for large-scale sequencing of cDNA by PCR amplification, *DNA Sequence*, *2*, **1991**, 137-144.

Raamsdonk, L.M. *et al.* A functional genomic strategy that uses metabolome data to reveal the phenotype of silent mutations. *Nat. Biotechnol. 19*, **2001**, 45-50.

Roessner, U., C. Wagner, J. Kopka *et al.*, Simultaneous analysis of metabolites in potato tuber by gas chromatography-mass spectrometry. *Plant J. 23*, **2000**, 131-142.

Schwikowski, B., P. Uetz, and S. Fields, A network of protein –protein interactions in yeast, *Nature Biotechnol.*, *18*, **2000**, 1257–1261.

Shatkay, H., E. Stepthen, B. Mark, Information retrieval meets gene analysis, *Intelligent system in biology*, March/April, **2002**, 45-53.

Staab, S., Mining Information for Functional Genomics, *IEEE Intelligent systems*, May/July, **2002**, 66-80.

Tamayo, P., D. Slonim, J. Mesirov *et al.*, Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation, *Proc. Nat. Acad. Sci. USA*, *96*, **1999**, 2907–2912.

Uetz, P. *et al.*, A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. *Nature*, *403*, **2000**, 623-627.

Uetz, P. *et al.*, A comprehensive analysis of protein–protein interactions in *Saccharomyces cerevisiae*, *Nature*, *403*, **2000**, 623–627.

Vasmatzis, G., M. Essand, et al., Discovery of three genes specifically expressed in

human prostate by expressed sequence tag database analysis, *Proc. Nat. Acad. Sci.*, *95*, **1998**, 300–304.
Wang, Z., Z.Z. Mei, Z.X. Sun *et al.*, Antisense nucleic acid—the valuable tool of study genomic function, *Foreign Medicine-Genetics*, 25(1), **2002**, 1-3.
Whitchurch, A.K., Gene expression microarrays, *Protentials*, Feb./Mar., **2002**, 30-34.
Wiechert, W., ^{13}C Metabolic flux analysis. *Metab. Eng. 3*, **2001**, 195-206.
Wolfberg, T.G. and D. Landsman, A comparison of expressed sequence tags (EST's) to human genomic sequences, *Nucleic Acids Res.*, *25*, **1997**, 1626-1632.
Yang, Q, J.Z Song *et al.*, A study on biocontrol mechanism of *Chaetomium* spp, Advanced study on plant pest biological comtrol. Heilongjiang science and technology press, **2000**, 110-115.
Yang, Q, L. Mei *et al.*, Methods of transforming resistance gene to benzimidazole funficides into *Trichoderma harzianum* and *Chaetomium globosum*, Biological control and bio-technology, Heilongjiang science and technology press, **2003**, 1-8.
Yujie, Chi, Q. Yang, Construction and identification of recombinant plasmid of TUB2 gene and transformation into *Chaetomium* sp., *High Technology Letters*, *2*, **2003**, 34-40.
Zhang, L., W. Zhou, V. E. Velculescu, *et al.*, Gene expression profiles in normal and cancer cells, *Science*, *276*(5316), **1997**, 1268-1272.
Zhao, J. H., X. Q. Wang *et al.*, Cotent and methods of functional genomics, *Prog. Biochem. Biophys.*, *27*(1), **2000**, 6-8.
Zhu, Y.Y., E.M. Michleder, *et al.*, Reverse transcriptase template switching: A SMART approach for full-length cDNA library construction, *30*(4), **2001**, 892-897.