

Genomic analysis

Pat S Lee and Kelvin H Lee

Advances in genomic analysis include improved technology for DNA sequencing, routine use of DNA microarray technology for the analysis of gene expression profiles at the mRNA level and improved informatic tools to organize and analyze such data. At the same time, new developments in chip-based analysis of samples and the emergence of models of gene networks hold promise for the future of the 'Genomic Era'.

Addresses

School of Chemical Engineering, Cornell University, Ithaca, NY 14853-5201, USA
Correspondence: Kelvin H Lee; e-mail: khlee@cheme.cornell.edu

Current Opinion in Biotechnology 2000, 11:171–175

0958-1669/00/\$ – see front matter
© 2000 Elsevier Science Ltd. All rights reserved.

Abbreviations

EST expressed sequence tag
Mb megabasepairs
PCR polymerase chain reaction
SNP single nucleotide polymorphism

Introduction

The 'Genomic Era' has been characterized by developments occurring at a tremendous pace. One could argue that the Genomic Era was initially characterized by the ability to sequence whole genomes effectively and relatively quickly. Sequence data is catalogued in online databases and the challenge of many computer scientists and biologists is to annotate this sequence information using statistical and biological analysis. Indeed, this phase of the Genomic Era is growing rapidly and continues to be vitally important. A more recent and emerging phase of the Genomic Era involves the use of microarray patterns of cDNA or mRNA sequences, which can provide information about the simultaneous and relative expression — at the mRNA level — of the genes expressed by a genome. The hope is that such information can be used to elucidate parts of the genetic network and regulation that define a particular organism or biological pathway. It has become clear that a microarray approach or other mRNA-based approach is necessary, but not sufficient, to accomplish the intended goals of such analysis. Experimental evidence clearly shows a disparity between the relative expression levels of mRNAs and their corresponding proteins [1,2**]. Furthermore, it has recently been proven mathematically, that expression information from both mRNA and proteins is required to understand a gene network [3**]. A discussion of the relative role of gene expression profiling at the protein level to that at the mRNA level is beyond the scope of this review and is contained in an accompanying article (see Dutt and Lee this issue, pp 176–179). In parallel with the effort to generate genomic sequence and expression data is the need to develop an informatic framework to organize the acquired data, as well as to extract as

much information as possible from the raw data. Such bioinformatic efforts, which are aggressively pursued by engineers, statisticians and computer scientists, rely on statistical analysis of the raw sequence data and on clustering analysis (the grouping together of genes that show similar expression profiles) to yield an added layer of information to the raw data. The bioinformatics effort can be extended further, however, to the analysis and design of gene networks based on the synthesis of mathematical descriptions of biological processes, which is the realm of the applied mathematician, physicist and engineer. Here, we distinguish this extended bioinformatic effort as the 'generation of knowledge from information', and highlight some interesting examples of this approach. Because of space limitations in this review, we focus only on technological developments in the areas of genomics and urge the reader to consider other sources. We cover key aspects in microarray analysis but will focus primarily on developments in the past 1–2 years.

Genome sequencing

The past two decades have seen nucleic acid sequencing (DNA and RNA) evolve from complicated laboratory procedures performed by skilled technicians to automated, commonplace techniques available to biologists and non-biologists alike. The application of these techniques to the study of biological processes has spawned the field of genomics. Many of the new technologies arising from this field are relevant to virtually all areas of biological research because these tools aid in the interpretation of the genetic blueprints of life. The nature and quantities of genomic information being generated has also led to a shift toward a holistic approach to the study of biological systems, requiring the development of a plethora of new analysis tools.

This revolution has largely arisen with the development of the high-throughput DNA sequencers. These machines amplify genomic DNA fragments and determine their sequences using slab gel or capillary electrophoresis methods. In 1995, The Institute for Genomic Research (TIGR; Rockville, MD, USA) used these methods to complete the nucleotide sequence of the genome of *Haemophilus influenzae*, the first fully genome-sequenced free-living organism [4], which contains 1.83 megabasepairs (Mb). Recently, these machines have been used to sequence entire genomes of several unicellular organisms, such as the yeast *Saccharomyces cerevisiae* (12.05 Mb) [5] and the eubacteria *Escherichia coli* (4.64 Mb) [6] (sequence information for this much-studied organism is also available for use at <http://www.genome.wisc.edu>) by sequencing overlapping fragments and reconstructing the genomic sequences. Efforts are underway to sequence a large number of organisms, including mouse, zebrafish, rice, and human — the sequence of the first human chromosome (number 22) is

nearly complete. Currently, reconstruction of the chromosomal DNA sequence from short overlapping segments, particularly for large genomes, is the biggest challenge for computational biologists because of the large segments of repeating sequences present in most complex organisms.

At the same time that existing technology is generating information, new technology is under development. Miniaturization promises to reduce sample sizes and reaction times. Burns *et al.* [7**] and other groups have developed prototype chip sequencers that appear to be well on their way toward genomic applications. These small chip sequencers are capable of PCR amplification and capillary electrophoresis of the DNA fragments with minimal sample preparation. Others are developing high-throughput methods for DNA sequencing by exploring new dyes, different polymeric matrix materials, and alternative formats for capillary electrophoresis [8], which may simplify detection, decrease run times, and lower costs.

Other approaches to genomic sequence determination are also being pursued. Notably, single molecule sequencing methods [9] use exonucleases to degrade DNA by cleaving individual nucleotides from the 5'-end of a DNA molecule and detecting them. The most common methods use DNA segments made from nucleotides where the four bases have been differentially tagged with fluorescent labels; when the tagged nucleotide is clipped, it flows past a laser-based fluorescence detector [10]. Because the exonuclease is capable of cleaving continuously from long stretches of DNA (~50,000 bases) these methods reduce the need to sequence overlapping fragments and eliminates the cumbersome task of ordering short sequences.

Sequencing efforts are also directed toward compiling expressed sequence tags (ESTs) to assemble databases such as dbEST (<http://www.ncbi.nlm.nih.gov/dbEST/>). ESTs are identified using reverse transcriptase (RT) to create cDNA sequences (which have higher stability than mRNA) from mRNA present in a cell, allowing genes expressed in different tissues or environmental conditions to be easily amplified by PCR for further study [11]. This approach, however, provides only a snapshot of the mRNA sequences expressed at the time of sampling, and intron and regulatory DNA sequences cannot be determined using this method. Because ESTs might be only gene fragments (typically 3' or 5'), they are more easily generated than entire sequence information [11]. The ease of this approach has resulted in exponential growth in the number of sequences in cDNA libraries; to date, well over three million ESTs have been compiled in dbEST. A major challenge in probing EST databases is the relatively large database size and the high frequency of redundant gene sequences.

Recently, more attention has been focused on single nucleotide polymorphisms (SNPs), the most common types of stable genetic variations, because it has been found that these point mutations can produce different phenotypes

and can be contributory factors for human disease [12**]. Studies show that SNPs can cause major changes in the resulting mRNA structural folds, which might affect cell regulation [13]. SNP detection methods are diverse, and a thorough accounting is beyond the scope of this review. One notable approach uses matrix-assisted laser desorption ionization mass spectroscopy (MALDI-MS) on a silicon chip for SNP identification [14*]; another system employs electronic circuitry on silicon microchips capable of changing electrical polarity to produce a fluorescent signal with SNP detection. A high-throughput system has been developed that combines gel-based sequencing and oligonucleotide chip technology to identify SNPs [12**]. Like most SNP detection procedures, these methods require PCR amplification of the target sequence, a costly and time consuming step [15]. Using these and several other analytical methods, large numbers of SNPs are being identified and compiled into databases for use [12**].

Functional genomics and gene expression patterns

Databases of genetic information include vast collections of information and the size of these databases is growing exponentially. Traditional methods of analysis have attempted to reduce biological processes to their smallest components, such as genetic sequence information; with the success of these methods, these reductionist methods have become insufficient to analyze and integrate the large amount of information being generated. Functional genomics is primarily concerned with the determination of protein function and structure from the genetic sequence and requires the development of new tools and techniques. One approach examines the cell's use of sequence information to relate function and structure by examining the cellular RNA content. Stanford researchers have developed DNA microarray methods using cDNA tags deposited onto a glass slide in known locations by high speed printing methods [16]. Reverse transcription of mRNA from different cell populations produces cDNA that can be labeled with different fluorescent tags. These tagged cDNA fragments are allowed to hybridize to the cDNA on the chip and differences in mRNA expression between the cell populations can be examined. ESTs of the entire yeast genome have been printed onto a single chip and used to study diauxic growth [16] and regulation of sporulation [17]. One innovative application of this technology is its use as a tool for cancer classification of human leukemias based on gene expression profiles [18*].

A similar oligonucleotide chip method (Affymetrix, Inc., Santa Clara, CA, USA) synthesizes the cDNA sequences directly onto the glass slide using photolithographic masks to produce the correct sequences [19]. Recent advances in these techniques have the potential to make this procedure less expensive by using a digital micromirror array to form virtual masks [20]. Efforts are also being made to increase the yield of mRNA isolation procedures, thus reducing the cell population sizes required for analysis

[21]. As these microarray methods become faster and more affordable, mRNA expression levels for cell populations exhibiting different phenotypes can be investigated more readily. Already, these techniques are being applied to analyze gene expression in cancer, sleep, stress responses and other areas. Analogous macroarray procedures using nylon membranes as substrates (Genosys Biotechnologies, The Woodlands, TX, USA) have also been used to compare mRNA expression patterns of *E. coli* grown in minimal and rich media [22]. These methods use protocols similar to Southern blot procedures and thus the same tools can be used for their analysis. Results of many genome-wide analyses of gene expression patterns are available on the Internet (<http://cmgm.stanford.edu/pbrown> [yeast] and <http://www.genetics.wisc.edu> [*E. coli*]).

Serial analysis of gene expression (SAGE) is another high-throughput method very different from microarray technology. cDNA made from cellular mRNA is treated to create a single tag from each cDNA. The tag sequence (10–14 basepairs) can uniquely identify each transcript, and the concentration of each tag sequence is proportional to the level of mRNA in the original sample [23]. Thus, this method is much more quantitative than standard microarray methods because it eliminates the sequence-to-sequence variations in translation rate inherent in PCR. The tag sequences are ligated to long multimers, cloned, and sequenced. The sequences are long enough to identify the transcript in database searches. This important method is being used to explore gene regulation in a diverse array of cell populations, including human gastrointestinal cancer cells [23] and yeast cell-cycle regulation [24]. Many of these results are also readily available on the Internet (<http://www.ncbi.nlm.nih.gov/SAGE> [human]). One drawback to this method is the high level of initial mRNA required; however, researchers are optimizing the protocols to reduce these amounts to alleviate this problem [25].

The success of these methods has led to the development of other approaches to the study of gene expression on a genome-wide scale. A promising method developed by Aurora Biosciences (San Diego, CA, USA) uses gene trapping to randomly insert a promoterless β -lactamase reporter gene into introns, disrupting mammalian host genes. The fraction of cells that produce β -lactamase create a library of tagged clones that can be used to study gene expression in living cells [26]. Changes in gene expression levels of transfected cells were detected using a cell-permeable fluorogenic β -lactamase substrate that changes fluorescence emission from green (520 nm) to blue (450 nm) upon cleavage by the reporter enzyme. High gene expression levels result in rapid changes in fluorescence emission, whereas lower levels often require hours for detection of colour change. This method has been used to study pathways involved in T-cell activation in human T-cell clones [26]. Advantages of this system include its ease of use with fluorescence-activated cell sorting (FACS) and its sensitivity.

Despite the utility of these methods for investigating gene expression, studies in yeast [27] and human liver cells [1] have shown that protein expression levels are not well-correlated with mRNA expression levels. Furthermore, mathematical analysis has shown that both mRNA and protein expression information are essential for the creation of even simple gene network models to predict dynamic cell behavior [32]. Because protein rather than mRNA levels determine phenotype, there are efforts underway to investigate this difference by analyzing the translation state of mRNA [27]. Active mRNA is usually associated with multiple ribosomes, forming polysome complexes, whereas inactive mRNA is usually bound to a single ribosome or sequestered in messenger ribonucleoproteins. These two mRNA states are readily separated using sucrose gradient centrifugation; the fractions can then be identified using labeled cDNA probes and used to interpret data based on mRNA expression to estimate protein levels [27]. This approach provides one method for the study of translation of nucleotide sequence into proteins; however, as proteins are the primary directors of cell activity, genomics, the study of genetic sequence on a whole-cell level, must ultimately lead to the holistic study of protein levels and proteomics.

Bioinformatics

Although an incredible amount of genetic sequence information is now available, many of the genes identified have yet to be assigned a function. To begin this daunting task, the field of bioinformatics has embraced computational methods to supplement experimental approaches in relating genotype to phenotype. The major method of functional assignment clusters genes based on multiple sequence alignment to known proteins [28]. Other approaches to 'mining' of the genetic sequence include comparative genomics and phylogenetic profiling to relate proteins across species and evolutionary paths [29]. Domain-fusion analysis finds proteins composed of two or more proteins found separately in other genomes; the two separate proteins are then identified as potentially interacting protein species [30,31]. A method that combines phylogenetic profiles, gene expression patterns, and domain-fusion information to identify protein function is also being pursued [28]. These algorithms are also being applied toward the prediction of protein structure and folding from sequences. Thus far, these approaches have enjoyed only limited success due to the complexity of protein patterns and folding, and annotation of gene banks using these methods sometimes leads to flaws in function assignments. Computational analysis of sequence information is also being used to study transcription events by identifying promoters and regulatory sites in *E. coli* on a genome-wide basis [32].

One approach to a holistic look at biological processes uses the genomic information in stoichiometric models. Palsson and co-workers [33] have created mathematical models by integrating extensive information on metabolic pathways with the genetic sequences of its components to describe cellular processes and to link genotype and phenotype. The

exclusion of kinetic information in building these models makes them unsuitable, however, for studying the dynamic evolution of the cell system [34]. Incorporation of kinetics into stoichiometric models is often unfeasible because of complexity, and many of the transcriptional regulation and enzymatic control pathways are not well understood. In cases where detailed enzymatic or gene expression information is available, models have been created that can address dynamic system responses. Yin and co-workers [35,36] have augmented a kinetic model of bacteriophage T7 function with a gene-shuffling algorithm to examine dynamic genotype and phenotype relationships. As knowledge of enzymatic and transcriptional control increases, these models will become more important in studying the dynamic response of cells to stimuli.

Conclusions

The Genomic Era thus far has experienced a significant increase in available genetic sequence information for a wide variety of species; the challenge for scientists and engineers will be to interpret it to relate genotype to phenotype. Genomics as a field has been primarily involved in the development of new technologies to both generate and to make sense of whole genome sequence information. The first phase of the Genomic Era, generation of genomic information, has been made possible by the development of high-throughput DNA sequencing methods. The nature and size of genomic databases has necessitated a shift towards a holistic approach to biological research and the development of new analysis tools. These tools have required the application of a diverse range of competencies toward understanding this information. Computational methods are being applied to interpret and link genes in these databases by comparing and identifying sequence homology, evolutionary pathways, gene expression data, and domain-fusion proteins. A plethora of experimental methods have been developed to study mRNA expression levels on a genome-wide basis, and thus increase our understanding of transcriptional regulation of the cell. Advances in these methods promise to reduce complexity and cost, allowing comparison of gene expression patterns for diverse cell populations on a nucleotide level. The partnership of such information with proteomic data will lead to improved mathematical descriptions of gene networks and their underlying regulation.

Acknowledgements

This work was supported in part by the National Science Foundation (BES-9874938), Intel (98-238), The New York State Science and Technology Foundation, and DuPont.

References and recommended reading

Papers of particular interest, published within the annual period of review, have been highlighted as:

- of special interest
- of outstanding interest

1. Anderson L, Seilhamer J: **A comparison of selected mRNA and protein abundances in human liver.** *Electrophoresis* 1997, **18**:533-537.

2. Gygi SP, Rochon Y, Franz A, Aebersold R: **Correlation between protein and mRNA abundance in yeast.** *Mol Cell Biol* 1999, **19**:1720-1730.

A large-scale investigation into the disparity of mRNA and protein levels in yeast.

3. Hatzimanikatz V, Lee KH: **Dynamical analysis of gene networks requires both mRNA and protein expression information.** *Metab Eng* 1999, **1**:275-281.

This paper uses nonlinear stability analysis to show that mRNA expression data alone is insufficient for the creation of simple gene network models. This emphasizes the need to study proteomics in addition to genomics.

4. Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, Kerlavage AR, Bult CJ, Tomb JF, Dougherty BA, Merrick JM *et al.*: **Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd.** *Science* 1995, **269**:496-512.

5. Goffeau A, Barrell BG, Bussey H, Davis RW, Dujon B, Feldman II, Galibert F, Hoheisel JD, Jacq C, Johnston M *et al.*: **Life with 6000 genes.** *Science* 1996, **274**:563-567.

6. Blattner FR, Plunkett G III, Bloch CA, Perna NT, Burland V, Riley M, Collado-Vides J, Glasner JD, Rode CK, Mayhew GF *et al.*: **The complete genome sequence of *Escherichia coli* K-12.** *Science* 1997, **277**:1453-1462.

7. Burns MA, Johnson BN, Brahma Sandra SN, Handique K, Webster JR, Krishnan M, Sammarco TS, Man PM, Jones D, Heldinger D *et al.*: **An integrated nanoliter DNA analysis device.** *Science* 1998, **282**:484-487.

Groundbreaking research proving the effectiveness and speed of miniaturization as applied to integrated DNA sequencing.

8. Schmalzing D, Koutny L, Salas-Solano O, Adourian A, Matsudaira P, Ehrlich D: **Recent developments in DNA sequencing by capillary and microdevice electrophoresis.** *Electrophoresis* 1999, **20**:3066-3077.

9. Weiss S: **Fluorescence spectroscopy of single biomolecules.** *Science* 1999, **283**:1676-1683.

10. Service RF: **Deconstructing DNA for faster sequencing.** *Science* 1999, **283**:1668-1669.

11. Boguski MS, Schuler GD: **ESTablishing a human transcript map.** *Nat Genet* 1995, **10**:369-371.

12. Wang DG, Fan J-B, Siao C-J, Berno A, Young P, Sapolsky R, Ghandour G, Perkins N, Winchester E, Spencer J *et al.*: **Large-scale identification, mapping and genotyping of single-nucleotide polymorphisms in the human genome.** *Science* 1998, **280**:1077-1082.

An innovative approach to the problem of high-throughput SNP identification and mapping.

13. Shen LX, Basilion JP, Stanton VP Jr: **Single-nucleotide polymorphisms can cause different structural folds of mRNA.** *Proc Natl Acad Sci USA* 1999, **96**:7871-7876.

14. Tang K, Fu D-J, Julien D, Braun A, Cantor CR, Köster H: **Chip-based genotyping by mass spectrometry.** *Proc Natl Acad Sci USA* 1999, **96**:10016-10020.

A novel approach combining miniaturization and mass spectrometry to create an effective method for the identification of SNPs on a large scale.

15. Gilles PN, Wu DJ, Foster CB, Dillon PJ, Chanock SJ: **Single nucleotide polymorphic discrimination by an electronic dot blot assay on semiconductor microchips.** *Nat Biotechnol* 1999, **17**:365-370.

16. DeRisi JL, Iyer VR, Brown PO: **Exploring the metabolic and genetic control of gene expression on a genomic scale.** *Science* 1997, **278**:680-686.

17. Chu S, DeRisi J, Eisen M, Mulholland J, Botstein D, Brown PO, Herskowitz I: **The transcriptional program of sporulation in budding yeast.** *Science* 1998, **282**:699-705.

18. Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA *et al.*: **Molecular classification of cancer: class discovery and class prediction by gene expression monitoring.** *Science* 1999, **286**:531-537.

A novel use of microarray technology as applied to the classification of human leukemia.

19. Wodicka L, Dong H, Mittmann M, Ho MH, Lockhart DJ: **Genome-wide expression monitoring in *Saccharomyces cerevisiae*.** *Nat Biotechnol* 1997, **15**:1359-1367.

20. Singh-Gasson S, Green RD, Yue Y, Nelson C, Blattner F, Sussman MR, Cerrina F: **Maskless fabrication of light-directed oligonucleotide**

- microarrays using a digital micromirror array. *Nat Biotechnol* 1999, **17**:974-978.
21. Mahadevappa M, Warrington JA: **A high-density probe array sample preparation method using 10- to 100-fold fewer cells.** *Nat Biotechnol* 1999, **17**:1134-1136.
 22. Tao H, Bausch C, Richmond C, Blattner FR, Conway T: **Functional genomics: expression analysis of *Escherichia coli* growing on minimal and rich media.** *J Bacteriol* 1999, **181**:6425-6440.
 23. Zhang L, Zhou W, Velculescu VE, Kern SE, Hruban RH, Hamilton SR, Vogelstein B, Kinzler KW: **Gene expression profiles in normal and cancer cells.** *Science* 1997, **276**:1268-1272.
 24. Velculescu VE, Zhang L, Zhou W, Vogelstein J, Basrai MA, Bassett DE Jr, Hieter P, Vogelstein B, Kinzler KW: **Characterization of the yeast transcriptome.** *Cell* 1997, **88**:243-251.
 25. Datson NA, van der Perk-de Jong J, van den Berg MP, de Kloet ER, Vreugdenhil E: **MicroSAGE: a modified procedure for serial analysis of gene expression in limited amounts of tissue.** *Nucleic Acids Res* 1999, **27**:1300-1307.
 26. Whitney M, Rockenstein E, Cantin G, Knapp T, Zlokarnik G, Sanders P, Durick K, Craig FF, Negulescu PA: **A genome-wide functional assay of signal transduction in living mammalian cells.** *Nat Biotechnol* 1998, **16**:1329-1333.
 27. Zong Q, Schummer M, Hood L, Morris DR: **Messenger RNA translation state: the second dimension of high-throughput expression screening.** *Proc Natl Acad Sci USA* 1999, **96**:10632-10636.
- The first known study of the translational state of mRNA. This paper represents an important step in understanding the link between gene expression and protein level.
28. Marcotte EM, Pellegrini M, Thompson MJ, Yeates TO, Eisenberg D: **A combined algorithm for genome-wide prediction of protein function.** *Nature* 1999, **402**:83-86.
 29. Pellegrini M, Marcotte EM, Thompson MJ, Eisenberg D, Yeates TO: **Assigning protein functions by comparative genome analysis: protein phylogenetic profiles.** *Proc Natl Acad Sci USA* 1999, **96**:4285-4288.
 30. Marcotte EM, Pellegrini M, Ng H-L, Rice DW, Yeates TO, Eisenberg D: **Detecting protein function and protein-protein interactions from genome sequences.** *Science* 1999, **285**:751-753.
 31. Enright AJ, Iliopoulos I, Kyrpides NC, Ouzounis CA: **Protein interaction maps for complete genomes based on gene fusion events.** *Nature* 1999, **402**:86-90.
 32. Thieffry D, Salgado H, Huerta AM, Collado-Vides J: **Prediction of transcriptional regulatory sites in the complete genome sequence of *Escherichia coli* K-12.** *Bioinformatics* 1998, **14**:391-400.
 33. Schilling CH, Edwards JS, Palsson BO: **Toward metabolic phenomics: analysis of genomic data using flux balances.** *Biotechnol Prog* 1999, **15**:288-295.
 34. Varner J, Ramkrishna D: **Mathematical models of metabolic pathways.** *Curr Opin Biotechnol* 1999, **10**:146-150.
 35. Endy D: **Development and application of a genetically structured simulation for bacteriophage T7 [PhD Thesis].** Hanover, NH: Dartmouth College; 1997.
 36. Endy D, Kong D, Yin J: **Intracellular kinetics of a growing virus: a genetically structured simulation for bacteriophage T7.** *Biotechnol Bioeng* 1997, **55**:375-389.