Bioinformatics and Functional Genomics: Challenges and Opportunities

Vassily Hatzimanikatis
Chemical Engineering Dept., Robert R. McCormick School of Engineering and Applied Sciences,
Northwestern University, Evanston, IL 60208

Biological systems—single cells, multicellular tissues, and multitissue organisms—are complex entities. From an engineering point of view, they consist of a large number of physicochemical and mechanical processes that operate in parallel and in series. Under-standing the function of the individual processes and their interactions is critical for advances in drug discovery, as well as in industrial and agricultural biotechnology. The combined annual R&D expenditures of these activities are over $10 billion accounting for more than 60% of those of the chemical and allied products sector (Ernst and Young, 1999; *C&EN*, 1999).

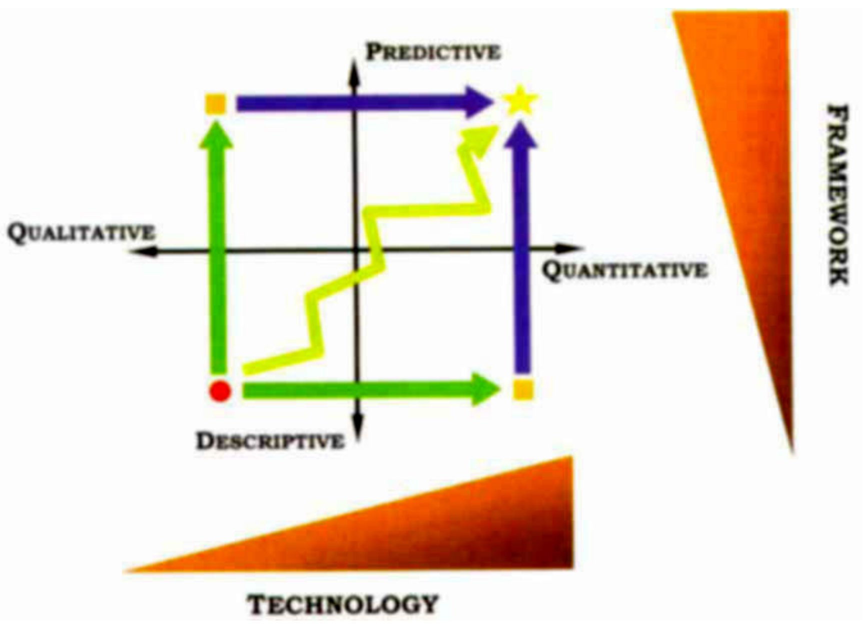

Figure 1: From qualitative description to quantitative prediction.
The ultimate goal of bioinformatics and functional genomics is a quantitative prediction of cellular function based on DNA sequence and a small number of experiments. The achievement of this goal will require developments in analytical technology and in analysis frameworks. As we move towards quantification of every cellular component and of the thermodynamic and kinetic properties of the various processes in a cell, technology becomes harder and innovative developments are required. The supporting mathematical and computational frameworks required for the integration of this information become more sophisticated as one moves from descriptive to predictive understanding. While the final objective is well-defined, there is not a unique path towards the accomplishment of this objective. Technological innovations will drive the development of analysis frameworks, and creative analysis frameworks will identify the technology needs.

Chemical engineers have been contributing to many aspects of biotechnology research and development. In his review of these contributions, Bailey (1995, 1998) suggested that particular virtues of our field have allowed successful "chemical engineering of cellular processes." Biological sciences and every field in biotechnology are experiencing an ongoing revolution. It is an *information* revolution driven by advances in analytical technology, biochemistry, nanotechnology, polymer chemistry, and material science. These technologies enable the precise and quantitative characterization of various molecules within a cell and the monitoring of many cellular processes simultaneously. This revolution in biology offers two main challenges to chemical engineers: contribution to technology development and meaningful analysis of the large-scale information generated by these technologies.

This column offers a perspective on the challenges posed by *bioinformatics* and *functional genomics. Bioinformatics* here is defined as the development and implementation of computational tools and frameworks for the management, analysis and interpretation of biological information, and *functional genomics* is defined as the integration of analytical technologies and bioinformatics for the elucidation of the function of cellular components. Capabilities and limitations of some of the current technologies used in studying cellular function are discussed, as well as qualitative similarities between problems in chemical engineering and bioinformatics.

Bioinformatics and functional genomics

Three central processes that support cellular function are transcription, translation, and biotransformation (Alberts et al., 1994), which are strongly interdependent. The products of one process serve as catalysts or reactants for another process, and the products or reactants of a process can regulate the performance of other processes in the cellular system. A fourth process, signal transduction, is responsible for the coordinated responses of the cellular function to environmental changes and stimuli. The structural and functional information on the components of various processes is encoded into polydeoxyribonucleic acid (DNA) sequences, which themselves are assembled into single molecules. The set of DNA molecules within a single cell is called a genome. *Functional genomics* has emerged as a research field that aims to map DNA sequences and the components they encode for to the function they perform within various cellular processes.

Table 1: Qualitatively similar problems in Chemical Engineering and Biology.

Chemical Engineering	Biology
Process Identification	Interpretation of mRNA and protein expression data
Process Control	Enzyme regulation Transcription and Translation regulation
Polymer Processes	mRNA synthesis Protein synthesis
Polymer Design	Protein Structure Protein Folding
Chemical reaction networks Atmospheric Chemistry Polymer Chemistry Catalysis	Cellular reaction networks Integrated cellular processes
Fluid mechanics Transport Phenomena	Development Morphogenesis Cell-to-cell communication
Thermodynamics	Physicochemical properties of cellular components Cellular energetics

Modern technologies have allowed high-throughput generation of information about the DNA sequence of the genomes of an organism, the quantitative monitoring of its RNA and protein molecules, the identification of protein-protein and DNA-protein interactions, and the mapping of the genetic variations within a population. Computer science, statistics, and biology have given birth to *bioinformatics* as a new discipline that is concerned with the efficient management and useful interpretation of large-scale biological information. Early efforts in bioinformatics were focused on the analysis of DNA sequence data. They involved the design and integration of DNA sequence databases, the alignment of protein and DNA sequences, the assembly of DNA sequence fragments into genomic maps, and the prediction of the function of a gene based on comparison of its sequence with sequences of genes with known function. The accelerating generation of data from various sources and for various cellular processes, the ever-changing analytical technologies, and the increasing computational power broaden the scope and objectives of bioinformatics research and development. Prediction of protein structure, image analysis, data visualization methods, analysis of gene and protein expression data, and simulation and dynamic analysis of integrated cellular processes are some of the growing activity areas of bioinformatics.

Every gene and its products, poly-ribonucleic acid (RNA) and protein, can be classified according to their *elementary functions* and their *systemic functions*. For example, the type of reaction, such as dehydrogenation, that a protein (enzyme) catalyzes is its elementary function, whereas its reactants and products characterize its systemic function, i.e., its function within a system of reactions. The identification of the elementary functions of genes has been a main field of study in biology and biochemistry research. Problems in drug discovery, metabolic engineering, agricultural biotechnology, and molecular biology have created the need for understanding the systemic properties of genes. Identification of these intertwined functions is essential, and bioinformatics offers an indispensable technology for function assignment.

Function assignment: technologies and constraints

The sequences of the whole genome of over 50 organisms have been published (http://www.ncbi.nlm.nih.gov), and the sequencing of the genome of more than 300 organisms is in progress. An increasing number of private companies are also investing in sequencing the genomes of their proprietary organisms, such as bacteria and plants, or of organisms involved in infectious diseases, such as pathogenic bacteria and viruses. The DNA sequence information is growing exponentially, and it is doubling almost every year. Each new DNA sequence is computationally analyzed to determine the regions of sequence that *might* encode for proteins (genes) and the regions that might be responsible for the regulation of transcription (Quellette and Baxevanis, 1998). The next step is prediction of function of the various genes. This is done using pairwise alignment algorithms that compare the DNA sequence of new genes with DNA sequences of known functions from other organisms. The output of these computational analyses is a *similarity index* between every new gene and every gene in the databases.

The extent of similarity between two sequences is based on the sequence identity (i.e., percent of invariant nucleotides or amino acids) and/or conservation (i.e., changes at a specific position of an amino acid that preserve the physicochemical properties of the original residue). Thus, if a new gene is similar to a gene of known function, the new gene *might* perform the same function. This genome-wide analysis that assigns a possible function to every new gene is called *genome annotation.* In every new genome, however, 30–40% of the genes cannot be assigned a function with confidence using the current computational approaches. Moreover, a significant percentage of the unassigned genes do not share significant similarity with any other gene in the databases. This implies that, as the sequence information is growing exponentially, the number of genes of unknown function is also growing exponentially.

A number of intelligent computational approaches have been developed to narrow this gap. One of them employs hidden Markov models for the identification and classification of DNA sequences within a gene that might be responsible for a specific protein domain (Durbin et al., 1998). In this approach, a hypothetical function can be assigned to a protein if it shares common domains with proteins of known function even if, based on pairwise comparison methods, it is not significantly similar with them. In another approach, proteins were clustered into families based on all pairwise comparisons among 18,000 proteins encoded in seven complete genomes (Tatusov et al., 1997). This protein classification approach has, as its basic unit, a group of descendants of a single ancestral gene and, since such a group is associated with a conserved, specific function, the inclusion of a protein in a cluster automatically entails functional prediction.

The approaches discussed above have been proven to be extremely useful in every aspect of research and development in biological sciences and biotechnology. They constitute the basis for annotation of every newly sequenced gene and genome. At their current state, however, there are limitations in their ability to predict the function of a gene and its product based on DNA sequence information. A combination of experimental and computational approaches is needed to overcome these limitations.

In most of the cases, computational analysis of DNA sequence provides information about the elementary function of a gene and its product. Recent developments in analytical methods allow the observation of a set of genes and their products "in action" providing information about their systemic function. One of the technologies used to infer function for various genes and to understand the complex interactions between cellular function and environment allows rapid and cell-wide monitoring of gene expression profiles (*Nature Genetics*, 1999). Using nanofabrication techniques, thousands of DNA probes are attached to microarrays, and they hybridize with fluorescent labeled complementary DNA (cDNA) that has been quantitatively produced from the messenger RNA (mRNA) content of whole cells. The relative intensity of hybridized elements allows rapid, reproducible and parallel quantification of the mRNA species within a cell. Data mining, clustering, and statistical analysis are used to classify the expression profiles and to identify sets of genes that share similar expression patterns. Comparison of the gene expression profiles between two experimental conditions over time can provide important information about groups of genes whose transcription is subject to the same regulatory rules and, therefore, they *might* serve similar cellular objectives. The power of these methodologies in studying gene expression has attracted considerable attention and has raised a lot of excitement. However, as mentioned in a recent review article, there are more review articles on gene expression technologies than primary research papers in this field (Bassett et al., 1999).

Proteomics refers to the array of technologies that focus on the identification of the systemic properties of proteins (Pandey and Mann, 2000). Two-dimensional gel electrophoresis allows the separation of cellular proteins on a polymer gel according to their molecular weight in one dimension and according to their isoelectric point in the second dimension. This technique allows the quantification of sets of cellular proteins and similarly the gene expression monitoring, and comparative studies provide invaluable information about the cellular function of proteins whose elementary function can either be known or unknown. Another set of proteomic technologies allows the identification of interactions between proteins and between proteins and DNA domains. These kinds of interactions suggest the involvement of the corresponding proteins in the regulation of signal transduction and transcription. Efforts in proteomic technologies address the key limits of the current technologies: the narrow spectrum and quantification of the protein properties that can be monitored simultaneously.

Bioinformatics and chemical engineering

The ultimate goal of functional genomics and bioinformatics is to integrate these large-scale data sets of the cellular processes toward a quantitative, and ultimately predictive, understanding of the function of individual cells and multicellular tissues (Figure 1). Current knowledge and analysis frameworks for most of the problems in biology allow a descriptive, qualitative understanding. A closer examination of the current technologies for cell-wide monitoring of the cellular processes suggests that, as we move from DNA sequence through transcription and translation to signal transduction and biotransformations, the corresponding technologies are limited. On the other hand, analysis frameworks that could be used to analyze these data and provide quantitative predictions are inadequate currently. These limitations present a number of challenges for technology development, data interpretation, and, ultimately, for integrating information from multiple levels of cellular function.

Chemical engineers can successfully contribute to resolving some of these problems in bioinformatics and functional genomics. There are, in fact, qualitative similarities between core chemical engineering research areas and problems addressed by bioinformatics (Table 1). Let us consider some of these analogies in more detail.

The identification of the function of various gene products based on mRNA and protein expression data is a *process identification* problem. However, the large number of components involved and errors associated with the technology-specific measurements is new to the existing methodologies for process identification. Biotransformations, as well as transcription and translation processes, are subject to a very elaborate regulatory scheme. While deciphering of these regulatory structures is another process identification problem, the discovery of design principles behind these structures will require engineering process control principles and methodologies.

Transcription and translation are similar in spirit to *polymerization processes.* They, however, are more complex than most of the processes studied in chemical engineering. Multiple mRNA and protein species are synthesized in parallel, and they compete for the same monomers, nucleic acids, and amino acids, respectively.

Moreover, they are autocatalytic processes since their products serve as catalysts for the synthesis of the monomers and for the polymerization itself. mRNA and protein expression are also subject to elaborate regulatory actions that are realized via protein-protein, protein-DNA, and protein-RNA interactions.

The resolution of the three-dimensional structure of proteins and RNA, based on crystallographic and NMR data and the prediction of these structures from the known nucleic acid and amino acid sequences, is analogous to *polymer design problems.* The issues here, however, are more complex. Proteins are composed of 20 distinct monomers, and their folding and structure depend strongly on the "nonideal" cellular environment.

Reaction-diffusion processes drive the complex processes of development and morphogenesis, during which a single cell, the fertilized egg, is evolving into a multitissue organism. These processes are very similar to problems studied in *fluid mechanics* and *transport phenomena.* The analysis of developmental processes, morphogenesis, and cell-to-cell communication will require novel theoretical and computational frameworks originating from fluid mechanics for the study of reaction and transport processes in the context of cellular environment.

Biotransformations, transcription, and translation are complex *physicochemical processes.* Their complexity and the uncertainty of their physicochemical properties resemble those of many problems common in combustion, atmospheric chemistry, polymer chemistry and thermodynamics of complex mixtures. The large number of components, the thermodynamic interactions among these components, the nonlinearity of the kinetic properties, the different time scales of various processes, and the spatial organization of the cellular environment are some of the qualitative similarities between the biological processes and processes whose study is central to chemical engineering.

Concluding remark

The construction of quantitative simulation models of living organisms, based on DNA sequence information and large-scale functional genomics data, is the seemingly Utopian goal of bioinformatics. Although it is hard to predict when such a task will be accomplished, it is the ultimate objective that drives developments in analytical technologies and bioinformatics. The very engineering nature of this endeavor promises an almost unimaginable array of challenges and opportunities for chemical engineers.

Literature cited

Alberts B., J. D. Watson, D. Bray, K. Roberts, J. Lewis, and M. Raff, *Molecular Biology of the Cell,* 3rd ed., Garland Publishing, New York (1994).

Bailey, J. E., "Chemical Engineering of Cellular Processes," *Chem. Eng. Sci.,* **50**, 4091 (1995).

Bailey, J. E., "Mathematical Modeling and Analysis in Biochemical Engineering: Past Accomplishments and Future Opportunities," *Biotechnol. Prog.,* **14**, 8 (1998).

Bassett, D. E., M. B. Eisen, and M. S. Boguski, "Gene Expression Informatics—It's All in Your Mine," *Nature Genetics,* **21**(supplement), 51 (1999).

"Double-Digit Growth for Industrial R&D," *C&EN,* **77**, 61 (1999).

Durbin, R., S. Eddy, A. Krogh, and G. Mitchison, *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids,* Cambridge University Press, New York (1998).

Ernst and Young, "Biotech 99: Bridging the Gap," 13th Biotechnology Industry Annual Report, available on the Web at http://www.ey.com/global/gcr.nsf/International/Biotech99summary, (1999).

Nature Genetics (Supplement:), "The Chipping Forecast," *Nature Genetics,* **21** (1999).

Pandey, A., and M. Mann, "Proteomics to Study Genes and Genomes," *Nature,* **405**, 837 (2000).

Quellette, B. F., and A. D. Baxevanis, eds, *Bioinformatics: A Practical Guide to the Analysis of Gene and Proteins,* Wiley, New York (1998).

Tatusov, R. L., E. V. Koonin, and D. J. Lipman, "A Genomic Perspective on Protein Families," *Science,* **278**, 631 (1997).