Controlled Biological Processes and Computational Genomics

James S. Schwaber

Daniel Baugh Institute for Functional Genomics/Computational Biology

Thomas Jefferson University College of Medicine

Philadelphia, PA 19107

Francis J. Doyle III and Daniel E. Zak

Department of Chemical Engineering

University of Delaware

Newark, DE 19716

Introduction

The purpose of the present paper is to point out opportunities for systems and process engineering approaches that appear to be arising as a consequence of the explosion of genomic data. In particular we consider our own and other's work in approaching biological systems with modeling and simulation studies as a guide to the contemporary landscape of such opportunities.

It has been impossible to miss the hoopla as the Human Genome Project has moved rapidly forward over the past year. The genome—i.e. the full set of human genes—is the code that guides the development and operation of the human organism. The excitement over the Human Genome Project is certainly justified, but gene sequences are only a first step towards understanding gene function (Schena, 1996). Genes function in highly interconnected, hierarchical, and nonlinear networks. Organismic states and characteristics are often not the result of the expression of single genes but rather the result of interactions of multiple genes, as in the case of some human cancers (Szallasi and Liang, 1998), as well as the past and present intracellular and extracellular environments.

A surprising result of the genome projects has been the similarity of genomes across species, even between man and yeast. This has led many to conclude that the genes alone can not account for species complexity and differences. Regulation of specific gene activities is crucial for creating the complexity of higher organisms. Consider DNA as a huge, flexible macromolecule containing genes and a large number of binding sites. The DNA molecule organizes simultaneous and interacting chemical binding with very large numbers of other molecules at these sites. These sites bind transcription factors that control gene activity with resulting control of gene and protein expression, regulating cell phenotype. The control is the result of inputs from the environment through receptors which trigger signals that guide gene activity. Cell states are dynamic, being constantly influenced by environmental and intercellular signals.

The control engineering approach to biosystems modeling and analysis offers an integrative perspective and brings unique insights and tools. The motivation for such an approach is provided by the richly interconnected feedback layers that underlie much of biological regulation. Integrative analyses across spatial and temporal orders of magnitude are essential for understanding and interpreting the underlying behavior.

Tools from control engineering are also relevant in light of emerging high throughput quantitative techniques, such as DNA microarrays and proteomic methods. These allow the measurement of thousands of intracellular factors, such as messenger RNA transcripts (that are indicative of gene activity) and proteins, in parallel and over time. The data from these methods present significant challenges and opportunities. The challenges are due to the enormity of the systems and their complexity. They present opportunities because never before has it been possible to so sensitively measure the conditions within cells.

Unfortunately it is still a common view that biological systems are beyond the scope of engineering approaches. It is true that there are huge gaps in knowledge and the systems are immensely complex. However, it is instructive to look at the results to date from modeling and simulation studies of biological systems because there are substantial positive results. Two of the most influential, Nobel Prize-winning pieces of biological work, however, are purely theoretical: the Hodgkin-Huxley formalism (Hodgkin and Huxley, 1952) and the Watson-Crick double helix model for DNA (Watson and Crick, 1953). To further illustrate the usefulness of computational approaches for complex biological processes, our successes in computational neuroscience are discussed next. This is followed by a detailed discussion of the opportunities and challenges in high throughput methods, particularly DNA microarrays, that are relevant to control engineering.

Computational Neuroscience

The greatest success for theoretical and modeling approaches in biology has been in Computational Neuroscience. Our own early work in this domain was with a focus on biological control systems. At that time the
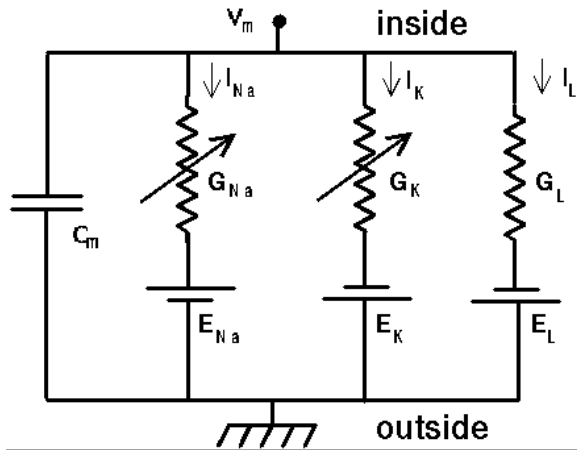
Figure 1: Hodgkin Huxley formalism.

Hodgkin-Huxley (HH) formalism was used in modeling studies of channel kinetics, one channel type at a time. Neuron models were sum and squash, non-spiking variety. We conceived of creating neuron models built up of multiple species of channel types using HH. The problem was computationally intractable, however, and the interaction within an extremely complex set of nonlinear, dynamical systems, including feedback, was certain to be extremely complex. We had excellent success, however, in creating complex, spiking neurons that were robust, occupying large, continuous and high-dimensional parameter spaces (Schwaber et al., 1993; Foster et al., 1993). We took these results to indicate that this class of biological system has evolved to be extremely robust, since its performance was not sensitive to parameter values. Parameter value variations are expected to result from varying environmental conditions in a real system.

We also were encouraged by the success at crossing levels of analysis, in this case from channel kinetics to whole cell behavior. We decided to extrapolate this work to additional levels, using several types of our HH neuron models to create neuronal networks with synaptic, chemical communication. We extended these network models to closed system models controlling peripheral organs. We were able to study the impact of manipulating cellular properties on system behavior in these models (Rybak et al., 1997a,b,c).

We have also been successful in connecting organism-level function (in this case, blood pressure regulatory behavior) with intracellular processing (second messenger pathways: Cheng et al., 1999, 1997; Hardwick et al., 1995; Parsons et al., 1987; Schwaber et al., 1993). Figure 2 shows the proposed reflex circuitry for short-term regulation of cardiovascular system. The local reflex architecture in Figure 2 is consistent with the experimental results in Cheng et al. (1997).

We have constructed a computer simulation model for the local cardiac reflex based on the anatomical exper-

imental results and physiological data in the literature. Simulation results indicate that the local cardiac reflex could be effecting attenuation of the nonlinearity of the relationship between cardiac vagal drive and the arterial blood pressure. We have explored the hypothesis that the functional role of dynamic neuromodulation by SIF cells is an input-output "linearizing" effect on the actuator (heart) dynamics (Vadigepalli et al., 2001). We employed coherence analysis to characterize the nonlinearity between the frequency of vagus nerve pulsatile input and the mean arterial pressure. We also investigated the role of modulatory synaptic transmitters in this function.

We also have driven the levels of analysis question to more molecular and chemical processes, including receptors and signaling cascades through which environmental signals affect system behavior (Brown et al., 1999; Kholodenko et al., 1999; Kholodenko, 2000; Vadigepalli et al., 2001). This latter work has raised the question of connecting from the biochemical level to the gene level.

The cell and system regulation we are studying must, in life, interact with the regulation occurring at the gene level. The question that arises next is, "What would be needed to create an equally successful Computational Genomics?"

Computational Genomics

New analytical techniques have been developed that allow the quantification of many intracellular factors (Gombert and Nielsen, 2000). These include DNA microarray technology (Schena et al., 1995), which allows the relative transcription levels of thousands of genes to be measured in parallel, and gel electrophoresis and mass spectrometry, which allow levels of hundreds of proteins to be quantified (Gygi et al., 2000). Many groups have applied these methods as well as information from the genome sequencing projects to explore gene function on a genomic scale (Fodor et al., 1993; Schena et al., 1995; Ross-Macdonald et al., 1999; Winzeler et al., 1999; Uetz et al., 2000). DNA chips which allow for genome-wide measurements of mRNA levels (Schena, 1996; De Saizieu et al., 1998; Eisen et al., 1998; Marshall and Hodgson, 1998; McKenzie et al., 1998; Ramsay, 1998; Spellman et al., 1998; Brown and Botstein, 1999; Jia et al., 2000).

It is known that the data obtained with these methods have limitations. For example, it has been observed that mRNA and protein levels in do not correlate well enough for relative mRNA transcription levels to be predictive of protein expression levels (Gygi et al., 1999). The value of microarray data, however, should not be understated. It is of great value for generating hypotheses and it is widely accepted that, "The mRNA levels sensitively reflect the state of the cell, perhaps uniquely defining cell types, stages, and responses. To decipher the logic of gene regulation, we should aim to be able to monitor the expression level of all genes simultaneously"
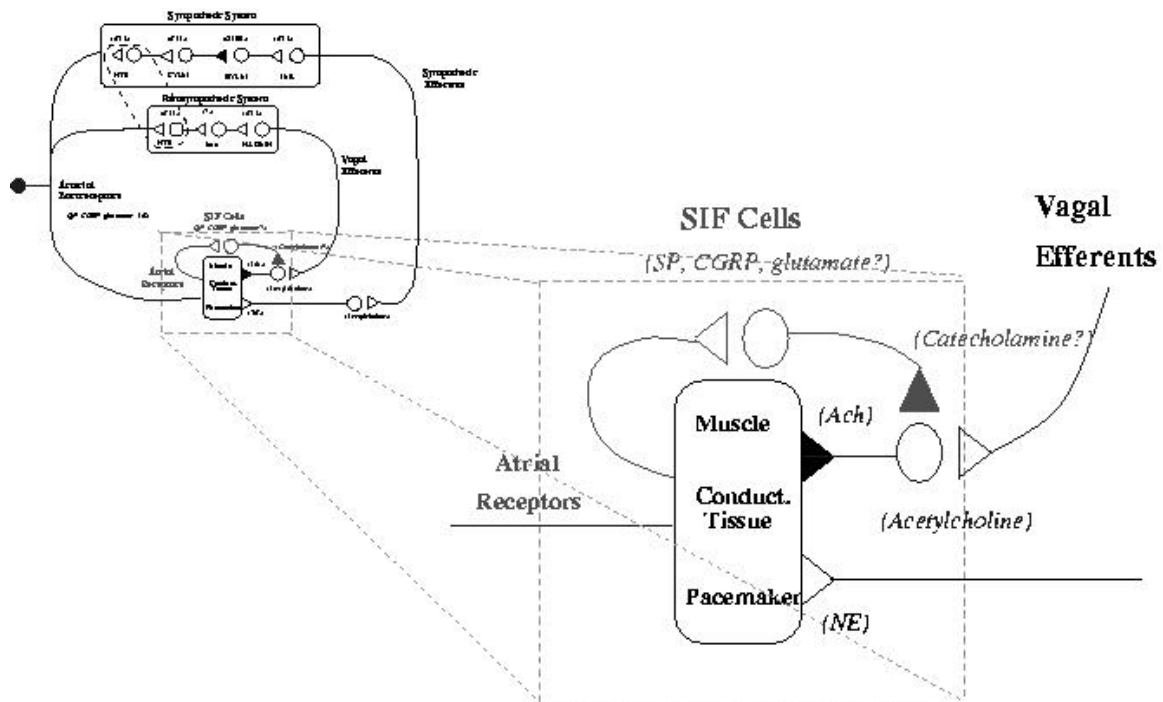
Figure 2: Schematic of local cardiac reflex control. The local cardiac reflex consists mainly of three components. (i) SIF cells receive sensory inputs from the atrial receptors and project to the principal neurons (PNs); (ii) PNs also receive input from vagal efferents from dorsal motor nucleus of the vagus; and (iii) PN activity has a phase-dependent effect on the SA node pacemaker and hence on the heart-beat cycle.

(Lander, 1996).

One motivation behind our group's interest in DNA microarrays and other high throughput quantitative techniques is their ability to provide insight into the genetic networks that dictate cellular responses to intercellular and enviromental stimuli. A simple example of a genetic network is the genetic switch modeled and synthesized by Gardner et al. (2000), shown below in Figure 3. Here, the product of one gene represses the transcription of the other gene, causing the system to have two stable steady states. A review of some of the complex behavior that can be observed for similar systems is given by Smolen et al. (2000). Gene networks in even the simplest organisms are expected to be complex. The bacteria E. Coli, for example, has approximately 2000 genes, with connectivities between the genes averaging between two and three (Thieffry et al., 1998).

Several factors make the determination of gene networks from DNA microarray data a challenge. These are related to the design of the experiments, interpreting the data and incorporating biological information, and determining model structures for gene networks. These are discussed below.
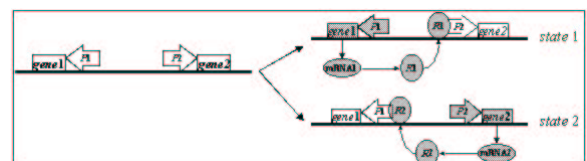

Figure 3: Schematic of the bistable genetic switch (Gardner et al., 2000).

Experimental Approach

The microarray experiments we perform in our lab involve the collection of transcript profiles over time while the organism is undergoing a systemic perturbation (chronic ethanol exposure, sleep deprivation, or chronic hypertension). By collecting array data over time, we expect to gain insights into the genetic basis for the organism's response to the perturbation. Our approach in designing these experiments is to apply the methods of system identification from systems engineering that follow a formal procedure of: (i) variable definition (i.e., which variables can be perturbed for maximum information), (ii) input sequence design (i.e., size and frequency of changes in the input variables), (iii) execution of input sequence, (iv) data cleaning (outlier detection, noise fil-
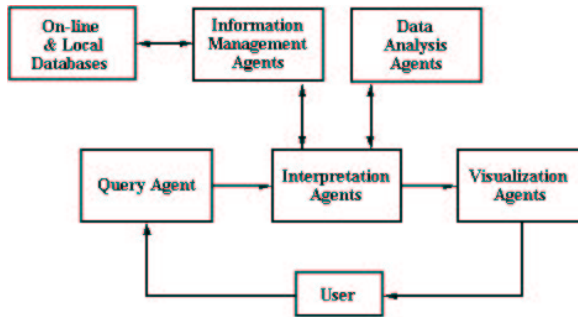
Figure 4: Architecture for analysis of genomic data using intelligent agents.

tering), (v) model building (using, e.g., correlation analysis or other statistical methods), and finally, and most importantly, (vi) validation of the model. Critical elements in this protocol for the biological problem at hand are the definition of suitable perturbation sources, the character of the perturbation signal, and the processing of the noisy data. A "rich" perturbation sequence will be designed to maximize the information content in the resulting signals. Richness in this context will be assessed by both the dynamic character of the forcing signal, as well as the nonlinear character of the forcing signal. The latter is especially important for the identification of nonlinear models, using such techniques as Laguerre kernel expansions (e.g., Marmarelis, 1993). A simple binary sequence (on-off) generates little or no information content at the output of a nonlinear system.

Data Issues

Interpretation of biological data consists of integrating data from several levels in the organismic hierarchy, including genomic sequences, microarray information, proteomic and metabolic information, and physiological information. This data will be analyzed through the use of various algorithms and statistical techniques. However, many of these techniques ignore domain knowledge of the system, which could lead to better and more efficient analysis of the data. Data analysis techniques like clustering are independent of whether they are applied on shopping data, weather data or biological data. We will incorporate domain knowledge from biology for improved and intelligent analysis of data. We have started some preliminary work in this area for the analysis of microarray data.

One of the analysis techniques that we have tested using simulator data is clustering. We will also compare clustering methods with other techniques and also with knowledge based clustering. In knowledge-based clustering, the implementation will use intelligent software agents to obtain knowledge automatically from databases. The intelligent agents will gather the knowl-

edge from different sources on the web such as pathway databases, etc., and will use this knowledge during data analysis. We propose to build a system for microarray data analysis that can be incorporated to related ongoing work in intelligent multi-agent systems for data analysis. Specific agents will be used to gather the knowledge relevant for data analysis. This knowledge will then be integrated with the analysis technique such as clustering. A proposed architecture for this is given in Figure 4 above.

Gene Network Model Structures

Determining network architecture from microarray data is nontrivial. This data can consist of relative transcription levels for thousands of genes over hundreds of time points. Mathematical models, combined with biological knowledge, are necessary to determine the relationships contained within this data. A discussion of some of the key considerations in building such models, notably data requirement, is given in Fuhrman et al. (1999).

Several approaches for building models of transcriptional regulatory gene networks from temporal microarray data have been described in the literature. Top-down modeling approaches have been primarily designed for elucidating network connections from temporal microarray expression data. As mentioned above, there is no clear correlation between protein levels and relative transcription levels and it is therefore not possible to determine network connections from microarray transcription data alone (Gygi et al., 1999). Microarray data can be used, however, to generate hypotheses that can direct future experiments. Also, the methods are generally applicable to any system of large numbers of interacting components, and are thus relevant for interpreting temporal protein level data from gel electrophoresis and mass spectrometry. There are few examples in the literature where these methods have actually been applied (Reinitz and Sharp, 1995; D'Haeseleer et al., 1999), and therefore the utility of each of these methods still requires verification. Four examples of top-down modeling approaches are logical, linear, "linear plus squashing", and differential.

In the logical approach, genes are either "on" or "off" and have a limited number of inputs from other genes (Kauffman, 1993). This approach is appealing because it may give basic structural information and has the smallest data requirement, of order $\log2(N)$ time points if the transcription of N genes are only influenced by two genes each (Akutsu et al., 1999). Its main limitations are that the number of regulatory inputs must be limited a priori and genes that can have intermediate expression levels or influence the transcription of other genes to varying degrees are neglected (Weaver et al., 1999). At the next level of complexity is the linear approach (D'Haeseleer et al., 1999), where the transcription levels of the genes

at one time point are linear combinations of the expression levels of all of the genes at the previous time point. The drawbacks of this approach are that it requires at least as many time points as genes since it has N2 parameters, that it poses no economy on interconnections, and that the process it describes is not linear. In spite of this, D'Haeseleer et al. (1999), has had some success with this approach when applying it to expression data for rat nervous system development. An improvement to the linear approach is the "linear plus squashing" approach (Reinitz and Sharp, 1995; Weaver et al., 1999). The input to a gene is still a linear combination of the expression levels of all of the other genes, but now the input and the gene expression level are related by a sigmoidal "squashing function." This is a more realistic model of gene expression. A fourth approach is the differential model proposed by Chen et al. (1999). The time rate of change of mRNA concentration is expressed as a linear combination of the protein concentrations minus a degradation term. The time rate of change of protein concentration is a linear combination of the transcript concentrations minus a degradation term. Since this model includes additional states for the protein concentrations, it is an improvement over the linear model. A drawback is that the number of empirical parameters for this model is nearly twice that of the linear model, giving it a significantly larger data requirement. The system is also not completely determined unless initial protein concentrations are known.

Conclusions

Clearly a new era for biology is emerging that can bring tremendous developments in medicine and understanding. Before these can be realized, however, computational approaches must be developed that can make full use of the data coming from the high throughput technologies. Given the highly regulated and interconnected nature of biological systems, methods from control engineering should be able to contribute significantly towards this goal.

References

Akutsu, T., S. Miyano, and S. Kuhara, "Identification of Genetic Networks from a Small Number of Gene Expression Patterns Under the Boolean Network Model," *Pac. Symp. Biocomput.*, **4**, 7–28 (1999).

Brown, P. O. and D. Botstein, "Exploring the new world of the genome with DNA microarrays," *Nat. Genet.*, **21**, 33–37 (1999).

Brown, L. J., G. E. Gonye, and J. S. Schwaber, "Non-linear PI Control Inspired by Biological Control Systems," *Advances in Neural Information Processing Systems*, **11**, 975–981 (1999).

Chen, T., H. L. He, and G. M. Church, "Modeling Gene Expression with Differential Equations," *Pac. Symp. Biocomput.*, **4**, 29–40 (1999).

Cheng, Z., T. L. Powley, J. S. Schwaber, and F. J. Doyle III, "Vagal afferent innervation of the rat heart atria reconstructed with laser confocal microscopy: A hypothesis of a local reflex circuit," *J. Comp. Neurol.*, **381**, 1–17 (1997).

Cheng, Z., T. L. Powley, J. S. Schwaber, and F. J. Doyle III, "Projections of the dorsal motor nucleus of the vagus to cardiac ganglia of rat atria: an anterograde tracing study," *J. Comp. Neurol.*, **410**, 320–41 (1999).

De Saizieu, A., U. Certa, J. Warrington, C. Gray, W. Keck, and J. Mous, "Bacterial transcript imaging by hybridization of total RNA to oligonucleotide arrays," *Nat. Biotechnol.*, **16**, 45–48 (1998).

D'Haeseleer, P., X. Wen, S. Fuhrman, and R. Somogyi, "Linear Modeling of mRNA Expression Levels During CNS Development and Injury," *Pac. Symp. Biocomput.*, **4**, 41–52 (1999).

Eisen, M. B., P. T. Spellman, P. O. Brown, and D. Botstein, "Cluster Analysis and Display of Genome-Wide Expression Patterns," *P. Natl. Acad. Sci. USA*, **95**, 14863–14868 (1998).

Fodor, S. P. A., R. P. Rava, X. C. Huang, A. C. Pease, C. P. Holmes, and C. L. Adams, "Multiplexed biochemical assays with biological chips," *Nature*, **364**, 555–556 (1993).

Foster, W. R., L. H. Ungar, and J. S. Schwaber, "Significance of conductances in Hodgkin-Huxley models," *J. Neurophysiol.*, **70**, 2502–18 (1993).

Fuhrman, S., P. D'haeseleer, and R. Somogyi, Tracing genetic information flow from gene expression to pathways and molecular networks, in DNA microarrays: the new frontier in gene discovery and gene expression analysis, Short course, Society for Neuroscience (1999).

Gardner, T. S., C. R. Cantor, and J. J. Collins, "Construction of a Genetic Toggle Switch in Escherichia coli," *Nature*, **403**, 339–342 (2000).

Gombert, A. K. and J. Nielsen, "Mathematical Modeling of Metabolism," *Curr. Opin. Biotech.*, **11**, 180–186 (2000).

Gygi, S. P., Y. Rochon, B. R. Franza, and R. Aebersold, "Correlation Between Protein and mRNA Abundance in Yeast," *Mol. Cell. Biol.*, **19**, 1720–1730 (1999).

Gygi, S. P., G. L. Corthals, Y. Zhang, Y. Rochon, and R. Aebersold, "Evaluation of Two-Dimensional Gel Electrophoresis-Based Proteome Analysis Technology," *P. Natl. Acad. Sci. USA*, **97**, 9390–9395 (2000).

Hardwick, J. C., G. M. Mawe, and R. L. Parsons, "Evidence for afferent fiber innervation of parasympathetic neurons of the guinea-pig cardiac ganglion," *J. Autonom. Nerv. Syst.*, **53**, 166–74 (1995).

Hodgkin, A. L. and A. F. Huxley, "A Quantitative Description of Membrane Current and its Application to Conduction and Excitation in Nerve," *J. Physiol.*, **117**, 500–544 (1952).

Jia, M. H., R. A. LaRossa, J. M. Lee, A. Rafalski, E. J. De Rose, G. E. Gonye, and Z. Xue, "Global Expression Profiling of Yeast Treated with Sulfometuron Methyl, an Inhibitor of Amino Acid Biosynthesis," *Physiol. Genomics*, **3**, 83–92 (2000).

Kauffman, S. A., *The Origins of Order*. Oxford University Press, New York (1993).

Kholodenko, B. N., O. V. Demin, G. Moehren, and J. B. Hoek, "Quantification of short-term signaling by the epidermal growth factor receptor," *J. Biol. Chem.*, **274**, 30169–30181 (1999).

Kholodenko, B. N., "Negative feedback and ultrasensitivity can bring about oscillations in the miogen- activated protein kinase cascades," *Eur. J. Biochem.*, **267**, 1583–1588 (2000).

Lander, E. S., "The new genomics: global views of biology," *Science*, **274**, 536–539 (1996).

Marmarelis, V. Z., "Identification of nonlinear biological systems using Laguerre expansions of kernels," *Ann. Biomed. Eng.*, **21**, 573–89 (1993).

Marshall, A. and J. Hodgson, "DNA chips: an array of possibilities," *Nat. Biotechnol.*, **16**, 27–31 (1998).

McKenzie, S. E., E. Mansfield, E. Rappaport, S. Surrey, and P. Fortina, "Parallel molecular genetic analysis," *Eur. J. Hum. Genet.*, **6**, 417–429 (1998).

Parsons, R. L., D. S. Neel, T. W. McKeon, and R. E. Carraway, "Organization of a vertebrate cardiac ganglion: a correlated biochemical and histochemical study," *J. Neurosci.*, **7**, 837–46 (1987).

Ramsay, G., "DNA chips: State-of-the art," *Nat. Biotechnol.*, **16**, 40–44 (1998).

Reinitz, J. and D. H. Sharp, "Mechanism of eve Stripe Formation," *Mech. Develop.*, **49**, 133–158 (1995).

Ross-Macdonald, P., P. S. Coelho, T. Roemer, S. Agarwal, A. Kumar, R. Jansen, K. H. Cheung, A. Sheehan, D. D. Symoniatis, L. Umansky, M. Heidtman, F. K. Nelson, H. Iwasaki, K. Hager, M. Gerstein, P. Miller, G. S. Roeder, and M. Snyder, "Large-scale analysis of the yeast genome by transposon tagging and gene disruption," *Nature*, **402**, 413–8 (1999).

Rybak, I. A., J. F. Paton, and J. S. Schwaber, "Modeling neural mechanisms for genesis of respiratory rhythm and pattern I. Models of respiratory neurons," *J. Neurophysiol.*, **77**, 1994–2006 (1997a).

Rybak, I. A., J. F. Paton, and J. S. Schwaber, "Modeling neural mechanisms for genesis of respiratory rhythm and pattern II. Network models of the central respiratory pattern generator," *J. Neurophysiol.*, **77**, 2007–26 (1997b).

Rybak, I. A., J. F. Paton, and J. S. Schwaber, "Modeling neural mechanisms for genesis of respiratory rhythm and pattern III. Comparison of model performances during afferent nerve stimulation," *J. Neurophysiol.*, **77**, 2027–39 (1997c).

Schena, M., D. Shalon, R. W. Davis, and P. O. Brown, "Quantitative Monitoring of Gene Expression Patterns with a Complementary DNA Microarray," *Science*, **270**, 467–470 (1995).

Schena, M., "Genome Analysis with Gene Expression Microarrays," *BioEssays*, **18**, 427–431 (1996).

Schwaber, J. S., J. S. Graves, and J. F. R. Paton, "Computational Modeling of Neuronal Dynamics for Systems Analysis: Applications to Neurons of Cardiorespiratory NTS in the Rat," *Brain Res.*, **604**, 126–141 (1993).

Smolen, P., D. A. Baxter, and J. H. Byrne, "Mathematical Modeling of Gene Networks," *Neuron*, **26**, 567–580 (2000).

Spellman, P. T., G. Sherlock, M. Q. Zhang, V. R. Iyer, K. Anders, M. B. Eisen, P. O. Brown, D. Botstein, and B. Futcher, "Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization," *Mol. Biol. Cell.*, **9**, 3273–3297 (1998).

Szallasi, Z. and S. Liang, "Modeling the Normal and Neoplastic Cell Cycle with Realistic Boolean Genetic Networks," *Pac. Symp. Biocomput.*, **3**, 66–76 (1998).

Thieffry, D., A. Huerta, E. Perez-Rueda, and J. Collado-Vides, "From specific gene regulation to genomic networks: a global analysis of transcriptional regulation in, E," *Coli. Bioessays*, **20**, 433–440 (1998).

Uetz, P., L. Giot, G. Cagney, T. A. Mansfield, R. S. Judson, J. R. Knight, D. Lockshon, V. Narayan, M. Srinivasan, P. Pochart, A. Qureshi-Emili, Y. Li, B. Godwin, D. Conover, T. Kalbfleisch, G. Vijayadamodar, M. M. Yang, M. Johnston, S. Fields, and J. M. Rothberg, "A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae," *Nature*, **403**, 623–627 (2000).

Vadigepalli, R., F. J. Doyle III, and J. S. Schwaber, Analysis and Neuronal Modeling of the Nonlinear Characteristics of a Local Cardiac Reflex in the Rat, Submitted to Neural Comput. (2001).

Watson, G. D. and F. H. C. Crick, "Molecular structure of nucleic acids: a structure for deoxyribose nucleic acid," *Nature*, **171**, 737 (1953).

Weaver, D. C., C. T. Workman, and G. D. Stormo, "Modeling Regulatory Networks with Weight Matrices," *Pac. Symp. Biocomput.*, **4**, 102–111 (1999).

Winzeler, E. A., D. D. Shoemaker, H. Astromoff, K. Liang, B. Anderson, B. Andre, R. Bangham, R. Benito, J. D. Boeke, H. Bussey, A. M. Chu, C. Connelly, K. Davis, F. Dietrich, S. W. Dow, M. El Bakkoury, F. Foury, S. H. Friend, E. Gentalen, G. Giaever, J. H. Hegemann, T. Jones, M. Laub, H. Liao, N. Liebundguth, D. J. Lockhart, A. Lucau-Danila, M. Lussier, N. M'Rabet, P. Menard, M. Mittmann, C. Pai, J. L. Rebischung, J. L. Revuelta, L. Riles, C. J. Roberts, P. Ross-MacDonald, B. Scherens, M. Snyder, S. Sookhai-Mahadeo, R. K. Storms, V. R. S. M. Voet, G. Volckaert, T. R. Ward, R. Wysocki, G. S. Yen, K. Yu, K. Zimmermann, P. Philippsen, M. Johnston, and R. W. Davis, "Functional Characterization of the S. cerevisiae Genome by Gene Deletion and Parallel Analysis," *Science*, **285**, 901–906 (1999).