Available online at www.sciencedirect.com
SCIENCE DIRECT®
Computers and Chemical Engineering 29 (2005) 547–563
Computers & Chemical Engineering
www.elsevier.com/locate/compchemeng

Unconventional systems analysis problems in molecular biology: a case study in gene regulatory network modeling☆

Daniel E. Zak[a,b], Rajanikanth Vadigepalli[b], Gregory E. Gonye[b], Francis J. Doyle III[c], James S. Schwaber[b], Babatunde A. Ogunnaike[a,*]

[a] *Department of Chemical Engineering, University of Delaware, Newark, DE 19716, USA*
[b] *Department of Pathology, Cell Biology and Anatomy, Daniel Baugh Institute for Functional Genomics and Computational Biology, Thomas Jefferson University, Philadelphia, PA 19107, USA*
[c] *Department of Chemical Engineering, University of California, Santa Barbara, CA 93106, USA*

Received 22 September 2003; received in revised form 7 January 2004
Available online 3 October 2004

Abstract

The broad conceptual postulate that systems engineering techniques developed for complex chemical processes may be applicable to complex cell biological processes is very compelling. However, a naïve, "direct" application of systems engineering techniques to biological problems of practical significance may be rendered virtually ineffective by fundamental differences between cell biology and chemical processes. These differences and the problems they pose are illustrated in this paper with an example problem: modeling a gene regulatory network involved in the yeast cell cycle. We demonstrate how the biological essence complicates a straightforward "process modeling/identification" problem and subsequently recommend an alternative approach. The approach—a middle ground between a direct, "off the shelf" application of systems engineering tools and a "one-at-a-time" ad-hoc development—incorporates fundamental knowledge of the mechanisms and constraints intrinsic to biological systems. The principles and implementation details of the approach are illustrated with the case study.

Keywords: Molecular biology; Yeast cell cycle; Gene regulatory network

1. Introduction

Cells are complex dynamical systems that are constantly remodeling themselves over time in response to changes in their internal and external environments. The emergence of technology for acquiring genome-wide gene expression and other global data sets has created an opportunity for understanding this process at the system-wide level. Given its scale and complexity, the problem of deciphering system-wide cellular regulation has naturally attracted computational approaches from the physical, engineering, and computer sciences. The net result is that many cell biologists are increasingly applying computational methods in their work while many computational scientists are also increasingly engaging in biological research problems. Just as individuals with training in the biological sciences must acquire sufficient domain knowledge before they can effectively apply advanced modeling, simulation, and analysis techniques, so must the computational scientist acquire sufficient domain knowledge to be effective in dealing with complex biological problems. Illustrating this latter point is the focus of the present work. Specifically we present the idiosyncrasies, complexities, and constraints intrinsic to biological systems that must be addressed for systems engineering tools and expertise to be successful in dealing with typical and yet non-trivial problems in biology. We take as an example the modeling and identification of gene regulatory networks from global gene expression data, specifically the regulatory network underlying the yeast cell cycle.

☆ Online appendix at http://www.dbi.tju.edu/dbi/publications/cache04.

* Corresponding author. Tel.: +1 302 831 4504; fax: +1 302 831 1048.
E-mail address: ogunnaik@che.udel.edu (B.A. Ogunnaike).

0098-1354/$ – see front matter © 2004 Elsevier Ltd. All rights reserved.
doi:10.1016/j.compchemeng.2004.08.016

1.1. Computational modeling in biology

Whereas a key objective of modeling and identification of chemical processes is model development for process control and optimization, currently the key objective of modeling biological systems is "*process understanding*" or "*reverse engineering*" (Csete and Doyle, 2002). Such understanding may be in terms of metaphors for signaling pathways that explain the "rationale" behind their complex architectures (Bhalla, 2003; Neves & Iyengar, 2002); it may also be understanding that serves to reduce complexity and scale, for example understanding the global gene expression response of a cell to a perturbation as the response of a few transcriptional regulators instead of the response of thousands of genes (Bussemaker, Li, & Siggia, 2001). The latter may be considered *integrative* understanding in that the thousands of observations are integrated into a concise explanation, and the consistent framework provided by modeling and identification for such integrative understanding may be one of the strongest arguments for applying such techniques in biology (Hartemink, Gifford, Jaakola, & Young, 2002; Ideker et al., 2001; Jarvis et al., 2002). Integration of complementary data types into computational models provides understanding that would not be possible if the data types were considered individually. Genomic sequences, gene expression profiles, protein–DNA interaction data, and transcript degradation constants, for example, may be integrated into computational models of system-wide transcriptional regulation that are far more interpretable with respect to cellular function than any of these data types are individually. Finally, computational models can be used to enhance experimental studies: they can be used to generate testable hypotheses and for designing experiments tailored for optimal extraction of the desired system information. When sufficient process understanding has been acquired, it may become possible to "forward engineer" biological systems to meet specific objectives (Yokobayashi, Collins, Leadbetter, Weiss, & Arnold, 2003).

2. Modeling and identification of gene regulatory networks

With the availability of global gene expression profiles, which, for almost any cellular system, quantify the activity of every gene in a genome, in parallel, over time, and in response to perturbations, it is now possible to attempt to identify gene regulatory networks (Brazhnik, de la Fuente, & Mendes, 2002; D'Haeseleer, Liang, & Somogyi, 2000). The objective of gene regulatory network identification is to enable the scientist to go beyond merely observing the qualitative changes in gene activities, and actually infer and quantify causal links between the genes that underlie physiological responses. Computational models of these causal links between genes can provide system-wide understanding of the regulation that is fundamental to all life processes, and accelerate efforts in deciphering the structure of complex biological systems. Ultimately, such models may be used for generating testable hypotheses about novel drug targets for prevention and treatment of complex diseases.

Mathematically, the gene network identification problem may be formulated as the identification of the vector function $\mathbf{f}(\cdot)$ and the estimation of the parameter vector $\mathbf{p}$ in Eq. (1), given measurements of the expression levels (mRNA levels) over time, $\mathbf{x}(t)$, and the external input perturbations, $\mathbf{v}(t)$, that initiated this observed response:

$$\frac{\mathrm{d}\mathbf{x}}{\mathrm{d}t} = \mathbf{f}(\mathbf{x}, \mathbf{v}, \mathbf{p}) \tag{1}$$

From this formulation, the gene regulatory network identification problem appears to be a straightforward modeling/identification problem. However, given the complexity of the biological processes Eq. (1) describes, and given the technical issues associated with the measurement of gene expression, the problem is not so straightforward. First is the issue of scale: the vector of expression levels $\mathbf{x}(t)$ at each sample point in time is of dimension N_g, where N_g is the number of genes in the genome of the organism being studied. For yeast, $N_g \sim 6000$, while for humans, $N_g \sim 25{,}000$, making Eq. (1) several orders of magnitude larger in scale than what is typical for chemical process identification problems (for a database of typical system identification problems, see: http://www.esat.kuleuven.ac.be/sista/daisy (De Moor, De Gersem, De Schutter, & Favoreel, 1997)). This issue of scale is further compounded as follows. In standard chemical process identification, the process model is based entirely on $\mathbf{x}(t)$ and $\mathbf{v}(t)$ data. However, for the biological systems considered presently, using only $\mathbf{x}(t)$ and $\mathbf{v}(t)$ data for system identification requires making allowance for all genes to interact with all genes, giving rise to a model with *at least* $N_g \times N_g$ parameters that must be estimated from data. Of course, the fully connected network is not realistic because biological networks are known to be sparsely interconnected (Arnone & Davidson, 1997; Jeong, Mason, Barabasi, & Oltvai, 2001; Ravasz, Somera, Mongru, Oltvai, & Barabasi, 2002) and thus the majority of the parameters will be zero. The identification of Eq. (1) is thus a problem in simultaneous structure and parameter identification on a scale that is extremely rare for chemical process systems. Lastly are the data requirement issues. The problems associated with the quality and quantity of gene expression data are extreme compared to what is typical with chemical process data. With the current state of the art, each sample of $\mathbf{x}(t)$ is highly contaminated with noise (Nadon & Shoemaker, 2002; Sebastiani, Gussoni, Kohane, & Ramoni, 2003); furthermore, because of the expense and difficulty in acquiring gene expression data, the number of time points N_m at which $\mathbf{x}(t)$ measurements are taken is relatively small ($N_m \ll N_g$). Thus, from the classical process identification perspective, the gene regulation network identification problem has the following characteristics: