Representing and Analyzing Biochemical Networks Using BioMaze

Yves Deville, Christian Lemer, and Shoshana Wodak

Summary

Systems biology aims at understanding the holistic behavior of biological systems. A very important step toward this goal is to develop a theoretical framework in which we can embed the detailed knowledge that biologists are accumulating at increasing speed, which will then allow us to compute the outcomes of the complex interplay between the myriad interactions that take place in the system. This chapter deals with important basic aspects of this theoretical framework that lie on the divide between systems biology and bioinformatics. In the first part, it discusses the conceptual models used for representing detailed knowledge on various types of biochemical pathways and interactions. As much of this knowledge deals with the complex networks of functional and physical interactions between the different molecular players, the second part of this chapter reviews the conceptual models and methods used to analyze various properties of these networks.

Key Words: Biochemical networks; network analysis; metabolic pathways; signal transduction; artificial intelligence; BioMaze.

1. General Introduction

The major challenge of the post-genomic era is the interpretation of the vast body of genomic sequence information in terms of the biological function of the gene products and the mechanisms of the cellular processes in which they are involved. This endeavor is driven in great part by the expectation that the gained understanding will lead to new ways of diagnosing and curing human diseases, and making our planet a better place to live.

But the task is daunting. The very notion of biological function is complex. The function of proteins, which are one type of gene product, essentially depends on the molecular interactions they make and on the cellular context in which they find themselves. Understanding function, thus, requires knowledge of how the different molecular players...
cooperate to produce the observed behavior of the living cell and of key processes therein. Acquiring this knowledge is the main object of systems biology, a field that has attracted renewed interest in recent years, and to which this volume is devoted.

A first key step in this endeavor is to acquire the information necessary to describe the system under study in a useful way. Major efforts are therefore being devoted worldwide to collecting such information by diverse means. Experimental procedures are used to measure gene expression profiles (1), transcription factor–gene interactions (2), and mRNA lifetimes (3) on the genome scale. Protein–protein interactions are characterized using high-throughput pull-downs or two-hybrid screens (4), and indirect “interactions” between genes are being probed by multiple gene deletions (5). In parallel, protein–protein interactions, sets of co-regulated genes (6,7), and metabolic pathways (8,9,10) are inferred using theoretical methods. These methods exploit information on protein and DNA sequences in related genomes, on protein three-dimensional (3D) structures, domain architecture, and gene order (11). Others use automatic procedures to extract links from texts of Medline abstracts (12).

All of these approaches yield very large bodies of valuable, but rather noisy, data, which systems biology research endeavors to exploit. Clearly, the bulk of the data pertains to the description of the circuitry of the cellular systems; the interaction, regulatory networks, and pathways (on gene regulation, metabolism and signal transduction), and provides limited information on the temporal sequence of events, or on their spatial organization. But obtaining detailed information on the circuitry is a key first step that can yield valuable clues on the system-level behavior (13), provided, however, that this information can be adequately validated and readily analyzed.

Currently, such analyses face various difficulties. The ability of accessing and manipulating the information is limited by the fact that it is distributed across heterogeneous databases. Also, our current knowledge of the various cellular processes (protein interaction, gene regulation, or signal transduction) is poorly structured and partial. The data can therefore be incomplete, inconsistent, or approximate. In addition, the size of the pathways and networks available for analysis can be very large, leading to problems of spatial and temporal computational complexity. All this makes the representation and analysis of pathways and interaction networks, which we denote here as biochemical networks, challenging problems in systems biology and bioinformatics.

This chapter describes strategies for addressing these challenges, with examples taken from our own work on the BioMaze system. Section 2 discusses data models for representing rich information on biochemical networks for archival and query purposes. That section starts with a short overview of existing models and proceeds with a description of an attractive integrative data model implemented in the BioMaze database. Section 3 deals with data models used for the purpose of performing computational analyses of biochemical networks. Section 4 reviews analysis methods that use standard graph-based techniques and presents some recent advances in the application of constraint satisfaction