CHAPTER 1

Introduction

Suddenly, systems biology is everywhere. What is it? How did it arise? The driving force for its growth is high-throughput (HT) technologies that allow us to enumerate biological components on a large scale. The delineation of the chemical interactions of these components gives rise to reconstructed biochemical reaction networks that underlie various cellular functions. Systems biology is thus not necessarily focused on the components themselves, but on the nature of the links that connect them and the functional states of the networks that result from the assembly of all such links. The stoichiometric matrix represents such links mathematically based on the underlying chemistry, and the properties of this matrix are key to determining the functional states of the biochemical reaction networks that it represents.

1.1 The Need for Systems Analysis in Biology

Biological parts lists

During the latter half of the 20th century, biology was strongly influenced by reductionist approaches that focused on the generation of information about individual cellular components, their chemical composition, and often their biological functions. Over the past decade, this process has been greatly accelerated with the emergence of genomics. We now have entire DNA sequences for a growing number of organisms, and we are continually delineating their gene portfolios. Although functional assignment to these genes is presently incomplete, we can expect that we will eventually have assigned and verified function for the majority of genes on selected genomes. Extrapolation between genomes will then most likely accelerate the definition of what amounts to a "catalog" of cellular components in a large number of organisms. Expression array and proteomic technologies

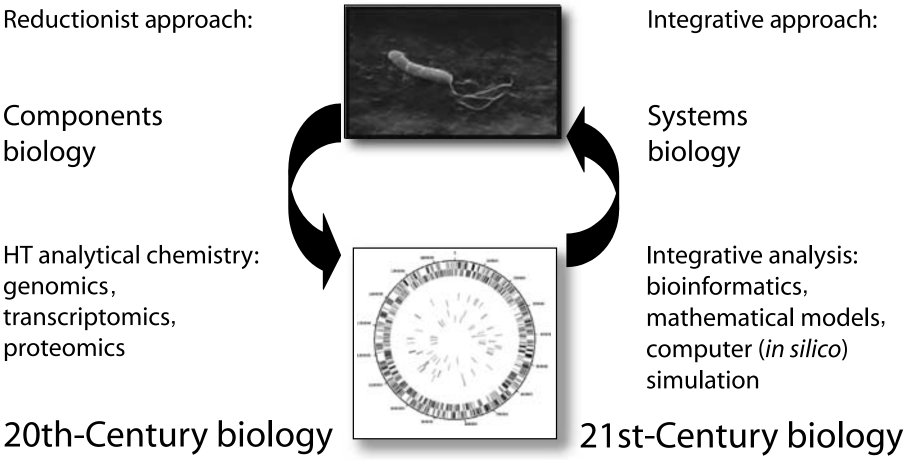
Reductionist approach:
Components biology
HT analytical chemistry: genomics, transcriptomics, proteomics
20th-Century biology
Integrative approach:
Systems biology
Integrative analysis: bioinformatics, mathematical models, computer (*in silico*) simulation
21st-Century biology

Figure 1.1: Illustration of a paradigm shift in cell and molecular biology from component to systems analysis. Redrawn from [152].

give us the capability to determine when a cell uses particular genes, and when it does not (left side in Figure 1.1). At the beginning of the 21st century, this process was unfolding at a rapid rate, driving a fundamental paradigm shift in biology.

Beyond bioinformatics

The advent of high-throughput experimental technologies is forcing biologists to view cells as systems, rather than focusing their attention on individual cellular components. Not only are high-throughput technologies forcing the systems point of view, but they also enable us to study cells as systems. What do we do with this developing list of cellular components and their properties? As informative as they are, these lists only give us basic information about the molecules that make up cells, their individual chemical properties, and when cells choose to use their components.

How do we now arrive at the biological properties and behaviors that arise from these detailed lists of chemical components? It is now generally accepted that the integrative analysis of the function of multiple gene products has become a critical issue for the future development of biology. Such integrative analysis will rely on bioinformatics and methods for systems analysis (right side of Figure 1.1). It is thus likely that over the coming years and decades biological sciences will be increasingly focused on the systems properties of cellular and tissue functions. These are the properties that arise from the whole and represent biological properties. These properties are sometimes referred to as "emergent" properties since they emerge from the whole and are not properties of the individual parts.

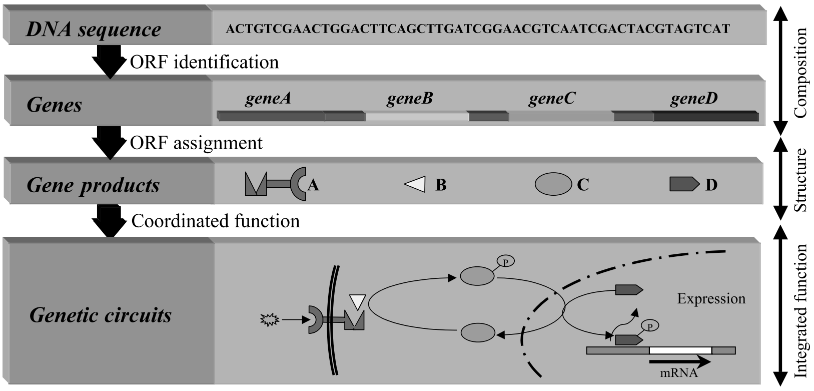

Figure 1.2: Genetic circuits. From sequence, to genes, to gene product function, to multicomponent cellular functions. Prepared by Christophe Schilling.

Genetic circuits

The relationship between genetics and cellular functions is hierarchical and involves many layers, some of which are illustrated in Figure 1.2. Gene sequences allow for the identification of open reading frames (ORFs). The base pair sequence of the ORFs in turn allows for the functional assignment of the defined gene. Although not always unambiguous, such assignments are being carried out with increasing accuracy, due to our expanding biological databases. Sequence is important and so is the functional assignment of ORFs. However, the interrelatedness of the genes may prove to be even more important. Establishing these relations and studying their systemic characteristics is now necessary.

Cellular functions rely on the coordinated action of the products from multiple genes. Such coordinated function of multiple gene products can be viewed as a "genetic circuit" (some synonyms that are commonly used are "cellular wiring diagrams" and modules). The term genetic circuit is used here to designate a collection of different gene products that together are required to execute a particular cellular function. The functions of such genetic circuits are diverse, including DNA replication, translation, the conversion of glucose to pyruvate, laying down the basic body plan of multicellular organisms, and cell motion. It is likely that we will view cellular functions within this framework and the physiological functions of cells and organisms as the coordinated or integrated functions of multiple genetic circuits. Consequently, we will need to develop the conceptual framework within which to describe and analyze these circuits.

Not all the properties of genetic circuits are clear at present, but some important ones are summarized in Table 1.1. For many of these characteristics, it is also clear what methodology is needed to describe and analyze

Table 1.1: Some of the characteristics of genetic circuits and the analysis methods required.

Characteristic	Analysis method
They are complex	Bioinformatics
They are autonomous	Control theory
They are robust	System science
They function to execute a physicochemical process	Transport and kinetic theory
They have ″creative functions″	Bifurcation analysis
They are conserved, but can adjust	Evolutionary dynamics

them. Genetic circuits tend to have many components; they are complex. From the standpoint of system science, they are “robust,” i.e., in many, but not all cases, one can remove their components without compromising their overall function.

Accepting the concept of a gene circuit seems straightforward. However, the implications of this acceptance are quite profound. We will view bioinformatics as a way to establish, classify, and cross-species correlate genetic circuits. The beginning of such classification is illustrated in Figure 1.3. Metabolism, information processing, and cellular fate processes represent some of the major categories of genetic circuits. Considerable unity in biology is likely to result in conceptualizing biological functions as genetic circuits. From this standpoint, gene therapy may no longer be viewed as replacing a “bad” gene, but instead fixing a “malfunctioning” genetic circuit. Evolution may be viewed as the “tuning” or “honing” of circuits to improve performance and chances of survival. Classifying organisms based on the types of genetic circuits they possess may lead to “genomic taxonomy.” *Ex vivo* “evolutionary” procedures for designing genetic circuit performance are emerging [99, 258]. Understanding the function of genetic circuits will become fundamental to applied biology, in fields as diverse as metabolic engineering and tissue engineering.

The concept of a genetic circuit as a multicomponent functional entity (either in time or space, or both) is an important paradigm in systems biology. It will be a fundamental component in our treatment of the relationship between genetics and physiology. the genotype–phenotype relationship. Individual genetic circuits do not operate in isolation, but in the context of other genetic circuits. The assembly of all such circuits found on a genome produces cellular and organismic functions and leads to hierarchical decomposition of complex cellular functions. Thus, the need for genome-scale analysis arises. This need in turn leads to viewing the genome as the “system.”

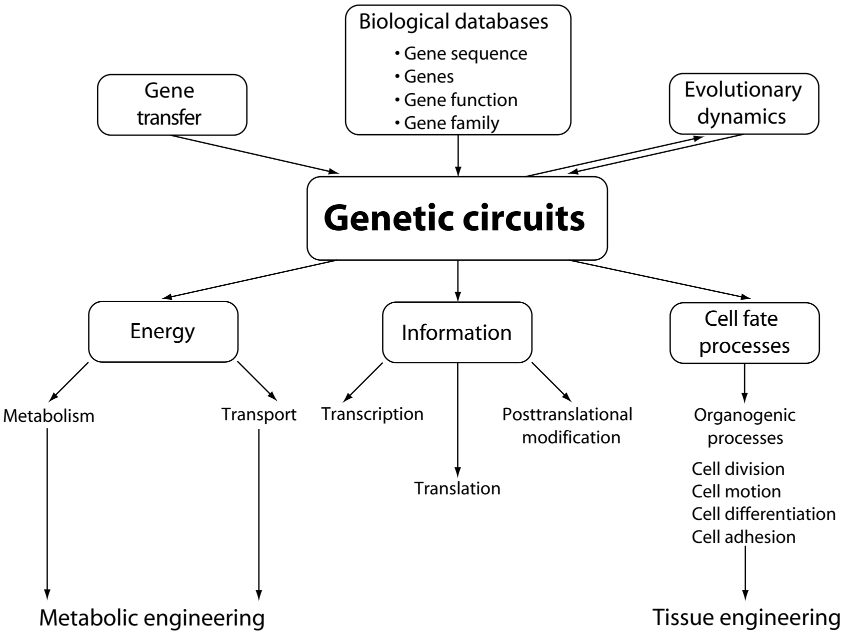

Figure 1.3: Coarse-grained classification of the types of genetic circuits that are found on genomes. Some major categories are indicated, in particular to indicate that some underlie the important metabolic and tissue engineering applications of cell and molecular biology. Prepared by Christophe Schilling.

1.2 The Systems Biology Paradigm

The ability to generate detailed lists of biological components, determine their interactions, and generate genome-wide data sets has led to the emergence of systems biology [101]. The process comprises four principal steps (Figure 1.4). First, the list of biological components that participate in the process of interest is enumerated. Second, the interactions between these components are studied and the "wiring diagrams" of genetic circuits are constructed. This process is one of biochemical reaction network reconstruction and is covered in detail in Part I of this text. Third, reconstructed network are described mathematically and their properties analyzed (Part II). Computer models are then generated to analyze, interpret, and predict the biological functions that can arise from the reconstructed networks (Part III). Fourth, the models are used to analyze, interpret, and predict experimental outcomes. Prediction essentially corresponds to generating specific hypotheses that can then be experimentally tested. These *in silico* models of reconstructed networks are then improved in an iterative fashion [152].

There is much creative work that has led to the development of high-throughput technologies (step 1). Many different mathematical methods

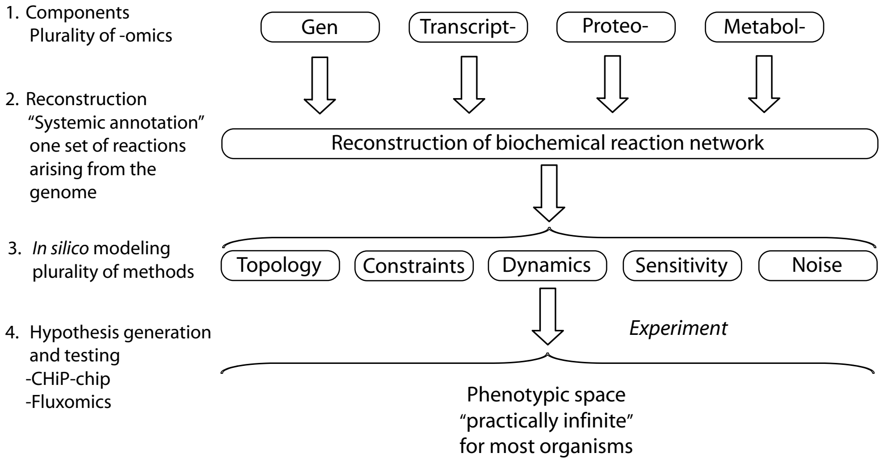

Figure 1.4: The four principal steps in the implementation of systems biology. Note that the second step is unique, while the others are diverse, and is the interface between high-throughput data and *in silico* analysis.

have been formulated for the analysis of biochemical reaction networks (step 3), and the phenotypic space explored by experimentation (step 4) is essentially infinite. In contrast, the reconstruction effort leads to one result. The reconstruction should culminate in the generation of the set of chemical reactions that take place inside a cell and underlie its function. Although systems biology is currently thought of as a cell-scale effort representing a genome-enabled science, it is likely that it will be conceptualized as a broader field as it develops. We will begin to talk about the systems biology of tissues, through networks of cellular interactions, and so on.

Systemic annotation

The unity represented by step 2 in Figure 1.4 leads to an effort to create a *two-dimensional* annotation of a genome (Figure 1.5). The classical component annotation of a genome leads to the identification of open reading frames, their location, and often the corresponding DNA regulatory sequences, a one-dimensional list of components. The open reading frames can then be assigned function based on homology searches of known genes. The two-dimensional annotation accounts for not only the components, but all their chemical states (represented as rows in the table in Figure 1.5) and the links between them. The latter are the columns in the table and ideally should represent the stoichiometric coefficients that correspond to the underlying chemical transformations that are possible between the components. This table represents the full genome-scale stoichiometric matrix for a genome.

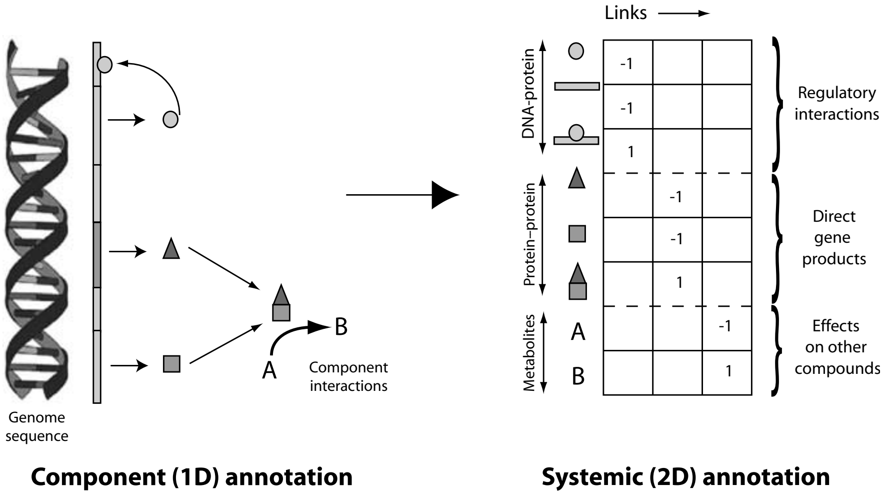

Figure 1.5: Systemic or two-dimensional annotation of genomes: the origin of the stoichiometric matrix. From [155].

Calling for the formulation of this matrix may represent as bold a statement as asking for the full base pair sequence of the human genome some 20 years ago. However, progress is being made. Genome-scale metabolic networks have been reconstructed for microorganisms. We are in the process of beginning to define signaling networks and transcriptional regulatory networks. Sometimes events in such networks are known chemically, but sometimes we only have causal relationships, which eventually will be converted into chemical equations once the underlying mechanisms are discovered.

Hierarchical thinking in systems biology

We are quite used to thinking hierarchically about DNA. We think about a base pair as the irreducible unit of DNA sequence. Then we talk about codons, introns, exons, alleles, and chromosomes, and other measures of DNA size. We will need to adapt similar hierarchical thinking about the genome-scale stoichiometric matrix. The irreducible elements in a network are the elementary chemical reactions. These can combine into reaction mechanisms, many reactions into modules or motifs, pathways can form, and sectors can be defined. Currently, such coarse graining of a network often relies on somewhat ill-defined notions of hierarchical structure.

Our understanding of how to hierarchically decompose a network is likely to improve as we begin to build genome-scale networks and are able to define their properties. Components that always function together in steady or dynamic states normally would fall into modules. Correlated

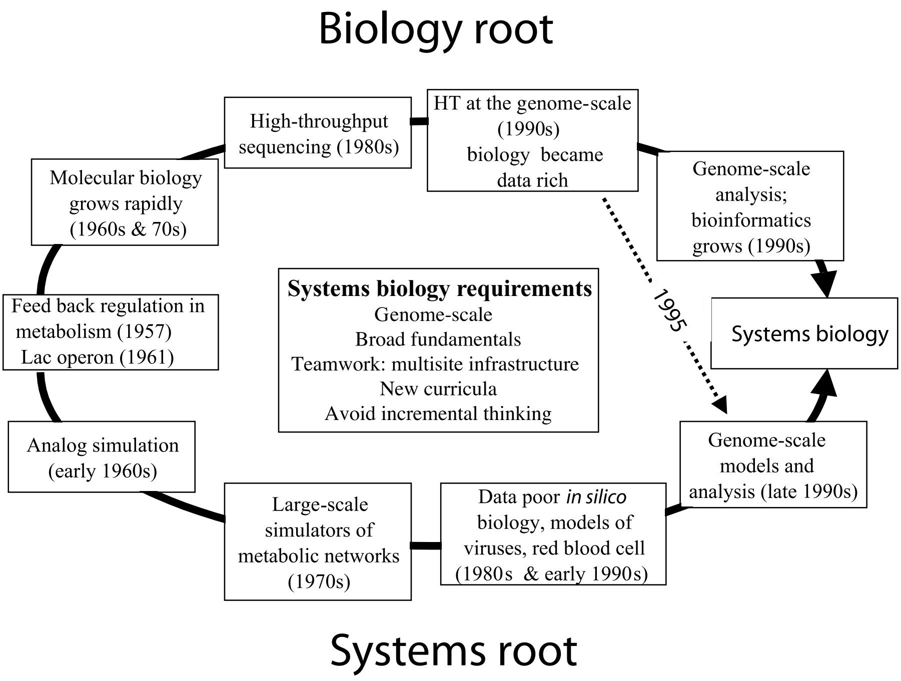
Biology root
High-throughput sequencing (1980s)
HT at the genome-scale (1990s) biology became data rich
Molecular biology grows rapidly (1960s & 70s)
Genome-scale analysis; bioinformatics grows (1990s)
Systems biology requirements
Genome-scale
Broad fundamentals
Teamwork: multisite infrastructure
New curricula
Avoid incremental thinking
1995
Feed back regulation in metabolism (1957) Lac operon (1961)
Systems biology
Analog simulation (early 1960s)
Genome-scale models and analysis (late 1990s)
Large-scale simulators of metabolic networks (1970s)
Data poor in silico biology, models of viruses, red blood cell (1980s & early 1990s)
Systems root

Figure 1.6: The two roots of systems biology.

subsets of reactions do appear in the delineation of steady-state properties of networks ([163], Chapter 9). Time-scale separation is often used for temporal decomposition of complex systems, and the stoichiometric matrix does seem to play a role in this formation of dynamics pools [109, 149] that represent dynamic course graining of a network.

Historical roots

Although it is often stated that reductionist thinking has characterized molecular biology, it does not mean that integrative thinking has not taken place. The first genetic circuits were indeed mapped out more than 40 years ago (Figure 1.6). The feedback inhibition of amino acid biosynthetic pathways was discovered in 1957 [225, 257], and the transcriptional regulation associated with the glucose–lactose diauxic shift led to the definition of the *lac* operon [12, 124]. These regulatory mechanisms began the unraveling of the molecular logic that underlies cellular processes.

In the decades following these discoveries, molecular biology blossomed as a field. In the 1980s we began to see the scale-up of some of the fundamental experimental approaches of molecular biology. Automated DNA sequencers began to appear and reached genome-scale sequencing in the mid-1990s. Automation, miniaturization, and multiplexing of various assays led to the generation of additional "omics" data types. The large volumes of data generated by these approaches led to a rapid growth of the field of bioinformatics. Although this effort was mostly focused on statistical models and object classification approaches in the late 1990s, it became recognized that a more formal and mechanistic framework was needed to systematically analyze multiple high-throughput data types [153]. This need led to calls for genome-scale model building.

Following the events of the late 1950s and early 1960s, efforts were initiated to formulate mathematical models to simulate the functions of the newly discovered genetic circuits. Even in the early 1960s, before digital computers became available, the function of such circuits was simulated on analog computers [78]. These efforts grew to the dynamic simulation of large metabolic networks in the 1970s [69, 123, 253, 255]. By the late 1980s and early 1990s, cell-scale models of the human red cell had appeared [106], genome-scale models of viruses were formulated, and large-scale models of mitosis appeared [146]. The advent of genome-scale sequencing led to the first genome-scale metabolic models for bacteria [50, 51].

These two roots of systems biology are illustrated in Figure 1.6. The upper branch had much greater presence in the scientific community, dazzling us with a never-ending stream of discoveries and exciting technologies. One might say that this was the "biology" root to systems biology. The lower branch never gained much notoriety, although, unlike in the United States, this activity was reasonably prominent in Europe. Systems modeling and simulation in molecular biology was seen as purely theoretical and not a contributor to understanding real biology. However, with biology now having become a "data-rich" field, the need for theory, model building, and simulation has emerged. One might think of this branch as the "systems" root to the emergence of systems biology.

These two branches must now merge to further the field. While there are many books and sources on the "biological root," few exist for the "systems root." This book is an attempt to meet this need, although real genome-scale analysis is still the material of cutting-edge research. Thus, this initial effort is conceptual and illustrative in nature with references to the genome-scale studies that have appeared.

1.3 About This Book

Purpose

The availability of annotated genome sequences in the mid to late 1990s enabled the reconstruction of genome-scale metabolic networks [37]. Similar reconstructions of signaling and transcriptional regulatory networks are now beginning to appear [85, 213]. The topological structure and functional properties of these networks can now be studied. More importantly, for the first time, we can analyze, interpret, and predict the phenotypic functions that such networks could produce. The stoichiometric matrix is a compact mathematical representation of biochemical reaction networks. It represents the interface between the HT data world and that of *in silico* analysis, e.g., Figure 1.4, and the two-dimensional annotation of a genome. The purpose of this book is to describe how the stoichiometric matrix is formed, what its basic properties are, and how it can be used to analyze the functional states of networks.

Approach

We will first outline some of the basic concepts of systems biology in Chapter 2. We will then divide the material into three parts.

Part I will briefly review three types of networks – metabolic, regulatory, and signaling – and show how they are comprised of underlying biochemical reactions. The efforts to reconstruct them are intensive in analyzing the data from various HT experimental technologies and legacy (bibliouric) data. Reconstructions basically culminate in the formation of a chemically, genetically, and genomically (BIGG) structured database that represents all the data types simultaneously. Once curated, a genome-scale reconstruction represents a BIGG structured integration of the available information about a cell or an organism.

Part II will describe the formation of the stoichiometric matrix, **S**, including its function as a mathematical mapping operation, the chemical constraints on its structure, and its topological properties. An understanding of basic linear algebra now becomes essential to the reader. The topological properties of the stoichiometric matrix are then outlined and methods for their analysis described. We then explore the more subtle and intricate properties of the stoichiometric matrix. To do so, we need to study its fundamental spaces associated with **S** and will thus require an intermediate-level understanding of linear algebra. The two null spaces of **S** contain systemically defined reaction pathways and concentration conservation quantities. The row and column spaces of **S** contain the dynamic flux vector and the time derivatives, respectively. These two spaces are thus key to studying the transient function and the underlying thermodynamics. The transition from Part I to Part II may be challenging for some life scientists, but it is important for mastering systems biology.

Part III will describe the mathematical methods that have been developed to interrogate the properties of reconstructed networks. The reconstructions and their associated information are not sufficient to completely define the state of a network. Flexibility in function exists, leading to constraint-based analysis. This approach is consistent with the