1 Introduction

E pluribus unum

Some 60 years ago, the promise of molecular biology held that if we knew and understood the function of the molecules that comprise cells, then we could understand cells and their functions. Although this was true in principle (and in practice in a few cases), the sheer number of molecules made it very difficult to comprehend so many simultaneous functions. The simultaneous measurement of the majority of these molecules became possible over the last 10–15 years through the development of many ingenious technologies. As a result, we now have a growing number of data sets that give us the composition of particular cells and organisms under certain conditions. The chemical interactions between many of these components are now known and this knowledge gives rise to reconstructed biochemical reaction networks on a genome-scale that underlie various cellular functions. Thus, enter (molecular) systems biology.

Systems biology is not necessarily focused on the components themselves, but on the nature of the links that connect them and on the functional states of the biochemical networks that result from the collection of all such links. These functional states of networks correspond to observable physiological or homeostatic states. Completing the relationship between all the chemical components of a cell, with their genetic bases, and its physiological functions is the promise of (molecular) systems biology. This undertaking represents the de facto construction of a mechanistic genotype–phenotype relationship.

1.1 The Genotype–Phenotype Relationship

The concept Through breeding experiments, Gregor Mendel discovered that there are discrete quanta of information passed from one generation to the next that determine the form and function of an organism. These quanta, or packets, of information are now generally referred to as genes. The collection of all the genes and the particular version of them found in a genome of an individual organism is referred to as its

genotype. The form and function of an organism is referred to as its phenotype. How the phenotype is related to the genotype represents the fundamental relationship of biology.

For monogenic traits, the genotype–phenotype relationship can be readily understood. One gene confers a phenotype. In the human population, there are now well over 100 of these traits documented and they can be diagnosed at the neonatal stage. However, most phenotypic traits involve coordinated functions of multiple gene products. This makes the genotype–phenotype relationship a challenge to reconstruct and understand. This challenge has two underlying issues. The first comes from the need to know what all the gene and gene products are, and the second comes from understanding the consequences of the complex interactions that can form among a large number of gene products. Today, the former can be addressed using omics data and the latter from the principles of systems analysis applied to biochemistry. The ability to address these challenges has developed over the last 10–15 years and it forms the basis for the bottom-up approach to molecular systems biology that, in turn, ultimately facilitates the realization of the promise of molecular biology.

Towards a mechanistic basis With the publication of the first full genome sequences in the mid 1990s [123], it became possible, in principle, to identify all the gene products that make up an organism. In practice, it has proven difficult to achieve complete or even comprehensive coverage of the genetic elements in simple genomes, but substantial progress has been made. The well-studied biochemistry of metabolic transformations made it possible to reconstruct, on a genome-scale, metabolic networks for a target organism in a biochemically detailed fashion [105,106]. Such metabolic network reconstructions can be converted into a mathematical format yielding mechanistic genotype–phenotype relationships for microbial metabolism [314].

The mathematical format of the underlying biochemical, genetic, and genomic (BiGG) knowledge facilitates the formulation of genome-scale models. Such models are not based on any biophysical theory, but simply represent the reconciliation of the known biochemical properties of the gene products expressed in an organism. Through such large-scale reconciliation, genome-scale models enable the computation of phenotypic traits based on the genetic composition of the target organism [314, 344]. Since the first metabolic genome-scale reconstruction in 1999 and *in silico* models thereof, many more have followed, including that for human metabolism [100]. The scope and content of network reconstructions continues to grow to include the entire transcription/translation apparatus of a cell, for instance [420], and the structural information about the metabolic enzymes [473].

We thus stand at an historical crossroads in the life sciences: the formulation of mechanistic genotype–phenotype relationships has become possible. Given the fundamental nature of this relationship, having mechanistic versions of it is foundational. Today, such relationships are being established for metabolic functions, with an increasing scope in biological content and coverage. Building mechanistic genotype–phenotype relationships and their use represent the scope and content of this book.

1.2 Some Concepts of Genome-scale Science

Paradigm shift The first full genome sequences emerged in the mid 1990s. At roughly the same time, mRNA expression profiling array and proteomic technologies gave us the capability to determine when a cell uses particular genes. These technologies allow us to achieve a genome-scale view of the contents of target organisms (left side in Figure 1.1). At the beginning of the twenty-first century, this process was unfolding at a rapid rate, driving a fundamental paradigm shift in biology.

The advent of high-throughput experimental technologies forced biologists to begin to view cells as systems, rather than focusing their attention on individual cellular components. Not only did the high-throughput technologies force a systems point of view, but they also enabled the study of cells as systems. What should one do with an available list of cellular components and their properties? As informative as they are, such lists only give basic information about the molecules that comprise cells, their individual chemical properties, and when cells choose to use their components. Such integrative analysis relies on bioinformatics and methods for systems analysis (right side of Figure 1.1).

Thus, at the turn of the century, molecular biology became focused on the systems properties of cellular and tissue functions. These are the properties that arise from the whole, and represent biological properties. In turn, genome-scale science emerged and started to grow.

Genetic circuits and molecular machines Cellular functions rely on the coordinated action of the multiple gene products. Such coordinated function results from what can

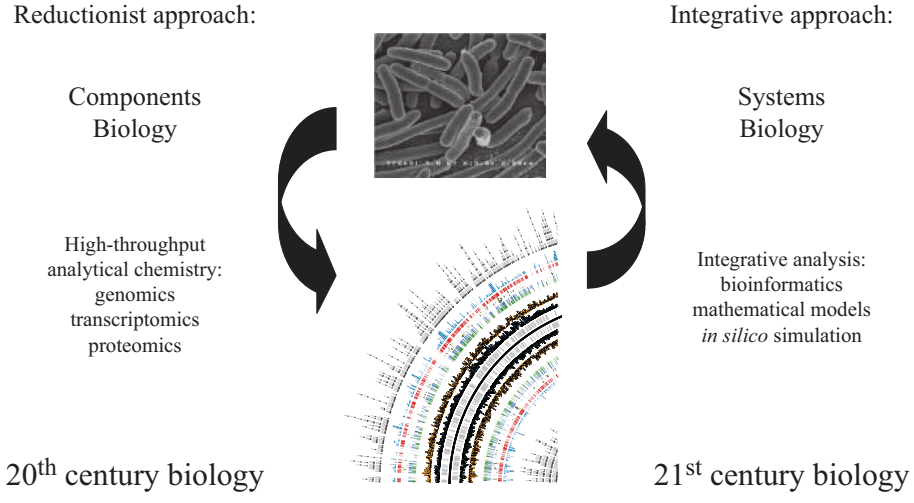

Figure 1.1 Illustration of a paradigm shift at the turn of the century in cell and molecular biology from components to systems analysis. Redrawn from [311]. Top image from Rocky Mountain Laboratories, NIAID, NIH. Bottom image courtesy of Byung-Kwan Cho.

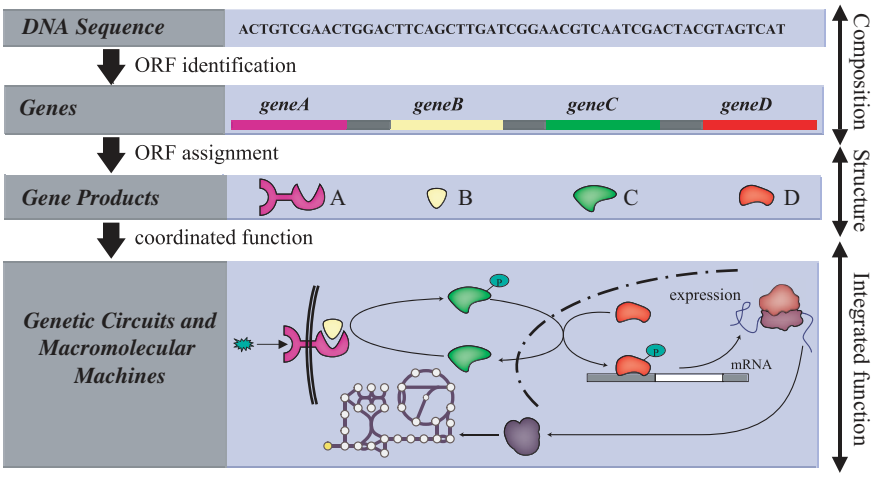

Figure 1.2 Genetic circuits. From sequence, to genes, to gene product function, to multi-component cellular functions. Prepared by Christophe Schilling and Nathan Lewis.

be called a *genetic circuit* (see Figure 1.2). The functions of genetic circuits are diverse, and include DNA replication, translation, the conversion of glucose to pyruvate, laying down the basic body plan of multicellular organisms, and cell motion. Cellular functions are increasingly viewed within this framework, and the physiological function of cells and organisms are viewed as the coordinated or integrated functions of multiple genetic circuits.

The concept of a genetic circuit as a multi-component functional entity – in time or space, or both – is important in systems biology. It is a fundamental factor in the establishment of genotype–phenotype relationships. 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 produce cellular and organismic functions, and leads to hierarchical decomposition of complex cellular functions and a multi-scale view of the genotype–phenotype relationship.

Genetic circuits function in the context of the entire organism A commonly used concept in biology is a pathway, an example of a genetic circuit. Figure 1.3 shows an amino acid biosynthetic pathway. One may be interested in various aspects of this pathway; its cellular localization, aggregation of the participating protein, debilitating genetic changes, etc. Such questions represent common pursuits in cellular and molecular biology.

A pathway does not function in isolation, however. It functions in the context of the entire network of interactions in the organism and may interact with many other cell processes. Such interactions can be weak or strong. With the advent of genome-scale network reconstructions, we can now place the function of pathways in the context of all the other known processes in a cell (Figure 1.3).

This genome-scale point of view has proved important in many settings. In the case of metabolic engineering, where new metabolic phenotypes are being built, it is

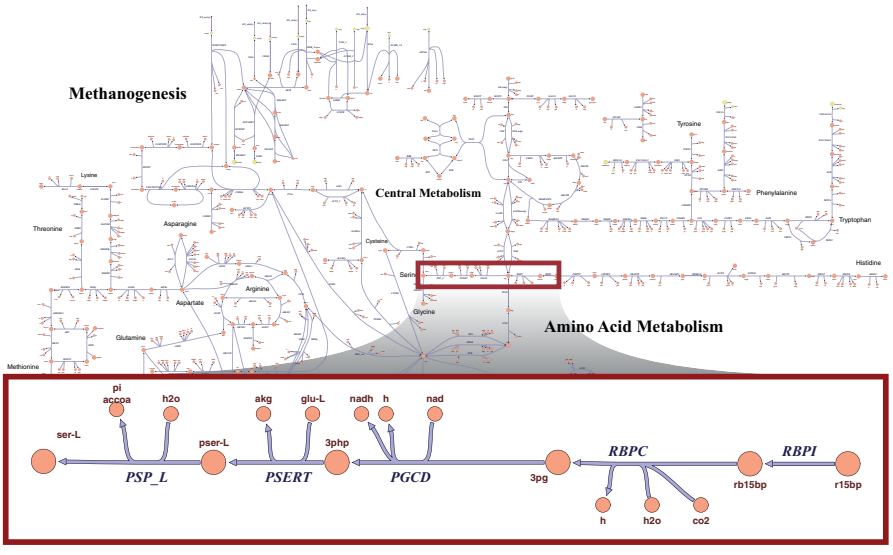

Figure 1.3 Illustration of the functions of a pathway in the context of a whole network. Prepared by Adam Feist.

not enough to identify and express all the genes of the pathway; one must also make sure it functions properly in the network as a whole. An over-expressed pathway may, for instance, drain key biosynthetic precursors, cause an imbalance in redox metabolism, or simply crowd out other cellular functions, leading to a sick or dead host cell if proper balancing of the function of the pathway relative to the whole network is not achieved.

Myriad constraints: the improbability of life The multiplicity of simultaneous molecular and genetic circuit functions occurring in a cell is mind-boggling. However, they all take place in a coherent, organized manner to produce functioning phenotypic states. In a growing bacterial cell, as many as 2.5 million protein molecules are functioning coherently. Ten thousand ribosomes and one hundred thousand tRNAs are busy synthesizing protein molecules. Thousands of RNA polymerases are synthesizing messages to be translated. And all of this happens in a volume of a cubic micron (Figure 1.4). Thus, countless constraints are placed on these functions.

We want to compute these functions with an *in silico* analog of the real cell through a mathematical model. It is popular to state that as the complexity of a mathematical model grows and the number of parameters increases, anything is possible. Inside a cell, nothing could be further from the truth. The numerical range of parameter values that allow these very complex networks to function coherently are severely restricted. Sometimes it is hard to believe that there is even one set of parameter values that allows the living process to take place. Life is indeed improbable.

Evolution and optimization Finding a functioning set of parameter values is the result of a long process of trial and error where the genetic elements change slightly

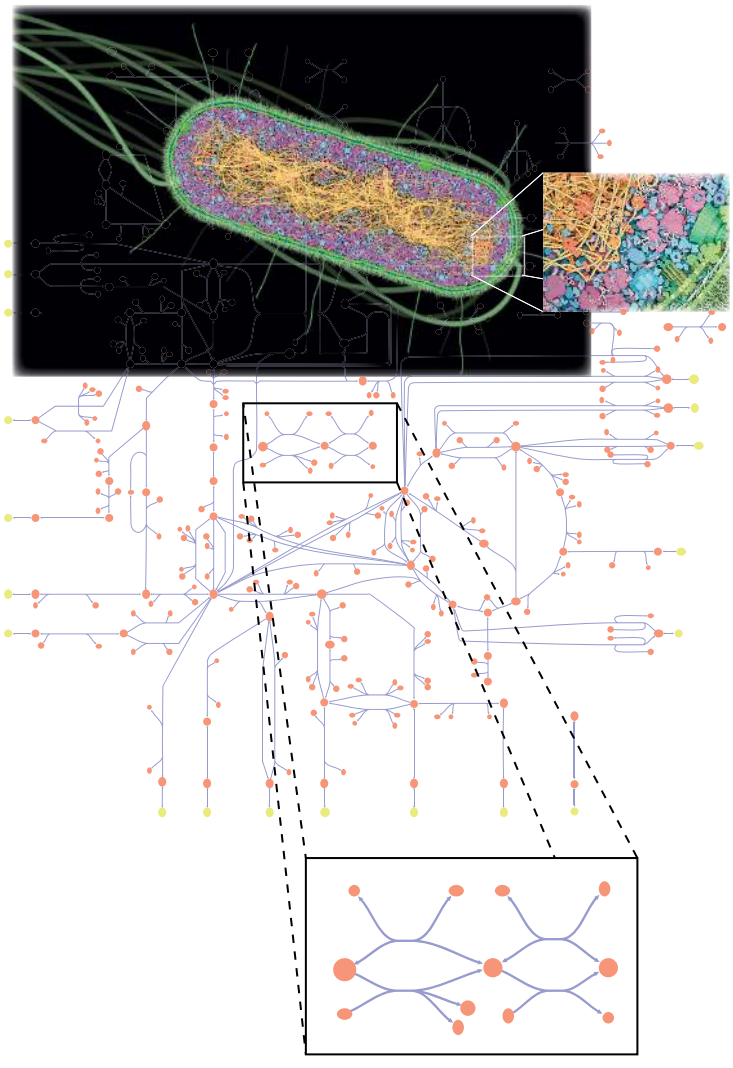

Figure 1.4 A cell operates in a confined space in which many components function simultaneously as a system to produce phenotypic states. Inserted image from [148] used with kind permission from Springer Science+Business Media B.V.

from generation to generation. Evolution is an ongoing optimization process that steadily hones the functions of the existing gene products and adds new ones through mechanisms such as gene duplication and horizontal gene transfer. This process of trial and error is never-ending. Poor choices are punished by extinction while more functional alternatives continue the process.

The change in organism properties with subsequent generations is called *distal causation*. Given the selection process that is at work, key components for describing distal causation are optimization principles. Clearly, one way to generate increased functionality and complexity is through the generation of hierarchy, where new

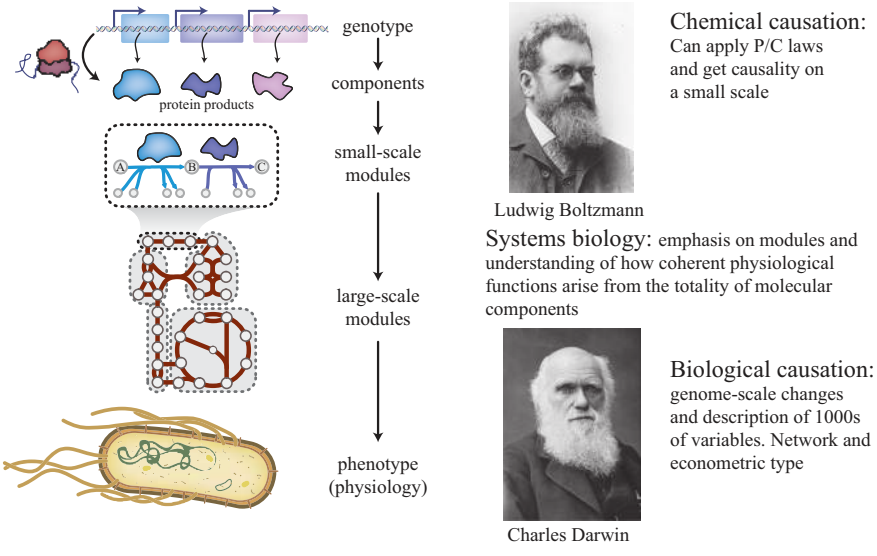

Figure 1.5 Hierarchy and modularity in biology. Prepared by Nathan Lewis.

functions are built around existing ones. Hierarchy in biology exists in space, time, component abundance, and other properties. Hierarchical organization is key to understanding the genotype–phenotype relationship.

Hierarchical thinking is important in systems biology We are quite familiar with thinking hierarchically about DNA. We think about base pairs as the irreducible unit of DNA sequence. Then we talk about codons, introns, exons, alleles, chromosomes, whole genomes, and other similar measures of DNA size. We can understand them readily, even if we have to scan over nine orders of magnitude in sequence length as in the case of the human genome.

We will need to adapt similar hierarchical thinking techniques to the properties of genome-scale networks. 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, as illustrated in Figure 1.5. Currently, coarse-graining of a network relies on objective or subjective definitions [325] of 'modules' that are used to conceptualize a hierarchical network structure.

Our understanding of how to decompose a network hierarchically is likely to improve as we gain a better understanding of the functions of genome-scale networks and our ability to define their properties. Components that always function together in steady or dynamic states normally would fall into modules. Correlated subsets of reactions do appear in the delineation of steady-state properties of networks (Chapter 12). 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 dynamic pools [198, 309] that represent the dynamic coarse-graining of a network. Thus, measures of multi-scale thinking are indeed developing in the field.

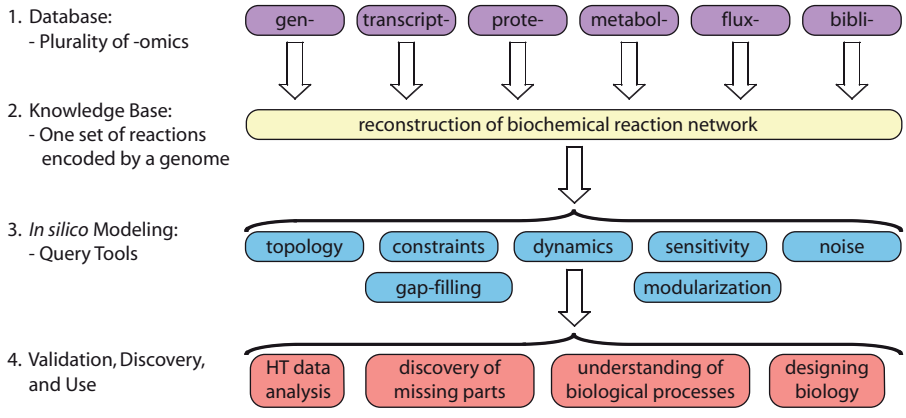

Figure 1.6 The four principal steps in the implementation of systems biology. Note that the second step is unique, while the others are diverse, and it is the interface between high-throughput data and *in silico* analysis. Modified from [118,314].

1.3 The Emergence of Systems Biology

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 a fundamental paradigm for systems biology [178]. It is composed of four principal steps (Figure 1.6):

- First, define and enumerate the list of biological components that participate in a cellular process.
- Second, the interactions between these components are studied, the 'wiring diagrams' of genetic circuits are reconstructed, and genome-scale maps are formed in a step-wise manner. This process is one of biochemical reaction network reconstruction.
- Third, reconstructed networks are converted into a mathematical format that formally describes the biological knowledge that underlies the reconstructed network. Computer models are then generated to analyze, interpret, and predict the biological functions that can arise from reconstructed networks.
- Fourth, the models are used in a prospective manner. Prediction entails generating specific hypotheses that can then be tested experimentally. These *in silico* models of reconstructed networks are then improved in an iterative fashion [311].

Much creative work has led to the development of high-throughput technologies (Step 1). Workflows now exist for network reconstruction (Step 2). Many different mathematical methods 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 need for genome-scale models It is a common experience of students of biochemistry to come across the same molecule in different chapters of their textbooks.

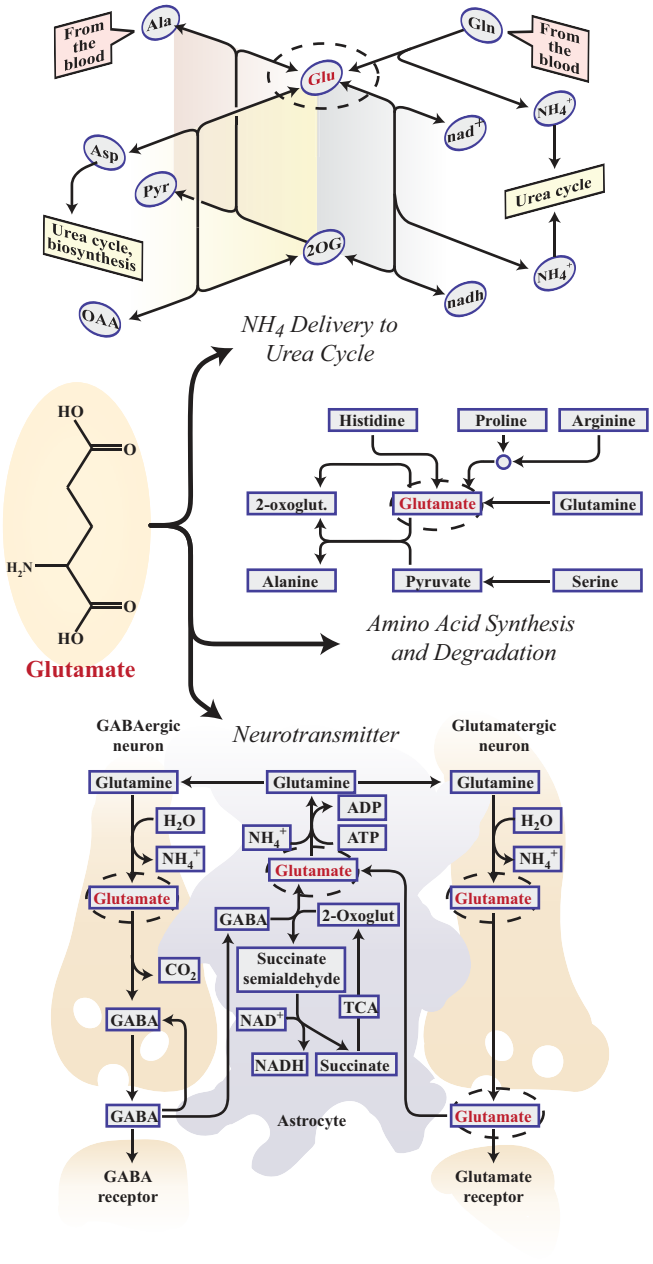

Figure 1.7 The many roles of glutamate. Images adapted from [216] (reprinted with permission). Prepared by Nathan Lewis.

After learning about one function of a molecule, another function arises in a different context. One example is given for glutamate in Figure 1.7. This figure shows the role of glutamate in the urea cycle, as a neurotransmitter, as well as its biosynthetic and degradative pathways. How can all of these functions be reconciled? Thoughtful students of biochemistry struggle with this question.

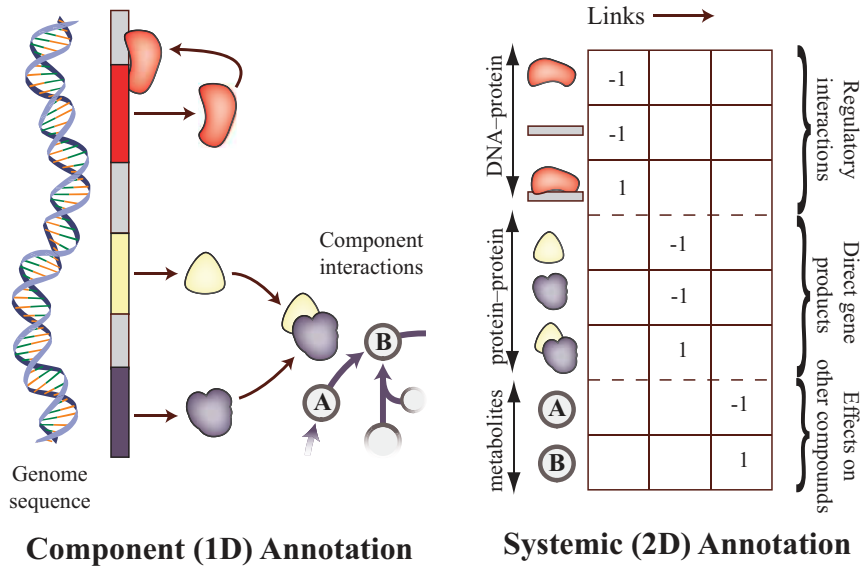

Figure 1.8 Systemic, or two-dimensional annotation of genomes: the origin of the stoichiometric matrix, **S**. Modified from [313].

Enter the power of systems science. It turns out that once a genome-scale network is built, all these functions can be reconciled simultaneously with a simple matrix equation. This ability may seem far-fetched at first, but it is achieved through relatively simple accounting principles. For this reason, bottom-up systems biology becomes a mathematical pursuit as its principal goal cannot be achieved without mathematics.

Two-dimensional genome annotation The unity represented by Step 2 in Figure 1.6 leads to an effort to create a *two-dimensional annotation* of a genome (Figure 1.8). The classical annotation of a genome leads to the identification of open reading frames, their location, and often the corresponding DNA regulatory sequences; basically, a one-dimensional list of components. The open reading frames can then be assigned function based on homology searches of known genes. If the function of a gene product is known, one can describe its interactions with other known gene products, resulting in components and the links between them; fundamentally, a two-dimensional description.

A two-dimensional annotation therefore not only accounts for the components, but all their chemical states (represented as rows in the table in Figure 1.8) and the links between them. The links are represented as columns in the table in Figure 1.8 and ideally should represent the stoichiometric coefficients that correspond to the underlying chemical transformations that are possible between the components. In principle, this table represents a full genome-scale stoichiometric matrix (**S**) for a genome.

Calling for the formulation of this matrix may represent as bold an undertaking as asking for the full base pair sequence of the human genome over 20 years ago.