Metabolomics

JONATHAN ARNOLD, H.-B. SCHUTTLER, D. BATTOGTOKH, JAMES
GRIFFITH, B. ARPINAR, S. BHANDARKAR, K. J. KOCHUT, E.
KRAEMER, J. A. MILLER, A. SHETH, G. STROBEL, T. TAHA, and B.
ALEMAN-MEZA
University of Georgia Athens
Georgia, U.S.A.

D. LOGAN, JEREME DOSS, LATREACE HARRIS, and ABASSI
NYONG
Clark Atlanta University Atlanta
Georgia, U.S.A.

S. DATTA
Georgia State University Atlanta
Georgia, U.S.A.

1. INTRODUCTION

1.1. Metabolomics

Metabolomics is the computing of emergent properties of biological systems such as development, biological clocks, and infection processes from kinetic models of DNA, RNA, and proteins. These kinetic models are used to guide the process of gene-validated product discovery to transform medicine, industry, and agriculture. The ultimate challenge of genetics is to predict global properties of the organism, properties not necessarily manifested by individual subcomponents within the cell. Some of these properties only make sense with respect to the organism as a whole (e.g., pathogenicity lifestyle [1]) or life itself. These complex traits controlled by many genes represent the ultimate challenge in seeking an explanation in terms of detailed molecular mechanisms in the cell.

From the standpoint of human health, an explanation is sought for how an organism such as the opportunistic fungal pathogen, Pneumocystis, changes from a benign commen-
sal in the mammalian lung to the major killer of AIDs patients through a lethal pneumonia [2]. From the standpoint of agriculture, one of the major challenges of peanut production is controlling an opportunistic pathogen, *Aspergillus flavus*, that causes aflatoxin contamination in peanuts. Controlling aflatoxin biosynthesis has consequences for human health, for the quality of a major U.S. crop, and for domesticated animals ingesting contaminated peanuts [3,4]. From the standpoint of industry, fungi are producers of chemical feedstocks and biologicals, ranging from ethanol to taka-amylase and citric acid [5,6]. From the standpoint of fundamental questions in biology, an explanation is needed for how a fungus programs the development of a conidiophore [7] or captures the diurnal cycle within the cell [8]. One approach to explaining these global processes is through the identification of a biochemical and regulatory network that rationalizes these processes with a mechanistic model [9]. Unlocking these regulatory and biochemical networks provides an opportunity for their manipulation either through targeting of critical steps in metabolism for the discovery of antifungals or through manipulation of pathways to overproduce needed compounds like penicillin.

1.3. Paradigm Shift in Biology

Biology is currently going through a paradigm shift driven by microbial systems. The discipline is becoming data-driven through the avalanche of genomics information being released on a variety of fungal systems [10]. The discipline of fungal biology has become high-throughput, with vast amounts of data robotically generated through the use of automated sequencing machines [11] and the use of microarrays for analysis of gene expression [12] as well as mass spectrometers for protein–protein interaction mapping [13]. These data are highly structured and hierarchically organized [14]. At the center of any biochemical and regulatory network, whether it be the *lac* operon [15] or the biological clock [16], is the central dogma describing the most fundamental flow of information in the cell from DNA → RNA → protein. Within a cell, this dynamic flow of information is hierarchically arranged. Functionally, reaction networks have structure [17]. At their highest level, they are organized into broad functional categories, such as energy metabolism, nucleotide metabolism, recombination, and DNA repair. At a lower functional level, within any one of these functional categories, there is a finer definition of function in terms of genes and their products involved in, for example, the Embden-Meyerhof pathway, Krebs cycle, and oxidative phosphorylation. At the lowest level in the functional hierarchy, there is a particular pathway [18,19].

This information flow also has a structural hierarchy. Genes and proteins do not work in isolation within a reaction network [20,21]. Rather, proteins form complexes that carry out the work of the cell, such as signaling. Signaling cascades of proteins may work in a coupled fashion to connect the surface of the cell with the nucleus in order to respond to different environmental conditions [22]. These signaling wires themselves are made up of shared components to allow a coupled response to environmental signals. In other parts of the cell, proteins form smaller aggregates to carry out a specific function, such as transcription, which in turn aggregate to form a “some” like the transcriptosome composed of more than 100 proteins [23]. New tools are being developed to identify these molecular machines [24,13]. These subcellular structures within the cell may have arisen from simpler precursors, and the structure of these molecular machines may in part reflect their history [25].
Metabolomics

The information in the cell is hierarchically organized through its history. The shared thread of the DNA links organisms into a reticulate structure, in which the history of genes traces out the organismal phylogeny linking all organisms in the tree of life. The appearance of each new mutation in the DNA can be viewed as the ticking of a molecular clock. This ticking can be used to link organisms into families, which can in turn linked into pedigrees, which in turn give rise to genera, which in turn radiate into the larger taxonomic branches. This evolutionary organization is played out at different levels by comparing genes evolving at different rates [26]. Through the consideration of the detailed mechanics of the cell, biology has thus become an information science from a functional, structural, and evolutionary standpoint.

Because of the avalanche of information resulting from the genomics revolution, biology has changed into a mathematical discipline. Extensive automation is required to capture the data [27] through laboratory information management systems [28]. The information needs to be stored, managed, and retrieved in sophisticated databases [28–30]. The information needs to be integrated with new algorithms [31–37] and with new tools such as the semantic web to make queries of the diverse resources now available for identifying reaction networks [38]. Models are being created to summarize the information [39,40]. This information must be analyzed to test hypotheses about the structure, function, and evolution of living systems [41]. Finally, the information needs to be visualized [42–44] to be understood and utilized. Computer scientists, mathematicians, and statisticians are engaged in all aspects of biology as an information science.

With the focus on complex traits involving many genes and their products, a new approach is needed that is more familiar to ecologists and neural biologists. This systems approach is at the heart of genomics. Measurements are taken on the system as a whole. The relative levels of all RNAs are measured [12]. The relative levels of all proteins are captured from crude protein extracts [45]. The response of the system as a whole is measured by capturing all RNA and protein levels in the cell. The ability to predict the global response of the system becomes the ultimate test of a biological hypothesis.

The promise is that by measuring the global response of the system, we can understand and predict complex traits [46]. Currently, this is only a promise by genomics, but it is a compelling challenge to move beyond Mendelian genetics. Most of the traits of interest, such as antibiotic production, pathogenicity, or clocks, are controlled by many genes and are tightly coupled to other processes [47].

In this section, metabolomics as a discipline has been defined, and connections made with metabolic engineering. In section 2, the origin of metabolomics is explored. In section 3, the models or ‘biological circuits’ behind metabolomics are sketched with their applications. In section 4, the process of discovery at the heart of metabolomics is considered. In section 5, the process of identifying biological circuits is considered along with the challenges. Lastly, in section 6, metabolomics is put in a larger context and summarized.

2. BIRTH OF METABOLOMICS AS A SYSTEMS SCIENCE

The challenge of genetics is formulating a detailed understanding of complex traits, particularly those that characterize the organism as a whole. Examples include high blood pressure, biological clocks, sex, development, and pathogenicity. Much of what we know about the biological clock, for example, comes from the study of a particular fungal system, Neurospora crassa [8]. Fig. 1 shows an example of this emergent property of N. crassa, the regular temporal sequence of conidiation by this organism growing in race tubes.
Figure 1  The promise of genomics is in understanding complex traits such as the circadian clock. An example of the effects of gene mutations (wc-1 and wc-2) on circadian oscillations in *Neurospora crassa* is redisplayed. *N. crassa* is shown growing in race tubes with the regular pattern of conidia being formed displaying the clock under varied conditions. (From Ref. 148. Copyright 1997 American Association for the Advancement of Science.)

When transcriptional activators such as white-collar 1 (wc-1) and white-collar 2 (wc-2) are knocked out, the organism loses its ability to tell time. The extent of the circuit is unknown, but a number of genes are now implicated in the functioning of the circadian oscillator. One goal is to be able to predict the oscillatory response from a detailed biological circuit specifying the function of genes and their products.

Traditionally, the subject of quantitative genetics has focused on complex traits [48]. The approach has been model-based with the hypothesis of several loci on chromosomes contributing to a particular complex trait. Assumptions about dominance, penetrance, and epistasis are made, and then predictions about the inheritance of the trait are calculated. The subject has given rise to powerful methods for identifying quantitative trait loci (QTLs) that affect particular complex traits [49–52]. When QTLs are integrated with other kinds of genomic information, precise predictions of the location of genes can be made [52,53]. Unfortunately, this approach is divorced from a detailed understanding of genes and their products. In the end, the explanation of a complex trait is only a location on a chromosome.

Both genomics and quantitative genetics have a common goal, the understanding of complex traits. The challenge is how to transform genomics from a data-driven discipline to hypothesis-driven science. One approach is to cross genomics and quantitative genetics. The result is metabolomics. Metabolomics at its outset embraces the model-driven approach of quantitative genetics and combines this with the data-driven discipline of genomics. Metabolomics thus becomes hypothesis-driven genomics.

Metabolomics begins with the data-rich foundation of genomics. The starting point is the entire DNA sequence of an organism. This resource is used to capture RNA and protein profiles (i.e., the cellular state) under varied conditions. Models of the complex
trait are introduced to explain the trait in terms of RNA levels, protein levels, and metabolite levels plus the organization of genes, their products, and substrates in the cell. The models serve to explain and predict, using the data-rich foundation of genomics. Predictions are made about the complex trait and the global state of the system from a detailed understanding of DNA, RNA, and proteins. The success or failure of a scientific hypothesis can be judged in this wider genomic context.

2.1. System State

One of the remarkable advances of genomics has been in obtaining a fairly complete description of what the cell is doing. It is now possible to measure all relative RNA and protein levels in microbial systems [12,45]. Varied strategies can be used to examine gene expression, including differential display, subtractive hybridization, and restriction fragment differential display. In particular, two technologies have come to the fore—microarray analysis [54] and serial analysis of gene expression (SAGE) [55]. Some comparisons of these approaches have been made [56], and the result is that each method identifies different subsets of the total RNA population. With microarray analysis, varied implementations exist.

2.1.1. RNA Profiling by Microarraying

One illustration of the approach developed by DeRisi et al. [12] is described in Chu et al. [57], in which microarray analysis is used to analyze an emergent property of all living systems, reproduction. RNAs are isolated from 10 different time points in sporulation and the meiotic cell cycle, reverse transcribed, labeled with a red or green chromophore, and the cDNAs (red) from each time point mixed with cDNA derived from the 0 time point (green). This cocktail is then probed against all 6000 genes in the yeast genome (Fig. 2) [58]. The advantages of this approach are the linearity of signal response, the presence of an internal control (by mixing the cDNAs from different sources into one probe), and the simple approach to visualizing the transcriptosome.

One limitation has been an interaction between the source of the RNA and color label (i.e., the red or green chromophore), which has led others to radiolabeling cDNAs [23]. Seven clusters of genes are differentially regulated during sporulation [41], and the genes are clustered by the similarity of their profiles as shown in Fig. 2 [58]. This information then becomes a resource for detailed hypotheses about the cell cycle [59].

2.1.2. RNA Profiling by Serial Analysis of Gene Expression

An alternate approach and the one first used to characterize the yeast transcriptosome is simply to sequence efficiently the resulting cDNAs from different cellular states and to count the RNAs present (i.e., SAGE [55]). With the ability to quantify expression of all genes, the next step in the information flow of the central dogma is capturing relative protein levels in the cell.

2.1.3. Protein Profiling by Isotope-Coded Affinity Tagging

Isotope-coded affinity tagging (ICAT) has been used to characterize the GAL cluster in yeast [45]. Protein profiling can help us to identify genes that are under translational control [16,60] as well as provide a more complete description of the cellular state. The ICAT reagent contains a sulfhydryl-specific reaction group (iodoacetamide) to label cysteines, an affinity ligand (biotin) to capture the protein, and a linker region that contains
either eight or zero deuterium atoms (D8 or D0) to label the cellular state. In the case of
yeast, Gygi et al. [45] compared proteins in cells grown on galactose or ethanol as a carbon
source as the two cellular states.

Using the ICAT reagent, they were able extract and identify more than 800 proteins
that responded differentially to change in carbon source. Data collection operated in two
modes on a mass spectrometer. In one mode, peaks coming off the column were used to
identify proteins from their BN-Y fragments. In the other mode, pairs of peaks were
captured separating the two labeled forms (D0 vs. D8) of each protein to quantify the
relative amounts of particular proteins in the two cellular states (labeled by D0 and D8).
The use of the cysteine label decreased the complexity of the protein mixture and thus
increased the opportunity to characterize more proteins in the cell. This is an attractive
approach because of the use of an internal control, dual peaks as a form of replication (D0
and D8), and the ability to analyze insoluble proteins in contrast to other mass spectrometric
methods, such as MALD-TOF-MS/MS [13]. The major limitation is resolving all the
proteins in a cell-free protein extract.

2.2. A Journey into the MudPIT

In multidimensional protein identification technology (MudPIT), the aim is to resolve all
proteins in the proteome. One approach that has been successful is to combine multidimen-
Metabolomics

Professional chromatography with electrospray ionization on a mass spectrometer [61,62]. Peptides are systematically separated by charge in one dimension and by hydrophobicity in another dimension. An SCX cation exchange column was used to separate by charge and preceded by a prefractionation step on hydrophobicity prior to the dual liquid chromatography LC/LC step. Wolters et al. [62] estimated that up to 23,000 peptides could be separated using this approach. Using protein extracts from *S. cerevisiae*, they were able to separate 5540 unique peptides (approximately 1484 proteins) from a complex mixture. The estimated dynamic range for detection varied 10,000 to 1 with a lower detection limit of 100 copies of a protein per cell. The major advance MudPIT provides is reaching the insoluble protein fraction and more than doubling the number of resolvable proteins with two-dimensional polyacrylamide gel electrophoresis (2D-PAGE) [13]. The major limitation is still the complexity of the protein mixture.

2.3. Who Counts the Small Molecules?

Having completed the review of large molecules, a segue to the characterization of the metabolite profiles is needed, and some initial efforts are reviewed [63,64]. Some of the new approaches to metabolite profiling are discussed in Chapter 14 [64]. A variety of separation procedures are being explored.

2.4. System Measurements

The basic measurements available on the system are the RNA and protein levels in different cellular states along with the levels of small molecules as available. These transcriptional and protein profiles become the resource to which the biological circuit is fitted.

2.5. Making Genomics Hypothesis-Driven

The process of making genomics hypothesis-driven is summarized in Fig. 3. The state of the system is captured in the RNA and protein profiles and whatever elements of metabolite profiles can be captured. These data are then used to identify a formal kinetics model to describe what genes and their products are doing. The classic elements of a biochemical reaction network are shown in Figure 3. The simplest kind of reactions are those that lead to Michaelis-Menten kinetics [65]. Once the table of reactions is specified, the profiling information can be used to identify the rate constants and initial conditions of the biochemical and gene regulatory reaction network.

One of the simplest kinds of reaction networks is a pair of coupled signaling cascades as shown in Fig. 3. A receptor protein (R or R\(^1\)) at the plasma membrane responds to an incoming signal, such as a pheromone or osmolarity [22]. The message is passed to a G-protein (G or G\(^1\)) which, through a signaling cascade, activates a transcriptional activator (E or E\(^1\)) to program the cell for an adaptive response. Even this simplest of systems can display emergent properties such as memory of the incoming signal, which the individual signaling wires by themselves do not manifest.

Once the kinetics model, or so-called biological circuit, is identified, familiar simulators like GEPASI, MIST, or SCAMP can be invoked to yield predictions about the workflow process, which is defined as an automated organizational process involving these tasks. Recently, laboratory information management systems (LIMS) systems have been introduced to design and manage these workflow processes [27]. An automated workflow management system (WFMS) is the collection of tools enabling workflow creation (which
Figure 3  Genomics is made hypothesis driven by utilizing a chemical reaction network to integrate genomics information and to make predictions about emergent properties of the system of interest.

3. MODELS

A variety of modeling approaches for biological circuits have been proposed. These include linear models with time as a factor [74], linear dynamic [75], Bayesian networks [76,77], neural networks [78], Boolean networks [79–81], and classic chemical-reaction networks.
Metabolomics 605

satisfying mass balance [39]. At one extreme, Boolean networks are draconian simplifications of chemical reaction networks satisfying mass balance, but they may be informative about crucial links in large reaction networks. At the other extreme are the stochastic formulations of reaction networks, in which the fate of individual molecules are tracked by applying a set of master equations summarizing the chemical reactions. Deterministic reaction networks strike a balance on this spectrum. The success of these competing approaches will ultimately be decided by the data. The focus here is on classic chemical-reaction network models because they are well grounded in physics and chemistry. These chemical reaction network models can either be deterministic [39] or stochastic [82,83].

For most reactions, enforcement of mass balance leads to specification of a system of differential equations to describe this reaction network [39]. Alemen-Meza et al. explain how the reaction network captured in the circuit of Fig. 4 is translated into a coupled system of nonlinear differential equations [40].

3.1. Water Models

Because most of the examples here are drawn from respiration, the modeling framework is illustrated with one of the simplest examples of combustion, the mixing of molecular oxygen and hydrogen, as shown in Fig. 4. This network diagram is the model. Circles denote reactions, and squares denote reactants or products. The arrows define the forward direction of a particular reaction. Incoming arrows lead from reactants, and outgoing arrows lead to reaction products. The end product is water, and we term this simplest of models, water model I (or the simple water model). Take reaction 1 as an example:

\[ R1: \text{H}_2 + \text{O} \rightarrow \text{H} + \text{OH} \] (with left to right as the forward reaction). Each such reaction has a pair of reaction constants: the forward reaction constant \((k_f)\) and backward reaction constant \((k_b)\). The net

![Figure 4](image-url) Water model 1 is a simple example of a chemical reaction network. Squares indicate chemical species, and circles indicate chemical reactions. Species with arrows pointing into a reaction are reactants, and species with arrows towards themselves and out of a reaction are products. In essence, the arrows define the forward reaction direction.
rate of production of species OH due to reaction R1 would be given by the simplest multiplicative kinetics by [65]:
\[
d\frac{[OH]}{dt} = k_f [H_2][O] - k_b[H][OH]
\]
where, for example, [OH] denotes the concentration of OH at time t

The total rate of production of a species is then obtained by summing over reactions, containing OH for instance:
\[
d\frac{[OH]}{dt} = \sum_r d\frac{[OH]}{dt}
\]

The system of six differential equations characterizing the behavior of the reaction network can be found in [40], and the reader is encouraged to try the simulator KINSOLVER for this simple reaction network, found at http://gene.genetics.uga.edu/stc. With the advent of simulators like KINSOLVER, the focus for biologists then simply becomes to identify the biological circuit. This is the model.

As chemists have accumulated more kinetics data, they found the initial reaction network was an oversimplification of what makes some rockets go up (i.e., H2 and O2). The reaction network or circuit needed to be refined to that in Fig. 5 (water model 2). This inclusion of additional species and/or reactions is typical of building a model that fits a biological system.

### 3.2. Carbon Metabolism

A slightly more complicated biological circuit can be constructed for one of the two early paradigms for eukaryotic gene regulation [84,85] along with the GAL cluster in *S. cerevis-

![Figure 5](image) Water model 2 is an elaboration of water model 1, as dictated by being able to predict the kinetics of the reactions.
Figure 6  A pictorial summary of what is known about QA metabolism or informal biological circuit for QA metabolism. There are seven genes in the qa cluster that are coordinately regulated. Four of the genes are thought to participate in QA metabolism. Two of the genes are regulators of QA metabolism. The gene qa-1F is a transcriptional activator; the gene qa-1S is a repressor that is hypothesized to bind to the activator to shut down the genes in the cluster. The qa-y gene is thought to encode a permease, allowing QA into the cell.
model between the repressor, qa-1Sp, and the transcriptional activator, qa-1Fp. Quinic acid in the cell is hypothesized to be the cell signal that disrupts the bound complex of qa-1Sp/qa-1Fp to favor induction by qa-1Fp [85]. This is summarized in Fig. 6. This story is converted into a formal biological circuit in Fig. 7. The top structure to the circuit is the central dogma. At the bottom of the circuit is a piece of a biochemical pathway metabolizing QA. The pathway feeds into the Krebs cycle. The qa-1Fp acts to create a feedback loop to activate the cluster and itself. When sucrose is added to the medium, a mechanism for catabolite repression is hypothesized, in which the presence of sucrose favors the binding of the repressor protein qa-1Sp to the transcriptional activator qa-1Fp. At this time, the boxes in the third and fourth rows are the observables. The circuit can be simulated over the web at http://gene.genetics.uga.edu/stc as described by Aleman-

Figure 7  A formal biological circuit for the qa cluster is presented. The top part of the circuit represents the central dogma, while the bottom part of the circuit is the biochemistry. The top row of squares represents the transcriptionally inactive forms of the genes. The second row of squares from the top is the set of transcriptionally active genes bound to the activator protein qa-1Fp. The third row of squares is the set of the cognate RNAs, which are translated in the fourth row into polypeptides. In the bottom half of the diagram, the polypeptides are carrying out their biochemical functions. There is a feedback loop created by the transcriptional activator. The repressor qa-1Sp is shown binding to qa-1Fp to inactivate same. Sucrose acts to facilitate this repression reaction, acting as a catabolite repressor. The end metabolic product shown is protocatechuic acid, which eventually leads into the Krebs cycle. The qa-yp polypeptide acts a transporter for QA.
Figure 8 The formal biological circuit specifies a system of ordinary differential equations describing the kinetics of all species. Each reaction contributes to the specification of the time rate of change of the species involved. In the first reaction, the transcriptional activator $qa-1Fp$ binds to the inactive gene to form the complex representing the transcriptionally active form of the gene. The forward reaction involving the collision of the $qa-2$ gene with $qa-1Fp$ occurs at a rate determined by the forward reaction rate $k_f$ and the product of the molar concentrations indicated in brackets to produce the transcriptionally active complex $qa-2/qa-1Fp$. In the backward reaction, the complex falls apart at a rate determined by the backward reaction constant $k_b$ and the molar concentration of the complex. Similarly, the instantaneous change in the $qa-3p$ protein from the reaction converting QA to DHQ can be computed. The reactants must collide as determined by the forward reaction constant and their concentrations, and the products must collide for the backward reaction to take place as determined by the backward reaction constant and the product of concentrations.

3.3. The lac Operon

The classic example of a biological circuit and the first one to be worked out is the lac operon [15]. The top structure of the circuit reflects the central dogma in Fig. 9. The transition from inactive transcriptional state (i.e., $lacY^0$) to an active transcriptional state (i.e., $lacY^+$) is coupled to the transition from active transcriptional (i.e., $lacZ^+$) to an inactive transcriptional state (i.e., $lacZ^0$), as the RNA polymerase is handed from one gene to the next to form a polycistronic message. The $lacF^p$ protein binds to the operator in the absence of lactose. The catabolite repressor protein, $crp^p$, acts as a positive activator (like $qa-1Fp$) by stabilizing the recruitment of the RNA polymerase to the promoter site. The biological circuit differs from the usual story told in texts by inclusion of the internal signaling cascade linking PEP in glycolysis to the glucose transporter [87]. This particular circuit is about twice the size of the $qa$ cluster circuit and still leaves out components of the...
Figure 9  The biological circuit for the lac operon is more elaborate than that of the qa cluster. The lacP repressor can bind to the operator to shut down the cluster through a negative feedback loop unless lactose is present to bind to lacP, thereby titrating out the repressor. There is also a positive feedback loop provided by the catabolite repressor protein crp, which aids in the recruitment of RNA polymerase to the promoter (lacP). The catabolite repressor protein is only active when bound to cAMP. The enzymes acp and pdp make cAMP from ATP and convert cAMP to AMP, respectively. An internal signaling cascade including (e1P, eII P, eIII P, and hP) is included to take a phosphate on phosphoenolpyruvate (PEP) to glucose to pump glucose into the cell as glucose-6-phosphate.

Embden-Meyerhof pathway linking glucose-6-phosphate to PEP. Again, the circuit can be simulated over the web at http://gene.genetics.uga.edu/stc.

3.4. The trp Operon

One other classic circuit illustrating translational control is the trp operon [88]. As summarized in Fig. 10, there are two configurations of the RNA: one in which the ribosome is stalled at a trp codon and one in which the ribosome is not stalled. When there is plenty of tryptophan, there is a feedback loop created in which the RNA forms a structure that leads to transcription termination such that no proteins are made. In the other configuration, the RNA polymerase transcribes the downstream structural genes, and the RNA gets translated. In addition, a feedback loop involving trpP which, in contrast to the lac operon,
Figure 10  The trp operon differs from the lac operon in having translational control. Tryptophan synthesis provides feedback to attenuate translation. If tryptophan is rare in the cell, then the message assumes one configuration efficient for translation. If tryptophan is at high levels in the cell, the message assumes an altered conformation with the ribosome not conducive to translation. In addition, there is a repressor trpRp acting on the operator to shut down the operon, and the repressor is activated in the presence of tryptophan as would be expected for a biosynthetic pathway.

is activated in the presence of the metabolite to shut down the trp operon. The pathway is at the bottom of the circuit.

3.5. Examples of Biological Circuits Relevant to Agriculture, Industry, and Medicine

A preliminary circuit can be constructed for aflatoxin biosynthesis from the 25 known components of the sterigmatocystin cluster [89] and the identification of a positive regulator of aflatoxin biosynthesis [90]. A preliminary circuit is available at http://gene.gene-genetics.uga.edu/stc, but it is much larger than the other circuits described here. Mechanisms of negative regulation of the pathway are yet to be identified. This A. flavus system is one of the few approved for USDA piloting of release of competing strains to displace those strains producing aflatoxin. In this case, the circuit could help to identify which mutations are likely to be most effective in knocking out aflatoxin biosynthesis in genetically engineered strains and in determining how aflatoxin biosynthesis is triggered.

Another important example is the penicillin gene cluster in Aspergillus nidulans and Penicillium chrysogenum [47]. The cluster, with its approximately three genes, is conserved in prokaryotes and eukaryotes [91] with a partially specified regulatory system and may have arisen in fungi by horizontal transfer from a prokaryote such as Streptomyces. The regulation of the penicillin cluster at first sight appears more complicated than the
paradigms like the qa cluster. The regulation of penicillin synthesis appears to be tied to biological circuits for carbon metabolism, pH sensing, and nitrogen metabolism as examples [47]. For example, Suarez and Penalva [92] present evidence that a pacC transcription factor involved in pH sensing may bind to an intergenic region between acrA and ipnA genes in the P. chrysogenum penicillin cluster. A hypothesis for the pathway describing biosynthesis of penicillin is well developed. A biological circuit is likely to contain several kinds of feedback loops to incorporate connections to other circuits. A genetic perturbation of this circuit is likely to interact with a process of amplification of the penicillin cluster by site-directed homologous recombination mediated by a conserved hexanucleotide sequence [93]. Relevant environmental perturbations include the carbon source and pH. Some kinetics models have already been tried.

The final example is drug discovery for P. carinii (Pc), the major killer of AIDS patients [2]. An ATP bioluminescent assay for in vitro screening of anti-Pc drugs has been developed [94]. With the resources of the Pneumocystis Genome Project [37,95], more than 2,000 distinct cDNAs have been generated and partially sequenced. This cDNA collection includes genes such as erg-9, erg-1, and erg-7 in sterol biosynthesis. A partial reaction network for sterol biosynthesis can be hypothesized by reference to KEGG [19]. The cDNA collection can be then exploited for transcriptional profiling to understand the mechanism of action of existing anti-Pcs (e.g., TMP-SMX, pentamidine, and atovaquone) for highlighting new potential drug targets in sterol biosynthesis and other critical pathways and for evaluating the proposed reaction network for sterol biosynthesis. The approach is to perturb the system with an array of potential protein inhibitors, observe the response with the ATP bioluminescent assay and transcriptional profiling, fit a hypothesized reaction network, evaluate the model, modify the model and perturbations, and repeat the cycle to discover drugs and their mechanism of action.

3.6. Simulating Arbitrary Reaction Networks Satisfying Mass Balance

A number of simulators now exist that simulate arbitrary reaction networks that satisfy mass balance. These include METAMODEL [96], GEAPASI [97,98], SCAMP [99], KIN-SIM [100,101], MIST [102], E-CELL [103], and KINSOLVER [40]. These packages differ in the diversity and type of numerical solution methods for the systems of differential equations illustrated in Fig. 8. The simulators also differ in their ability to be used on different types of computers or over the web. Lastly, the simulators differ in their capability to examine many reaction networks at once relevant to a particular system [104].

3.7. Steady-State Approximations to Simplify Biological Circuits

A classic approach to simplifying the reaction network is to make steady-state approximations to obtain simplified kinetics [65]. The classic example is Michaelis-Menten (MM) kinetics derived from the reactions in Fig. 3 by making a steady-state approximation with respect to the level of enzyme complex (ES). With a general purpose simulator, this is not necessary and can in some cases be positively misleading. For example, the MM formulation tends to break down when there are multiple substrates for the enzyme.

With this caveat, it may be possible to simplify the kinetics model by steady-state approximations to reduce the number of parameters and to gain interpretability of the model (i.e., heuristic appeal). One example is shown in Fig. 11.
Figure 11  Steady-state approximations to the levels of some species can be used to reduce the number of model parameters in a biological circuit. After assuming that the levels of transcriptionally active genes are in steady state, the system of ordinary differential equations for the full biological circuit in Fig. 6 can be approximated by the reduced model specification below. It is enough to describe the message levels denoted by m and protein levels denoted by p. The $\alpha$'s denote basal transcription rates, the $\delta$'s, the QA inducible transcriptional rates, the $\gamma$'s, the repressor effects, and the $\beta$'s, the rates of protein decay. The subscripts f, s, and sg denote the qa-1F, qa-1S, and structural genes in the qa cluster.

The deterministic model in Fig. 11 is a steady-state approximation to the full model in Fig. 7 in which the velocities for the concentrations of the bound-state of the genes are assumed approximately constant (i.e., $d[qa - x]/dt = C$). In this model there are two sets of promoters: one set that is QA inducible and one set that is not QA inducible [105]. QA-independent rates of transcription of the activator (f), repressor (s), and structural genes (sg) are denoted by $\alpha_f$, $\alpha_s$, and $\alpha_{sg}$. In contrast, the rate of production of message induced by QA is proportional to the level of inducer and activator protein. The rates of QA-inducible transcription of activator, repressor, and structural genes are denoted by $\delta_f$, $\delta_s$, and $\delta_{sg}$, respectively. The repressor interacts with the activator, and the effect of the repressor on transcriptional activation is captured in the repressor effects $\gamma_f$ and $\gamma_{sg}$. Message levels ($m_i$) decay at the same rate in proportion to their level. The Hill Coefficient, n, is introduced as a shape parameter for the cooperative effect of repressor polypeptides on message levels. In this model, there is no posttranscriptional regulation. All messages are translated at the same rate, and protein levels ($p_i$) have different constant rates of decay of $\beta_f$ for the activator, $\beta_s$ for the repressor protein, and $\beta_{sg}$ for the structural gene proteins. The number of parameters is reduced to 42, and the model in Fig. 11 is analytically tractable.

3.8. Stochastic Circuits

McAdams and Arkin [106] have presented evidence that stochastic factors play an important role in the $\lambda$ switch response. Kepler and Elston [107] have also demonstrated that
stochastic factors can play an important role in transcriptional control through the recruitment of RNA polymerase to the promoter. Abastado et al. [108] have made the case for stochastic factors in translational initiation by ribosome scanning of the uORFs upstream to GCN4 [109]. The extent to which such stochastic factors play a role in most biological circuits is unknown. Gillespie [82] established a framework for stochastic kinetic models, which under certain regularity conditions, converge to the deterministic circuits satisfying mass balance described in the previous modeling sections.

The formulation of the model, as with deterministic models, begins by writing down the circuit diagram or, equivalently, the tables of hypothesized reactions as in Fig. 12.

**Figure 12** Part of the list of master equations for a stochastic circuit with the same structure as Fig. 6 is listed. Here $m_{F_1}$ and $m_{S_1}$ are the basal number of mRNAs for qa-1F and qa-1S; $m_{F_2}$ and $m_{S_2}$ are the number of induced mRNAs for qa-1F and qa-1S; and $m_{F_R}$ is the number of transcriptional activators bound to a repressor protein. The quantities $Z_i$ represent the number of product molecules, and the constants $k_i$ are reaction rates. The sources A,B are the qa cluster DNA and assumed constant.
The formulation of a stochastic circuit is illustrated with the qa gene cluster circuit. From a microscopic point of view, binding of a free inducer molecule (i.e., quinic acid in the cell), activator, and repressor to the activator, gene, or activator, respectively, is very likely to be a random process because of the low concentrations of the reactants in the cell [107,110].

In Fig. 12, time can be taken to advance in discrete steps due to random collisions of molecules participating in the reactions, where $m_{F_{nd}}$ and $m_{S_{nd}}$ are the basal number of mRNAs for $qa-1F$ and $qa-1S$; $m_{T_{1}}$ and $m_{T_{2}}$ are the number of induced mRNAs for $qa-1F$ and $qa-1S$; $qa-1F^p$ and $qa-1S^p$ are the number of protein molecules encoded by $qa-1F$ and $qa-1S$; and $m_{R}$ is the number of transcriptional activators bound to a repressor protein. The quantities $Z_i$ represent the numbers of product molecules, and the constants $k_i$ are reaction rates. The sources A,B are the qa cluster DNA and assumed constant. As in Fig. 12, similar reactions can be written down to specify the role of the structural genes in the reaction scheme. The model is a discrete-time denumerable Markov chain [111]. A formal relation among the parameters in Fig. 12 and the reaction rates can be established following Gillespie [82]. For example, $\alpha_0 = k_1A\tau V/m_{\text{max}}$, where $\tau$ is the time-scaling parameter, $m_{\text{max}}$ is a concentration-normalization coefficient, and $V$ is a volume factor. Recently, Kierzek [83] developed methodologies for simulating stochastic networks.

3.9. Limitations of Reaction Network Models

3.9.1. Too Many Parameters and Too Few Data

With each new species, a new parameter—its initial concentration—is added. With each new reaction, two new parameters—the forward and backward reaction constants—are added. In general, only a subset of the species are observed over time. The major problem is identifying one model that fits one reaction network with limited and noisy profiling data. To address this problem will require novel fitting procedures.

3.9.2. We May Not Have All the Pieces

To overcome this problem, any modeling, fitting, and model-evaluation framework must be general enough so that discovery of new species during profiling or new topological features during protein–protein interaction mapping can be included in the circuit. For example, a general purpose simulator KINSOLVER [40] is required.

3.9.3. Stochastic Factors May Play a Significant Role

Stochastic factors may play a significant role in the reaction network [106,112]. As a consequence, it is important to build on the work of Gillespie [82] and Kierzek [83] to generalize a deterministic simulator for an arbitrary reaction network satisfying mass balance, as Tomita et al. [103] have begun to do.

3.9.4. The Cell Is Well Stirred

That the cell is well stirred is a basic assumption of the family of models proposed. Weng et al. [113] point out that consideration needs to be given to cellular compartments, scaffolding, and reaction channeling. Compartmentalization can be handled in part by simulators like KINSOLVER by indexing the species by the compartment containing them [114]. Similarly, scaffolding and channeling can be represented by allowing for additional concentration variables and corresponding reactions for chemical species participating in
a protein scaffold or reaction channel. Another option is the approach of E-CELL [103], which is to introduce another table describing the compartmentalization of reaction species.

3.9.5. Higher-Order Kinetics May Come into Play

The formal model is based on collision dynamics determining the right-hand side (RHS) of the coupled differential equations, like those in Fig. 8. Any number of reactants or products can be introduced into a particular reaction, allowing higher-order kinetics. The more standard nonmultiplicative MM kinetics can be derived as steady-state approximations to the full reaction network as in those based on collision dynamics [65], as was done in Fig. 11.

4. PERTURBATION, PREDICTION, AND OBSERVATION

4.1. Emergent Properties as a Predicted Response

Once a systems approach is embraced, an experimental framework is needed to study the global response of the system. One approach is to perturb the system experimentally and then to measure the global system response. Predictions are made about the effects of various system perturbations and then compared to the observed state of the cell through profiling. Experiments are designed to test the predictions. In a systems approach, the goal is to understand and recover the behavior of the entire system. The system is not take-it-apart, but rather it is perturbed, and it and its total response measured. The hope is to be able to predict its system-wide behavior.

System perturbations can fall into three broad classes as illustrated with respect to the qa cluster in Fig. 13. They can be genetic in nature, such as gene mutations or more specifically, gene knockouts. A gene mutation in the qa-2 gene removes its function in Fig. 13. Perturbations can be chemical in nature, such as adding a protein inhibitor to the medium to inhibit qa-3p. This kind of perturbation would characterize the search for drugs to inhibit essential activities in organisms such as Pneumocystis. Finally, a perturbation can be environmental in nature, such as a change in carbon source (i.e., sucrose for quinic acid). In each case, the response is predicted from the simulation and compared to that observed to validate the circuit.

Figure 13  Three kinds of system perturbations are illustrated for the qa cluster: (1) genetic; (2) chemical, as in a drug; or (3) environmental.
4.2. Genetic Perturbations Can Be a Challenge

The major challenge for perturbation experiments is carrying out targeted gene knockouts in *N. crassa* and other fungi with a low rate of homologous recombination. High-throughput strategies for directed and random signature tagged mutagenesis (STM) using transposons have been developed in bacteria and yeast [115–121]. Recently Hamer et al. [121] have successfully utilized an STM strategy on a close relative of *N. crassa*, the rice blast fungus *Magnaporthe grisea*.

The STM strategy used by Hamer et al. [121] shares many of the common elements of all STM strategies originally developed by Hensel et al. [116] and Burns et al. [115]. Loss of function mutations are generated with a transposon. A tag is introduced into the mutation. The tag contains a marker that can be selected for in the target system after transformation. Strategies differ on whether or not the mutations are generated in a targeted way [119] or randomly [118] and whether or not they exploit homologous recombination present in the organism. They can also differ on the nature of the tag and whether or not the collection of mutants is ultimately generated by a negative selection or screen. As knockout technology has progressed, there has also been a shift away from knockouts to conditional mutations and adding further functionality to the insertion cassette [122].

Hamer et al. [121] began the mutagenesis process with an engineered transposon cassette (containing a hygromycin resistance gene) that could be mobilized *in vitro* to mutagenize a large insert clone, such as a cosmid or BAC. The mutagenized cosmid or BAC is then transformed into the target organism such as *M. grisea* by selecting for hygromycin resistance. Polymerase chain reaction can then be used to screen for homologous recombination events. In this way, the researchers were able to generate 25,179 insertion mutants. A total of 33% of these insertion mutants were identified to have homologues in public databases. One example of an insertion mutant included insertions in the pathogenicity gene MAC1. The STM approach has also been used successfully to isolate pathogenicity islands in *Candida glabrata* [123] and *Cryptococcus neoformans* [124].

A simple example is shown in Fig. 14. The control perturbation involves growing wildtype *N. crassa* on quinic acid alone, and the main product, protocatechuic acid (PCA), is graphed using the simulator [40]. The system is perturbed by introducing a mutation into the *qa*-2 gene. The predicted result in Fig. 14 is no PCA, a block in QA metabolism with no growth on QA alone.

With each perturbation, one of several responses might be observed. A transient response may be predicted. As in the case of the *qa* cluster, a transient response may be initiated by the environmental signal of QA, but once the signal is removed, the whole circuit may shut down again. In contrast, even relatively simple circuits can display emergent properties [39]. For example, Gardner et al. [125] built a simple toggle switch that may mimic many coupled signaling pathways. The product of gene A represses gene B, and the product of gene B represses gene A. Such a simple system has a biphasic response [125]—i.e., memory of a previous signal even after the signal is removed. Another example of an emergent property manifested by a circuit is an oscillatory response. The classic example is the biological clock [16], but a simpler circuit called the *repressilator* has been engineered in *Eschericia coli* that oscillates [126]. Whatever the response of the biological circuit, if the model correctly predicts the emergent property, this serves as a validation of the model.
4.3. Observation by Profiling

An example of this prediction, perturbation, and observation process is given for the qa gene cluster. A quinic acid (QA)-inducible cDNA library was initially characterized. The QA-inducible cDNA library of 33 plates was robotically arrayed on one membrane [127]. Twelve replicates of the arrayed library were stamped, and one membrane was probed with an AatII-fragment of the H123E02 cosmid containing only the qa cluster [127]. Two of the positives in the cDNA library were sequenced and confirmed to be derived from messages of qa-1F and qa-3.

Transcriptional profiling allowed us to examine the outcome of an environmental perturbation and of a genetic perturbation [12]. The WT and mutant 246-89601-2A (a mutation in the qa-2 gene) were shifted to 0.3% QA (with aromatic supplements only for the mutant) [128,129], and RNAs were isolated from WT and mutant 246-89601-2A at 4 time points after induction (30, 90, 120, or 240 minutes). These RNAs were reverse transcribed and radiolabeled to probe the cDNA arrays [127] as in Fig. 15.

As time progresses from left to right, more spots (genes) appear, and the intensities of the spots increase. The membranes are double-stamped so the spots appear symmetrically about the middle of each figure. A total of 12 genes (spots) appear to be QA-inducible by 240 minutes. Two of these genes were confirmed by end-sequencing to be qa-1Fand qa-3. The remaining 10 genes did not hybridize to an AatII fragment of H123E02 (containing the entire qa cluster) at high stringency [127]. If so, this implies there are other genes outside the qa cluster that need to be considered in QA metabolism. For example, some of these 12 genes may be involved in a starvation response since QA is not a preferred carbon source. To distinguish these two hypotheses, the experiment needs to be repeated with a shift to a medium with no carbon source or starvation for an aromatic amino acid.

The experiment was repeated with a qa-2/aro-9 mutant, and the transcriptional profile at
Figure 15  Transcriptional profiling: N. crassa was shifted from 1.5% sucrose to 0.3% quinic acid. A cDNA library derived from cells induced in quinic acid was robotically arrayed on nylon membranes (133). RNA was extracted from cells by grinding under liquid nitrogen with the High Pure RNA Isolation kit (Roche, Inc.). Simultaneous cDNA synthesis and $^{33}$P radiolabeling were performed according to manufacturer directions. Unincorporated $^{33}$P was removed by spin columns (Sigma, Inc.). Arrays were probed with $^{33}$P-labeled cDNAs derived from three time points after the shift to quinic acid. Images were collected on a Packard Instant Imager over a 26-minute period. The $qa-2/arо-9$ doublemutant is also shown expressing the sametranscripts at 240 minutes, but it does not grow. 240 minutes appeared identical to WT, although it did not grow when shifted to QA [128]. The experiment was replicated once with the same findings. The profiling experiment was modified with a genetic perturbation. An $aro-9/qa-2$double mutant was transformed with the $qa-2$ gene [130]. Transformant’s RNA profile was obtained as previously described, except that the exposure time was increased to 1 hour on the Packard Instant Imager. The same 12 genes came up, but also an additional seven or eight cDNAs were positive. None of the additional positives matched to known $qa$ cluster cDNAs. The findings here are likely to be typical of studies that reexamine classic stories from a genomic perspective [9]. In Fig. 16, the expression of the 12 genes responding to quinic acid, or possibly a starvation response, is shown. Only two of these genes appear to be part of the $qa$ cluster. The remaining genes are unknown at this time. There are other genes that need to be included in the circuit in Fig. 7 because their response is not accounted for. New tools are being developed to permit scientists to explore relationships between genes uncovered in RNA profiling [131].
4.4. Protein–Protein Interactions: Observing the Links in a Circuit

Protein–protein interaction mapping is being pursued in a number of model systems [13,20,21,132]. Early approaches use the two-hybrid system to detect protein–protein interactions [20,21,132]. Two classes of strategies are being used to create the maps.

4.4.1. Clone-by-Clone Strategy

In a clone-by-clone strategy, one prey clone interrogates a robotically arrayed bait library by mating the yeast strain with the prey clone and each yeast strain with a bait clone. Interaction mating is achieved by robotically pinning the prey strain to all of the arrayed bait strains [21]. With yeast, this means that about 6000 potential interactions out of the $6000^2$ can be examined at one time. While this approach is slow, the sensitivity to detect interactions is about three times that of the high-throughput screens.

4.4.2. High-Throughput Strategy

In a high-throughput strategy, some pooling scheme is employed. Ito et al. [20] pooled both bait and prey in pools of 96 and mated the pools. The resulting positives can be picked and sequenced to identify the interactors. Each interaction mating allowed the examination of 96$^2$ potential interactions. They found 183 interactions after scanning about 10 percent of the proteome. By extrapolation, about 2000 interactions are expected in the yeast proteome. Uetz et al. [21] created a 96-well plate of bait clones and then mated...
them with a strain containing a whole prey library. In this way, each experiment resulted in the examination of 96 × 6000 potential interactions. With their high-throughput strategy, they identified 957 potential protein–protein interactions out of about 2000 expected [20].

Their map can be visualized by tools like those of Fang et al. [44] as a protein mobile as shown in Fig. 17. In this protein mobile, nodes are proteins and edges are detected interactions. This protein mobile then becomes a search grid on which the scientist refines a biological circuit. Possible links in the circuit are, in part, guided by the links reported in the protein mobile presented in resources like BIND [149].

There are a number of limitations of two-hybrid screens. Numerous false positives and false negatives occur as evidenced by the lack of overlap between screens conducted by Ito et al. [20] and Uetz et al. [21]. Also, promiscuous proteins show up, repeatedly interacting with other proteins. Something may be missed in the original target system to make an interaction go in the *S. cerevisiae* or *E. coli* detection system. As a consequence, other approaches to building protein–protein interaction maps are being pursued.

Gavin et al. [13] describes how to use tandem affinity purification (TAP) in conjunction with MALDI-TOF mass spectrometry to characterize protein complexes in *S. cerevisiae*. By this method, they were able to identify 232 distinct protein complexes. A total of 58 of these complexes had not been previously reported. The major limitation of their approach was the use of two-dimensional protein gels to separate proteins, thereby setting aside insoluble proteins. Ho et al. [133] described a related approach.

4.5. Protein–DNA Interactions: Observing the Links in a Circuit

Other resources are needed for systematically reconstructing a biological circuit. One is a map of all protein–DNA interactions [134]. At this time, all that is available is the equivalent of a clone-by-clone screen of protein–DNA interactions.
4.6. Web Services to Unite the Bioinformatics Pool of Data Sources

Data sources for profiling data, protein–protein interaction, and protein-DNA interaction data, and metabolic pathway information for reaction network modeling have too many differences that inhibit unified access and interoperability of data sources [10,37]. For example, bioinformatics databases such as BIND, KEGG, NCBI, Ensembl, FlyBase, SGD, WormBase, and UCSC all provide relevant data, but they are use a wide range of different systems and formats [135]. Researchers wishing to integrate these data need to write hundreds and even thousands of different programs to carry out the integration of data sources without any assurance of correct enactment of bridging services. To unify these bioinformatics services, providers may adopt a Web services model [136].

A Web service is any piece of software that makes itself available over the Internet and uses a standardized XML messaging system. A Web service can have a public interface, defined in a common XML grammar. The interface describes all the methods available to clients and specifies the signature for each method. Currently, interface definition is accomplished via the Web Service Description Language (WSDL). Furthermore, if a web service is created, there should be a simple mechanism to publish it. There should also be a mechanism for interested parties to locate the service and its public interface. The most prominent directory of web services is currently available via Universal Description, Discovery, and Integration (UDDI). In this web services model, the data providers register their services in a formalized service registry, and researchers’ scripts no longer need to be concerned with the interface details of the different databases. This model may represent a unification platform needed in bioinformatics [135].

A number of bioinformatics services are currently available [137]. For example, the OmniGene project [138] from MIT aims to create an open source web services platform for bioinformatics. Additionally, the Distributed Annotation Service (DAS) provides a distributed platform for aggregating genome annotation data from multiple sources [139]. Lastly, the BioMOBY project aims to provide distributed access to multiple bioinformatics services and provides a centralized registry for finding new services. All of these projects are likely to see much growth in the near future.

5. FITTING BIOLOGICAL CIRCUITS

The profiling information together with protein–protein and protein—DNA interaction maps provide the information necessary to identify biological circuits. After system perturbation, the profiling information either agrees with the predictions of the circuit or does not. A figure of merit can then be used to guide the selection of a biological circuit that is consistent with the profiling data from the system in different cellular states. The information about links in the circuit can be used both to constrain the fitting process and to guide the comparison of new models evaluated for fit relative to the existing best model. We will describe the standard fitting approach for reaction networks:

Let the parameters in the biological circuit be denoted by the M-tuple, $\theta: = (\theta_1, \cdots, \theta_M)$. In the case explored here, the parameters are the rate constants $k_f$ and $k_r$, for all reactions $r = 1, 2, \cdots, M_R$, as in the reaction network of Fig. 11 and the initial concentrations $[s]_{t=0}$ for all intracellular species $s = 1, 2, \cdots, M_S$. The number of parameters is $M = M_S + 2M_R$. For the deterministic model in Fig. 11, the rate constants and initial conditions are:
In the following, this parameter vector shall be referred to as as "model\( \theta \)".

Next, let \( Y := (Y_1, \ldots, Y_D) \) represent the D-tuple of all experimental observables which have been measured in one experiment or a series of time-dependent profiling experiments. Suppose that in a series of \( E \) experiments, labeled by \( e \in E = \{1, \cdots, E\} \), each experiment the concentrations \([s]\) of certain species \( s \) are measured at time points \( t \). Different experiments would be distinguished by externally controlled and quantitatively known experimental conditions which include, the carbon source and its concentrations, feeding/starvation schedules, choice of measurement time points, and functional presence or absence of certain genes or proteins, as controlled by mutations or protein inhibitors. The data vector \( Y \) would then comprise components:

\[
Y := Y_{s,t,e} := ([s]_t,e/[s]_0) \quad \text{[with \( l := (s,t,e) \)]}
\]

with some (known or unknown) reference concentration \([s]_0\) (ref), if for example, some linear measure of concentration is used or:

\[
Y := Y_{s,t,e} := \ln ([s]_t,e/[s]_0) \quad \text{[with \( l := (s,t,e) \)]}
\]

if log-induction ratios [12] are recorded. Here \( l := (s,t,e) \) and \( s \in S' \) labels the \( M_S' \) different molecular species, with \( S' \) denoting the subset of all species whose time-dependent concentrations actually have been observed. Note that in general, \( S' \) is only a subset (generally a small one) of the set \( S \) of all \( M_S \) participating species in the biological circuit. With \( t \in \{t_1, \cdots, t_M\} \) labeling the \( M_T \) different time points at which species concentrations have been measured, the dimensionality of the data vector \( Y \) is then:

\[
D = M_S' \times M_T \times E.
\]

For the present mRNA profiling data set in Fig. 15 for the \( qa \) cluster, \( E = 1, M_S' = 6, M_T = 7, \) and \( D = 42 \).

Now, let \( F(\theta) = [F_1(\theta), \cdots, F_D(\theta)] \) denote the corresponding predicted values for these observables \( Y \) for a given model \( \theta \). For the previously described set of observables \( Y_{s,t,e} \), the predicted values \( F_l(\theta) = F_{s,t,e}(\theta) \) [with \( l := (s,t,e) \)] are calculated from \( \theta \) by solving the network’s system of rate equations for the rate constants and initial conditions comprised in \( \theta \) using the simulator KINSOLVER [40] and then calculating from that solution the log-concentration ratio \( \ln([s]_t,e/[s]_0) \) or the respective linear concentration measure for each observed species \( s \) at each observation time point \( t \) in each experiment \( e \).

It is reasonable to assume that the probability distribution \( P(Y;\theta) \) of the data is representable as a multivariate Gaussian, with error correlations only between data \( Y_l \) taken at the same time point. Hence, the following likelihood function will be used as the figure of merit:

\[
P(Y;\theta) = \text{const} \times \exp \left[ -\chi^2/2 \right] \quad \text{with} \quad \chi^2 = [Y - E(Y)]' \Sigma^{-1} [Y - E(Y)]
\]

where \( E(Y) \) and \( \Sigma \) denote the mean and variance–covariance matrix of the observation vector \( Y \) for model \( \theta \).
When multiple realizations of each profiling experiment are performed, then the variance–covariance matrix can be estimated. In the fitting reported in Fig. 18, a univariate Gaussian with $\sigma/E(y) \times 100 = 20\%$ has been assumed and the observed single-experiment concentrations have been used as the data vector $Y$ with the link function $E(Y) = F(\theta)$ [140]. To date, heteroscedasticity has been reportedly not an issue [141].

The fit is then obtained by maximizing the figure of merit $P(Y; \theta)$ with respect to the model parameters. A number of tools exist to carry out this fitting process [142,143]. A model from the family in Fig. 11 is displayed that fits the RNA profiling data of the $qa$ cluster quite well in Fig. 18. Profiles were obtained for six out of seven of the $qa$ genes. The RNA profile for $qa-1F$ peaks at 4 hours and then drops after that point with another rise at 6 hours. The remaining profiles track that of $qa-1F$ message levels. The simulator also yielded predictions about the protein profiles which are now testable [45].

5.1. Too Many Parameters, Too Few Data

The major limitation of current fitting procedures is that they do not address the major problem of too many parameters and too few data. In the example in Figure 11, after making steady-state assumptions, there were 42 parameters and 42 data points. This situation is not likely to change even with the availability of genome-wide RNA and protein profiling technologies. The reaction network in a cell is large and interconnected, and it is not clear at this time in studying a particular process such as carbon metabolism what other components of the reaction network need to be considered. For example, QA metabolism is intimately connected to aromatic amino acid metabolism through the $aro$ cluster in $N. crassa$ [144]. This raises the question of how QA metabolism is linked to general control [145]. Even in well-studied circuits involved in antibiotic production, it may not be safe to decouple secondary metabolism from, for example, energy metabolism. New fitting procedures are needed that directly address the problem of too many parameters and too few data [146].
5.2. A Stochastic Alternative

It is not clear at this time what role stochastic factors play in biological circuits. Patel and Giles [110] estimated that the number of qa-1F messages is on the order of 0.1 to 1 RNA per nucleus. This granularity within the cell may mean that a transcription factor finding a small 17 kbp stretch of DNA on the smallest chromosome in N. crassa may not be guaranteed. As a consequence, the stochastic formulation in Fig. 12 was simulated with the results for the qa-1F message shown in Fig. 19.

In this case, the four stochastic realizations have the same basic mountain shape observed for the real profile. The number of RNA molecules rises to about 400 molecules per cell. It is likely that the stochastic circuit will provide a description similar to the deterministic circuit. Either formulation leads to a similar story relative to the observed profiles. The challenge is comparing stochastic vs. deterministic circuits with the same circuit structure. Under some circumstances, deterministic circuits can be viewed as limits of the underlying master equations in Fig. 12 describing the stochastic circuit [82], but inference problems arise in distinguishing stochastic circuits vs. their limiting deterministic relative when one model (i.e., a deterministic one) lives on the boundary of the parameter space for a larger class of models (i.e., the stochastic ones).

6. CONCLUSION

Metabolomics is a process of discovery that promises a mechanistic understanding for interesting biological processes. This mechanistic understanding is captured in a kinetics model well grounded in physics and chemistry. The discovery process itself is more akin to approaches used in systems ecology or neural biology. For 60 years, biologists have been taking biological systems apart to find their components. Now the process is about to reverse. With the complete genomic blueprint now in hand, the challenge is to reassemble the pieces. The adjectives describing metabolomics are hypothesis-driven, integrative, and reconstructive.

The most basic question in metabolomics is, What is a living system? [27]. One approach to answering this question is reconstructive and rooted in an approach originally...
adopted by Beadle and Tatum [147]: ‘‘From the standpoint of physiological genetics the
development and functioning of an organism consist essentially of an integrated system
of chemical reactions controlled in some manner by genes.’’ To identify this hypothesized
reaction network requires an integrative approach.

The flow of the reconstruction process can be summarized simply in Fig. 20. The
fungal system is perturbed. In the case of drug discovery, cells are treated with potential
drugs, as an example, or in the case of industrial fermentation, genetically engineered
strains are selected to increase production. The system is observed through RNA, protein,
and metabolite profiling to compare the response with a control. The cells may die or
may produce more of a desired product, such as penicillin. The profiling data are used to
identify kinetics models or ‘‘biological circuits’’ to predict the response of the system. In
many cases, the profiling step identifies additional genes and their products that must be
included in the biological circuit. The fitted model then allows predictions about the total
response of the system. The response of the system can sometimes be surprising when
pathways are coupled and enlarged to explain the profiling data. Possible emergent proper-
ties include memory and a cyclical response. The model is reevaluated and tested for good
fit. Current tests for better alternatives are limited and need to be developed.

A better model is selected and a new perturbation is selected. Choosing an informa-
tive perturbation is a challenging problem. The cycle completes and starts over. The result
is a process of discovery and refinement. In each cycle the model serves to integrate
available information on sequence, profiling, protein–protein interactions, protein–DNA
interactions, and protein–lipid interactions. This discovery process ultimately will be auto-
mated into an adaptive control process to speed the process of gene-validated product
discovery [28].

ACKNOWLEDGMENT

This work was supported by USDA-2002-35300-12475.

REFERENCES

1745–1749.
2. Cushion, M.T. Pneumocystis carinii. In Molecular Epidemiology of Infectious Disease.


73. Hall, R.D.; Miller, J.A.; Arnold, J.; Kochut, K.J.; Sheth, A.P.; Weise, M.J. Un. GA. See
74. Zhao, L.P.; Prentice, R.; Breeden, L. Statistical modeling of large microarray data sets to
76. Murphy, K.; Mian, S. Modeling gene expression data using dynamic Bayesian networks.
77. Friedeman, N.; Linial, M.; Nachman, I.; Pe’er, D. Using Bayesian network to analyze expres-
78. Weaver, D.C.; Workman, C.T.; Stormo, G.D. Modeling regulatory networks with weight
79. Huang, S. Gene expression profiling, genetic networks and cellular states: an integrating
80. Shmulevich, I.; Yliharja, O.; Astola, J. Inference of genetic regulatory networks under the
best-fit extension paradigm. In Proceedings of the IEEE-EuRASIP Workshop on Nonlinear
81. Shmulevich, I.; Dougherty, E.R.; Kim, S.; Zhang, W. Probabilistic Boolean networks: a rule-
1977, 81, 2340–2361.
83. Kierzek, A.M. STOCKS: STOChastics kinetic simulations of biochemical systems with Gil-
and regulation in the qa (Quinic Acid) gene cluster of Neurospora crassa. Microbiol. Rev.
Giles, N.H. DNA sequence, organization and regulation of the qa gene cluster of Neurospora
86. Johnston, M. A model fungal regulatory mechanism: the GAL genes of Saccharomyces
88. Yanofsky, C. Advancing our knowledge in biochemistry, genetics, and microbiology through
T.H.; Leonard, T.J. Twenty-five coregulated transcripts define a sterigmatocystin gene cluster
in Aspergillus nidulans. PNAS USA 1996, 93, 1418–1422.
Molecular characterization of aflR, a regulatory locus for aflatoxin biosynthesis. Appl Envi-
91. Smith, D.J.; Burnham, M.K.R.; Bull, J.H.; Hodgson, J.E.; Ward, J.M.; Browne, P.; Brown,
J.; Barton, B.; Earl, A.J.; Turner, G. B-lactam antibiotic biosynthetic genes have been con-
92. Suarez, T.; Penalva, M.A. Characterization of a Penicillium chrysogenum gene encoding a
PacC transcription factor and its binding sites in the divergent pcbAB-pcbC promoter of the
gene cluster is amplified in tandem repeats linked by conserved hexanucleotide sequences.
PNAS USA 1995, 92, 6200–6204.
94. Chen, F.; Cushion, M.T. Use of ATP bioluminescent assay to evaluate viability of Pneumo-
5655-X—An—Ch22—R1—03-16-04 10:52:18—

Metabolomics 631


117. Wach, A.; Brachat, A.; Alberti-Segui, C.; Rebischung, C.; Philippson, P. Heterologous HIS3
and GFP reporter modules for PCR-targeting in Saccharomyces cerevisiae. Yeast 1997, 13,
1065–1075.
R.; Benito, R.; Boeke, J.; Bussey, H.; Chu, A.; Connelly, C.; Davis, K.; Dietrich, F.;
Jones, T.; Laub, M.; Liao, H.; Liebenduth, N.; Lockhart, D.; Lucau-Danila, A.; Lussier,
M.; M’Rabet, N.; Menard, M.; Mittmann, M.; Pai, C.; Revuelta, J.; Riles, L.;
Roberts, C.; Ross-Macdonald, P.; Scherens, B.; Snyder, M.; Sookha-Mahadeo, S.; Storms,
Zimmermann, K.; Philippson, S.; Johnston, M.; Davis, R.W. Functional characterization of
120. Giaever, G.; Chu, A.; Ni, Li; Connelly, C.; Riles, L.; Veronneau, S.; Dow, S.; Lucau-
Danila, A.; Anderson, K.; Andre, B.; Arkin, A.; Astromoff, A.; Bakkoury, M.; Bangham,
Flaherty, P.; Foury, F.; Garfinkel, D.; Gerstein, M.; Gotte, D.; Guldener, U.; Hegemann,
Ool, S.; Revuelta, J.; Roberts, C.; Rose, M.; Ross-Macdonald, P.; Scherens, L.;
Schlammack, G.; Shafer, B.; Shoemaker, D.; Sookha-Mahadeo, S.; Storms, R.; Strathern,
J.N.; Valle, G.; Voet, M.; Volckaert, G.; Wang, C.; Ward, T.; Wilhelmy, J.; Winzeler,
E.A.; Yang, Y.; Yen, G.; Youngman, E.; Yu, K.; Bussey, H.; Boeke, J.; Snyder, M.; Phil-
ippson, P.; Davis, R.W. Johnston, M. Functional profiling of the Saccharomyces cerevisiae
121. Hamer, L.; Adachi, L.; Montegegro-Chamorro, M.V.; Tanzer, M.M.; Mahanty, S.K.; Lo,
Lampe, D.J.; Slater, T.M.; Ramamurthy, L.; DeZwaan, T.M.; Nelson, G.H.; Shuster, J.R.;
Woessner, J.; Hamer, J.E. Gene discovery and gene functional assignment in filamentous
122. Ross-Macdonald, P.; Sheehan, A.; Shirleen Roeder, G.; Snyder, M. A multipurpose
transposon system for analyzing protein production, localization, and function in Saccharo-
123. Cormack, B.; Ghor, N.; Falkow, S. An adhesin of the yeast pathogen Candida glabrata
pathogen Cryptococcus neoformans using signature-tagged mutagenesis. Genetics 2001, 157,
935–947.
126. Elowitz, M.B.; Leibler, S. A synthetic oscillatory network of transcriptional Regulators. Na-
ture 2000, 403, 335–338.
Orbach, M.J.; Sachs, M.S.; Wagner, R.; Weise, M.J.; Wunderlich, J.; Arnold, J. The Neuros-
128. Studies on the Genetic Control of the Inducible Quinate-Shikimate Catabolic Pathway in
Neurospora crassa. 129. Chaleff, R.S. The inducible quinate-shikimate catabolic pathway in Neurospora crassa: genetic
Metabolomics


137. Cerami, E. Web Services for Bioinformatics, O’Reilly Net May, 2002.


AUTHOR QUERIES:

AQ1: Au: Do you mean “sum”? 
AQ2: Au: Please run through and make sure that the section numbers you refer to are still accurate per our renumbering. 
AQ3: Au: Rewording still accurate? 
AQ4: Au: Worded correctly? 
AQ5: Au: Spelling correct? 
AQ6: Au: Instead of story, can we use the word “example”? 
AQ7: Au: Should this asterisk be here? 
AQ8: Au: Here again, can you use a word other than “story”? Makes it sound like fiction. 
AQ9: Au: Rewording still accurate? 
AQ10: Au: Should this be “site directed”? 
AQ11: Au: Not sure where you’re referring with the word “below.” A particular figure? Please clarify. 
AQ12: Au: Here again, can “stories” be changed to “examples”? 
AQ13: Au: Are colons correctly inserted in these equations? 
AQ14: Au: References 71,72,73, and 75 are no longer cited in your manuscript. Thus, we have deleted them from your reference list here and renumbered accordingly.