Metabolomics

Jonathan Arnold[*], H.-B. Schuttler[+], D. Logan[^], D. Battogtokh[+], James Griffith[*], B. Arpinar[$], S. Bhandarkar[$], S. Datta[@], K. J. Kochut[$], E. Kraemer[$], J. A. Miller[$], A. Sheth[$], G. Strobel[+], T. Taha[+], B. Aleman-Meza[$], Jereme Doss[^], LaTreace Harris[^], Abassi Nyong[^],

[*]Department of Genetics, University of Georgia, Athens, GA 30602

+Department of Physics & Astronomy, University of Georgia, Athens, GA 30602

$Department of Computer Science, University of Georgia, Athens, GA 30602

^Department of Biological Sciences, Clark Atlanta University, Atlanta, GA 30314

[@]Department of Mathematics and Statistics, Georgia State University, Atlanta, GA 30303

[*] Corresponding author

Key words: fungal genomics

Corresponding Author: Jonathan Arnold

Genetics Department

University of Georgia

Athens, GA 30602

(706) 542-1449

fax: (706) 542-3910

email: arnold@uga.edu

I. INTRODUCTION

I.A 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 being used to guide the process of gene-validated product discovery transforming 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 commensual 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*, causing aflatoxin contamination in peanuts. Controlling aflatoxin biosynthesis has consequences for human health, the quality of a major US 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 anti-fungals or through manipulation of pathways to overproduce needed compounds like penicillin.

I.B Paradigm Shift in Biology. Biology is 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 use of automated sequencing machines (11) and the use of microarrayers for analysis of gene expression (12) and 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).

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 of the clock can be used to link organisms into families, which can be 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 (29, 28, 30). The information needs to be integrated with new algorithms (31, 32, 33, 34, 35, 36, 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). The information has to be analyzed to test hypotheses about the structure, function, and evolution of living systems (41), and, finally, the information needs to be visualized (42, 43, 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, an approach 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 are made with metabolic engineering. In Section II the origin of metabolomics is explored. In Section III the models or "biological circuits" behind metabolomics are sketched with their applications. In Section IV the process of discovery at the heart of metabolomics is considered. In Section V the process of identifying biological circuits is considered along with the challenges. In the final section metabolomics is put in a larger context and summarized.

II. 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). In Figure 1 is shown an example of this emergent property of *N. crassa*, the regular temporal sequence of conidiation by this organism growing in race tubes. When transcriptional activators like 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 an area where there are now powerful methods for identifying quantitative trait loci (QTLs) that affect particular complex traits (49,50,51,52). When QTLs are integrated with other kinds of genomic information, precise predictions of the location of genes can be made (53, 52). 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 make Genomics as a data-driven discipline into hypothesis-driven science. One approach is to cross Genomics and Quantitative Genetics. The resulting child 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.

II.A 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.

II.A.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).

II.A.2. RNA profiling by SAGE. 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.

II.A.3. Protein profiling by ICAT. Isotope-coded affinity tagging (ICAT) has been used to characterize the *GAL* cluster in yeast (45). Protein profiling will help us to identify genes that are under translational control (16,60) as well as providing 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 8 deuterium atoms (D8) or 0 deuterium atoms (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. The combination 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) make this an attractive approach. The major limitation is resolving all the proteins in a cell-free protein extract.

II.B. 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 multidimensional 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 of MudPIT 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.

II.C. Who counts the small molecules? Having completed the story 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 by (64). A variety of separation procedures are being explored.

II.D. 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.

II.E. Making genomics hypothesis-driven. The process of making genomics hypothesis-driven is summarized in Figure 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 as shown (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 Figure 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 what we will term a "biological circuit" is identified, then familiar simulators like GEPASI, MIST, or SCAMP can be invoked to yield predictions about the system (66,67,68), which can then be validated experimentally.

II.F. Role of Hypothesis-Driven Genomics. The advantage of this framework for hypothesis-driven genomics is several fold. This framework leads to a natural integration framework for genomics information. The profiling data are used for fitting the reaction network. Protein-protein and protein-DNA interactions enter to constrain the search for a better reaction network topology (21, 69). Available information on pathways is incorporated into the chemical reaction network from BIND, KEGG, EcoCyc and other sources (70,19,18). The fundamental structure of the reaction network reflects the information flow and information hierarchy in the cell. The reaction network or biological circuit summarizes the available information in an intelligible framework familiar to biologists, chemists, and physicists; moreover, the resulting kinetics model makes predictions about the system as a whole, which are then subject to stringent assessment against the measured states of the system.

Finally, the consideration of a biochemical and gene regulatory reaction network generalizes and regularizes the process of hypothesis-generation. Questions about how genes are regulated, whether or not translational control exists, or how one circuit involved in secondary metabolism is coupled to other circuits in carbon metabolism can be directly addressed. More importantly, the framework of a reaction network is open-ended and begs a genomic perspective, *i.e.*, fleshing out the wiring diagram. The

experimentalist is forced to continue to incorporate new genes and their products until the system-wide behavior can be recovered according to a more stringent standard of correctly predicting system response. It is no longer enough to get the story qualitatively right, but the levels of reactants and products must be correctly predicted.

This framework with the biological circuit at its heart provides a search tool for better hypotheses. As more relevant genes are found to be responding together at the RNA and protein level or as more genes are found to have the same upstream sequences binding to a bait DNA-binding protein (69), they can be incorporated into the circuit. In essence, protein-protein interaction maps (20, 21, 13) and protein-DNA interaction maps (69, 71, 72) provide a search grid for alternative reaction networks.

II.G Automated workflows to implement the process of hypothesis-driven genomics. The process of perturbing the system, measuring the response through profiling, fitting a hypothesized biological circuit, evaluating fit, modifying the model, and selecting the next system perturbation involves many automated and human tasks. This process can be modeled and managed as an automated workflow (28). LIMS have become a standard part of genomic sequencing (27, 73, 74*)*. The task of identifying a reaction network is more complicated. The data sources are disparate, and the experiments required to identify such a biological circuit will typically involve over 100,000 task executions. Thus, the human and automated tasks to identify a reaction network are complex, evolving, and involve managing data accumulated over a period of years. A simple diagram summarizing the tasks in a workflow is given in Figure 20.

A workflow is defined as an activity involving the coordinated execution of multiple tasks, which can be performed by people, programs, or machines (28), while a

workflow *process* 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 includes design), workflow enactment, and management of workflow processes. The goal of a workflow management system is to enforce inter-task dependencies, scheduling, data management, and reliable execution. Workflow management systems can play a central role in monitoring and enforcing quality of service (QoS), such as sequence quality (76, 77). Genomics workflow systems require adaptability and ensured QoS.

Several workflow management systems are available (78, 79, 80). One such system is METEOR which provides four kinds of services: workflow builder, workflow repository, workflow enactment, and workflow manager. Enactment has been implemented in two versions, WebWork (81), which is entirely web-based and is suited for workflows that do not dynamically change their architecture and OrbWork (82), which supports dynamic modifications to a workflow.

METEOR has already been used to support portions of a workflow to identify reaction networks. A workflow like that in Figure 17 includes a subflow for sequencing, which has been implemented (74). A larger subflow for protein-protein interaction mapping has also been created (28), which again is a subcomponent of circuit identification. Workflow management systems will be essential for integration of genomic and bioinformatic projects aimed at identifying biological circuits.

III. MODELS

A variety of modeling approaches for biological circuits have been proposed. These include linear models with time as a factor (83), linear dynamic (84), Bayesian networks (85, 86), neural networks (87), Boolean networks (88, 89, 90), and classic chemical reaction networks satisfying mass balance (39). At one extreme Boolean networks are Draconian simplifications of chemical reaction networks satisfying mass balance, but 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 as in (39) or stochastic (91, 92).

For most reactions, enforcement of mass balance leads to specification of a system of differential equations to describe this reaction network (39). An example of how the reaction network captured in the circuit of Figure 4 is translated into a coupled system of nonlinear differential equations is given in (40).

III.A. **Water Models**. In that 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 Figure 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: $H_2 + O \quad H + OH$

(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 rate of production of Species OH due to reaction R1 would be given by the simplest multiplicative kinetics by (65):

$$d[OH]/dt = k_f [H2][O] - k_b[H][OH],$$

where, *e.g.*, [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 , say, OH:

$$d[OH]/dt = \Sigma_r \, d[OH]/dt$$

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 accumulated more kinetics data, they found the initial reaction network was an oversimplification of what makes some rockets go up (*i.e.*, H_2 and O_2). The reaction network or circuit needed to be refined to that in Figure 5 (Water Model 2). This inclusion of additional species and/or reactions is typical of building a model that fits a biological system.

III.B. **Carbon metabolism**. A slightly more complicated biological circuit can be constructed for one of the two early paradigms for eukaryotic gene regulation (93, 94) along with the *GAL* cluster in *S. cerevisiae* (95, 9). Specifying the model in Figure 6 begins by writing down the chemical reactions of the known participants in quinic acid (QA) metabolism. The circles on the wiring diagram denote reactions; boxes denote reactants. Arrows are used to indicate the reactants entering a reaction, and outgoing arrows, the products of a reaction. Some reactions have no outgoing arrows, and they (the lollipops) are decay reactions. At the top of the circuit, reactants include the 7 genes in the *qa* cluster (*qa-x*, qa-2, *qa-4*, *qa-3*, *qa-y*, *qa-1S*, *qa-1F*) (94). These genes can be either in an unbound or a bound state with a transcription factor produced by the *qa-1F* gene as indicated by a superscript, 0 or 1, respectively. These genes are, in turn, transcribed into messenger RNAs (superscripted with an r) which, in turn, are translated into **proteins superscripted with a p (a slight departure from convention).** A total of 4 out of the 7 protein products participate on a known biochemical pathway at the bottom of the diagram. In the circuit, there are two hypothesized cellular states for quinic acid, extracellular (denoted with an e) or intracellular quinic acid (QA). One of the genes, *qa-y* , is thought to encode a permease, $qa\text{-}y^p$, which may be involved in the transport of quinic acid into the cell. One hypothesized protein-protein interaction exists in the model between the repressor, $qa\text{-}1S^p$, and the transcriptional activator, $qa\text{-}1F^p$. Quinic acid in the cell (QA) is hypothesized to be the cell signal that disrupts the bound complex of $qa\text{-}1S^p/qa\text{-}1F^p$ to favor induction by $qa\text{-}1F^p$ (94). This story is summarized in Figure 6.

This story can be converted into a formal biological circuit in Figure 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-1F*^{p} 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-1S*^{p} to the transcriptional activator *qa-1F*^{p}. At this time the boxes in the 3rd and 4th rows are the observables. The circuit can be simulated over the web at http://gene.genetics.uga.edu/stc as described in (40). Some examples of the equations describing the 50 reactions can be found in Figure 8. In this example transcriptional regulation is a reaction in which the transcriptional activator *qa-1F*^{p} binds to the inactive gene like *qa-2*. One of the metabolic reactions is shown in which the enzyme *qa-3*^{p} converts intracellular QA* into dehydroquinate (DHQ).

III.C. **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 Figure 9. The transition from inactive transcriptional state (*i.e.*, *lacY*^{0}) to an active transcriptional state (*i.e.*, *lacY*^{1}) is coupled to the transition from active transcriptional (*i.e.*, lacZ^{1}) 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 *lacI*^{p} protein binds to the operator in the absence of lactose. The catabolite repressor protein, *crp*^{p}, acts as a positive activator (like *qa-1F*^{p}) 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 (96). This particular circuit is about twice the size of the *qa* cluster circuit and still leaves out components of the 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.

III.D. ***trp* operon**. One other classic circuit illustrating translational control is the *trp* operon (97). In this story, summarized in Figure 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 $trpR^{p}$ which, in contrast to the *lac* operon, is activated in the presence of the metabolite to shut down the *trp* operon. The pathway is at the bottom of the circuit.

III.E. 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 (98) and the identification of a positive regulator of aflatoxin biosynthesis (99). A preliminary circuit is available at http://gene.genetics.uga.edu/stc, but it is much larger than the circuits described above. 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 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 ~3 genes is conserved in

prokayotes and eukaryotes (100) with a partially specified regulatory system and may have arisen in fungi by horizontal transfer from a prokaryote like 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 (101) 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. An 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 sited directed homologous recombination mediated by a conserved hexanucleotide sequence (102). 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 (103). With the resources of the Pneumocystis Genome Project (104, 37), over 2000 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-*Pc*s like 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.

III.F. 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 (105); GEPASI (106,66); SCAMP (68), KINSIM (107, 108), MIST (67), E-CELL (109), and KINSOLVER (40). These packages differ in the diversity and type of numerical solution methods for the systems of differential equations illustrated in Figure 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 (110).

III.G. **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 Figure 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 Figure 11.

The deterministic model in Figure 11 is a steady-state approximation to the full model in Figure 7 in which the velocities for the concentrations of the bound-state of the genes are assumed approximately constant *(i.e*., $d[qa\text{-}x^1]/dt = C$). In this model there are two sets of promoters, one set being QA-inducible and one set not being QA-inducible (111). QA independent rates of transcription of the activator (f), repressor (s), and structural genes (sg) are denoted by α_f, α_s, and α_{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 δ_f, δ_s, and δ_{sg}, respectively. The repressor interacts with the activator, and the effect of the repressor on transcriptional activation is captured in the repressor effects γ_f and γ_{sg}. Message levels (m_x) 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 post-transcriptional regulation. All messages are translated at the same rate, and protein levels (p_x) have different constant rates of decay of β_f for the activator, β_s for the repressor protein, and β_{sg} for the structural gene proteins. The number of parameters is reduced to 42, and the model in Figure 11 is analytically tractable.

III.H. Stochastic circuits. McAdams and Arkin (112) have presented evidence that stochastic factors play an important role in the λ switch response. Kepler and Elston (113) have also demonstrated that stochastic factors can play an important role in transcriptional control through the recruitment of RNA polymerase to the promoter, and Abastado *et al.* (114) have made the case for stochastic factors in translational initiation by ribosome scanning of the uORFs upstream to *GCN4* (115). The extent to which such stochastic factors play a role in most

biological circuits is unknown. Gillespie (91) 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 Figure 12. 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 (113, 116).

In Figure 12, time can be taken to advance in discrete steps due to random collisions of molecules participating in the reactions, where mF_{unI} and mS_{unI} are the basal number of mRNAs for *qa-1F* and *qa-1S*; where mF_I and mS_I are the number of induced mRNAs for *qa-1F* and *qa-1S*; where $qa\text{-}1F^p$ and $qa\text{-}1S^p$ are the number of protein molecules encoded by *qa-1F* and *qa-1S*; and mF_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 Figure 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 (117). A formal relation among the parameters in Figure 12 and the reaction rates can be established following Gillespie (91). For example, $\alpha_\phi = k_1(A\tau V)/m_{fmax}$, where τ is the time scaling parameter, m_{fmax} is a concentration normalization coefficient, and V is a volume factor. Recently Kierzek (92) developed methodologies for simulating stochastic networks.

III.I. Limitations of reaction network models described. *III.I.1. There are 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.

III.I.2. We may not have all the pieces. To overcome this problem any modeling, fitting and model evaluation framework must be general enough 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.

III.I.3. Stochastic factors may play a significant role in the reaction network (112, 118). As a consequence, it will be important to build on the work of Gillespie (91) and Kierzek (92) to generalize a deterministic simulator for an arbitrary reaction network satisfying mass balance, as Tomita *et al*. (109) have begun to do.

III.I.4. The cell is well-stirred is a basic assumption of the family of models proposed. Weng *et al.* (119) 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 (120). 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 (109), which is to introduce another table describing the compartmentalization of reaction species.

III.I.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 Figure 8. Any number of reactants or products can be introduced into a particular reaction, allowing higher-order kinetics. The more standard non-multiplicative 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 Figure 11.

IV. PERTURBATION, PREDICTION, AND OBSERVATION

IV.A. 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 Figure 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 Figure 13. Perturbations can be chemical in nature, such as adding a protein inhibitor to the medium to inhibit $qa\text{-}3^{p}$. 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.

IV.B. 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 (121,122,123,124,125,126,127). Recently Hamer *et al*. (127) have successfully utilized a STM strategy on a close relative of *N. crassa*, the rice blast fungus *Magnaporthe grisea*.

The STM strategy used by Hamer *et al.* (127) shares many of the common elements of all STM strategies originally developed by Hensel *et al.* (122) and Burns *et al.* (121). 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 (125) or randomly (124) 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 (128).

Hamer *et al.* (127) 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, they were able to generate 25,179 insertion mutants. A total of 33% of these insertion mutants were identified to have homologs 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* (129) and *Cryptococcus neoformans* (130).

A simple example is shown in Figure 14. The control perturbation involves growing wild type *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 Figure 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 quinic acid, 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*. (131) 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 (131), *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 *E. coli* that oscillates (132). Whatever the response of the biological circuit, if the model correctly predicts the emergent property, this serves as a validation of the model.

IV.C. 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 (133). 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 (133). 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% quinic acid (with aromatic supplements only for the mutant) (134,135), 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 (133) as in Figure 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 min. Two of these genes were confirmed by end-sequencing to be *qa-1F* and *qa-3*. The remaining 10 genes did not hybridize to an *AatII* fragment of H123E02 (containing the entire *qa* cluster) at high stringency (133). 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 240 min appeared identical to WT, although it did not grow when shifted to quinic acid (134). The experiment was replicated once with the same findings.

The profiling experiment was modified with a genetic perturbation. An a*ro-9/qa-2* double mutant was transformed with the *qa-2*$^{+}$ gene (136). Transformant's RNA profile was obtained as above except that the exposure time was increased to 1 hr on the Packard Instant Imager. The same 12 genes came up, but also an additional 7-8 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 Figure 16 the expression of the 12 genes responding to quinic acid, or possibly a starvation response, is shown; only 2 of these genes appear to be part of the *qa* cluster. What the remaining genes are is unknown at this time. There are other genes that need to be included in the circuit in Figure 7 because their response is not accounted for. New tools are being developed which permit scientists to explore relationships between genes uncovered in RNA profiling (137).

IV.D. Protein-protein interactions: observing the links in a circuit. Protein-protein interaction mapping is being pursued in a number of model systems (20, 21, 138, 13). Early approaches use the 2-hybrid system to detect protein-protein interactions (20, 21, 138). Two classes of strategies are being used to paint the maps.

IV.D.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 3 times that of the high-throughput screens now described.

IV.D.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 x 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 Figure 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 (70).

There are a number of limitations of 2-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 reported before. The major limitation of their approach was the use of 2D protein gels to separate proteins, thereby setting aside insoluble proteins. Ho *et al.* (139) describe a related approach.

IV.E. Protein-DNA interactions, observing the links in a circuit. Other resources are needed for systematically reconstructing a biological circuit. One of these is a map of all protein-DNA interactions (69). At this time all that is available is the equivalent of a clone by clone screen of protein-DNA interactions.

IV.F. Web services to unite the bioinformatics nation 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 like BIND, KEGG, NCBI, Ensembl, FlyBase, SGD, WormBase, and UCSC all provide relevant data, but they are using a wide range of different systems and formats (140). 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 the services of the "bioinformatics nation", providers may adopt a Web services model (141).

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 UDDI, or Universal Description, Discovery, and Integration. 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 (140).

A number of bioinformatics services are currently available (142). For example,the OmniGene project (143) 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 (144). 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.

V. FITTING BIOLOGICAL CIRCUITS.

The profiling information together with protein-protein interaction maps 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, and 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 guide the comparison of new models evaluated for fit relative to the existing best model. The standard fitting approach for reaction networks is now described.

Let the parameters in the biological circuit be denoted by the M-tuple, $\theta := (\theta_1, ..., \theta_M)$. In the case explored here the parameters are the rate constants k_f and k_r, for all reactions r = 1,2, …, M_R as in the reaction network of Figure 11 and the initial concentrations $[s]_{t=0}$ for all intra-cellular species s = 1,2,…,M_S. The number of

parameters is $M = M_S + 2M_R$. For the deterministic model in Figure 11, the rate constants and initial conditions are $\theta = (\alpha_f, \alpha_{s,} \alpha_{qa\text{-}2}, \alpha_{qa\text{-}3}, \alpha_{qa\text{-}4}, \alpha_{qa\text{-}x}, \alpha_{qa\text{-}y}, \beta_f, \beta_{s,} \beta_{qa\text{-}2}, \beta_{qa\text{-}3}, \beta_{qa\text{-}4}, \beta_x, \beta_{qa\text{-}y}, \gamma_f,$

$\gamma_{qa\text{-}2}, \gamma_{qa\text{-}3}, \gamma_{qa\text{-}4}, \gamma_{qa\text{-}x}, \gamma_{qa\text{-}y}, \delta_f, \delta_{s,} \delta_{qa\text{-}2}, \delta_{qa\text{-}3}, \delta_{qa\text{-}4},$

$\delta_{qa\text{-}x}, \delta_{qa\text{-}y}, \kappa_0, Q_0, m_{f,0}, m_{s,0}, m_{qa\text{-}2,0}, m_{qa\text{-}3,0}, m_{qa\text{-}4,0}, m_{qa\text{-}x,0}, m_{qa\text{-}y,0}, p_{f,0}, p_{s,0,} p_{qa\text{-}2,0}, p_{qa\text{-}3,0},$

$p_{qa\text{-}4,0}, p_{qa\text{-}x,0}, p_{qa\text{-}y,0})$ with unit Hill coefficient. In the following, this parameter vector shall be referred to as as the "model θ".

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 $e \in E = \{1,.E\}$ experiments, in 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, for example, 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

$$Y_l := Y_{s,t,e} := ([s]_{t,e}/[s]_0) \text{ [with } l := (s,t,e)]$$

with some (known or unknown) reference concentration $[s]_e^{(ref)}$, if, *e.g.*, some linear measure of concentration is used or

$$Y_l := Y_{s,t,e} := \ln ([s]_{t,e}/[s]_0) \text{ [with } l := (s,t,e)]$$

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 ε $\{t_1,\ldots, t_{Mt}\}$ 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 Figure 15 for the *qa* cluster, E = 1, $M_S' = 6$, $M_T = 7$ and D = 42.

Now, let $F(\theta) := (F_1(\theta),\ldots,F_D(\theta))$ denote the corresponding predicted values for these observables Y for a given model θ. For the above-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 θ by solving the network's system of rate equations for the rate constants and initial conditions comprised in θ using the simulator KINSOLVER (40) and then calculating from that solution the log-concentration ratio $\ln([s]_t/[s]_0$ 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;θ) 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:

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

and E(Y) and Σ denote the mean and variance-covariance matrix of the observation vector Y for model θ. When multiple realizations of each profiling experiment are performed, then the variance-covariance matrix can be estimated. In the fitting reported in Figure 18, a univariate Gaussian with $\sigma/E(y_l) \times 100 = 20\%$ has been assumed and the observed single-experiment concentrations have been used as the data vector Y with the

link function $E(Y) = F(\theta)$ (145). To date, heteroskedasticity has been reported not an issue (146).

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 (147, 148). A model from the family in Figure 11 is displayed that fits the RNA profiling data of the *qa* cluster quite well in Figure 18. Profiles were obtained for 6 out of 7 of the *qa* genes. The RNA profile for *qa-1F* peaks at 4 hrs and then drops after that point with another rise at 6 hrs. The remaining profiles track that of *qa-1F* message levels. The simulator also yielded predictions about the protein profiles which are now testable (45).

V.A. 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, quinic acid metabolism is intimately connected to aromatic amino acid metabolism through the *aro* cluster in *N. crassa* (149). This raises the question of how QA metabolism is linked to general control (150). 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 (151).

V.B. A stochastic alternative. It is not clear at this time what role stochastic factors play in biological circuits. Patel and Giles (116) 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 Figure 12 was simulated with the results for the *qa-1F* message shown in Figure 19.

In this case, the 4 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 Figure 12 describing the stochastic circuit (91), 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).

VI. 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 (152), "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 Figure 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 like 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 will identify additional genes and their products which will have to 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 properties include memory and a cyclical response. The model

is re-evaluated and tested for goodness of 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 informative 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 automated into an adaptive control process to speed the process of gene-validated product discovery (28).

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

LITERATURE CITED

1. EJ Strauss & S Falkow. Microbial pathogenesis: genomics and beyond. Science 276: 1745-1749, 2000.
2. MT Cushion. *Pneumocystis carinii*. In Molecular Epidemiology of Infectious Disease, A Thompson (ed.), Kluwer Academic and Lippincott Raven Publishers, 1998.
3. GA Payne. Aflatoxin in maize. Critical Reviews in Plant Science: 423, 1996.
4. GA Payne. The aflatoxin biosynthetic pathway. Chapter 25. In Handbook of Industrial Mycology. Z. An (ed.), Marcel Dekker, NY, NY, 2002.
5. WE Timberlake & MA Marshall. Genetic engineering of filamentous fungi. Science 244: 1313-1317, 1989.
6. F Pelaez et al. Biological activities of fungal metabolites. Chapter 3. In Handbook of Industrial Mycology. Z. An (ed.), Marcel Dekker, NY, NY , 2002.
7. WE Timberlake. Molecular genetics of Aspergillus development. Annu. Rev. Genet 24: 5-6, 1990.
8. JC Dunlap. Molecular bases for circadian clocks. Cell 96: 271-290, 1999.
9. T Ideker, V Thorsson, JA Ranish, R Christmas, J Buhler, JK Eng, R Bumgarner, DR Goodlett, R Aebersold, & L Hood**.** Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. Science 292: 929-934, 2001.
10. JW Bennett & J Arnold. Genomics of fungi. The Mycota VIII. Biology of the Fungal Cell. Howard/Gow (eds.), Springer-Verlag, NY, NY. pp. 267-297, 2001.
11. A Goffeau, BG Barrell, H Bussey, RW Davis, B Dujon, H Feldmann, F Galibert, JD

Hoheisel, C Jacq, M Johnston, EJ Louis, HW Mewes, Y Murakami, P Philippsen, H Tettelin and SG Oliver. Life with 6000 genes. Science 274: 546-567, 1996.

12. JL DeRisi, VR Iyer & PO Brown. Exploring the metabolic and genetic control of gene expression on a genomic scale. Science 278: 680-686, 1997.
13. A-C Gavin, M Bosche, R Krause, P Grandi, M Marzioch, A Bauer, J Schultz, JM Rick, A-M Michon, C-M Cruciat, M Remor, C Hofert, M Schelder, M Brajenovic, H Ruffner, A Merino, K Klein, M Hudak, D Dickson, T Rudi, V Gnau, A Bauch, S Bastuck, B Huhse, C Leutwein, M-A Heurtier, RR Copley, A Edelmann, E Querfurth, V Rybin, G Drewes, M Raida, T Bowmeester, P Bork, B Seraphin, B Kuster, G Neubauer, & G Superti-Furga. Functional organization of the yeast proteome by systematic analysis of protein complexes. Nature 415: 141-147, 2002.
14. Gene Ontology Consortium**.** Gene ontology: tool for the unification of biology. Nature Genet 25: 25-29, 2000.
15. F Jacob, & J Monod. Genetic regulatory mechanisms in the synthesis of proteins. J. Mol. Biol. 3: 318-356, 1961.
16. Lee, K, JJ Loros, & JC Dunlap. Interconnected feedback loops in the Neurospora Circadian system. Science 289: 107-110, 2000.
17. H Jeong, B Tombor, R Albert, ZN Oltvai, and AL Barabasi**.** The large-scale organization of metabolic networks. Nature 407: 651-654, 2000.
18. PD Karp, PD, M Riley, SM Paley, A Pellegrini-Toole & M Krummenacker. Eco Cyc: Encyclopedia of *Escherichia coli* genes and metabolism. Nucl. Acids Res 27: 55-58, 1999.
19. M Kanehisa, & S Gota. KEGG for computational genomics. In Current Topics in

Computational Molecular Biology. T Jiang, Y Xu and MQ Zhang (eds.), pp. 301-315, MIT Press, Cambridge, MA, 2002.

20. T Ito, K Tashiro, S Muta, R Ozawa, T Chiba, M Nishizawa, K Yamamoto, S Kuhara, & Y Sakaki. Toward a protein-protein interaction map of the budding yeast: a comprehensive system to to examine two-hybrid interactions in all possible combinations between the yeast proteins. PNAS USA 97: 1143-1147, 2000.
21. P Uetz, L Glot, G Cagney, TA Mansfield, RS Judson, JR Knight, D Lockshon, V Narayan, M Srinivasan, P Pochart, A Qureshi-Emili, B Godwin, D Conover, T Kalbfleish, G. Vijayadamodar, M. Yang, M. Johnston, S Fields, & JM Rothberg. A comprehensive analysis of protein-protein interactions in *Sacharomyces cerevisiae* Nature 403: 623-627, 2000.
22. MC Gustin, J Albertyn, M Alexander, & K Davenport**.** MAP kinase pathways in the yeast *Saccharomyces cerevisiae*. Microbiology and Molecular Biology Reviews 62: 1264-1300, 1998.
23. FCP Holstege, EG Jennings, JJ Wyrich, TI Lee, CJ Hengartner, MR Green, TR Golumb, ES Lander, & RA Young. Dissecting the regulatory circuitry of a eukaryotic genome.Cell 95: 717-728, 1998.
24. H Zhu, M Bilgin, R Bangham, D Hall, A Casamayor, P Bertone, N Lan, R Jansen, S Bidlingmaier, T Hoafek, T Mitchell, P Miller, RA Dean, M Gerstein, & M Snyder. Global analysis of protein activities using proteome chips. Science 293: 2101-2105, 2001.
25. A Wagner. The yeast protein interaction network evolves rapidly and contains few redundant duplicate genes. Mol. Biol. Evol 18: 1283-1292, 2002.

26. JC Avise. Molecular Markers, Natural History, and Evolution. Chapman and Hall, NY, NY, 1994.

27. JC Venter *et al*. The sequence of the human genome. Science 291: 1304-1351, 2001.

28. KJ Kochut, J Arnold, A Sheth, JA Miller, E Kraemer, B Arpinar, & J Cardoso. INTELLIGEN: a distributed workflow system for discovering protein-protein interactions. Parallel and Distributed Databases, 13, 43-72, 2003.

29. KJ Kochut,, J Arnold, JA Miller & WD Potter. Design of an object-oriented database for reverse genetics, p. 234-242. *In* L. Hunter and D. Searls and J. Shavlik (ed*.),* Proceed. First Internl Conference on Intelligent Systems for Molecular Biology. AAAI Press, Menlo Park, CA, 1993.

30. SB Davidson, SB, J Crabtree, B Brunk, J Schug, V Tannen, C Overton, and C Stoeckert. K2/Kleisli and GUS: experiments in integrated access to genomic data sources. IBM Systems Journal (March issue) 40: 512-530, 2001.

31. JD Kececioglu & EW Myers. Combinatorial algorithms for DNA sequence assembly. Algorithmica 13: 7-51, 1995.

32. M Borodowsky & J McInich. GeneMark: parallel gene recognition for both DNA strands. Computers and Chemistry 17: 123-133,1993.

33. SM Bhandarkar, S Machaka, S Chirravuri, & J Arnold. Parallel computing for chromosome reconstruction. Parallel Computing 24: 1177-1204, 1998.

34. SM Bhandarkar, SA Machaka, SS Shete & R Kota. Parallel computation of a maximum-likelihood estimator of a physical map. Genetics 157: 1021-1043, 2001.

35. RD Hall, S Bhandarkar, & J Wang. ODS2: a multiplatform software application for

creating integrated physical and genetic maps. Genetics 157: 1045-1056, 2001a.

36. E Kraemer, J Wang, J Guo, S Hopkins, & J Arnold 2001. An analysis of gene-finding approaches for *Neurospora crassa*. Bioinformatics 17: 901-912, 2001.
37. Z Xu, B Lance, C Vargas, B Arpinar, E Kraemer, KJ Kochut, JA Miller, JR Wagner, MJ Weise, JK Wunderlich, J Stringer, G Smulian, MT Cushion, & J Arnold. Mapping by sequencing the Pneumocystis genome using the ODS3 tool, Genetics, in press, 2003.
38. K Shah & A Sheth**.** InfoHarness: an information integration platform for managing distributed, heterogeneous information. IEEE Internet Computing, November-December, p. 18-28, 1999.
39. US Bhalla & R Iyengar. Emergent properties of networks of biological signaling pathways. Science 283: 381-387, 1999.
40. B Aleman-Meza, HB Schuttler, J Arnold, & TR Taha. KINSOLVER: a simulator for biochemical and gene regulatory networks. Bioinformatics, submitted (attached as preprint), 2002. Also found in B Aleman-Meza. Advances in numerical solution of kinetics reaction equations. M.A.M.S. Thesis, University of Georgia, Athens, GA. 2001.
41. S Datta, & J Arnold. Some comparisons of clustering and classification techniques Applied to transcriptional profiling data. In Advances in Statistics, Combinatorics, and Related Areas. World Scientific, Singapore, 2002.
42. E Kraemer, & TE Ferrin. Molecules to maps: tools for visualization and interaction in support of computational biology. Bioinformatics 14: 764-771, 1998.
43. EM Marcotte, M Pellegrini, MJ Thompson, TO Yeates, and D. Eisenberg. A

combined algorithm for genome-wide prediction of protein functon. Nature 402: 83-86, 1999.

44. X Fang, JA Miller, and J Arnold. J3DV: a Java-based 3D database visualization tool. Software – Practice and Experience 32: 443-463, 2002.

45. SP Gygi, B Rist, SA Gerber, F Trecek, MH Gelb, & R Aebersold. Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. Nature Biotechn. 17: 994-999.

46. E Lander & NJ Schork. Genetic dissection of complex traits. Science 265: 2037-2038, 1994.

47. AA Brakhage. Molecular regulation of beta-lactam biosynthesis in filamentous fungi. Microbiol. And Molecul. Biol Rev 62, 1998.

48. M Lynch, & B Walsh . Genetics and analysis of quantitative traits. Sunderland, MA, 1998.

49. E Lander & D Botstein. Mapping Mendelian factors underlying quantitative traits with RFLP linkage maps. Genetics 121: 185-199, 1989.

50. Z-B Zeng. Precision mapping of quantitative trait loci. Genetics 136: 1457-1468, 1994.

51. RW Doerge, Z-B Zeng, & BS Weir**.** Statistical issues in the search for genes Affecting quantitative traits in experimental populations. Statistical Science 12: 195-220, 1997.

52. ML Wayne & LM McIntyre. Combining mapping and arraying: a novel approach to candidate gene identification. PNAS USA, in press, 2002.

53. M Xiong & SW Guo. Fine-scale mapping of quantitative trait loci using historical

recombinations. Genetics 145: 1201-1218, 1997.

54. RA Young. Biomedical discovery with DNA arrays. Cell 102: 9-15, 2000.
55. VE Velculescu, L Zhang, B Vogelstein, KW Kinzler. Serial analysis of gene expression. Science 270: 484-486, 1995.
56. M Prosniak, DC Hooper, B Dietzschold, and H Koprowski. Effect of rabies virus Infection on gene expression in mouse brain. PNAS USA 98: 2758-2763, 2001.
57. S Chu, J DeRisi, M Eisen, J Mulholland, D Botstein, PO Brown, and I Herskowitz. The transcriptional program of sporulation in budding yeast. Science 282: 699-705, 2000.
58. T Wu. An extensible framework for developing visualization software for gene expression data. M.S., Department of Computer Science, University of Georgia 2001.
59. A Sveiczer, A Csikasz-Nagy, B Gyorffy, JJ Tyson, & B Novak. Modeling the fission yeast cell cycle: quantized cycle times in *wee1- cdc25D* mutant cells. PNAS USA 97: 7865-7870, 2000.
60. Sattlegger, E, AG Hinnebusch, & IB Barthelmess. *cpc-3*, the *Neurospora crassa* homologue of yeast *GCN2*, enclodes a polypeptide with juxtaposed eIF2a kinase and histidyl-tRNA synthetase-related domains required for general amino acid control. J. Biol. Chem. 273: 20404-20416, 1998.
61. MP Washburn, D Walters, JR Yates. Large-scale analysis of the yeast proteome by multidimensional protein identification technology. Nature Biotech. 19: 242-247, 2001.
62. DA Wolters, MP Washburn, & JR Yates, III. An automated multidimensional

protein identification technolgy for shotgun proteomics. Analytical Chemistry, in press, 2002.

63. O Fiehn. Metabolomics - the link between genotypes and phenotypes. Plant Molecular Biology 48: 155-171: 2002.

64. J Friesvad. Fungal metabolites profiling by chemometrics methods. In Handbook of Industrial Mycology. Z An (ed.), Marcel Dekker, NY, NY, 2002.

65. IH Segel. *Enzyme Kinetics*. Wiley, NY, 1975.

66. P Mendes. Biochemistry by numbers: simulation of biochemical pathways with Gepasi 3. Trends in Bioch. Sci. 22: 361-363, 1997.

67 M Ehlde & G Zacchi**.** MIST: a user-friendly metabolic simulator. Comput. Appic. Biosci. 11: 201-207, 1995.

68. HM Sauro. SCAMP: a general-purpose simulator and metabolic control analysis program. Comput. Applic. Biosci. 9: 441-450, 1993.

69. B Ren, F Robert, JJ Wyrick, O Aparicio, EG Jennings, I Simon, J Zeitlinger, J Schreiber, N Hanett, E Kanin, TL Volkert, CJ Wilson, SP Bell, & RA Young. Science 290: 2306-2309, 2000.

70. G Bader, I Donaldson, C Wolting, BF Francis Ouellette, T Pawson, & CWV Hogue. BIND - the biomolecular interaction network database. Nucl. Acids. Res. 29: 242-245, 2001.

71 . JD Lieb, X Liu, D Botstein, & P.O. Brown Promoter-specific binding of Rap1 revealed by genome-wide maps of protein-DNA association. Nature Genetics 28: 327-334, 2001.

72. VR Iyer, CE Horak, CS Scafe, D Botstein, M Snyder, & PO Brown. Genomic

binding sites of the yeast cell-cycle transcription factors SBF and MBF. Nature 409: 533-538, 2001.

73. International Human Genome Sequencing consortium. Initial sequence and analysis of the human genome. Nature 409: 860-921, 2001.

74. RD Hall, JA Miller, J Arnold, KJ Kochut, AP Sheth, & MJ Weise. Using workflow to build an information management system for a geographically distributed genome sequencing initiative. In "Genomics of Plants and Fungi", R.A. Prade and H.J Bohner (eds.), Marcel Dekker, Inc. NY, NY. In Hall, RD 1999.
New Computational Tools for Genome Mapping. Ph.D. Dissertation. Un. GA. See http://gene.genetics.uga.edu/workflow for demonstation.

75. RD Hall, SM Bhandarkar, J Arnold, & T Jiang. Physical mapping with automatic capture of hybridization data, Bioinformatics 17: 205-213, 2001.

76. A Sheth, W Aalst, & I Arpinar. Processes driving the networked economy. IEEE Concurrency 7: 18-31, 1999.

77. J Cardoso, JA Miller, A Sheth, & J Arnold. Modeling quality of service for workflows and web processes. Very Large Database Journal, in press, 2002.

78. D Georgakopoulos, M Hornick, & A Sheth**.** An overview of workflow management: from process modeling to infrastructure for automation. Distributed and Parallel Databases Journal 3: 119-153, 1995.

79. L Dogac, A Kalinechenko, T Ozsu & A Sheth. Workflow management systems and interoperability. NATO ASI Series F, Vol 164, 524 pages, 1998.

80. W Aalst & K Hee . Workflow Management: Models, Methods, and Systems. MIT Press, 2002, Cambridge, MA

81. JA Miller, D Palaniswami, A Shet, K Cochut, & H Singh. WebWork: METEOR's web-based workflow management system. J. Intell. Inform. Systems (JIIS) 10: 186-215, 1998.

82. KJ Kochut, AP Sheth, & JA Miller. Optimizing workflows. Component Strategies 1: 45-57, 1999.

83. LP Zhao, R Prentice, & L Breeden. Statistical modeling of large microarray data sets to identify stimulus-response profiles. PNAS USA 98: 5631-5636, 2001.

84. R Heinrich, & S Schuster**.** The Regulation of Cellular Systems. Chapman & Hall, NY, 1996.

85. K Murphy & S Mian. Modeling gene expression data using dynamic Bayesian networks. Technical Report. Berkeley, 1999.

86. N Friedman, M Linial, I Nachman, & D Pe'er**.** Using Bayesian network to analyze expression data. J. Comput. Biol 7:601-620, 2000.

87. DC Weaver, CT Workman, & GD Stormo. Modeling regulatory networks with weight matrices. Pac. Symp. Biocomput 4: 112-123, 1999.

88. S Huang. Gene expression profiling, genetic networks and cellular states: an integrating concept for tumorigenesis and drug discovery. J. Mol. Med. 77: 469-480, 1999.

89. I Shmulevich, O. Yliharja, & J Astola. Inference of genetic regulatory networks under the best-fit extension paradigm. In Proceedings of the IEEE-EuRASIP Workshop on Nonlinear Signal and Image Processing (NSIP-01), June 3-6, Baltimore, Maryland, 2001.

90. I Shmulevich, ER Dougherty, S Kim, and W Zhang. Probabilistic Boolean

networks: a rule-based uncertainy model for gene regulatory networks. Bioinformatics 18: 261-274, 2002.

91. DT Gillespie. Exact stochastic simulation of coupled chemical reactions. J. Phys. Chem 81:2340-2361, 1977.

92. AM Kierzek. STOCKS: STOChastics kinetic simulations of biochemical systems with Gillespie algorithm. Bioinformatics 18: 470-481, 2002.

93. NH Giles, ME Case, J Baum, R Geever, L Huiet, V Patel, & B Tyler. Gene organization and regulation in the *qa* (Quinic Acid) gene cluster of *Neurospora crassa*. Microbiol. Rev. 49: 338-358, 1985.

94. RF Geever, L Huiet, JA Baum, BM Tyler, VB Patel, BJ Rutledge, ME Case, & NH Giles. DNA sequence, organization and regulation of the *qa* gene cluster of *Neurospora crassa*. J. Mol. Biol. 207: 15-34, 1989.

95. M Johnston. A model fungal regulatory mechanism: the *GAL* genes *of Saccharomyces cerevisiae*. Microbiological Reviews 51: 458-476, 1987.

96. D Voet & JG Voet. *Biochemistry*. Wiley, NY, 1990.

97. C Yanofsky. Advancing our knowledge in biochemistry, genetics, and microbiology through studies of tryptophan metabolism. Annu. Rev. Biochem. 70: 1-37, 2001.

98. DW Brown, JH Yu, HS Kelkar, M Fernandes, TC Nesbitt, NP Keller, TH Adams & TJ Leonard**.** Twenty-five coregulated transcripts define a sterigmatocystin gene cluster in *Aspergillus nidulans*. PNAS USA 93: 1418-1422, 1996.

99. CP Woloshuk, KR Foutz, JF Brewer, D Bhatnagar, TE Cleveland & GA Payne. Molecular characterization of aflR, a regulatory locus for aflatoxin biosynthesis. Appl Environ. Microbiol 60: 2408-2414, 1994.

100. DJ Smith,, MKR Burnham, JH Bull, JE Hodgson, JM Ward, P Browne, J Brown, B. Barton, AJ Earl & G Turner. B-lactam antibiotic biosynthetic genes have been conserved in clusterin prokaryotes and eukaryotes. EMBO J. 9: 741-747, 1990.

101. T Suarez, & MA Penalva. Characterization of a *Penicillium chyrsogenum* gene encoding a *PacC* transcription factor and its binding sites in the divergent *pcbAB-pcbC* promoter of the penicillin biosynthetic cluster. Molec. Microb. 20: 529-540, 1996.

102. F Fierro, JL Barvedo, B Diez, S Gaterrez, FJ Fernandez, & JF Martin 1995. The penicillin gene cluster is amplified in tandem repeats linked by conserved hexanucleotide sequences. PNAS USA 92: 6200-6204, 1995.

103. F Chen & MT Cushion**.** Use of ATP bioluminescent assay to evaluate viability of *Pneumocystis carinii* from rats. J. Clin. Microbiol 32: 2791-2800, 1994.

104. AG Smulian, T Sesterhenn, R Tanaka, & MT Cushion. The *ste3* pheromone receptor gene of *Pneumocystis carinii* is surrounded by a cluster of signal transduction genes. Genetics 157: 991-1002, 2001.

105. A Cornish-Bowden and JH Hofmeyer**.** MetaModel: a program for modeling and control analysis of metabolic pathways on the IBM PC and compatibles. Comput. Applic. Biosci. 7: 89-93, 1991.

106. P Mendes. GEPASI: a software package for modeling the dynamics, steady states and control of biochemical and other systems. Comput. Applic. Biosci. 9: 563-571, 1993.

107. BA Barshop, RF Wrenn & C Frieden. Analysis of numerical methods for computer simulation of kinetic processes: development of KINSIM - a flexible, partable

system. Anal. Biochem 130: 134-145, 1983.

108. Q Dang and C Frieden**.** New PC versions of the kinetic-simulation and fitting programs, KINSIM and FITSIM. Trends Biochem Sci 22: 317, 1997.

109. M Tomita, K Hashmoto, K Takahasi, TS Shimizu, Y Matsuzaki, F Miyoshi, K Saito, Tanida, K Yugi, JC Venter, & CA Hutchison. E-CELL: software environment for whole-cell simulation. Bioinformatics 15: 72-84, 1999.

110 R Alves & MA Savageau . Systemic properties of ensembles of metabolic networks: application of graphical and statistical methods to simple unbranched pathways. Bioinformatics 16: 534-547, 2002.

111.BM Tyler, RF Geever, ME Case, & NH Giles. Cis-acting and trans-acting regulatory mutations defin two types of promoters controlled by the *qa-1F* gene of Neurospora. Cell 36: 493-502, 1984.

112.HM McAdams, & A Arkin. Stochastic mechanisms of gene expression. PNAS USA 94: 814-819, 1997.

113.TB Kepler & TC Elston. Stochasticity in transcriptional regulation: origins, consequences,and mathematical representations. Biophys. J. 81: 3116-3136, 2001.

114.J-P Abastado, PF Miller, & AG Hinnebusch. A quantitative model for translational control of the *GCN4* gene of *Saccharomyces cerevisiae*. New Biol 3: 511-524, 1991.

115.AG Hinnebusch. Mechanism and regulation of initiator methionyl-tRNA binding to ribosomes. In Translational Control of Gene Expression. CSHL Press, NY. pp. 185-243, 2000.

116.VB Patel & NH Giles. Autogenous regulation of the positive regulatory *qa-1F* gene in *Neurospora crassa*. MCB 5: 3593-3599, 1985.

117.JG Kemeny, JL Snell, & AW Knapp . *Denumerable Markov Chains*. Springer-Verlag, NY, 1976.

118.MA Gibson & J Bruck. A probabilistic model of a prokaryotic gene and its regulation. In Computational Modeling of Genetic and Biochemical Networks. MIT Press, Cambridge, MA 2001, pp. 49-71.

119.G Weng, US Bhalla, & R Iyengar. Complexity in biological signaling systems. Science 284: 92-96**,** 1999.

120.P Mendes & DB Kell. MEG (Model Extender for Gepasi): a program for the modelling of complex, heterogeneous cellular systems. Bioinformatics 17: 288-289, 2001.

121.N Burns, B. Grimwade, PB Ross-Macdonald, E-Y, K Finberg, GS Roeder, & M Snyder . Large-scale analysis of gene expression, protein localization, and gene disruption in *Saccharomyces cerevisiae*. Genes and Development 8: 1087-1105, 1994.

122.M Hensel, JE Shea, C Gleeson, & DW Holden. Simultaneous identification of bacterial virulence genes by negative selection. Science 269: 400-403, 1995.

123.A Wach, A Brachat, C Alberti-Segui, C Rebischung, & P Philippsen. Heterologous *HIS3* and GFP reporter modules for PCR-targeting in *Saccharomyces cerevisiae*. Yeast 13: 1065-1075, 1997.

124. P Ross-Macdonald, PSR Coelho, T Roemer, S Agarwal, A Kumar, R Jansen, K-H Cheung, A Sheehan, D Symoniatis, L Umansky, M Heldtman, FK Nelson, H Iwasaki, K Hager, M Gerstein, P Miller, SS Roeder, & M Snyder**.** Large-scale analysis of the Yeast genome by transposon tagging and gene disruption. Nature

402: 413-418, 1999.

125. E Winzeler, DD Shoemaker, A. Astromoff, H Liang, K Anderson, B Andre, R Bangham, R Benito, JD Boeke, H. Bussey, AM Chu, C Connelly, K Davis, F Dietrich, SW Dow, ME Bakkoury, F. Foury, SH Friend, E. Gentalen, G. Giaever, JH Hegemann, T. Jones, M. Laub, H. Liao, N. Liebundguth, DJ. Lockhart, A Lucau-Danila, M Lussier, N M'Rabet, P. Menard, M. Mittmann, C Pai, C Rebischung, JL Revuelta, L. Riles, C.J. Roberts, P. Ross-MacDonald, B Scherens, M Snyder, S Sookhai-Mahadeo, RK Storms, S Veronneau, M Voet, G Volckaert, TR Ward, R Wysocki, GS Yen, K Yu, K Zimmermann, P Philippsen, M Johnston, & RW Davis. Functional characterization of the *S. cerevisiae* genome by gene deletion and parallel analysis. Science 285: 901-906, 1999.

126. G Glaever, AM Chu, Li Ni, C Connelly, L Riles, S Veronneau, S Dow, A Lucau-Danlla, K Anderson, B Andre, AP Arkin, A Astrmoff, ME Bakkoury, R Bangham, R Benito, S Brachat, S Campanaro, M Curtiss, K Davis, A Deutschbauer, K-F Entlan, P Flaherty, F Foury, DJ Garfinkel, M Gerstein, D Gotte, U Guldener, JH Hegemann, S Hempel, Z Herman, DF Jaramillo, DE Kelly, SL Kelly, P Kotter, D LaBonte, DC Lamb, N Lan, H Lian, H Liao, L Liu, C Luo, M Lussler, R Mao, P Menard, SL Ool, JL Revuetta, CJ Roberts, M Rose, P Ross-Macdonald, B Scherens, G Schlmmack, B Shafer, DD Shoemaker, S Sookha-Mahadeo, RK Storms, JN Strathern, G Valle, M Voet, G Volckaert, C-Y Wang, TR Ward, J Wilhelmy, EA Winzeler, Y Yang, G Yen, E Youngman, K Yu, H Bussey, JD Boeke, M Snyder, P Philippsen, RW Davis, & M Johnston. Functional profiling of the *Saccharomyces cerevisiae* genome. Nature 418: 387-391, 2002.

127. L Hamer, LK Adachi, MV Montegegro-Chamorro, MM Tanzer, SK Mahanty, C Lo, RW Tarpey, AR Skalchunes, RW Heiniger, SA Frank, BA Darveaux, DJ Lampe, TM Slater, L Ramamurthy, TM DeZwaan, GH Nelson, JR Shuster, J Woessner, and JE Hamer. Gene discovery and gene functional assignment in filamentous fungi. PNAS USA 98: 5110-5115, 2001.

128.PR Ross-MacDonald, A Sheehan, G Shirleen Roeder, & M Snyder. A multipurpose transposon system for analyzing protein production, localization, and function in *Saccharomyces cerevisiae*. PNAS USA 94: 190-195, 1997.

129. BP Cormack , N Ghori, & S Falkow**.** An adhesin of the yast pathogen *Candida glabrata* mediating adherence to human epithelial cells. Science 285: 578-582, 1999.

130. RT Nelson, J Hua, B Pryor & JK Lodge. Identification of virulence mutants of the fungal pathogen *Cryptococcus neoformans* using signature-tagged mutagenesis. Genetics 157: 935-947, 2001.

131. TS Gardner, T.S., C.R. Cantor, & JJ Collins. Construction of a genetic toggle switch in *Escherichia coli*. Nature 403: 339-342, 2000.

132. MB Elowitz & S Leibler**.** A synthetic oscillatory network of transcriptional Regulators. Nature 403: 335-338,2000.

133. HS Kelkar, J Griffith, ME Case, SF Covert, RD Hall, CH Keith, JS Oliver, MJ Orbach, MS Sachs, JR Wagner, MJ Weise, J Wunderlich, & J Arnold **.** The *Neurospora crassa* genome: cosmid libraries sorted by chromosome. Genetics 157: 979-990, 2001.

134. RS Chaleff. Studies on the Genetic Control of the Inducible Quinate-Shikimate

Catabolic Pathway in *Neurospora crassa*. Ph.D. dissertation, Yale University, 1972.

135.RS Chaleff. The inducible quinate-shikimate catabolic pathway in *Neurospora crassa*: genetic organization. J. General Microbiology 81: 337-355, 1974.

136. ME Case, M Schweizer, SR Kushner, & NH Giles. Efficient transformation of *Neurospora crassa* by utilizing hybrid plasmid DNA. PNAS USA 76: 5259-5263, 1979.

137. S Datta. Exploring relationships in gene expression: a partial least squares approach. Gene Expression 9: 257-264, 2001.

138. AJM Walhout, R Sordella, X Lu, JL Hartley, GF Temple, MA Brasch, N Thierry-Mieg, & M. Vidal**.** Protein interaction mapping in *C. elegans* using proteins involved in vulval development. Science 287: 116-122, 2000.

139. Y Ho, A Gruhler, A Helibut, GD Bader, L Moore, S-L Adams, A Millar, P Taylor, K Bennett, K Boutilier, L Yang, C Wolting, I Donaldson, S Schandorff, J Shewnarane, M Vo, J Taggart, M Goudreault, B Muskat, C Alfarano, D Dewar, Z Lin, K Michalickova, AR Willems, H Sassi, PA Nielsen, KJ Rasmussen, JR Andersen, LE Johansen, LH Hansen, H Jespersen, A Podtelejnikov, E Nielsen, J Crawford, V Paulsen, BD Serensen, J Mathlesen, RC Hendrickson, F Gleeson, T Pawson, MF Moran, D Durocher, M Mann, CWV Hogue, D Flgeys, & M Tyers**.** Systematic identification of protein complexes in Saccharamoyces cerevisiae by mass spectrometry. Nature 415: 180-183, 2002.

140. L Stein. Creating a bioinformatics nation: A web-services model will allow biological data to be fully exploited, Nature 417: 119-120, 2002.

141. S Chandrasekaran, JA Miller, G. Silver, IB Arpinar, and A Sheth. Composition,

Performance Analysis, and Simulation of Web Services UGA Computer Science Technical Report, September 2002.

142. E Cerami. Web Services for Bioinformatics, O'Reilly Net, May, 2002.

143. OmniGene Standardizing Biological Data Interchange Through Web Services Technology, http://omnigene.sourceforge.net/index.html, 2002.

144. DAS 2002 http://www.biodas.org/.

145. P McCullagh, & JA Nelder. Generalized Linear Models. Chapman & Hall, London,1983.

146. MK Kerr & GA Churchill. Bootstrapping cluster analysis: assessing the reliability of conclusions from microarray experiments. PNAS USA 98: 8961-8965, 2001.

147. P Mendes & DB Kell. Non-linear optimization of biochemical pathways: applications to metabolic engineering and parameter estimation. Bioinformatics 14: 869-883, 1998.

148. N Draper & H Smith. Applied Regression Analysis. Wiley, NY, 1981.

149. ME Case, NH Giles, & CH Doy. Genetical and biochemical evidence for further interrelationships between the polyaromatic synthetic and the quinate-shikimate catabolic pathways in *Neurospora crassa*. Genetics 71: 337-348, 1972.

150. GH Braus. Aromatic amino acid biosynthesis in the yeast *Saccharomyces cerevisiae*: a model system for the regulation of a eukaryotic biosynthetic pathway. Microbiol. Rev 55: 349-370, 1988.

151. D Battogtokh, DK Asch, ME Case, J Arnold, & HB Schuttler. An ensemble method for identifying regulatory circuits with special reference to the *qa* gene cluster of *Neurospora crassa*. PNAS USA, in press, 2003.

152.GW Beadle & EL Tatum. Genetic control of biochemical reactions in *Neurospora*. Proceedings of the National Academy of Sciences USA 27:499-506, 1941.

153.SC Crosthwaite, SC, JC Dunlap, & JJ Loros**.** Neurospora *wc-1* and *wc-2*; transcription, photoresponses, and the origins of circadian rhythmicity. Science 276: 763-769, 1997.

Figure 1. The promise of genomics is 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 reprinted with permission from (Crossthwaite, SC, JC Dunlap, & JJ Loros (153). Neurospora *wc-1* and *wc-2*; transcription, photoresponses, and the origins of circadian rhythmicity. Science 276: 763-769). Copyright 1997 American Association for the Advancement of Science. *N. crassa* is shown growing in race tubes with the regular pattern of conidia being formed displaying the clock under varied conditions.

Figure 2. A display of relative RNA levels under 10 conditions (time points) during the cell cycle of *S. cerevisiae*. The 6000 genes are displayed vertically by the similarity in their RNA profiles using UPGMA as implemented in Wu (58). Increasing red indicates increased expression relative to the beginning time point, and increasing green indicates decreased expression relative to the beginning time point.

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.

Figure 4. 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 out of

a reaction are products. In essence, the arrows define what is the forward reaction direction.

Figure 5. Water model 2 is an elaboration of Water Model 1, an elaboration 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 7 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, letting QA into the cell.

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 $qa\text{-}1F^{p}$. The third row of squares is the set of the cognate RNAs which are translated into 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 $qa\text{-}1S^{p}$ is shown binding to $qa\text{-}1F^{p}$ 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-y*^{p} 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-1F*^{p} 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-1F*^{p} 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-1F*^{p} . 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-3*^{p} 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.

Figure 9. The biological circuit for the *lac* operon is more elaborate than that of the *qa* cluster. The *lacI*^{p} repressor can bind to the operator to shut down the cluster through a negative feedback loop unless lactose is present to bind to *lacI*^{p}, thereby titrating out the repressor. There is also a positive feedback loop provided by the catabolite repressor

protein crp^p which aids in the recruitment of RNA polymerase to the promoter ($lacP^0$). The catabolite repressor protein is only active when bound to cAMP. The enzymes ac^p and pd^p make cAMP from ATP and convert cAMP to AMP, respectively. An internal signaling cascade including ($e1^p$, eII^p, $eIII^p$, and hPr^p) is included to take a phosphate on phosphoenolpyruvate (PEP) to glucose to pump glucose into the cell as Glucose-6-phosphate.

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 $trpR^p$ 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.

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 Figure 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 α 's denote basal transcription rates, the δ 's, the QA inducible transcriptional rates, the γ 's, the repressor effects, and the β '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.

Figure 12. Part of the list of master equations for a stochastic circuit with the same structure as Figure 6 is listed. Here mF_B and mS_B are the basal numer of mRNAs for *qa-1F* and *qa-1S*; mF_I and mS_I are the number of induced mRNAs for *qa-1F* and *qa-1S*; and mF_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.

Figure 13. Three kinds of system perturbations are illustrated for the *qa* cluster: 1) genetic; 2) chemical as in a drug; or 3) environmental.

Figure 14: Response of PCA (protochatechuic acid) level with and without a *qa-2* gene knockout over time as simulated in KINSOLVER for the biological circuit in Figure 6 (40).

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 (Roche, Inc.). Unincorporated ^{33}P was removed by spin columns (Sigma, Inc.). Arrays were probed with ^{33}P labeled cDNAs derived from 3 time points after the shift to quinic acid.

Images were collected on a Packard Instant Imager over a 26 min period. The *qa-2/aro-9* double mutant is also shown expressing the same transcripts at 240 min, but it does not grow.

Figure 16. Twelve genes appear to respond to a shift from 1.5% sucrose to 0.3% quinic acid. The counts of 12 genes (from Figure 15) recorded by the Packard Instant Imager are graphed as a function of time. Only two of these genes appear to be part of the *qa* cluster.

Figure 17. Protein mobile of protein-protein interaction map presented by Ito *et al.* (20). Nodes represent proteins, and edges represent interactions. Graphic was generated by software described in part in Fang *et al.* (44).

Figure 18. Measured trajectories over time of RNA levels for six of the seven *qa* cluster genes. Solid dots are the data (151). Smooth curves are those of a fitted model as in Figure 11 chosen to maximize the likelihood $P(Y; \theta)$

Figure 19. Counts of *qa-1F* message in a cell over time for 5 independent realizations of the stochastic alternative to the circuit in Figure 6, *i.e* . Figure 12

Figure 20. An example of Hypothesis-driven Genomics or the process of metabolomics.