MOLECULAR NETWORKS IN MODEL SYSTEMS

Timothy Galitski
Institute for Systems Biology, Seattle, Washington 98103;
email: tgalitski@systemsbiology.org

Key Words  data integration, graph, module, phenotype, genetics

Abstract  Model organisms, especially the budding yeast, are leading systems in the transformation of biology into an information science. With the availability of genome sequences and genome-scale data generation technologies, the extraction of biological insight from complex integrated molecular networks has become a major area of research. Here I examine key concepts and review research developments. I propose specific areas of research effort to drive network analysis in directions that will promote modeling with increasing predictive power.

INTRODUCTION

Biology is becoming a quantitative information science (18). Yeast was proposed as a model organism for eukaryotic genomics (6). The yeast system has continued to yield pioneering developments in genome-scale analysis and the understanding of biological systems. In many ways, it is the leading system in the transformation of biology. This transformation of biology raises the need for enhanced interdisciplinary understanding.

This review discusses key developments in the analysis of complex molecular networks in model systems, primarily the budding yeast. Whereas dynamic network simulations are currently typically carried out in networks known in considerable biochemical detail (13), the focus here is on large-scale networks in which data quality and coverage are variable. The hope is that by examining conceptual foundations and suggesting specific areas for research, this will guide the analysis and experimental investigation of complex networks in directions that will accelerate the modeling of biological systems of higher complexity with increasingly explicit predictive power. The emphasis here is complemented by reviews focusing on the collection of genome-scale data and the mapping of interaction networks (1, 3).
DATA INTEGRATION

Network-Data Integration in a Graph

The understanding of complex cell properties is an inherently integrative problem. Cell properties such as pathogenicity, growth control, morphology, and metabolic capabilities arise from the interactions of many diverse elements of molecular networks. Cell properties and biological processes involve changes in gene activity at multiple levels, including mRNA and protein levels, molecular modifications, changes in localization, etc. Integrating diverse data enables biological discovery (1). Graphs are a frequently and increasingly used framework for genome-scale data integration and network analysis.

Graph-based methods provide a general solution for data integration, visualization, and computational analysis of disparate data. Conceptually, a graph contains a set of nodes (vertices) connected by edges. For biomolecular networks, these typically represent molecules and interactions, respectively. Nodes and edges can be assigned arbitrary attributes, such as, molecule type or interaction type, expression pattern, annotation, and directionality, to name a few. Quantitative and qualitative attributes can be represented as visual (nonverbal) cues (for example, the shape, color, and size of graph elements).

As a basic example, Figure 1 illustrates data integration and the visualization of a complex global metabolic-network response of budding yeast to filamentous-form growth conditions. On solid growth media with limiting nitrogen source, diploid budding yeast shift from the familiar yeast form to growth in filamentous chains of elongated, adhesive, invasive cells. This dimorphism is a model system for cell differentiation and fungal pathogenesis (22, 23). Data are represented in the visual attributes of graph elements. Square and triangular nodes represent molecule types, protein–enzymes, and small-molecule metabolites, respectively. The color of each square represents the filamentous-form/yeast-form mRNA expression log ratio (27); shades of red are positive (induced in the filamentous form), blues are negative. Triangle color communicates the localization of metabolite pools. Yellow is external; green is mitochondrial; white is other (mostly cytosolic). Each gray edge, occurring between a metabolite and a protein, represents a metabolic interaction, indicating that the metabolite is a substrate or product of the protein enzyme. Reaction data and metabolite pool localizations are from Forster et al. (11). The graph enables the visualization and analysis of the global organization and response of the network.

Information on cell properties can be extracted from complex networks of integrated genome-scale data sets and enables biologists to formulate testable hypotheses. The graph in Figure 1 contains subnetworks with information on specific metabolic properties of the yeast filamentous form. Figure 2 shows two examples that lead to the extraction of insight on yeast filamentous-form metabolism. Figure 2a shows the pathway of external allantoin (ATNxt) uptake and degradation to utilizable nitrogen in the form of urea and NH₃, and carbon in the form of acetyl-CoA
(ACCOA) and malate (MAL). The entire pathway is induced in filamentous-form cells. Allantoin is a product of purine catabolism. This raises the hypothesis that fungal pathogens take up and utilize degraded purines, possibly from host tissues, as a nitrogen-rich nutrient source. Figure 2b shows a subnetwork involved in ammonium uptake. Note that the three homologous ammonium permeases, Mep1, Mep2, and Mep3, show different expression responses, although they all form part of the cell’s ammonium uptake system. Mep3 is a low-affinity high-capacity transporter, whereas Mep2 and Mep1 are high-affinity low-capacity transporters, with Mep2 having the highest affinity (24). The proteins of the subnetwork do not show the same response, but rather show different expression changes that are a collectively appropriate response to the low-ammonium growth conditions of the yeast diploid filamentous form. More generally, genes involved in the same process or pathway often exert positive or negative roles in complex ways that depend on input from other processes. These considerations suggest that methods such as gene-expression clustering may not suffice to reveal many biological insights because genes involved in a biological process often show very different expression responses that nonetheless serve a collective function. The above examples, and the complexity of the global metabolic network shown in Figure 1, highlight the need for automated approaches to identify informative subnetworks. These subnetworks inspire biological insight and hypothesis generation by the biologist.

**Computational Extraction of Biological Information**

Beyond the highly intuitive integration and visualization of large amounts of disparate data, a graphical approach offers another major advantage. The formalism of the graph allows the application of algorithms to computationally extract information that is not available in individual graph elements or single data types. Information derived from such analyses can be assigned to graph elements and visualized or analyzed further. Rather than an enumeration of methods, an example using commonly available data types may illustrate the automated extraction of information present in multiple graph elements and their associated data. The active-subnetworks identification method of Ideker et al. (19) can be applied to a network of interacting genes/proteins with gene-expression ratios and significances from a set of experimental conditions. The purpose is to identify information flows, metabolic paths, and biological processes that are active in one or more specific experimental conditions. The method combines statistical scoring of subnetwork activity with a simulated annealing algorithm to find connected regions of the network that show significant changes in expression (not necessarily correlated changes) over particular subsets of conditions. This method has been demonstrated using an interaction network with gene-expression data from perturbations of the yeast galactose regulon (19). Also, it has been applied successfully to an interaction network with integrated gene-deletion-phenotype data.
to identify pathways and genes involved in the response of yeast to mutagenic agents (5).

In addition to the abundances, localizations, interactions, etc. of molecules, it is possible to explicitly model information transmission using the framework of a graph, especially in subnetworks for which data quality and coverage are high. By incorporating information-transfer functions directly into a graph model, the graph moves beyond a static existence and acquires behaviors. Biological molecules and interactions transmit information in complex ways. A molecule’s activity likely depends on interactions with other molecules. For example, a signaling molecule may positively regulate a pathway. However, this regulation may require the binding of an upstream activator to the signaling molecule and the absence of an inhibitor. In this case, the activity function of the signaling molecule is a Boolean function of the inputs it receives from two other molecules; it is active when the activator is active and the inhibitor is inactive. By assigning Boolean functions, or other types of more complex functions, to network nodes, one can generate explicit predictions of the cascading consequences of molecular perturbations. Probabilistic Boolean Networks (PBNs) (31) are examples of models with dynamical properties. PBNs incorporate the appealing determinism of Boolean logic and practical uncertainties, both in the data and in the selection of Boolean functions.

Note that graph edges can represent molecular relationships more complex than the presence of an interaction. In Bayesian genetic-network models, edges represent probabilistic dependencies of gene-expression patterns (14). Key advantages of these models are the ability to rigorously score the fit of models to data, the accommodation of hidden variables, and the ability to describe arbitrarily complex (more than pair-wise) relationships. Approaches such as those cited above can drive the exploration of poorly characterized complex networks to reach a level of detail allowing biochemical simulations. Gilman & Arkin (13) reviewed biochemical modeling in detail.

Naturally, together with computational methods and algorithms, software is proliferating. Continuing needs are the development of general-purpose network analysis and modeling platforms and standards for network model communication. Candidates are emerging. Two prominent general-purpose software platforms for biological network integration, visualization, and analysis are Osprey (7) and Cytoscape (30). Osprey simplifies data access and visualization (http://biodata.mshri.on.ca/osprey/). Cytoscape has the advantages of a highly flexible and generalized approach to data integration, network visualization, and analysis, as well as open-source development and distribution (http://www.cytoscape.org). To enable the exploitation of an increasing diversity of computational methods and implementations, and to facilitate collaboration and the communication of results, standards are needed for the transmission of network models. Cell Markup Language (CellML) (16) and Systems Biology Markup Language (SBML) (17) are evolving machine-readable eXtensible Markup Language (XML)-based standards for the communication of network models.
NETWORK STRUCTURE AND FUNCTION

Local Network Motifs and Hierarchical System Organization

The analysis of molecular network structures has been a valuable approach to the extraction of biological insight. This approach aims to identify highly nonrandom network structural patterns that reflect function and the processes that created complex networks. These processes may include the action of selection driving either structural convergence or conservation.

In both protein-DNA and protein-protein interaction networks, one can find repeated local structural units, motifs, that occur much more often than expected by random chance. Lee et al. (21) collected a global data set of transcription–factor–binding interactions with target genes in yeast. This transcription network contained several repeating local structures (Figure 3). These transcriptional regulatory motifs can be associated with regulatory functions or behaviors, for example, the positive autoregulation of Ste12. Conant & Wagner (8) investigated the evolution of transcriptional regulatory motifs in yeast and E. coli. They show that the instances of these motifs evolved independently, rather than by duplication and divergence of one or a small number of ancestral circuits. In other words, the repeated instances of each motif are the result of evolutionary convergence on the motif structure. These results support the suggestion that the collective function of a transcriptional regulatory motif can be found in the decision-making behavior encoded in the structure, per se. This evolutionary convergence contrasts the results suggesting the evolutionary conservation of motif constituents in the yeast protein-protein interaction network (38). Furthermore, specific protein-protein interaction motifs show overrepresentation of specific functional classes of yeast proteins. Apparently, in the protein-protein interaction network, the collective functions of network motifs are more closely associated with specific cellular tasks, for example, transport facilitation.

Network motifs may be common core components of molecular modules. Modules are groups of preferred molecular partners that interact to perform some collective function (15). Evidence for the existence of modules comes from various groups (4, 20, 28, 29, 33, 34). The modular organization of molecular systems may be a consequence of the advantages of modules in the evolutionary process (2). Engineered modules provide the advantage of reuse of well-tested units performing complex functions. Rather than designing completely new systems to solve new problems, the engineer can use pre-existing modules. Similarly, in changing environments, biological systems that are able to readily reconfigure themselves to adapt will be afforded an evolutionary advantage. This logic may explain the hierarchical organization of molecular systems (Figure 4). In this hierarchy, molecules interact to form modules. Modules interact to form the complex networks of a cell. We collect data at the levels of cell properties (phenotypes) and molecular measurements (for example, expression and interaction). A central challenge is to quantitatively understand cell properties in terms of the activities of many
Hierarchical organization of biomolecular systems. Molecules interact in hierarchical organizational levels of increasing complexity. Data can be collected at the molecular and phenotype levels. The organizational hierarchy can be used as a framework to discover quantitative relationships between interacting molecules and the complex properties of cells.

Module identification and abstraction can be applied to specific biological responses to gain insight into the control of complex cell properties. Prinz et al. (27) developed and applied this strategy to the molecular network controlling yeast cell differentiation to filamentous-form growth. They assembled an integrated response network from genes implicated by either significant filamentous-form/yeast-form expression change or a filamentous-growth phenotype. Then they integrated protein-protein interactions if they connected a pair of implicated proteins. Prinz et al. added metabolic interactions, and the associated metabolites, if they involved an implicated protein. Figure 5 shows the results of automated abstraction of the filamentation network. Modules were identified by clustering network nodes based on network topology. Modular units, representing clusters of interacting molecules, are circular nodes with an area proportional to the number of molecular members. Molecules falling outside modules are shown as rectangles (genes/proteins) or triangles (metabolites). Blue edges indicate protein-protein interactions. Green edges are protein-metabolite interactions, which indicates that the metabolite is a substrate or product of the protein enzyme. Expression ratios are mapped as node colors. Red nodes are induced in the filamentous form; blue nodes are repressed. Module-node color reflects the average among member genes. Module-node names are from a member node of greatest intramodule connectivity, the “module organizer” (29). Significant expression change coordination of cluster comembers and automated annotation of clusters with significantly overrepresented gene annotations in the clusters support the modular nature of these network clusters. The annotations, expression changes, and connections among molecules. The built-in organizational hierarchy of cells may provide a framework to enable this.

![Diagram of hierarchical organization of biomolecular systems](image)

**Figure 4** Hierarchical organization of biomolecular systems. Molecules interact in hierarchical organizational levels of increasing complexity. Data can be collected at the molecular and phenotype levels. The organizational hierarchy can be used as a framework to discover quantitative relationships between interacting molecules and the complex properties of cells.
modules in this network enable formulating and testing hypotheses on the control of yeast filamentous growth (27).

From an engineering perspective (9), the nature of modules facilitates hypothesis generation and testing. Modules have unique independent identities. For example, a mitogen-activated protein–kinase cascade is a signal amplifier no matter what its context. This property of modules imparts several advantages. (a) Because one can consider modules as units of biological organization, they allow network abstraction or simplification (29). Instead of relating the activities and interactions of hundreds of molecules to complex cell properties, one can study the determination of cell properties by a much smaller number of interacting modular units. (b) The principle of molecular guilt-by-association (25) applies to modules as well. In a specific biological response network, the modules to which a given module is connected provide information about the role of its collective function in determining cell properties. (c) If cell properties are associated with specific modules, those properties become associated with their component molecules. This forms a basis for molecular hypotheses. (d) Module functions can be modified somewhat independently to test hypotheses and potentially to engineer new networks.

The identification and abstraction of network motifs and modules with collective properties is advantageous in two important ways. The first is that these network units have collective functions and behaviors that isolated components do not have. Such collective properties emerge from the interactions of the components. Second, cell properties arise from the action of hundreds or thousands of molecules and interactions. Network abstraction will make the quantitative understanding of cell properties more accessible. Identifying and abstracting motifs and modules is a prelude to the analysis and modeling of highly complex networks of abstracted network units.

System Genetics

A quantitative understanding of the hierarchical organization of cells requires collecting data at different levels of the hierarchy. This includes the high-throughput collection of quantitative phenotype data, direct measurements of complex cell properties. In addition to adding phenotypes as quantitative variables, systematic or strategic network perturbation and phenotyping links cell properties to molecular network elements.

The design and execution of screens and selections to identify genetic perturbations (and their respective genes) affecting a complex biological response is a classical genetic approach. Recently, systematic methods for gene perturbation and phenotype measurement were developed and applied. Prominent examples are the yeast deletion strain set with sequence “bar codes” marking each genotype (32), and RNAi, a versatile gene knockdown technology that can be used in metazoans (10). These technologies enable one to rapidly identify genes mediating a biological response. Phenotype-based genetic approaches complement genomic expression
profiling. For example, Giaever et al. (11a) studied the response of yeast to several environmental conditions. For each condition, they applied genomic expression profiling to identify genes showing increased mRNA levels. Also, using the “bar code” deletion strains, they identified genes required for optimal growth under each condition. There was little overlap (a few percent) between the expression-implicated genes and the phenotype-implicated genes. Thus, genomic expression profiling and genetic approaches implicate different sets of genes. Phenotype-implicated genes may be more likely than expression-implicated genes to be involved in key regulatory mechanisms controlling a biological response.

Given a set of implicated genes and associated phenotypes, a genetic approach to understand the system is to observe how perturbations of those genes interact to affect phenotypes. Genetic interactions are fundamentally different from molecular interactions. A genetic interaction occurs between two genetic perturbations. Each single perturbation affects a reference phenotype (wild type). This information, plus the phenotype of the double perturbation, defines a genetic interaction. Essentially, a genetic interaction contains information about how two perturbations interact at a functional level. The functional information in genetic interaction data constrains network mechanisms, logic, and directions of information flow, and connects interacting network elements to phenotypes. Genetic-interaction types such as additive effects, suppression, and epistasis convey unique information about the functional relationships of network elements and paths. For example, epistasis analysis has been used successfully for decades to order pathway molecules. Suppressor screens reveal secondary perturbations that reverse or short-circuit the effects of the suppressed perturbation.

Boone and colleagues (36) have advanced genetic interaction studies with the systematic high-throughput detection of synthetic-lethal effects of yeast gene deletions. Synthetic growth defects often occur among functionally related genes. Also, members of the same pathway tend to show similar patterns of synthetic-lethal interaction. This approach has been highly successful. A challenge will be to generalize genetic interaction analyses to include all possible modes of genetic interaction. Genetic interactions of types other than synthetic defects occur commonly, and they are uniquely informative, as mentioned above.

Ultimately, the objective of system-level genetics is to drive the development of network models with more explicit and far-reaching predictive power. To this end, networks of integrated genetic interactions and physical interactions should be studied with a few specific goals in mind. The first is to add a functional dimension to physical interactions. For example, on its face, the observation of a protein-protein interaction lacks information about the direction of information flow and the functional consequence of this flow (activation, inhibition, etc.). This information is revealed by genetic interactions. The second purpose is to probe the complex functional interactions among physically distant pathways and modules. This will enable the building of module-level network models with genetic information. Lastly, patterns of genetic interaction can suggest the presence of unobserved physical interactions and information-flow paths.
Moving forward, system genetics research will benefit from progress in several key areas. (a) One is the development and broad application of systematic genetic-perturbation systems. Also, chemical perturbations will likely prove highly versatile and informative (12, 26). (b) Another is the generation of quantitative high-throughput phenotype assays and data sets involving single perturbations and combinations of perturbations (37). Quantitative image analyses may be instrumental in this regard (35). (c) A third area of progress is the development of generalized quantitative frameworks for the detection, visualization, and analysis of the full range of genetic interactions in high-throughput phenotype–data sets. The various modes of genetic interaction (epistasis, suppression, synthesis, etc.) convey unique insights on information processing in a molecular network. (d) A fourth area is the computational integration and study of genetic interaction data and molecular interaction data to derive network models with physical and functional dimensions. Success in these four areas will enable scientists to derive network models with greater power to make increasingly explicit predictions of the behaviors of network circuits and complex cell properties.

ACKNOWLEDGMENTS

The author thanks Susanne Prinz and Iliana Avila-Campillo for their contributions to the manuscript, and Tony Lee and Rick Young for permission to republish a figure. The author is a recipient of a Burroughs Wellcome Fund Career Award in the Biomedical Sciences.

The Annual Review of Genomics and Human Genetics is online at http://genom.annualreviews.org

LITERATURE CITED

10. Fire A, Xu S, Montgomery MK, Kostas
Figure 1  Data integration and visualization of a complex network. The graph represents the global metabolic-network response of budding yeast under filamentous-form growth conditions. See text for a detailed description.
Figure 2  Biologically informative subnetworks. The global network of Figure 1 contains subnetworks that convey information on the metabolic responses of filamentous-form yeast cells. Examples show (a) the induction of a pathway for the uptake and catabolism of purine-degradation products, and (b) the adaptation of the expression of ammonium-uptake genes to nitrogen-limited growth.
Figure 3 Motifs in the yeast protein-DNA interaction network. The global yeast protein-DNA interaction network contains several repeated structures. Blue circles indicate transcription-factor proteins; red squares indicate genes. Dashed arrows indicate the encoding of a protein by a gene. Solid arrows indicate the observed binding of a transcription factor to a gene. Reprinted with permission (21).
Figure 5: Modular abstraction of a complex molecular network controlling biological responses. Groups of interacting molecules in an integrated molecular network controlling yeast filamentous-form growth were computationally identified and abstracted as modular units (circular nodes). See text for details. Reprinted with permission (27).
CONTENTS

GENETIC TESTING IN PRIMARY CARE, Wylie Burke 1

COMPARATIVE GENOMICS, Webb Miller, Kateryna D. Makova, Anton Nekrutenko, and Ross C. Hardison 15

GENETIC SCREENING: CARRIERS AND AFFECTED INDIVIDUALS, Linda L. McCabe and Edward R.B. McCabe 57

NUTRITIONAL GENOMICS, Jose M. Ordovas and Dolores Corella 71

AFRICANS AND ASIANS ABROAD: GENETIC DIVERSITY IN EUROPE, Guido Barbujani and David B. Goldstein 119

FINDING PROSTATE CANCER SUSCEPTIBILITY GENES, Elaine A. Ostrander, Kyriacos Markianos, and Janet L. Stanford 151

MOLECULAR NETWORKS IN MODEL SYSTEMS, Timothy Galitski 177

GENETICS OF ATHEROSCLEROSIS, Aldons J. Lusis, Rebecca Mar, and Päivi Pajukanta 189

MEDICAL GENETICS IN DEVELOPING COUNTRIES, Arnold Christianson and Bernadette Modell 219

PROTEOMICS, Carmen L. de Hoog and Matthias Mann 267

POPULATION GENETICS, HISTORY, AND HEALTH PATTERNS IN NATIVE AMERICANS, Connie J. Mulligan, Keith Hunley, Suzanne Cole, and Jeffrey C. Long 295

VARIATION IN HUMAN MEIOTIC RECOMBINATION, Audrey Lynn, Terry Ashley, and Terry Hassold 317

COMPARATIVE PRIMATE GENOMICS, Wolfgang Enard and Svante Pääbo 351


MAMMALIAN CIRCADIAN BIOLOGY: ELUCIDATING GENOME-WIDE LEVELS OF TEMPORAL ORGANIZATION, Phillip L. Lowrey and Joseph S. Takahashi 407

PLANT GENOMICS: THE THIRD WAVE, Justin O. Borevitz and Joseph R. Ecker 443

EPIGENETICS AND HUMAN DISEASE, Yong-hui Jiang, Jan Bressler, and Arthur L. Beaudet 479
CONTENTS

INDEXES

Subject Index 511
Cumulative Index of Contributing Authors, Volumes 1–5 537
Cumulative Index of Chapter Titles, Volumes 1–5 539

ERRATA

An online log of corrections to Annual Review of Genomics and Human Genetics chapters (if any) may be found at http://genom.annualreviews.org/