

Chapter 1

An Overview of Systems Biology

Réka Albert^{1,2}, Sarah M. Assmann³

¹Department of Physics, The Pennsylvania State University, University Park, PA 16802

² Huck Institutes of the Life Sciences, The Pennsylvania State University, University Park, PA 16802

³ Department of Biology, The Pennsylvania State University, University Park, PA 16802

Keywords: systems biology, molecular networks, interactome, network inference, graph analysis, dynamic modeling

Abstract

This chapter provides an overview of three crucial aspects of systems biology: constructing biological networks, analyzing and modeling the structure of biological networks, and modeling the dynamics of biological networks. We describe the types of intracellular networks most often studied, and the “omic” information available to synthesize these networks, with a special focus on plant biology. We review the computational methods used to construct or infer (reverse engineer) intracellular networks. We present the graph theoretical measures most useful for understanding the organization of biological networks, from the single node level to the global properties of the whole network. A representative sample of biological network models is provided, ranging from static models to dynamic models that incorporate how the status of the nodes changes in time. Throughout the chapter we focus on the biological predictions possible by combining experimental, theoretical and computational methods.

1.1. Systems Theory and Biology

It is increasingly recognized that in order to understand the dynamics and function of a cell, as well as higher levels of biological organization, we need to know: 1) the components that constitute it; 2) the relations and interactions of these components, and 3) their dynamic behavior, i.e. how the biological entities behave over time under various conditions¹. Ultimately, this information can be combined in a model that is not only consistent with current knowledge, but provides new insights and predictions. These topics and their integration is the purview of the field of systems biology.

The origins of systems biology can be traced back to the beginning of the twentieth century to the work of Aleksandr Bogdanov^{2,3}, Ludwig von Bertalanffy⁴ and others. Systems thinking is used in a variety of scientific and technological fields, and concepts such as independent and dependent variables, feedback, modularity, robustness, sensitivity, and control have been extensively studied in the fields of systems theory^{5,6} and control theory⁷. Systems theory is a line of inquiry based on the assumptions that (i) all phenomena can be viewed as a web of relationships among elements (that is, a system) and (ii) all systems, whether technological or biological, have common properties and behaviors and can be handled by a common set of methods. Control theory is extensively used in designing engineered systems, and it seeks to identify and modulate the mechanisms that systematically control the state of a system to minimize malfunctions and deviations from the optimal dynamic behavior.

In the context of biology, Biochemical Systems Theory⁸ and Metabolic Control Theory⁹, developed between 1965-75, proposed general mathematical models of biological systems at and around a steady state (equilibrium). However, these and other attempts at systems level understanding of biology suffered from inadequate data on which to base their theories and models. It was the advent of genomic technologies that brought an abundance of data on system elements, interactions and states, and enabled the integration of knowledge across different levels (molecular, cellular, tissue, organ, etc.) of biological organization.

1.2. Graph Elements and Network Attributes

Every biological system includes a network of interactions, and, according to our definition, all systems biology includes network analysis. In some cases, the structure of the network and its interpretation are straightforward (e.g. a linear chain of interactions), while in other cases, more detailed analysis is required to understand how information can propagate through the network. Therefore, this chapter begins with some definitions of commonly used terms and measures in network analysis and graph theory. Many of these terms are illustrated in the hypothetical network depicted in Figure 1.1.

A **network** (or graph)² is used to represent a system of elements that interact. A graph has two basic parts: the elements of the system are portrayed as **graph nodes** (also called vertices) and the interactions are portrayed as **edges**, i.e. lines connecting pairs of nodes. Multi-partite graphs contain different classes of node (such as mRNA and protein). Directed edges emanate from a source (starting node) to a sink (ending node) and represent unidirectional flow of material or information. Non-directed edges are used to represent mutual interactions, e.g. physical interaction between two proteins of unknown function, or interactions where the directional flow of information is not known. Signs representing activation or inhibition, or weights indicating confidence levels or strength can be imposed on edges to enhance the information content of the network.

The organizational features of interaction graphs can be quantified by network measures denoting the importance (centrality) of individual nodes, the connectivity (reachability) among nodes, and the homogeneity or heterogeneity of the network in terms of a given property. Three of the most often used network measures, in an increasing order of locality, are the node degree, the clustering coefficient and the path length.

The **degree** of a node is the number of edges pointing toward or emanating from that node (Figure 1.1). A node's total degree is the sum of its **in-degree** and **out-degree**, which respectively quantify the number of incoming and outgoing edges of the node. In a weighted

graph, one can also define a **node strength**, the sum of the weights of the edges into and out from the node. The local information on the degree of each node can be combined to yield a global description of the network known as a **degree distribution**, $P(k)$, which gives the fraction of nodes with degree k . In directed networks one can similarly define in- and out-degree distributions. A large number of cellular networks have been shown to be **scale-free**, meaning that there are many different node degrees, such that one cannot validly describe the network in terms of a “typical” node degree (reviewed in ¹⁰). Scale-free networks are characterized by a degree distribution that is close to a power-law: $P(k) \cong Ak^{-\gamma}$, where A is a normalization constant, and where the degree exponent is typically $2 < \gamma < 3$ ¹⁰.

The first **neighborhood** of a node consists of the nodes connected to it by a single edge, and the edges among those nodes (if any). If this neighborhood is a completely connected subgraph, it is known as a **clique**¹¹. The **clustering coefficient** of a node is the ratio of the number of edges among the first neighbors of the node and the number of edges among them if the node’s first neighborhood were a clique. Thus a clique has a clustering coefficient of 1; conversely, when there are no edges among the first neighbors, the clustering coefficient is zero. Large average clustering coefficients have been observed for protein-protein interaction networks¹² and metabolic networks¹³ indicating topological redundancy and biological cohesiveness.

Two nodes of a graph are **connected** if they are linked by a sequence of adjacent nodes and edges, a **path**¹⁴. For example, a path could signify a biosynthetic pathway, or it could represent a cascade of events in a signal transduction chain. The **distance (path length)** between any two nodes in a network is the number of edges in the shortest path connecting those nodes (Figure 1.1). In a **weighted network**, the distance between two nodes is the sum of edge weights along the path for which this sum is a minimum¹⁵. Many networks’ average path length scales with the natural logarithm of the number of nodes, $d \sim \ln(N)$, indicating that path lengths of even very large networks remain small. This **small world**¹¹ property has been observed for metabolic, protein interaction, and signal

transduction networks^{12,16,17} and facilitates rapid spread of information in response to inputs. Another important global property related to paths is **path redundancy**, or the availability of multiple paths between a pair of nodes¹⁸. Either by allowing multiple channels of information from input to output or as alternate routes when the preferred pathway is disrupted, path redundancy promotes the robust functioning of cellular networks by reducing reliance on individual pathways.

Networks having paths between every pair of nodes are **connected**. A directed network can be **strongly connected** if all of its node pairs are connected in both directions; alternatively the network can have one or several strongly-connected subgraphs. Each strongly-connected subgraph is associated with an **in-component** (nodes that can reach the strongly-connected subgraph, but that cannot be reached from it) and an **out-component** (the converse) (Figure 1.1). The nodes of each subgraph may share a specific task within a given network. In signal transduction networks, for example, the nodes of the in-component tend to be involved in ligand-sensing; the nodes of the strongly-connected subgraph form a central signaling subnetwork; and the nodes of the out-component are responsible for the transcription of target genes, or for phenotypic outcomes^{17,19}.

The number, directionality, and strength of connections associated with a given node can be synthesized into measures of that node's **centrality**. The sources (nodes with only outgoing edges) and sinks (nodes with only incoming edges) of a directed network represent initial and terminal points of the flow of material or information. In a metabolic network describing a biosynthetic pathway, for example, the initial precursor is the source, and the final product is the sink. For nodes other than sources and sinks, the **betweenness centrality** -- the number of shortest paths from node s to node t passing through the node, divided by the total number of shortest st -paths (Figure 1.1) -- indicates the importance of that node to the flow of information or materials through the network^{20,21}. Betweenness centrality is often, but not obligately, correlated with degree. For example, in metabolic networks of microorganisms, the most ubiquitous substrates tend to have the highest

betweenness centralities but not the highest degrees²², and some low-degree metabolites are as critical to the overall network function as high-degree metabolites²³.

In addition to the general graph concepts and measures used to quantify the organization of biological networks, a number of specific terms are often invoked to reflect their functional constraints.

Hubs: In scale-free networks, small-degree nodes are most common, and the node degrees are highly heterogeneous such that the highest-degree nodes have degrees that are orders of magnitude higher than the average degree. Such highest-degree nodes are commonly referred to as **hubs**, although there is no explicit definition of the boundary between hubs and non-hubs. Consequently in scale-free networks random node disruptions do not lead to a major loss of connectivity, whereas the loss of the hubs causes the breakdown of the network into isolated clusters¹⁰. This point has been experimentally verified in *S. cerevisiae*, where the severity of a gene knockout has been shown to correlate with the number of interactions in which the gene's products participate^{24,25}. Indeed, as much as 73% of the *S. cerevisiae* genes are non-essential, i.e. the knockout has no phenotypic effects²⁶, and this confirms cellular networks' robustness in the face of random disruptions. Conversely, the likelihood that a gene is essential (that is, that its knockout is lethal) is greater for high-degree nodes^{24,25}. This confirms the intuitive prediction that cell viability is more likely to be compromised by loss of highly interactive nodes such as hubs.

Modularity: Cellular networks are expected to be decomposable to subnetworks (pathways) corresponding to specific biological functions²⁷. These subnetworks or modules should be distinguishable within interaction networks by the fact that they have dense intra-module connectivity but sparse inter-module connectivity. Methods to identify functional modules are based on the physical location or function of network components²⁸, the topology of the interaction network^{29,30}, or the evolutionary conservation of the nodes^{31,32}. The demonstrated overlap and cross-talk between pathways³³ presents a challenge to module-detecting

algorithms, as is the observation of hierarchical modularity, in which modules are made up of smaller and more cohesive modules³⁴.

Motifs: Cellular networks contain recurring interaction motifs, which are small subgraphs that have well-defined topologies. Interaction motifs such as autoregulation (usually a negative feedback) and feed-forward loops have a higher abundance in transcriptional regulatory networks than expected based on the degree distribution alone^{35,36}. In general, protein interaction motifs such as small cliques are both abundant³⁷ and evolutionarily-conserved³⁸, partly because many of them represent subgraphs of protein complexes. Feed-forward loops, positive and negative feedback loops and triangles of scaffolding (protein) interactions are over-represented in signal transduction networks¹⁷. Interaction motifs are proposed to form functionally-separable building blocks of cellular networks³⁹. For example, the abundance of negative feedback loops in the early steps of signal transduction networks and of positive feedback loops at later steps suggest mechanisms to filter weak or short-lived signals and to amplify strong and persistent signals¹⁷.

1.3. Building Biological Networks: Identifying Nodes and Mapping Interactions

Genome-level information concerning cellular networks is often described using five 'omes': genome, transcriptome, proteome, metabolome and interactome. During the last decade, the respective omics approaches have produced an incredible quantity of expression and interaction data, providing extensive, albeit still incomplete, knowledge regarding the nodes and edges of biological networks^{40,41}.

1.3.1. Identifying Nodes

Transcriptome data convey the identity of each expressed gene and its level of expression for a given cell type, tissue, organ, or organism. High-throughput mRNA data are obtained by serial analysis of gene expression (SAGE)⁴², DNA microarrays^{43,44}, massively parallel signature sequencing (MPSS)^{45,46} and 454 sequencing^{47,48}. For model plant species such as

Arabidopsis, much of the microarray information has been compiled in databases such as AtGenExpress at The Arabidopsis Information Resource (TAIR) and in NASCArrays at the European Arabidopsis Stock Centre (NASC), as well as at the Gene Expression Omnibus (GEO) site at NCBI (See Table 1.1 for a summary of the websites mentioned in this chapter). These databases provide genome-wide information on transcript levels in individual tissues or organs, or under specific stress or developmental conditions, and thus provide information on which transcripts are co-expressed and, potentially co-regulated. Genevestigator and the Botany Array Resource (BAR) both provide pictorial summaries of spatial patterns of gene expression.

Information about transcription factor binding motifs is available from the Transcription Factor Database (TRANSFAC)⁴⁹, the Regulon Database (RegulonDB)⁵⁰, and the Kyoto Encyclopedia of Genes and Genomes (KEGG). To identify novel motifs, one can download promoter sequences from TAIR and evaluate them using motif-finding algorithms such as AlignACE⁵¹, MEME⁵², MotifSampler⁵³, and Weeder⁵⁴, the latter two of which provide the option of using background models derived from Arabidopsis-specific datasets.

Proteome data provide information on global protein expression patterns in the sample of interest. The multitude of possible post-translational modifications dictates that the complexity of the proteome is much greater than that of the transcriptome, and proteomics methods are still rapidly evolving⁵⁵. In addition, currently there are fewer proteomics studies than transcriptomics studies⁵⁶. For these reasons, proteomics databases based on actual protein data (as opposed to predictions from genomic sequence) are far from complete, and this is especially true for multicellular organisms. Plant proteomics databases are available at NASC, at the Arabidopsis page in the ExPASy (**Expert Protein Analysis System**) proteomics server of the Swiss Institute of Bioinformatics, and at GABIPD, although they are currently sparsely populated. The Subcellular Proteomic database (SUBA) hosts information on subcellular localization of plant proteins, based on GFP tagging and proteomics methods. Other, more specific databases, are described in ⁵⁷.

Metabolome data. Metabolome analysis includes the study of both metabolites and the enzymes that catalyze their production, and methods for metabolome analysis include gas chromatography mass spectrometry (GC-MS), liquid chromatography mass spectrometry (LC-MS), and nuclear magnetic resonance (NMR)⁵⁸. Information on metabolic pathways is available at KEGG and MetaCyc. It has been estimated that plants produce between 100,000 and 200,000 secondary metabolites⁵⁹. Primary and secondary metabolites have highly diverse functions, e.g. as enzyme cofactors, hormones, and signaling agents. Although only a fraction of all the plant metabolites has been characterized, many of the known secondary metabolites have important functions in human biology, serving as medicines, flavorings, perfumes, colorants, etc. A metabolic database for Arabidopsis is available at Aracyc in TAIR and a database for the secondary metabolites of tomato fruits, MoTo DB, was recently developed⁶⁰. Metabolite databases with an emphasis on primary metabolites are available at the Golm Metabolome Database and at <http://fiehnlab.ucdavis.edu/projects/>.

1.3.2. Mapping Interactions

Advancements in molecular biology techniques have increased the resolution, type, and scale of interactions that can be detected. These improvements promise to change the focus of biology from an understanding of local, binary interactions to an understanding of the integration of these interactions into a functional system. At the genome level, transcription factors and chromatin structure promote or inhibit gene transcription. Since transcription factors and chromatin modulating agents are themselves products of genes, the ultimate effect is that genes regulate each other's expression as part of gene regulatory networks. Proteins participate in diverse post-translational processes as well as binding to form protein scaffolds, modulate subcellular localization or alter enzyme activity; the totality of these processes is called a protein-protein interaction network. The biochemical reactions of cellular metabolism can likewise be integrated into a metabolic network whose fluxes are

regulated by enzymes catalyzing the metabolic reactions. Often these different interactions types are intertwined, as occurs when an external signal triggers a cascade of interactions that involves biochemical reactions and metabolite fluxes, as well as transcriptional regulation.

Transcriptional regulatory maps link two types of nodes -- transcription factors and mRNAs --, and have two types of directed edges, corresponding to transcriptional regulation and translation⁶¹, where the regulatory edges can have two types of signs, corresponding to activation or repression. Transcriptional regulatory maps exist for *E. coli*³⁵ and *S. cerevisiae*⁶¹⁻⁶³. A given transcription factor usually regulates multiple genes, and this is reflected in the approximately scale-free out-degree distribution of these maps. Unidirectional regulation is prevalent, thus these networks do not have large strongly connected components, in contrast to protein interaction maps (see below).

In Arabidopsis, a transcriptional regulatory map has been created for cold signaling mediated by the ICE1 transcription factor⁶⁴. The source – sink distances are small (4-5 edges) in this network as well as in the *E. coli* and yeast networks, suggesting that this may be a common feature of transcriptional regulatory maps.

In **protein interaction graphs**, the nodes are proteins, and two proteins are connected by a directed edge if the direction of information flow during their interaction is known. Two proteins are connected by a non-directed edge if there is strong evidence of their physical interaction or association without, however, evidence for a directionality of interaction. Non-directional protein interaction networks are commonly generated by assays of interaction in yeast-based systems (yeast two-hybrid and split ubiquitin assays), assessment of co-immunoprecipitation, and mass spectrometry-based identification of the composition of protein complexes. There are many protein interactome databases, including the Database of Interacting Proteins (DIP)⁶⁵, yeast and mammalian protein interaction databases at the Munich Information Center for Protein Sequences (MIPS)⁶⁶ and the Human Protein Reference Database (HPRD)⁶⁷. Plant-specific global interaction databases derived from wet bench data were not available at the time of this writing. However, the Search Tool

for Interacting Proteins (STRING) will, upon user query, output predicted interactions based on orthology, as well as known interactions.

Protein-protein interaction maps have been constructed for a variety of prokaryotes and eukaryotes^{37,68-72}. Although these maps are incomplete, and some assays for protein interaction have a high rate of false positives⁷³, extant maps nevertheless are revealing common attributes, including an approximately scale free degree distribution^{12,24,37} and a large connected subgraph with short distances^{12,37}. This latter finding suggests why pleiotropy is commonly observed, since perturbations of a single gene or protein can propagate through the network, and have seemingly unrelated effects.

In plants, interaction maps have been experimentally defined for homo- and heterodimerization within two large classes of transcription factors: the MADS box transcription factors^{74,75} and the MYB transcription factor family⁷⁶. In addition, based on known protein-protein interactions in other species, interaction of the homologous proteins in Arabidopsis has been predicted on a global scale⁷⁷. A database of these “interologs” is available at www.interolog.gersteinlab.org.

Metabolic networks link two types of nodes, metabolites and enzymes, by edges that represent enzyme-catalyzed chemical reactions. An idealized metabolic network and its simplified representations as a metabolite graph or a reaction graph¹³ are given in Figure 1.2. Because some nodes of metabolic networks, e.g. ATP, participate in a multitude of reactions while other nodes, e.g. most enzymes, participate only in one specific reaction⁷, there is a range of node degrees in metabolic networks⁷⁸, resulting in broad-tailed or scale-free degree distributions^{13,16,79}. The path length of known metabolic networks is small^{13,16} although its value depends on the network representation used⁸⁰.

In plants, metabolite profiling has revealed characteristic metabolomes of different Arabidopsis ecotypes⁸¹ and has illustrated that metabolite signatures can be used in a predictive fashion to identify recombinant inbred Arabidopsis lines that accumulate biomass more rapidly⁸². Another major emphasis of plant metabolomics currently is in the use of combined metabolomic and transcriptomic datasets^{55,83}. Maps have been created for e.g.

carbon nitrogen regulatory networks⁸⁴ and responses to sulfur and nitrogen deficiency⁸⁵. A combination of metabolomics and transcriptomics has also been used to predict functions of previously uncharacterized genes involved in metabolic processes, including sulfur metabolism⁸⁶, anthocyanin synthesis⁸⁷ and volatile formation⁸⁸⁻⁹⁰.

Signal transduction networks connect signals to effectors with (typically) directed edges, and can encompass all the node types discussed above: RNAs, proteins, and metabolites. The largest reconstructed signal transduction network to date, for a brain neuronal cell type, consists of 545 nodes and 1259 interactions¹⁷, yet the average source-to-sink path length is only four, suggesting a rapid response capability, which is also supported by the observation that 60% of the nodes are strongly-connected. By contrast, the largest reconstructed cellular signal transduction network to date for plants, for stomatal closure induced by the plant hormone abscisic acid, has only ~40 nodes⁹¹, indicating that plant biologists are still amassing the types of information required to build comprehensive cellular network models.

Finally, information on gene co-expression⁹², gene co-occurrence⁹³, or genetic interactions⁹⁴ can be used to construct **networks of gene functional relationships**. For example, such analyses in Arabidopsis have yielded information on co-regulation of genes involved in cell wall biosynthesis⁹⁵ and primary and secondary metabolism (see above references). Genetic interactions and functional relationships are often complementary and only sometimes overlapping with physical interactions⁹⁴. Functional relationships and associations between genes are often inferred from gene expression information, and methods of inference are briefly reviewed in the next section.

1.4. Building Biological Networks: Computational Methods for Network Inference

Computational inference (also referred to as reverse engineering) is an approach to infer causal relationships within and between the transcriptome, proteome, and metabolome when direct experimental determination of these causal relationships has not been performed.

Data-mining schemes typically extract relationships between two entities based on their statistical co-occurrence, for example, their shared inclusion in journal articles^{96,97}. Algorithms of this nature have been used extensively to infer protein-protein interactions based on their genes' co-occurrence in the same chromosomal neighborhood⁹⁸, shared evolutionary pattern⁹⁹ or co-expression. Search tools such as the Search Tool for the Retrieval of Interacting Proteins (STRING) employ data-mining methods for the inference of protein-protein interactions in eukaryotes and prokaryotes^{98,100}.

Inference of functional relationships among gene products based on their mRNA, protein or metabolite expression profiles frequently invokes **statistical methods** such as clustering^{101,102} and Bayesian networks^{103,104}. **Clustering** aims to find groups of genes that respond in a similar manner to varying conditions, and that might therefore be co-regulated^{102,105-107}. In clustering algorithms, genes or proteins with statistically similar expression profiles¹⁰² are grouped using hierarchical clustering algorithms¹⁰⁸, self-organizing maps¹⁰⁹, K-mean clustering¹¹⁰ or topological measures of the inferred networks^{107,111}. Because of the strong evidence of correlations between co-expression of a pair of genes and interactions among the pair of proteins encoded by these two genes^{102,112}, clustering methods have also been used to augment protein-protein interaction networks¹¹². **Bayesian networks** aim to infer a directed, acyclic graph that summarizes the dependency relationships among variables in the system, and a set of local joint probability distributions that statistically convey these relationships¹¹³. The initial links of the dependency graph are established either randomly or based on prior knowledge, and the network is refined by an iterative search-and-score algorithm in which multiple candidate networks are scored against experimental observations and against one another¹¹⁴. Bayesian networks have been employed to sort yeast proteins into functional groupings¹⁰³ and to infer protein interactions in the *S. cerevisiae* cell cycle¹¹³.

Several methods proposed for the **inference of gene-regulatory networks** from time-course gene expression data seek to relate the change in the expression level of a given gene with the levels of other genes' transcripts in the network by describing it as a

differential equation^{115,116} or a discrete relationship^{117,118}. The resulting system of equations is typically underdetermined because there are more unknowns than experimental time points, for this reason additional assumptions (e.g. maximum parsimony) are also invoked¹¹⁶. Deterministic methods based on systems of linear differential equations have, for example, been used to infer gene-regulatory networks in *B. subtilis*¹¹⁶ and in the rat central nervous system¹¹⁵. Deterministic Boolean methods approximate gene expression levels with binary variables (e.g. by using a threshold), and describe gene regulation with logical functions (using the Boolean operators “and”, “or” and “not”)^{117,119}. Each gene’s logical function is found by a systematic search for the minimum set of regulator nodes whose combined expressions explain the experimentally observed state of the gene. Probabilistic Boolean methods^{117,120} incorporate uncertainty and fluctuations in expression levels by assigning several alternative logical functions to each gene^{117,120}; a machine learning algorithm^{120,121} then selects the most probable logical function at each time point. This method was used to infer the regulatory networks involved in embryonic segmentation and muscle development in *D. melanogaster*¹²².

Metabolic pathway reconstruction from known stoichiometric information is usually performed by constraint-based deterministic methods¹²³ such as Flux Balance Analysis^{124,125} or S-systems¹²⁶. Recently, a linear optimization strategy that first selects a subset of a predetermined set of possible metabolic reactions, and then optimizes the metabolic flux distribution, was proposed¹²⁷ and used to find the changes in an *E. coli* genome-scale metabolic model that are needed to minimize the discrepancy between model predictions and experimentally-measured flux data¹²⁸.

The majority of network inference methods presented in this section use node-level (expression) information to infer causal relationships. There also exists a complementary approach of **reverse inference**: inferring interactions from indirect causal relationships. Indeed, experimental information about the involvement of a protein in a process is often restricted to evidence of differential responses to a stimulus in the wild-type organism versus an organism in which the respective protein’s expression or activity is disrupted. These

observations can be incorporated as two intersecting paths (denoting the stimulus- response and protein – response indirect relationships) in an incompletely mapped interaction network. The inference algorithm must integrate indirect and direct evidence to find a network consistent with all experimental observations⁹¹. This inference approach is incorporated in the software NET-SYNTHESIS^{129,130}, and we expect it will play an increasing role when integrating information from disparate data sources.

1.5. Biological Network Models: Data Integration

Once a network has been derived or inferred, it is encapsulated in a graph. The graph measures and analysis techniques described in section 1.2 can then be used to quantify the global connectivity (reachability) among nodes, the densely interconnected clusters (modules), and the importance (centrality) of individual nodes. These graph measures, alone or combined with additional information regarding the network nodes (such as the functional annotation of the corresponding genes/proteins), provide testable biological predictions on several scales, from single interactions to functional modules.

The predictive power of biological network reconstructions can be substantially enhanced by **integrating** several types of interactions and functional associations. Composite networks superpose protein-protein and protein-DNA interactions¹³¹, protein-protein interactions, genetic interaction, transcriptional regulation, sequence homology, and expression correlation¹³² or metabolic reactions and transcriptional regulation of metabolic genes^{133,134}. Alternatively, gene-gene linkages can be defined as probabilistic summaries of physical and functional associations such as protein interaction, mRNA coexpression, synthetic lethal interactions, and comparative genomics¹³⁵.

The **functions of unannotated proteins** can be inferred on the basis of the annotation of their first neighbors in the protein interaction network^{135,136}. **New protein interactions** can be predicted based on the presence of conserved interaction motifs within the network¹³⁷. New protein functions and interactions can be inferred through global alignment between protein interaction networks in different species¹³⁸. Conversely, the relatively well mapped

protein interaction networks of *S. cerevisiae* and *Drosophila melanogaster* allowed the determination of the closest pair of functionally orthologous proteins, one in *S. cerevisiae* and one in *Drosophila*, by modeling the orthology relation as a probabilistic function of the orthology relations of the immediate network neighbors of each member of the pair¹³⁹.

The *S. cerevisiae* multi-networks allowed the identification of **regulatory themes** supported by several data types such as two interacting transcription factors regulating the same target gene or one transcription factor regulating several genes whose protein products are members of the same protein complex^{131,132}. There is great interest in identifying high confidence (multi)network subgraphs corresponding to components working toward a particular cellular function or within a common **pathway**. Frequently encountered subgraphs include linear or branching paths of interaction (suggestive of pathways), densely connected clusters (suggestive of functional protein complexes), and parallel clusters in which the proteins in one cluster are associated with the proteins in the other cluster by orthology or genetic interactions¹⁴⁰. The connected subgraphs of the probabilistic gene-gene linkage network have been used to identify highly connected **gene clusters** (modules). The highly coherent functional annotation of genes within each cluster allowed the annotation of unknown proteins that are part of a cluster¹³⁵.

Finally, the integrated transcriptional and metabolic network allowed **global predictions** of growth phenotypes and qualitative gene expression changes in *E. coli*¹⁴¹ and yeast¹³⁴. The creative (modeling) aspects of the definition of these (multi)networks and the variety of predictions enabled by them demonstrate that they are not simply compilations of data but **qualitative biological network models**. A recently developed database, CellCircuits, offers a repository of network models spanning yeast, worm, fly, *Plasmodium falciparum*, and human, and four types of interaction¹⁴⁰. It was proposed that qualitative network models may develop into something akin to a Biological Information System, incorporating components, interactions, and causal relationships described by a small group of verbs such as “promote”, “inhibit”, “bind” etc.¹⁴².

1.6. Biological Network Models: From Network Structure to Dynamics

The nodes of cellular interaction networks represent populations of proteins or other molecules. The abundances of these populations can range from a few copies of an mRNA to hundreds or thousands of molecules per cell, and they vary in time and in response to external or internal stimuli. To capture these changes, the interaction network needs to be augmented by quantitative variables indicating the expression, concentration, or activity - in short, the state - of each node, and by a set of equations indicating how the state of each node changes in response to changes in the state of its regulators.

Dynamic network models have as input information (i) the interaction network summarizing the regulatory relationships among components, (ii) the equations indicating how the state of a node depends on the state of its regulators, and (iii) the initial state of each component in the system. The model's output is the time evolution of the state of the system, for example the system's response to the presence or absence of a given signal. A dynamic model that correctly captures experimentally observed normal behavior allows researchers to track the changes in the system's behavior due to perturbations. It is easier to use a model to search for perturbations that have a significant effect on system behavior than it is to perform comparable experiments on the living system; for example, models can predict multiple small perturbations that produce large effects when combined. In the following we briefly outline the main mathematical frameworks for modeling the dynamics of cellular networks, and give examples of their applications.

Continuous and deterministic models characterize node states by concentrations and describe the rate of production or decay of all components by differential equations based on mass-action (or more general) kinetics¹²⁶. With sufficiently thorough knowledge of the elementary biochemical reactions and fluxes comprising a system and the associated reaction rates, it is possible to accurately reproduce the system's dynamics and to explore the effect of perturbations. For example, a differential equation-based model of an 11-node signaling network responsible for programmed cell death after infection of *Arabidopsis thaliana* with *Pseudomonas syringae* led to significant refinement of the signaling circuitry

(by discounting two previously proposed negative feedback loops) and of the kinetic parameters¹⁴³. The **stochasticity** (non-determinism) of biological processes is usually taken into account by appending stochastic (noise) terms to differential equations. **Discrete events** (such as the initiation of transcription) and low abundances for certain molecules are incorporated by characterizing the node states by the copy number of each molecule and describing the time evolution of the probabilities of each of a system's possible states¹⁴⁴⁻¹⁴⁶. A recent model of the ethylene signaling pathway and its gene response in *Arabidopsis thaliana* combines mass-action kinetics for signaling proteins with a probabilistic description of the target genes' states¹⁴⁷. This model reproduces the experimentally observed differential responses to different ethylene concentrations and predicts that the pathway filters rapid stochastic fluctuations in ethylene availability.

Boolean models characterize network nodes by one of two binary states corresponding to, e.g., an expressed or not expressed gene, open or closed channel, or above-threshold or below-threshold concentration of a molecule. The change in state of each regulated node is usually described by a logical function having as inputs the state of the node's regulators. Boolean models predict dynamic trends in the absence of detailed kinetic parameters; for example, such models have been used successfully to describe wild type and mutant behavior in the development of floral organs¹⁴⁸ and the process of abscisic acid-induced stomatal closure⁹¹ in *Arabidopsis thaliana*. **Hybrid dynamic models** meld a Boolean description of combinatorial regulation with continuous synthesis and decay by describing each node with both a continuous variable (akin to a concentration) and a Boolean variable (akin to activity)¹⁴⁹⁻¹⁵¹. For example, a hybrid model of the transcriptional regulation of the Endo16 sea urchin gene revealed that its spatial control during embryonic development is mediated by a cis-regulatory switch¹⁴⁹.

1.7. Perspectives

Systems biology develops through an ongoing dialogue and feedback among experimental, computational and theoretical approaches. High-throughput experiments reveal, or allow the

inference of, the edges of global interaction networks. Graph-theoretical analysis of these networks enables general insight into the topological and functional organization of cellular regulation. Comparative network analysis feeds back to network inference^{107,111,137,152}, and expands the tools of graph theory to incorporate the diversity of molecular interactions. While genome-level interaction maps help us in understanding regulatory design features, dynamic modeling of systems with a less than genome-wide scope and specified inputs and outputs allows the identification of key regulatory components or parameters.

Biochemical reactions within and between cells take place on timescales spanning several orders of magnitude¹⁵³, and these timescales are modulated by the spatial aspects of the interactions, the types of biomolecules or complexes that are interacting, and the environmental conditions to which the system is subjected^{33,36}. Although the use of extensive dynamic modeling is limited by the incomplete availability of detailed transfer functions and kinetic parameters, emergent qualitative and hybrid modeling techniques that map the propagation of signals through a network^{17,91} give hope that even when exhaustive knowledge of parameters is unreachable, predictive modeling of biological processes will still be possible. Similarly augmenting the currently available directionless interactome networks with information regarding the sources (signals) and outputs of the network and the cause-and-effect (directional) relationships along the edges will significantly enhance their functional information content.

In addition to the dynamic changes in the state of network nodes, the topology of biological networks itself is shaped by dynamic events whose impact occurs on multigenerational (e.g. imprinting) and evolutionary (e.g. gene duplications and point mutations), timescales. Integration of epigenetic and evolutionary aspects with transcriptional, metabolic, and signal transduction networks represents the “final frontier” of systems biology.

It is arguably of primary importance to the systems biologist to discern whether, and how, the graph properties of biological networks reflect their functional and evolutionary constraints. However, the general architectural features of many biological networks

described so far have been found to be shared to a large degree by other complex systems, ranging from technological networks to social networks. This universality is intriguing and should make systems biology of great interest to other fields, ranging from engineering to sociology.

References

1. Kitano, H. Systems biology: a brief overview. *Science* **295**, 1662-4 (2002).
2. Bogdanov, A. *Tektologiya: Vseobshchaya Organizatsionnaya Nauka*, (Berlin and Petrograd-Moscow, 1922).
3. Bogdanov, A. *Essays in Tektology: The General Science of Organization*, (Intersystems Publications, Seaside, CA, 1980).
4. von Bertalanffy, L. *General System Theory: Foundations, Development, Applications*, (George Braziller, New York, 1968).
5. Francois, C. Systemics and cybernetics in a historical perspective. *Systems Research and Behavioral Science* **16**, 203-219 (1999).
6. Weinberg, G. *An Introduction to General Systems Thinking*, (Wiley-Interscience, 1975).
7. Franklin, G., Powell, J. & Emami-Naeimi, A. *Feedback Control of Dynamic Systems*, (Prentice Hall, New Jersey, 2002).
8. Voit, E.O. *Computational Analysis of Biochemical Systems*, (Cambridge University Press, Cambridge, 2000).
9. Heinrich, R. & Schuster, S. *The regulation of cellular systems*, (Chapman & Hall, New York, 1996).
10. Albert, R. & Barabási, A.L. Statistical mechanics of complex networks. *Reviews of Modern Physics* **74**, 47-97 (2002).
11. Watts, D. & Strogatz, S.H. Collective dynamics of 'small-world' networks. *Nature* **393**, 440-442 (1998).
12. Yook, S.H., Oltvai, Z.N. & Barabási, A.L. Functional and topological characterization of protein interaction networks. *Proteomics* **4**, 928-42 (2004).

13. Wagner, A. & Fell, D.A. The small world inside large metabolic networks. *Proceedings of the Royal Society of London Series B-Biological Sciences* **268**, 1803-1810 (2001).
14. Bollobás, B. *Graph theory: an introductory course*, (Springer-Verlag, New York, 1979).
15. Dijkstra, E.W. A note on two problems in connection with graphs. *Numerische Math.* **1**, 269-271 (1959).
16. Jeong, H., Tombor, B., Albert, R., Oltvai, Z.N. & Barabási, A.L. The large-scale organization of metabolic networks. *Nature* **407**, 651-4 (2000).
17. Ma'ayan, A. et al. Formation of regulatory patterns during signal propagation in a Mammalian cellular network. *Science* **309**, 1078-83 (2005).
18. Papin, J.A. & Palsson, B.O. Topological analysis of mass-balanced signaling networks: a framework to obtain network properties including crosstalk. *J Theor Biol* **227**, 283-97 (2004).
19. Ma'ayan, A., Blitzer, R.D. & Iyengar, R. Toward predictive models of mammalian cells. *Annu. Rev. Giophys. Biomol. Struct.*, 319-349 (2004).
20. Anthonisse, J.M. The rush in a directed graph. in *Technical Report BN 9/71 Stichting Mathematicsh Centrum* (Amsterdam, 1971).
21. Freeman, C.L. A set of measures of centrality based on betweenness. *Sociometry* **40**, 35 (1977).
22. Holme, P., Huss, M. & Jeong, H. Subnetwork hierarchies of biochemical pathways. *Bioinformatics* **19**, 532 (2003).
23. Mahadevan, R. & Palsson, B.O. Properties of metabolic networks: structure versus function. *Biophys J* **88**, L07-9 (2005).
24. Jeong, H., Mason, S.P., Barabási, A.L. & Oltvai, Z.N. Lethality and centrality in protein networks. *Nature* **411**, 41-2 (2001).
25. Said, M.R., Begley, T.J., Oppenheim, A.V., Lauffenburger, D.A. & Samson, L.D. Global network analysis of phenotypic effects: protein networks and toxicity

- modulation in *Saccharomyces cerevisiae*. *Proc Natl Acad Sci U S A* **101**, 18006-11 (2004).
26. Giaever, G. et al. Functional profiling of the *Saccharomyces cerevisiae* genome. *Nature* **418**, 387-91 (2002).
 27. Hartwell, L.H., Hopfield, J.J., Leibler, S. & Murray, A.W. From molecular to modular cell biology. *Nature* **402**, C47-52 (1999).
 28. Rives, A.W. & Galitski, T. Modular organization of cellular networks. *Proceedings of the National Academy of Sciences of the United States of America* **100**, 1128-1133 (2003).
 29. Spirin, V. & Mirny, L.A. Protein complexes and functional modules in molecular networks. *Proc Natl Acad Sci U S A* **100**, 12123-8 (2003).
 30. Guimera, R. & Nunes Amaral, L.A. Functional cartography of complex metabolic networks. *Nature* **433**, 895-900 (2005).
 31. von Mering, C. et al. Genome evolution reveals biochemical networks and functional modules. *Proc Natl Acad Sci U S A* **100**, 15428-33 (2003).
 32. Sharan, R. et al. Conserved patterns of protein interaction in multiple species. *Proc Natl Acad Sci U S A* **102**, 1974-9 (2005).
 33. Han, J.D. et al. Evidence for dynamically organized modularity in the yeast protein-protein interaction network. *Nature* **430**, 88-93 (2004).
 34. Ravasz, E., Somera, A.L., Mongru, D.A., Oltvai, Z.N. & Barabási, A.L. Hierarchical organization of modularity in metabolic networks. *Science* **297**, 1551-5 (2002).
 35. Shen-Orr, S.S., Milo, R., Mangan, S. & Alon, U. Network motifs in the transcriptional regulation network of *Escherichia coli*. *Nat Genet* **31**, 64-8 (2002).
 36. Balázsi, G., Barabási, A.L. & Oltvai, Z.N. Topological units of environmental signal processing in the transcriptional regulatory network of *Escherichia coli*. *Proc Natl Acad Sci U S A* **102**, 7841-6 (2005).
 37. Giot, L. et al. A protein interaction map of *Drosophila melanogaster*. *Science* **302**, 1727-36 (2003).

38. Wuchty, S., Oltvai, Z.N. & Barabasi, A.L. Evolutionary conservation of motif constituents in the yeast protein interaction network. *Nat Genet* **35**, 176-9 (2003).
39. Mangan, S. & Alon, U. Structure and function of the feed-forward loop network motif. *Proceedings of the National Academy of Sciences of the United States of America* **100**, 11980-11985 (2003).
40. Pandey, A. & Mann, M. Proteomics to study genes and genomes. *Nature* **405**, 837-46 (2000).
41. Caron, H. et al. The human transcriptome map: clustering of highly expressed genes in chromosomal domains. *Science* **291**, 1289-92 (2001).
42. Wang, S.M. Understanding SAGE data. *Trends Genet* **23**, 42-50 (2007).
43. DeRisi, J.L., Iyer, V.R. & Brown, P.O. Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science* **278**, 680-6 (1997).
44. Cho, R.J. et al. A genome-wide transcriptional analysis of the mitotic cell cycle. *Mol Cell* **2**, 65-73 (1998).
45. Rensink, W.A. & Buell, C.R. Microarray expression profiling resources for plant genomics. *Trends Plant Sci* **10**, 603-9 (2005).
46. Reinartz, J. et al. Massively parallel signature sequencing (MPSS) as a tool for in-depth quantitative gene expression profiling in all organisms. *Brief Funct Genomic Proteomic* **1**, 95-104 (2002).
47. Margulies, M. et al. Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437**, 376-80 (2005).
48. Kanehisa, M. & Goto, S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* **28**, 27-30 (2000).
49. Wingender, E., Dietze, P., Karas, H. & Knuppel, R. TRANSFAC: a database on transcription factors and their DNA binding sites. *Nucleic Acids Res* **24**, 238-41 (1996).
50. Huerta, A.M., Salgado, H., Thieffry, D. & Collado-Vides, J. RegulonDB: a database on transcriptional regulation in *Escherichia coli*. *Nucleic Acids Res* **26**, 55-9 (1998).

51. Hughes, J.D., Estep, P.W., Tavazoie, S. & Church, G.M. Computational identification of cis-regulatory elements associated with groups of functionally related genes in *Saccharomyces cerevisiae*. *J Mol Biol* **296**, 1205-14 (2000).
52. Grundy, W.N., Bailey, T.L. & Elkan, C.P. ParaMEME: a parallel implementation and a web interface for a DNA and protein motif discovery tool. *Comput Appl Biosci* **12**, 303-10 (1996).
53. Aerts, S. et al. Toucan: deciphering the cis-regulatory logic of coregulated genes. *Nucleic Acids Res* **31**, 1753-64 (2003).
54. Pavesi, G., Mereghetti, P., Mauri, G. & Pesole, G. Weeder Web: discovery of transcription factor binding sites in a set of sequences from co-regulated genes. *Nucleic Acids Res* **32**, W199-203 (2004).
55. Glinski, M. & Weckwerth, W. The role of mass spectrometry in plant systems biology. *Mass Spectrom Rev* **25**, 173-214 (2006).
56. Rossignol, M. et al. Plant proteome analysis: a 2004-2006 update. *Proteomics* **6**, 5529-48 (2006).
57. Komatsu, S. Plant proteomics databases: Their status in 2005. *Current Bioinformatics* **1**, 33-36 (2006).
58. Hall, R.D. Plant metabolomics: from holistic hope, to hype, to hot topic. *New Phytologist* **169**, 453-468 (2006).
59. Oksman-Caldentey, K.M. & Inze, D. Plant cell factories in the post-genomic era: new ways to produce designer secondary metabolites. *Trends in Plant Science* **9**, 433-440 (2004).
60. Moco, S. et al. A liquid chromatography-mass spectrometry-based metabolome database for tomato. *Plant Physiology* **141**, 1205-1218 (2006).
61. Lee, T.I. et al. Transcriptional regulatory networks in *Saccharomyces cerevisiae*. *Science* **298**, 799-804 (2002).
62. Guelzim, N., Bottani, S., Bourguin, P. & Kepes, F. Topological and causal structure of the yeast transcriptional regulatory network. *Nature Genetics* **31**, 60-63 (2002).

63. Luscombe, N.M. et al. Genomic analysis of regulatory network dynamics reveals large topological changes. *Nature* **431**, 308-312 (2004).
64. Benedict, C., Geisler, M., Trygg, J., Huner, N. & Hurry, V. Consensus by democracy. Using meta-analyses of microarray and genomic data to model the cold acclimation signaling pathway in Arabidopsis. *Plant Physiology* **141**, 1219-1232 (2006).
65. Xenarios, I. et al. DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions. *Nucleic Acids Res* **30**, 303-5 (2002).
66. Mewes, H.W. et al. MIPS: analysis and annotation of proteins from whole genomes. *Nucleic Acids Res* **32**, D41-4 (2004).
67. Peri, S. et al. Human protein reference database as a discovery resource for proteomics. *Nucleic Acids Res* **32**, D497-501 (2004).
68. McCraith, S., Holtzman, T., Moss, B. & Fields, S. Genome-wide analysis of vaccinia virus protein-protein interactions. *Proc Natl Acad Sci U S A* **97**, 4879-84 (2000).
69. Rual, J.F. et al. Towards a proteome-scale map of the human protein-protein interaction network. *Nature* **437**, 1173-8 (2005).
70. Rain, J.C. et al. The protein-protein interaction map of Helicobacter pylori. *Nature* **409**, 211-5 (2001).
71. Uetz, P. et al. A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. *Nature* **403**, 623-7 (2000).
72. Li, S. et al. A map of the interactome network of the metazoan C. elegans. *Science* **303**, 540-3 (2004).
73. Hart, G.T., Ramani, A.K. & Marcotte, E.M. How complete are current yeast and human protein-interaction networks? *Genome Biol* **7**, 120 (2006).
74. Immink, R.G.H. et al. Analysis of the petunia MADS-box transcription factor family. *Molecular Genetics and Genomics* **268**, 598-606 (2003).
75. de Folter, S. et al. Comprehensive interaction map of the Arabidopsis MADS Box transcription factors. *Plant Cell* **17**, 1424-33 (2005).

76. Zimmermann, I.M., Heim, M.A., Weisshaar, B. & Uhrig, J.F. Comprehensive identification of *Arabidopsis thaliana* MYB transcription factors interacting with R/B-like BHLH proteins. *Plant J* **40**, 22-34 (2004).
77. Yu, H. et al. Annotation transfer between genomes: protein-protein interologs and protein-DNA regulogs. *Genome Res* **14**, 1107-18 (2004).
78. Arita, M. The metabolic world of *Escherichia coli* is not small. *Proc Natl Acad Sci U S A* **101**, 1543-7 (2004).
79. Tanaka, R. Scale-rich metabolic networks. *Phys Rev Lett* **94**, 168101 (2005).
80. Ma, H.W. & Zeng, A.P. The connectivity structure, giant strong component and centrality of metabolic networks. *Bioinformatics* **19**, 1423-30 (2003).
81. Keurentjes, J.J. et al. The genetics of plant metabolism. *Nat Genet* **38**, 842-9 (2006).
82. Meyer, R.C. et al. The metabolic signature related to high plant growth rate in *Arabidopsis thaliana*. *Proc Natl Acad Sci U S A* **104**, 4759-64 (2007).
83. Oksman-Caldentey, K.M. & Saito, K. Integrating genomics and metabolomics for engineering plant metabolic pathways. *Curr Opin Biotechnol* **16**, 174-9 (2005).
84. Gutierrez, R.A. et al. Qualitative network models and genome-wide expression data define carbon/nitrogen-responsive molecular machines in *Arabidopsis*. *Genome Biol* **8**, R7 (2007).
85. Hirai, M.Y. et al. Integration of transcriptomics and metabolomics for understanding of global responses to nutritional stresses in *Arabidopsis thaliana*. *Proc Natl Acad Sci U S A* **101**, 10205-10 (2004).
86. Hirai, M.Y. et al. Elucidation of gene-to-gene and metabolite-to-gene networks in *Arabidopsis* by integration of metabolomics and transcriptomics. *J Biol Chem* **280**, 25590-5 (2005).
87. Tohge, T. et al. Functional genomics by integrated analysis of metabolome and transcriptome of *Arabidopsis* plants over-expressing an MYB transcription factor. *Plant J* **42**, 218-35 (2005).

88. Goossens, A. et al. A functional genomics approach toward the understanding of secondary metabolism in plant cells. *Proc Natl Acad Sci U S A* **100**, 8595-600 (2003).
89. Guterman, I. et al. Rose scent: genomics approach to discovering novel floral fragrance-related genes. *Plant Cell* **14**, 2325-38 (2002).
90. Mercke, P. et al. Combined transcript and metabolite analysis reveals genes involved in spider mite induced volatile formation in cucumber plants. *Plant Physiol* **135**, 2012-24 (2004).
91. Li, S., Assmann, S.M. & Albert, R. Predicting essential components of signal transduction networks: a dynamic model of guard cell abscisic acid signaling. *PLoS Biol* **4**, e312 (2006).
92. Stuart, J.M., Segal, E., Koller, D. & Kim, S.K. A gene-coexpression network for global discovery of conserved genetic modules. *Science* **302**, 249-55 (2003).
93. Valencia, A. & Pazos, F. Computational methods for the prediction of protein interactions. *Curr Opin Struct Biol* **12**, 368-73 (2002).
94. Tong, A.H. et al. Global mapping of the yeast genetic interaction network. *Science* **303**, 808-13 (2004).
95. Persson, S., Wei, H., Milne, J., Page, G.P. & Somerville, C.R. Identification of genes required for cellulose synthesis by regression analysis of public microarray data sets. *Proc Natl Acad Sci U S A* **102**, 8633-8 (2005).
96. Marcotte, E.M., Xenarios, I. & Eisenberg, D. Mining literature for protein-protein interactions. *Bioinformatics* **17**, 359-63 (2001).
97. Stapley, B.J. & Benoit, G. Biobibliometrics: information retrieval and visualization from co-occurrences of gene names in Medline abstracts. *Pac Symp Biocomput*, 529-40 (2000).
98. Snel, B., Lehmann, G., Bork, P. & Huynen, M.A. STRING: a web-server to retrieve and display the repeatedly occurring neighbourhood of a gene. *Nucleic Acids Res* **28**, 3442-4 (2000).

99. Pellegrini, M., Marcotte, E.M., Thompson, M.J., Eisenberg, D. & Yeates, T.O. Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc Natl Acad Sci U S A* **96**, 4285-8 (1999).
100. von Mering, C. et al. STRING 7--recent developments in the integration and prediction of protein interactions. *Nucleic Acids Res* (2006).
101. Dougherty, E.R. et al. Inference from clustering with application to gene-expression microarrays. *J Comput Biol* **9**, 105-26 (2002).
102. Qian, J., Dolled-Filhart, M., Lin, J., Yu, H. & Gerstein, M. Beyond synexpression relationships: local clustering of time-shifted and inverted gene expression profiles identifies new, biologically relevant interactions. *J Mol Biol* **314**, 1053-66 (2001).
103. Drawid, A. & Gerstein, M. A Bayesian system integrating expression data with sequence patterns for localizing proteins: comprehensive application to the yeast genome. *J Mol Biol* **301**, 1059-75 (2000).
104. Jansen, R. et al. A Bayesian networks approach for predicting protein-protein interactions from genomic data. *Science* **302**, 449-53 (2003).
105. Alon, U. et al. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc Natl Acad Sci U S A* **96**, 6745-50 (1999).
106. Gargalovic, P.S. et al. Identification of inflammatory gene modules based on variations of human endothelial cell responses to oxidized lipids. *Proc Natl Acad Sci U S A* **103**, 12741-6 (2006).
107. Horvath, S. et al. Analysis of oncogenic signaling networks in glioblastoma identifies ASPM as a molecular target. *Proc Natl Acad Sci U S A* **103**, 17402-7 (2006).
108. Wen, X. et al. Large-scale temporal gene expression mapping of central nervous system development. *Proc Natl Acad Sci U S A* **95**, 334-9 (1998).
109. Toronen, P., Kolehmainen, M., Wong, G. & Castren, E. Analysis of gene expression data using self-organizing maps. *FEBS Lett* **451**, 142-6 (1999).

110. Tavazoie, S., Hughes, J.D., Campbell, M.J., Cho, R.J. & Church, G.M. Systematic determination of genetic network architecture. *Nat Genet* **22**, 281-5 (1999).
111. Gupta, A., Maranas, C.D. & Albert, R. Elucidation of directionality for co-expressed genes: predicting intra-operon termination sites. *Bioinformatics* **22**, 209-14 (2006).
112. Ge, H., Liu, Z., Church, G.M. & Vidal, M. Correlation between transcriptome and interactome mapping data from *Saccharomyces cerevisiae*. *Nat Genet* **29**, 482-6 (2001).
113. Friedman, N., Linial, M., Nachman, I. & Pe'er, D. Using Bayesian networks to analyze expression data. *J Comput Biol* **7**, 601-20 (2000).
114. Yu, J., Smith, V.A., Wang, P.P., Hartemink, A.J. & Jarvis, E.D. Advances to Bayesian network inference for generating causal networks from observational biological data. *Bioinformatics* **20**, 3594-603 (2004).
115. Chen, T., He, H.L. & Church, G.M. Modeling gene expression with differential equations. *Pac Symp Biocomput*, 29-40 (1999).
116. Gupta, A., Varner, J.D. & Maranas, C.D. Large-scale inference of the transcriptional regulation of *Bacillus subtilis*. *Computers and Chemical Engineering* **29**, 565-576 (2005).
117. Shmulevich, I., Dougherty, E.R., Kim, S. & Zhang, W. Probabilistic Boolean Networks: a rule-based uncertainty model for gene regulatory networks. *Bioinformatics* **18**, 261-74 (2002).
118. Smolen, P., Baxter, D.A. & Byrne, J.H. Mathematical modeling of gene networks. *Neuron* **26**, 567-80 (2000).
119. Liang, S., Fuhrman, S. & Somogyi, R. Reveal, a general reverse engineering algorithm for inference of genetic network architectures. *Pac Symp Biocomput*, 18-29 (1998).
120. Dougherty, E.R. & Shmulevich, I. Mappings between probabilistic Boolean networks. *Signal Processing* **83**, 799-809 (2003).

121. Dougherty, E.R., Kim, S. & Chen, Y. Coefficient of determination in nonlinear signal processing. *Signal Processing* **80**, 2219-2235 (2000).
122. Zhao, W., Serpedin, E. & Dougherty, E.R. Inferring gene regulatory networks from time series data using the minimum description length principle. *Bioinformatics* (2006).
123. Famili, I., Forster, J., Nielsen, J. & Palsson, B.O. *Saccharomyces cerevisiae* phenotypes can be predicted by using constraint-based analysis of a genome-scale reconstructed metabolic network. *Proc Natl Acad Sci U S A* **100**, 13134-9 (2003).
124. Famili, I. & Palsson, B.O. The convex basis of the left null space of the stoichiometric matrix leads to the definition of metabolically meaningful pools. *Biophys J* **85**, 16-26 (2003).
125. Reed, J.L., Vo, T.D., Schilling, C.H. & Palsson, B.O. An expanded genome-scale model of *Escherichia coli* K-12 (iJR904 GSM/GPR). *Genome Biol* **4**, R54 (2003).
126. Irvine, D.H. & Savageau, M.A. Efficient solution of nonlinear ODE's expressed in S-system canonical form. *SIAM Journal of Numerical Analysis* **27**, 704-735 (1990).
127. Burgard, A.P., Pharkya, P. & Maranas, C.D. Optknock: a bilevel programming framework for identifying gene knockout strategies for microbial strain optimization. *Biotechnol Bioeng* **84**, 647-57 (2003).
128. Herrgard, M.J., Fong, S.S. & Palsson, B.O. Identification of genome-scale metabolic network models using experimentally measured flux profiles. *PLoS Comput Biol* **2**, e72 (2006).
129. Albert, R. et al. A novel method for signal transduction network inference from indirect experimental evidence. *J Comput Biol* **14**, 927-49 (2007).
130. Kachalo, S., Zhang, R., Sontag, E., Albert, R. & DasGupta, B. NET-SYNTHESIS: a software for synthesis, inference and simplification of signal transduction networks. *Bioinformatics* **24**, 293-5 (2008).

131. Yeager-Lotem, E. et al. Network motifs in integrated cellular networks of transcription-regulation and protein-protein interaction. *Proc Natl Acad Sci U S A* **101**, 5934-9 (2004).
132. Zhang, L.V. et al. Motifs, themes and thematic maps of an integrated *Saccharomyces cerevisiae* interaction network. *J Biol* **4**, 6 (2005).
133. Covert, M.W., Schilling, C.H. & Palsson, B. Regulation of gene expression in flux balance models of metabolism. *J Theor Biol* **213**, 73-88 (2001).
134. Herrgard, M.J., Lee, B.S., Portnoy, V. & Palsson, B.O. Integrated analysis of regulatory and metabolic networks reveals novel regulatory mechanisms in *Saccharomyces cerevisiae*. *Genome Res* **16**, 627-35 (2006).
135. Lee, I., Date, S.V., Adai, A.T. & Marcotte, E.M. A probabilistic functional network of yeast genes. *Science* **306**, 1555-8 (2004).
136. Vazquez, A., Flammini, A., Maritan, A. & Vespignani, A. Global protein function prediction from protein-protein interaction networks. *Nat Biotechnol* **21**, 697-700 (2003).
137. Albert, I. & Albert, R. Conserved network motifs allow protein-protein interaction prediction. *Bioinformatics* **20**, 3346-52 (2004).
138. Kelley, B.P. et al. PathBLAST: a tool for alignment of protein interaction networks. *Nucleic Acids Res* **32**, W83-8 (2004).
139. Bandyopadhyay, S., Sharan, R. & Ideker, T. Systematic identification of functional orthologs based on protein network comparison. *Genome Res* **16**, 428-35 (2006).
140. Mak, H.C., Daly, M., Gruebel, B. & Ideker, T. CellCircuits: a database of protein network models. *Nucleic Acids Res* **35**, D538-45 (2007).
141. Covert, M.W., Knight, E.M., Reed, J.L., Herrgard, M.J. & Palsson, B.O. Integrating high-throughput and computational data elucidates bacterial networks. *Nature* **429**, 92-6 (2004).
142. Endy, D. & Brent, R. Modelling cellular behaviour. *Nature* **409**, 391-5 (2001).

143. Agrawal, V., Zhang, C., Shapiro, A.D. & Dhurjati, P.S. A dynamic mathematical model to clarify signaling circuitry underlying programmed cell death control in Arabidopsis disease resistance. *Biotechnol Prog* **20**, 426-42 (2004).
144. Morton-Firth, C.J. & Bray, D. Predicting temporal fluctuations in an intracellular signalling pathway. *J Theor Biol* **192**, 117-28 (1998).
145. Andrews, S.S. & Arkin, A.P. Simulating cell biology. *Curr Biol* **16**, R523-7 (2006).
146. Rao, C.V., Wolf, D.M. & Arkin, A.P. Control, exploitation and tolerance of intracellular noise. *Nature* **420**, 231-7 (2002).
147. Diaz, J. & Alvarez-Buylla, E.R. A model of the ethylene signaling pathway and its gene response in Arabidopsis thaliana: pathway cross-talk and noise-filtering properties. *Chaos* **16**, 023112 (2006).
148. Mendoza, L. & Alvarez-Buylla, E.R. Dynamics of the genetic regulatory network for Arabidopsis thaliana flower morphogenesis. *J Theor Biol* **193**, 307-19 (1998).
149. Yuh, C.H., Bolouri, H. & Davidson, E.H. Cis-regulatory logic in the endo16 gene: switching from a specification to a differentiation mode of control. *Development* **128**, 617-29 (2001).
150. Chaves, M., Sontag, E.D. & Albert, R. Methods of robustness analysis for Boolean models of gene control networks. *Syst Biol (Stevenage)* **153**, 154-67 (2006).
151. Glass, L. & Kauffman, S.A. The logical analysis of continuous, non-linear biochemical control networks. *J Theor Biol* **39**, 103-29 (1973).
152. Christensen, C., Gupta, A., Maranas, C.D. & Albert, R. Inference and graph-theoretical analysis of Bacillus Subtilis gene regulatory networks. *Physica A* **373**, 796-810 (2007).
153. Papin, J.A., Hunter, T., Palsson, B.O. & Subramaniam, S. Reconstruction of cellular signalling networks and analysis of their properties. *Nat. Rev. Mol. Cell. Biol.* **6**, 99-111 (2005).

Figure Legends

Table 1.1 URL addresses of databases cited in Chapter 1.

Figure 1.1. Hypothetical networks illustrating graph terminology. The node **degree** (k) quantifies the number of edges that start at or end in a given node, for example node K has both an in-degree and an out-degree of two. The **clustering coefficient** (C) characterizes the cohesiveness of the first neighborhood of a node, for example the clustering coefficient of node F is zero because its two first neighbors are not connected, and the clustering coefficient of node C is 1, indicating that it is a part of the BCD clique. The **graph distance** (d) between two nodes is defined as the number of edges in the shortest path between them. For example, the distance between nodes L and K is one, the distance between nodes K and L is two (along the KML path), and the distance between nodes J and I is infinite because no path starting from J and ending in I exists. The **betweenness centrality** (b) of a node quantifies the number of shortest paths in which the node is an intermediary (not beginning or end) node. For example, the betweenness centrality of node C is zero because it is not contained in any shortest paths that do not start or end in C, and the betweenness centrality of node G is one and a half because it is an intermediary in the FGH shortest path and in one of two alternative shortest paths between E and F (EGF and EDF). The **degree distribution**, $P(k)$ ($P(k_{in})$ and $P(k_{out})$ in directed networks) quantifies the fraction of nodes with degree k . For example, in panel (a), one node (A) has a degree of one; three nodes (C, F, H) have a degree of two, three nodes (B, E, G) have a degree of three and one node (D) has a degree of four; the corresponding fractions are obtained by dividing by the total number of nodes (eight). The **clustering coefficient distribution** $P(C)$ denotes the fraction of nodes with clustering coefficient C . For example, in panel (a), one node (A) has an undefined clustering coefficient because it only has a single first neighbor, one node (F) has a clustering coefficient of zero, one node (D) has a clustering coefficient of one sixth, three

nodes (B,D,G) have a clustering coefficient of one third, and two nodes (C,H) have a clustering coefficient of one. The distance distribution $P(d)$ denotes the fraction of node pairs having the distance d . The betweenness centrality distribution $P(b)$ quantifies the fraction of nodes with betweenness centrality b .

(a). This undirected graph is **connected**, has a range of degrees from one to four, clustering coefficients between zero and one, a range of pairwise distances from one to four, and node betweenness centralities between zero and twelve and a half. The BCD and FGH **subgraphs** are **cliques** (completely connected subgraphs) of three nodes.

(b) This directed graph contains two source nodes (I and O, both with $k_{in}=0$), one sink node (J, with $k_{out}=0$), one feed-forward loop (OLN; both O and L feed into N) and two feedback loops (MLN and MLK). The nodes K,L,M and N form the graph's **strongly connected** subgraph. The **in-component** of this subgraph contains the **source** nodes I and O, while its **out-component** consists of the **sink** node J.

Figure 1.2. Hypothetical metabolic network illustrating different possible representations. (a) Directed and weighted tri-partite graph representation whose three types of node are metabolites (circles), reactions (ovals), and enzymes (squares), and whose two types of edge represent mass flow (solid lines) and catalytic regulation (dashed lines), respectively. Mass flow edges connect reactants to reactions and reactions to products, and are marked by the stoichiometric coefficients of the metabolites; enzymes catalyzing the reactions are connected by regulatory edges to the nodes signifying the reaction. (b) Metabolite (substrate) graph, whose nodes are metabolites, joined by a non-directed edge if they occur in the same chemical reaction¹³. (c) Reaction graph, whose nodes are reactions, connected by a non-directed edge if they share at least one metabolite.

Table 1.1.

Database	URL
The Arabidopsis Information Resource (TAIR)	www.arabidopsis.org
The European Arabidopsis Stock Centre	http://arabidopsis.info/
AtGenExpress	http://www.arabidopsis.org/info/expression/ATGenExpress.jsp
NASCArrays	http://affymetrix.arabidopsis.info/narrays/experimentbrowse.pl
Gene Expression Omnibus (GEO)	http://www.ncbi.nlm.nih.gov/geo/
Arabidopsis Gene Expression Database	http://www.arexdb.org/
Genevestigator	https://www.genevestigator.ethz.ch/at/
Botany Array Resource	http://bbc.botany.utoronto.ca/
TRANSFAC	http://www.gene-regulation.com/
RegulonDB	http://regulondb.ccg.unam.mx/
Kyoto Encyclopedia of Genes and Genomes (KEGG)	http://www.genome.jp/kegg/
Aligns Nucleic Acid Conserved Elements (AlignACE)	http://arep.med.harvard.edu/mrnadata/mrnasoft.html
MEME	http://meme.sdsc.edu/meme/meme.html
MotifSampler	http://homes.esat.kuleuven.be/~thijs/BioDemo/MotifSampler.html
Weeder	http://159.149.109.16:8080/weederWeb/howto.html
ExpASy	http://ca.expasy.org/
GABIPD	http://gabi.rzpd.de/projects/Arabidopsis_Proteomics/
SUBA	http://www.suba.bcs.uwa.edu.au
MetaCyc	http://metacyc.org
Aracyc	http://www.arabidopsis.org/biocyc/index.jsp
MoTo Database	http://appliedbioinformatics.wur.nl
The Golm Metabolome Database	http://csbdb.mpimp-golm.mpg.de/csbdb/gmd/gmd.html
Metabolomics Fiehn Lab	http://fiehnlab.ucdavis.edu/projects/
Database of Interacting Proteins	http://dip.doe-mbi.ucla.edu/
Munich Information Center for Protein Sequences	http://mips.gsf.de/
Human Protein Reference Database	http://www.hprd.org/
Search Tool for the Retrieval of Interacting Proteins	http://string.embl.de/

Figure 1.1.

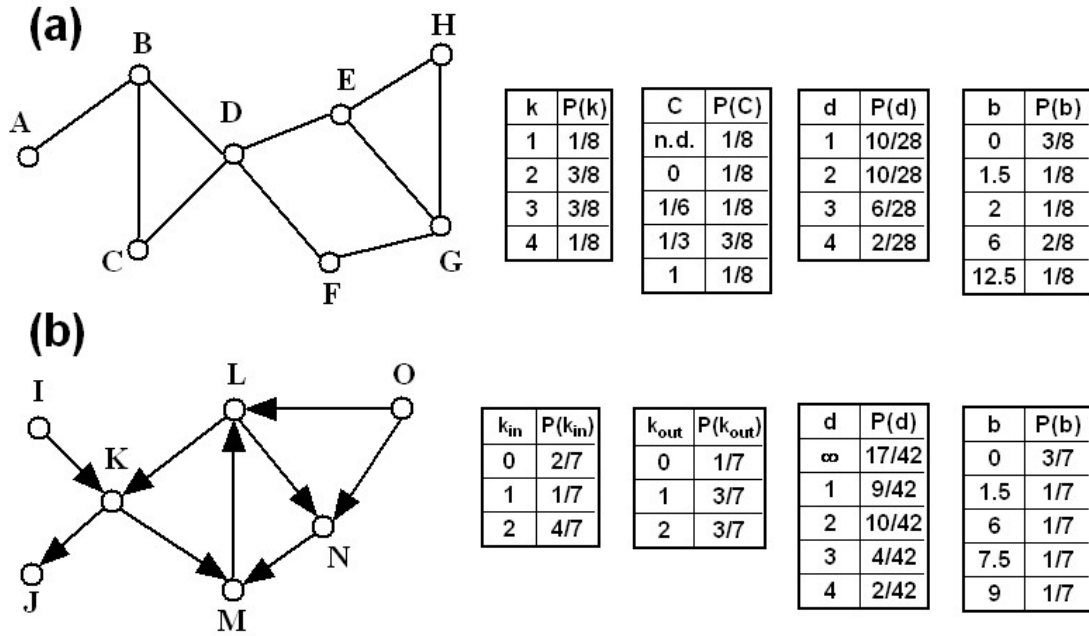


Figure 1.2.

