Chapter 1

AN INTRODUCTION TO SYSTEMS BIOLOGY FOR MATHEMATICAL PROGRAMMERS

Eivind Almaas
Microbial Systems Biology
Biosciences & Biotechnology Division
Lawrence Livermore National Laboratory
7000 East Avenue, P.O. Box 808, L-441
Livermore, CA 94551
almaas@llnl.gov

Allen Holder
Trinity University
Department of Mathematics
One Trinity Place
San Antonio, TX 78212
Allen.Holder@trinity.edu

Kevin Livingstone
Trinity University
Department of Biology
One Trinity Place
San Antonio, TX 78212
Kevin.Livingstone@trinity.edu

Abstract Many recent advances in biology, medicine and health care are due to computational efforts that rely on new mathematical results. These mathematical tools lie in discrete mathematics, statistics & probability, and optimization, and when combined with savvy computational tools and an understanding of cellular biology they are capable of remarkable results. One of the most significant areas of growth is in the field of systems biology, where we are using detailed biological information to

construct models that describe larger entities. This chapter is designed to be an introduction to systems biology for individuals in Operations Research (OR) and mathematical programming who already know the supporting mathematics but are unaware of current research in this field.

Keywords: Systems Biology, Computational Biology, Mathematical Programming, Operations Research, Optimization, Complex Networks

1. Introduction

The field of systems biology represents a new, exciting collaboration between biology, mathematics, and computer science. In broad collaborations such as this, it is usually not the case that a single discipline benefits to the exclusion of the others, but rather each discipline is rewarded from the inventions of the interaction. Classic examples of similar interactions involved mathematics and physics, which lead to the invention of Calculus, and the interaction between agronomists and statisticians that led to advances in experimental design, analysis of small sample sizes, and the development of analysis of variance. Many have argued that current problems in cellular biology are playing a similar role in mathematics and computer science today. In particular, the nexus of high-throughput data generation in biology and increasingly sophisticated mathematical and computational tools makes systems biology an exciting and innovative field of study.

Broadly speaking, biologists want to answer overarching questions related to how organisms work. The complexity of life and the difficulties inherent to experimental science have traditionally led biologists to adopt a reductionist approach, working for example in a single species to find and characterize single causative factors. Subsequent research then finds factors that interact with the first factors, and so on. The reductionist approach has shed light on many individual components of an organism, but for all our work, we only know a small percentage of how organisms work.

The painstaking progress of the reductionist approach is now being accelerated, however, by new "high-throughput" technologies. The most reductionist level of an organism is its DNA sequence, and it is almost inconceivable that the structure of DNA was discovered approximately 50 years ago, and that less than 20 years ago researchers labored to hand-sequence genes a few hundred bases at a time. Now our goal is to produce affordable, personal sequences of the 3 billion bases of an individual human genome in a matter of days or weeks, rather than the years it took to complete the first human genome sequence.

Although completing the human genome represented a pinnacle of achievement, it did not provide all the information needed to holistically model life's processes. To build a functional model of an organism, we need to know which proteins are actually made, at what times, in response to what environmental cues, and how these proteins interact either physically with other proteins and/or in metabolic pathways to create a static trait or dynamic response. Being able to characterize these higher levels of complexity is crucial: if anything, our reductionist studies have taught us that the whole is more than the sum of the parts.

Advances in technology similar to those seen in the sequencing arena are now also expanding our understanding of these higher-order questions. The current difficulty is how best to deal with the embarrassment of riches in biological information. On the whole, most biologists have not been trained in model building, data management, and computational skills. Experts in Operations Research, however, are trained in exactly these fields and are well positioned to accelerate this exciting area of research. What OR professionals lack is an understanding of the underlying biology and how it transforms into familiar research topics. This tutorial is intended to fill this educational gap.

In the end, our goal is to have quantitative, predictive models that describe systems from cells to entire organisms. In pursuing this goal, it is important to remember that our interest is not solely focused on understanding *Homo sapiens*. While it is true that much of our research on the bacterium *E. coli*, the single-celled eukaryotic yeast *S. cerevisiae*, and the millimeter-sized roundworm *C. elegans* and fly *D. melanogaster* is undertaken using these as surrogates, our interest also extends to a myriad of other organisms that provide food, fiber, fuel, pharmaceuticals, etc. It now appears that collaborations between biology, mathematics, and computer science in the field of systems biology are the way by which progress towards this goal will be made.

2. General Background

This chapter discusses the three levels of whole-cell modeling based on interactions between genes, proteins and metabolites. A thorough discussion of each whole-cell model exceeds the capability of this introductory chapter, so our goal in each section is to focus on key aspects of the underlying biology and the network representation, and then provide a summary of some of the insights this representation has provided.

To operate in modern biological terms, we need to understand the basic premises that support the research. This section is divided into two subsections: one that explains the guiding principle that dictates

the related biological research, called the "Central Dogma" of molecular biology, and another that defines the fundamental terms of the network analysis used by systems biologists.

2.1 Basic Biological Definitions

The Central Dogma, elaborated by Francis Crick soon after his co-discovery of the structure of DNA, states that biological information flows from deoxyribonucleic acid (DNA) to messenger ribonucleic acid (mRNA) to proteins. The DNA molecule that serves as the main repository of biological information is a pair of directional polymers whose monomers are denoted A, T, G, and C. Each of these monomers has a conserved portion that forms the backbone of the polymer and the variable portion that makes it an A, T, G, or C. The DNA double helix is comprised of two polymers that are oriented in opposite directions and held together by interactions between the variable parts of the monomers, A-T pairs and G-C pairs, see Figure 1.1(a). A DNA sequence is usually represented by the list of letters (ATGC) read directionally along one strand, the complementary strand being implied. Each of the 46 chromosomes inside a human cell is a double helix with about 10^7 to 10^8 base pairs, and a gene is a known stretch of hundreds or thousands of bases of the double helix with a defined function, usually encoding a protein.

The DNA is used as a template to make an mRNA polymer by a process called transcription. The monomers of mRNA have a conserved portion that forms the backbone of the polymer, although slightly different from the corresponding DNA monomers, and four variable portions denoted A, U, G, and C. Construction of an mRNA molecule involves partial unwinding of the DNA molecule and then the exposed ATGC bases of the DNA dictate the sequence of the mRNA through interactions similar to those described above, except that where there is an A in the DNA there will be a U in the mRNA, see Figure 1.1(b). The mRNA molecule is also directional and is represented by a string of AUGC.

The mRNA intermediate of a gene is used as a template to make proteins through a process called translation. During translation, cellular machinery reads an mRNA three monomers at a time from a defined starting position, see Figure 1.1(c). Each triplet determines one of the 20 amino acid monomers found in a protein or a message to stop protein synthesis. While proteins are represented by their primary sequence using an alphabet of 20 letters, protein function is ultimately determined by the protein's three-dimensional structure, which may not be predictable based on the primary sequence alone.

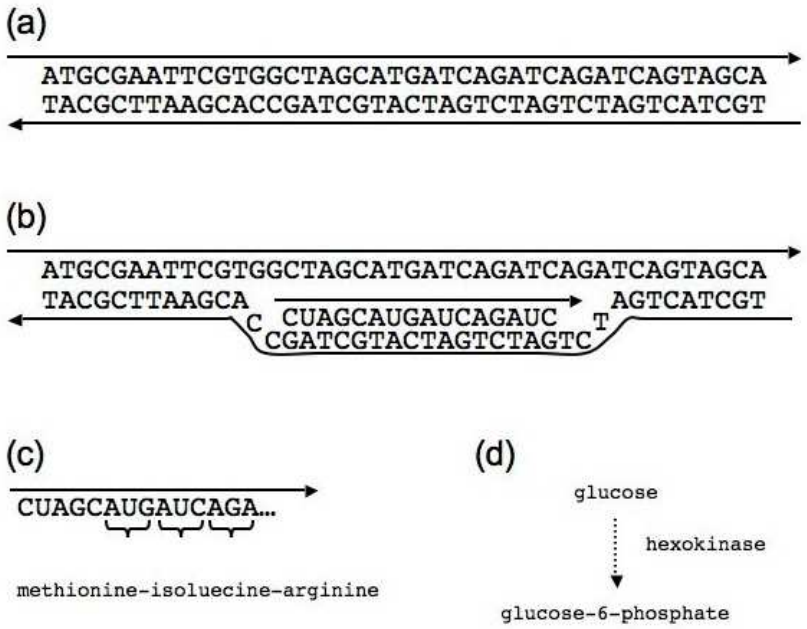

Figure 1.1. The Central Dogma of molecular biology. **(a)** DNA, **(b)** mRNA being made from a DNA template (transcription), **(c)** protein synthesis specification by mRNA (translation), **(d)** the enzyme hexokinase acts on the metabolite glucose in the metabolic pathway that breaks down glucose for energy production.

Proteins are both the structural and functional workhorses of a cell that convert information stored in the DNA (the genotype) to the visible characteristics of the cell or organism (the phenotype). In this chapter we focus on the protein's functional aspects as enzymes that take molecules and convert these to the products needed for cellular functions. These molecules are called metabolites, and Figure 1.1(d) gives an example of how metabolites and enzymes are organized into metabolic pathway.

Exceptions to the Centra Dogma exist, but these subtleties and variations cannot be addressed in this presentation. Detailed descriptions of all the biological processes we describe and exceptions can, however, be found in any current biology or genetics textbook. Despite these exceptions, the Central Dogma does appropriately model most of the information flow within a living system, so we will operate on the simplified Central Dogma throughout.

2.2 Basic Network & Mathematical Definitions

Networks are used in systems biology to model the relationships between cellular entities. Networks are familiar to those in OR, and this section specifies the common notation used throughout. In cases where terms vary between the two disciplines, we mention both terms but use those that are common to the field of systems biology. This convention helps those in OR understand the language of systems biology.

A network is a directed graph (V, E), where the elements of V are called *nodes* or *vertices* and the elements of $E \subseteq V \times V$ are called *arcs* or *links*. In network analysis the direction of an arc is important, and we distinguish between (v_1, v_2) and (v_2, v_1). If we are instead referring to the graph (V, E), then direction is not important and there is no distinction between (v_1, v_2) and (v_2, v_1). In this case the elements of E are called *edges*. We assume throughout that $|V| = N$ and $|E| = M$. The nodes v_1 and v_2 are *adjacent* if $(v_1, v_2) \in E$, and we further say that v_1 is *incident* to the edge (not arc) (v_1, v_2).

If $E = V \times V$, then (V, E) is complete in the sense that it contains as many arcs or edges as possible. Such graphs are defined by the size of V and are called *complete* and denoted K_N. We say that (V', E') is a *subnetwork* or *subgraph* of (V, E) if $V' \subseteq V$, $E' \subseteq E$, and $E' \subseteq V' \times V'$. A *clique* of a network or graph is a complete subnetwork or subgraph.

A network's structure is often referred to as the *topology* of the network, which is a bit awkward for mathematicians. A graph's topology is often described by the adjacency matrix $A = [a_{ij}]$, where $a_{ij} = 1$ if nodes i and j are adjacent and zero otherwise. For networks $a_{ij} = 1$ if $(v_i, v_j) \in E$ and $a_{ij} = -1$ if $(v_j, v_i) \in E$. In a graph, the *neighborhood* of a node is $N(v_i) = \{v_j : (v_i, v_j) \in E\}$ and the *degree* of the node is $\deg(v_i) = |N(v_i)|$. This concept naturally extends to a network where we discuss out-degree and in-degree. Much of the analysis considered by systems biologists is based on how well a graph is connected, and for this reason, the $\deg(v_i)$ is often called the *connectivity* of node i. Instead of $\deg(v_i)$, we denote the $\deg(v_i)$ as k_i, and for graphs we have

$$\deg(v_i) = k_i = \sum_j a_{ij}.$$

For an understood probability distribution, we let $P(x)$ be the probability of observing x. We use the typical big-O notation and write $f(x) = O(g(x))$ if there is a λ such that $f(x) \leq \lambda g(x)$. The vector of ones is denoted by e, where length is decided by the context of its use. Other notation is introduced as needed. All terms dealing with opti-

mization agree with those defined in the Mathematical Programming Glossary [33].

3. Gene Regulatory Networks

Complex organisms exhibit dramatic differences in cellular phenotypes (characteristics). Examples of these differences are fixed differences between cell types (e.g., brain cells and liver cells) or temporarily induced differences due to environmental stimuli (e.g., increased production of melanin by skin cells after UV exposure). In general, all the cells in an organism have the same DNA, so the cause of these phenotypic differences is variation in the amount and types of proteins present in the cell. Gene expression is the general term for this conversion of the information in the inert DNA into the functional proteins, and tight control over gene expression is what allows for different cellular phenotypes.

The majority of the control over gene expression occurs at the level of initiation of transcription (the making of the mRNA intermediate). One of the primary tools used to understand a cell, therefore, is characterization of what is called the transcriptome, the set of all genes expressed under defined conditions. Biologists can detect the levels of different mRNA molecules with precision, and new microarray technology even allows for the simultaneous measurement of the levels of all mRNAs in a cell. As will be seen later, presence of an mRNA does not always imply the presence of a functional protein, but mRNA production is a necessary first step and the correlation between mRNA and protein levels is strong enough to make mRNA quantitation a meaningful first measure for most gene expression studies.

Initiation of transcription for a gene is dependent on two factors. Production of mRNA requires a large group of proteins that unwind the DNA and facilitate the polymerization of the mRNA, and these proteins must bind to the DNA of the gene at locations called regulatory regions. Because multiple genes may have similar regulatory regions, coordinate gene expression can occur when the proteins in a cell increase the expression of all these target genes simultaneously. Coordinate repression of genes may also occur when binding of a protein to regulatory regions prevents transcription. Coordinate regulation allows groups of genes to be acted on as a unit, which is important given that many cellular actions require multiple types of proteins.

3.1 Network clustering

Genetic interactions are frequently represented as networks, where the nodes correspond to genes, and a (possibly directed) link is introduced

between genes A and B if the presence or absence of gene A's encoded protein enhances or suppresses the expression of gene B, or vice versa. The local properties of a gene-regulatory network are measured by how closely they resemble a clique [77]. The clustering coefficient c_i of a node, defined as

$$c_i \;=\; \frac{2}{k_i(k_i-1)} \sum_{j,l} a_{ij} a_{il} a_{jl}, \tag{1.1}$$

measures the degree to which the neighborhood of a node resembles a complete subgraph built from triangles, and is the ratio of the actual number of triangles to possible triangles, for which node i is a member. The average clustering coefficient $\langle C \rangle \;=\; (1/N) \sum_i c_i$ provides information on the global distribution of links. A value of $\langle C \rangle$ close to unity indicates a high level of modularity, or cohesiveness of triangles, in the network, while a value close to zero indicates a lack of modularity. It is customary to test the significance of a particular $\langle C \rangle$-value by comparing it to a random-network model with the same number of nodes and edges [2]. Typical random graphs have an average clustering coefficient of $\langle C \rangle_{rand} = 2M/N^2$.

Assuming that a network has a non-zero $\langle C \rangle$, we further investigate the network's large-scale modularity structure by studying the average clustering as function of degree k [19],

$$C(k) = \frac{\sum_{\{i:k_i=k\}} c_i}{\sum_{\{i:k_i=k\}} 1}. \tag{1.2}$$

If the network shows a hierarchical modularity [64], the clustering $C(k) \sim 1/k$. In this case, nodes with few neighbors tend to have network-neighborhoods with high clustering, while the highly connected nodes act as bridges tying the network together.

3.2 Network motifs

It has long been argued that biological systems are functionally modular [36], and understanding how this modularity is reflected in biological network is a primary goal. Given this modularity, additional questions arise, for example what network modules, or partitions, carry functional information, and how does the functional modularity depend on the environmental conditions and the dynamic states of a gene-regulatory network? An interesting possibility was suggested in [47, 48, 74], introducing the idea of network "motifs" as the functional building blocks of a gene-regulatory network. They suggest that these networks contain particular sub-graphs, many with easily identifiable functions such as feed-forward loops, at a significantly higher frequency than should be

expected by chance alone. The enrichment of biological networks with functional motifs is seen as a result of the evolutionary processes shaping the system [48]. However, the recent results in [75] indicate that caution is needed to determine if a motif is over expressed. By designing random networks that matched the experimental results, they found that certain sub-graphs occur at higher frequencies than in random networks without this restriction.

4. Protein Interaction Networks

As stated above, the majority of the structural and functional macromolecules in a cell are proteins, and the presence of these proteins is tightly regulated by the cell mainly through initiation of transcription. Even after translation of an mRNA into a protein, however, the protein may not be functional. The activity of many proteins is influenced by modifications such as the addition of chemical groups by other proteins, binding of cofactors (which may include other proteins), or cleavage by other proteins, to name a few. These mechanisms allow for rapid cellular responses by relying on quick modifications of existing proteins rather than *de novo* production. Another advantage is that modification allows for coordination and amplification of a signal if a single protein can interact with many other proteins. The protein interaction network (PIN) thus forms another level of biological organization that influences the cell.

The data needed to characterize the PIN include determination of the set of proteins present in a cell (the proteome), the state or location of those proteins if variable, and how these proteins interact. High-throughput methods to provide these data are developing, albeit more slowly than methods used to identify mRNA levels. This disparity is due to the lack of means to artificially increase the amount of any particular protein in a sample. The technique of polymerase chain reaction allows biologists to harness the natural process of DNA replication to make millions of copies of any known DNA or mRNA molecule in a biological sample, but no comparable technology exists for proteins.

When working with biological samples, specfic proteins can be detected using either antibodies or techniques that separate proteins based on biochemical properties and then calculate a molecular weight and compare that to a database of known protein weights. These techniques can sometimes reveal whether or not a protein has a phosphate group attached, for example, or whether it is in a particular subcellular location. Interactions between proteins can be determined (1) by assays that use antibodies to pull proteins out of cellular extracts and look

for proteins that are associated with the protein removed; (2) by assays that use synthetic hybrid proteins that produce a visible result if two proteins physically interact; or (3) by *in vitro* or *in vivo* experimentation with cells and/or organisms that have had specific genes mutated. Bioinformatics provides another method whereby protein function may be inferred by comparing the sequence of a protein to genes with known function. There are functional regions of proteins, called domains, that occur in many different proteins, which can be detected in the DNA or protein sequence. If a protein contains a sequence similar to a known functional domain, the protein is also assumed to have that functionality.

In constructing a graph to represent the PIN, the individual proteins are the nodes, and the existence of an interaction between a pair of proteins corresponds to an edge between the nodes. As seen above, there are many ways in which proteins may physically interact. Relations between proteins may also be established by examining mRNA profiles, for example. If the mRNA profiles of two proteins have a high correlation, we assume the corresponding proteins are related and include the edge even if there is not a physical interaction. Each of these techniques provides different information, and combinations thereof are thus important for a more complete characterization of the proteome and the protein-protein interactions that occur.

4.1 Connectivity distribution

Analyzing systems as disparate as the World Wide Web and a PIN has revealed surprising similarities in their structural organization. One simple characterization is the average number of nearest neighbors, or average degree. In a PIN, this corresponds to an average protein's number of interaction partners.

The average degree is simply $\langle k \rangle = (1/N) \sum_{ij} a_{ij}$. However, this measure does not provide detailed insight into the structure of a network. To gain further insight into the structure of a PIN, we study the connectivity, or degree, distribution $P(k)$, which is the number of nodes of degree k. From this measure, we determine the variation in connectivities on the network. Such distributions were studied by Erdös and Rényi [12], and they showed that random graphs lead to a Poisson distribution. However, for many real networks, $P(k)$ does not have a Poisson-type behavior as predicted by the Erdös-Rényi random graph model. Instead, $P(k)$ frequently adheres to a heavy-tailed distribution often modeled as a power-law $P(k) \sim k^{-\alpha}$ [2]. This is the case for the PIN of the yeast *S. cerevisiae*, the nematode *C. elegans*, and the fruitfly *D. melanogaster* in Figure 1.2 (see also Table 1.1).

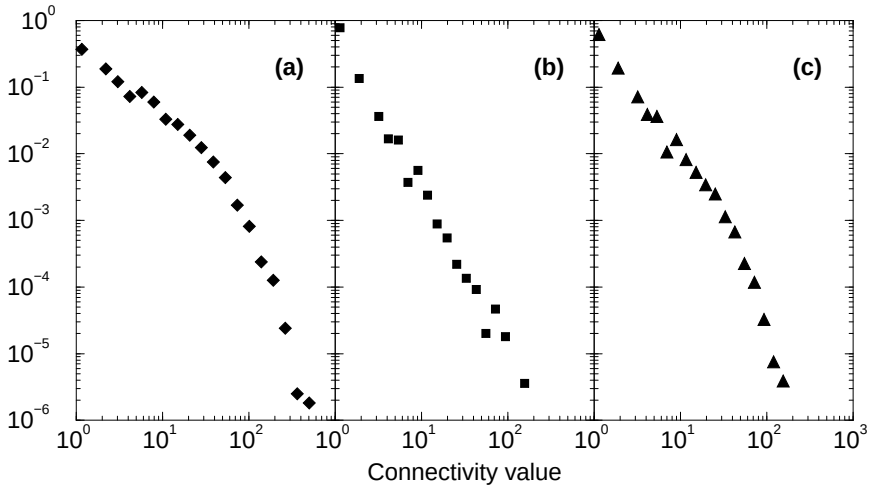

Figure 1.2. Connectivity distribution $P(k)$ for the protein-interaction networks of **(a)** the yeast *S. cerevisiae*, **(b)** the nematode *C. elegans*, and **(c)** the fly *D. melanogaster* [10].

It is interesting to note that if the connectivity distribution had been single-peaked, such as Poisson or Gaussian, the notion of a "typical" node, as described by the average degree $\langle k \rangle$, would have been valid. However, this is not the case for a heavy-tailed PIN. In these networks the majority of the nodes only have a few interaction partners while they coexist with nodes that participate in hundreds of interactions. Consequently, there is no typical node. Such networks are typically called "scale-free," and nodes with a large number of interactions are called "hubs." Hub proteins often have biological properties that are significantly different from non-hub proteins.

Organism	Nodes	$\langle k \rangle$	S	$\langle C \rangle$	$\langle C_{rand} \rangle$	ρ
S. cerevisiae	5298	19.04	5294	0.154	0.0036	-0.040
C. elegans	2774	3.14	2551	0.020	0.0011	-0.159
D. melanogaster	7490	6.67	7372	0.030	0.00089	-0.039

Table 1.1. Properties of three whole-organism protein-interaction networks available from [10]. For each network, we have indicated size, average node connectivity $\langle k \rangle$, size of the giant component S, average clustering $\langle C \rangle$, average clustering for a comparable Erdös-Rényi random network $\langle C_{rand} \rangle$, and assortativity ρ, which is defined momentarily.

One of the most popular network models to capture the heterogeneity of the connectivity distribution was proposed by Barabási and Albert [6]. It is similar to the network model by Price [63] (see [53] for a detailed discussion). These models are based on the notion that in a growing network, new nodes are not connected with uniform probability to already existing nodes. Instead, new nodes have a higher chance of connecting to those with many neighbors than to nodes with few. This is often called the "rich get richer" effect or "preferential attachment." If the chance of connecting to an already existing node i is linearly proportional to the degree, the resulting connectivity distribution is a power-law with an exponent of 3 [2, 53].

4.2 Network assortativity

In many real networks, properties of adjacent nodes are correlated. In particular, it is often the case that the connectivities of neighboring nodes are correlated, making $P(k_i, k_j) \neq P(k_i)P(k_j)$. Several methods have been developed [46, 51, 52, 60] to measure these connectivity correlations, and we highlight two such methods.

The first method of [60] measures connectivity correlations by calculating the average nearest-neighbor degree:

$$k_{nn,i} \;=\; \frac{1}{k_i}\sum_j k_j a_{ij} \tag{1.3}$$

Consequently, $k_{nn,i}$ measures the affinity with which a node i connects to other nodes of either high or low degrees. In Figure 1.3 we have plotted $k_{nn}(k)$, which is defined by

$$k_{nn}(k) = \frac{\sum_{\{i:k_i=k\}} k_{nn,i}}{\sum_{\{i:k_i=k\}} 1}. \tag{1.4}$$

So, $k_{nn}(k)$ is the average neighborhood degree for nodes with connectivity k, If $k_{nn}(k)$ is an increasing function of k, the network shows an *assortative* mixing and high-degree nodes preferentially tend to be connected to other high-degree nodes. For the opposite situation, where $k_{nn}(k)$ is a decreasing function of k (as in Figure 1.3(b)), low-degree nodes tend to be connected to high-degree nodes, and the network is *disassortative*. This is typically the case for computer networks, where a limited number of servers are connected to a large number of individual computers [60].

The second method of measuring degree-degree correlations collapses the distribution $P(k)$ into a single value called the *assortativity* of the

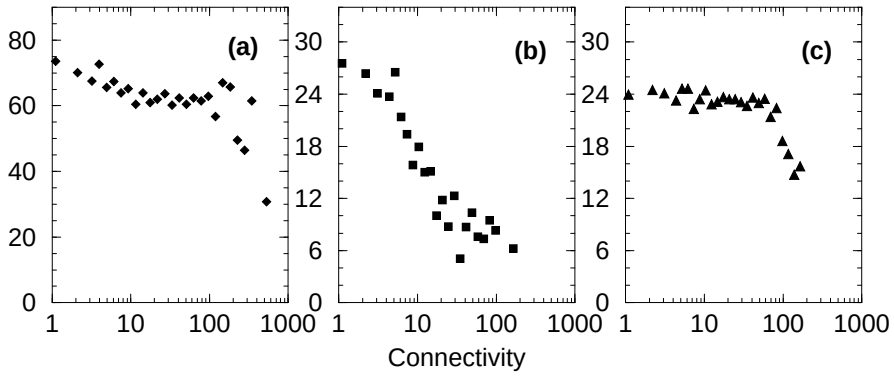

Figure 1.3. Average nearest neighbor connectivity $k_{nn}(k)$ for the protein-interaction networks of **(a)** *S. cerevisiae* **(b)** *C. elegans*, and **(c)** *D. melanogaster* [10].

graph [51]. This index is a Pearson correlation in nearest neighbor degrees, defined as

$$\rho \;=\; \frac{M\sum_i j_i k_i - \left(\frac{1}{2}\sum_i (j_i + k_i)\right)^2}{\frac{1}{2}M\sum_i (j_i^2 + k_i^2) - \left(\frac{1}{2}\sum_i (j_i + k_i)\right)^2}, \tag{1.5}$$

where the sums are over edges, and the numbers j_i and k_i are the connectivities of the two nodes connected by edge i. The distribution $k_{nn}(k)$ and the assortativity index ρ are related as follows. If $k_{nn}(k)$ is uniform, then $\rho = 0$. However, if $k_{nn}(k)$ is increasing or decreasing, then ρ is positive or negative, respectively. The magnitude of ρ indicates the strength of the correlation. It is straightforward to develop similar expressions for directed networks [52].

The last column of Table 1.1 shows the assortativity ρ for three whole-organism PINs. As expected, the trends displayed in Figure 1.3 agree with the assortativity correlations calculated from equation (1.5). In particular, panels (a) and (c) show no clear increasing or decreasing trend in $k_{nn}(k)$, which agrees with the calculated assortativity values close to zero. Taken together, these two methods offer detailed insights into the connectivity correlations of a network.

4.3 Community finding

The network properties just discussed are based on characteristics of individual nodes, such as clustering, average degree and connectivity. As stated previously, a longstanding hypothesis is that biological systems are modular, meaning that they consist of separable functional units. The idea of a community is different from the previous properties since

it considers the entire network. By carefully analyzing a network, we identify modules as collections of nodes that are tightly connected when compared to the full network. These modules are often biologically significant. For instance, since proteins that exist in a cell as a complex are commonly members of the same functional class, we expect a tightly connected region to indicate a single functional class [61, 62, 72, 79].

Several methods are currently available to detect community structures. Many of these were developed by sociologists, who have long been interested in community analysis. Unfortunately, these methods typically were designed for small networks and are not tractable on networks consisting of several thousands of nodes. Many of these methods are related to a measure called betweenness-centrality (BC), which is related to shortest paths [14, 26, 50, 76].

The study of shortest paths on networks is the source of the term 'small-world' [77]. The length of the average shortest path, ℓ, between two nodes can be calculated using a breadth-first search, which has complexity $O(NM)$. In a random network, such as that of the Erdös-Rényi model, the average shortest distance scales with the network size as $\ell \sim \ln(N)$ [12]. The betweenness-centrality of an edge or node is the fraction of shortest paths that pass through the node or edge (see [50] for a detailed discussion).

A typical algorithm based on BC is to recursively remove the edge with the largest BC value, followed by recalculating the BC values for the remaining network. The complexity of such an algorithm is $O(N^3)$. Approximations where the BC values are only calculated for the initial network are much faster, but the gain in computational run-time reduces accuracy.

There are alternatives to the BC approach, and we discuss two such methods. These techniques have the advantage of rapidly identifying communities on large networks with high accuracy. The first method is due to Newman [54] and is described as agglomerative hierarchical clustering. Let Q be the following measure of network modularity for any node partition

$$Q = \sum_i \left(e_{ii} - \left(\sum_j e_{ij} \right)^2 \right), \tag{1.6}$$

where e_{ij} is the fraction of edges in the network connecting nodes from module i to those of module j. This measures the number of inter-community links relative to that of a random occurrence. A value near 0 suggests that there is little information in the chosen partition, whereas a value greater than 0.3 indicates significant modularity [54].

Newman suggests optimizing Q heuristically by starting with N communities (one for each node) and joining the two that render the highest value of Q, which may increase or decrease the current value. When all nodes have been joined into a single module, the algorithm is finished and the optimal value of Q indicates a collection of communities. This approach is $O(N^2)$ and has been successfully applied to systems with more than $50,000$ nodes. Furthermore, it is possible to generalize this community detection algorithm to incorporate varying link-strengths.

The second alternative to the BC method is called *k-clique percolation* [57]. Unlike the method just described, this technique does not require that each node belong to a unique community. For many networks this is favorable. For example, a protein may have multiple functions and naturally belong to many communities.

This method is based on the observation that a community often decomposes into nearly complete subgraphs that share nodes. Consequently, this method is based on the k-clique. A network module is defined as the union of all k-cliques (for a fixed k) that share $k-1$ nodes, and thus are adjacent on the network. An alternative description is that of a "rolling" k-clique, only moving one node at the time.

A further benefit of k-clique percolation is that it allows a higher-level representation of a network. We may collapse the graph so that each community is a node, and two communities are connected if they have a non-empty intersection. This makes it possible to introduce a scalable map of the network that represent the communities at different levels of magnification, with the highest magnification corresponding to the actual nodes, the second level to communities, the third level to communities of communities, etc.

4.4 Biology and topology

So far we have discussed topological properties of PINs without emphasizing the connection between network representations and biological information. The first indication that a PIN might carry biological information arose from questions of robustness [3], which demonstrated that networks with heavy-tailed connectivity distributions were robust against random failures yet fragile when an attack occurred at a highly connected node.

Molecular biology techniques allow for the experimental disruption of single genes, and examination of the phenotypes of these modified organisms can reveal whether the disrupted gene is essential for survival of the organism under a set of defined conditions. In fact, a large-scale experimental study in *S. cerevisiae* shows that only 18.7% of the to-

tal number of genes are essential on disruption or removal [29], while a study on *E. coli* found 13.7% of the genes are essential [27]. Motivated by these experimental observations of network fragility, Barabási and coworkers investigated the possibility of correlations between a protein's connectivity and phenotypic essentiality, discovering an increased likelihood for highly connected proteins to be essential [43]. In other words, a protein that has a large number of interaction partners is more likely to be involved in an essential cellular function, often called the "centrality-lethality" rule. Although recently debated, the centrality-lethality result is considered robust [9].

A recent study suggests that this increased lethality of highly connected proteins can be explained by a simple mechanism [37]. The idea is to support the centrality-lethality rule by assuming essential nodes and *links* are randomly distributed on the network. The function of an essential link is carried out by the interaction of the incident proteins, and both nodes are essential. This model generates the centrality-lethality rule through the simple fact that it is more likely for a hub to be part of an essential link than a low degree node. By choosing the essential link and node fractions appropriately, it is possible to fit the observed centrality-lethality rule within experimental error bars [37].

Since highly connected proteins occupy a special role in the network, it is interesting to ask whether hub proteins evolve at a different pace from proteins with only a few interaction partners. The rationale for this question is that change to hub proteins might be constrained due to their interactions. While initial results were contradictory [18], a recent more decisive study [9] showed these results could be explained by subtle biases in the methods used to generate the PINs. After accounting for the equal density of active domains in hub and non-hub proteins, it was shown that there are not significant differences in mean rate of protein evolution. The hub proteins of *S. cerevisiae* did, however, contain a higher number of phosphorylation sites than non-hub proteins and showed a marked trend of being encoded by mRNA's with short half-lives. Taken together, this indicates that highly connected proteins are subject to much tighter control, being part of a dynamic, short-lived protein complex [9].

We have focused on static aspects of a PIN, but proteins are constantly produced and degraded and many interactions occur in specific cellular locations, such as the cellular membrane. A more realistic depiction would address the temporal and spatial aspects of the situation. Whole-organism protein-expression arrays are currently unavailable, and the chosen substitute has been mRNA expression arrays. The recent analysis in [35] indicates that highly connected nodes in the *S. cerevisiae*

PIN can be either "date-hubs", binding to their partners at different times or locations, or "party-hubs" interacting with most of their neighbors simultaneously. Including temporal aspects such as this allows us to investigate information flow since the temporal activation of protein transcription is reflective of evolved regulatory mechanisms that ensure proper cellular responses to external stimuli.

5. Metabolic Networks

Life depends on the ability to import molecules from the environment and convert these to the needed metabolites. These conversions are carried out by enzymes that catalyze (facilitate) specific conversions of starting molecules (reactants) into products. There may be several intermediary steps from intial reactants to the ultimate product, each carried out by a different enzyme, and the set of all these component reactants, products, reactions, and enzymes forms a metabolic pathway. Metabolic pathways can be classified as either anabolic pathways that construct needed molecules or catabolic pathways that break down molecules to provide necessary reactants.

The different reactions and catalyzing enzymes vary tremendously. As seen in the previous section, the enzymes may or may not be active depending on the presence of cofactors, modification state, etc. Another difference between enzymes is in their rates of catalysis, which may vary over orders of magnitude. Variation in these reaction rates affects the overall rate of flow (flux) of metabolites in a particular pathway.

From the reactant perspective, a particular type of molecule may participate in only one reaction or be used in several different reactions. A reaction may require one or more reactants, and the ratios (stoichiometry) of those reactants may vary. Finally, while for the most part metabolic pathways can be assumed to be one-way, there are cases of reversible reactions in a cell and cyclic reaction pathways that take a reactant through a series of intermediates but end up regenerating the initial reactant.

A cell's metabolism is the sum of all the reactions it carries out. It is important to recognize that while a cell has the potential to carry out many reactions, the actual reactions that are being carried out at any one time depend heavily on the cell's environment. For example, differential gene regulation in a bacterial cell will lead to different enzymes being present under aerobic (oxygen present) vs. anaerobic (oxygen absent) conditions or when glucose or lactose are present as the main carbon source.

5.1 Metabolic network structure

To represent a cell's metabolism with a network we need to assign meaning to the nodes and links. The network abstraction is not unique, and Figure 1.4 depicts several representations of a simple metabolic network. The three reactions of the metabolism are found in Figure 1.4(a). In the first reaction $A + B \rightarrow C + D$, we say that A and B are *reactants* and C and D are *products*. The most common representation of this metabolism is represented in Figure 1.4(c), where metabolites are nodes that are connected with an undirected link if they participate as reactant and product in a reaction. Note that a link does not represent a single reaction, as two metabolites may appear in multiple reactions. An example is shown in Figure 1.4(a), where metabolites A and D co-occur in reactions R1 and R3, and the edge or arc between A and D corresponds to both reactions. Furthermore, one reaction appears as multiple edges or arcs (see Figure 1.4).

An alternative representation that is particularly important for the discussions that follow is a bipartite network in which the nodes represent either metabolites or reactions. Allowing the set of reactions to be R and the set of metabolites to be M, we are interested in the bipartite network (R, M, E), where $(i, r) \in E$ if metabolite i is a reactant of reaction r and $(r, i) \in E$ if metabolite i is a product of reaction r. A depiction is seen in Figure 1.9.

Different network representations have different statistical properties. Using the bacterial metabolism in *E. coli* as an example, Figure 1.5 shows the differences in the connectivity distribution, $P(k)$, for the three network representations detailed in Figure 1.4. Note that $P(k)$ is heavy-tailed in Figure 1.5; however, the result is not as simple when using a bipartite network representation. In this case, it is possible to distinguish metabolites and enzymes. For the metabolites, the connectivity distribution is still heavy tailed, while the enzyme distribution is exponential. This is not surprising, as cofactors such as ATP or NADP may participate in hundreds of reactions while an enzyme has a limited number of active domains. To further contrast and compare biases of different network representations, Table 1.2 shows the average clustering coefficient $\langle C \rangle$ and the assortativity index ρ for three organisms using the representations in Figure 1.4(b) and (c). The clustering and assortativity corresponding to 1.4(b) is significantly higher than that of 1.4(c) since it introduces a fully connected subgraph for each reaction.

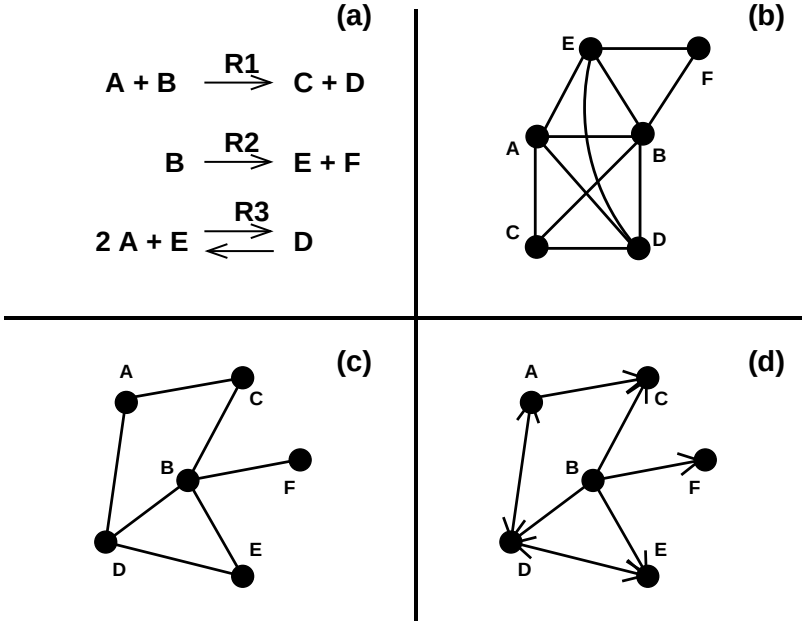

Figure 1.4. Cellular metabolism can be represented as a network. **(a)** A simplied metabolic reaction set. Network description of this reaction set: **(b)** connecting all metabolites in a single reaction with undirected links; **(c)** substrates are only connected to products with undirected links; **(d)** same as **(c)** with directed links.

5.2 Weighted metabolic networks

The majority of network studies have focused on topological properties and not on the rate of metabolic activity, which can vary significantly from reaction to reaction. This important function is not captured by topological approaches, and to develop an understanding of how the structure of a metabolic network affects metabolic activity, it is necessary to include this information in the network description. A meaningful understanding requires us to consider the intensity (strength) between metabolites, the direction (when applicable), and the temporal aspects

Organism	N	M_b	M_c	$\langle C \rangle_b$	$\langle C \rangle_c$	ρ_b	ρ_c
H. pylori	489	4058	1920	0.72	0.28	-0.285	-0.261
E. coli	540	3753	1867	0.66	0.20	-0.251	-0.217
S. cerevisiae	1064	6941	4031	0.67	0.23	-0.182	-0.150

Table 1.2. Average clustering and assortativity for three organismal metabolic networks using the network representations described in panels 1.4**(b)** and **(c)** - network model indicated with a subscript.

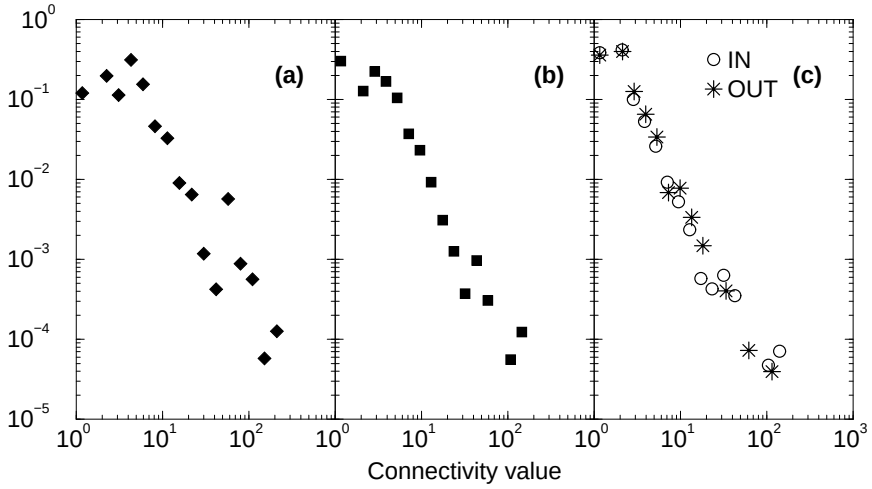

Figure 1.5. Connectivity distributions $P(k)$ of *E. coli* metabolism using the three metabolic network representations in Figure 1.4. **(a)** corresponds to panel (b); **(b)** to panel (c); **(c)** to panel (d).

of the interactions. Although not much is known about the temporal aspects of metabolic activity, recent results [11, 16, 17, 23, 24, 25, 31, 68] have provided information about the relative intensities of the interactions in single-cell metabolism, which we incorporate by considering weighted links. A natural, although not unique, measurement of interaction strength is the amount of substrate being converted to a product per unit time, the flux of the reaction.

A linear optimization approach, called flux-balance analysis (FBA), enables us to calculate the flux rate for each reaction in a whole-cell metabolic network. The FBA method assumes that the concentration of all metabolites that are not subject to transport across the cell membrane are in a steady state. Let $[A_i]$ be the concentration of metabolite i and S_{ir} be the stoichiometric coefficient of metabolite i in reaction r. For example, if reaction r is $3A_1 + 2A_2 \rightarrow 2A_3$, then $S_{1r} = -3$, $S_{2r} = -2$ and $S_{3r} = 2$. If metabolite i does not appear in reaction r, we assume that $S_{ir} = 0$. Allowing ν_r be the flux of reaction r, we have that the steady state assumption requires

$$\frac{d[A_i]}{dt} = \sum_r S_{ir}\nu_r = 0. \tag{1.7}$$

Any flux values satisfying this equation correspond to a stoichiometrically allowed state of the cell. To select flux values that are biologically relevant, we optimize for cellular growth. Experiments support this hy-

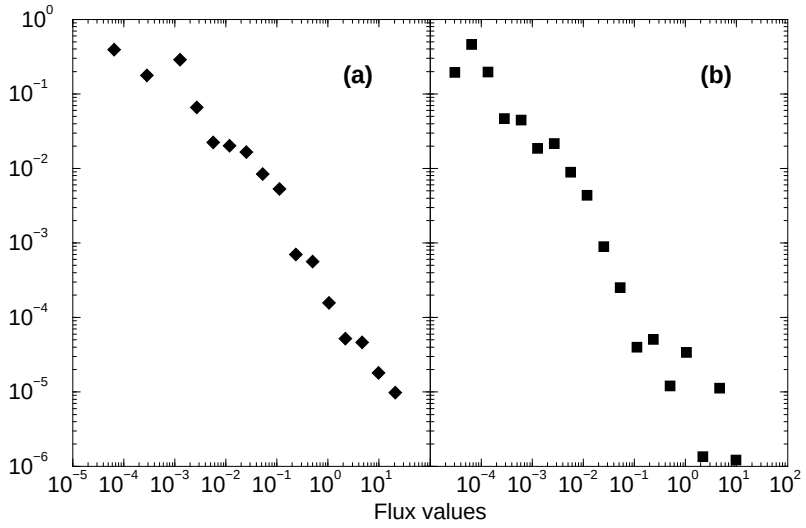

Figure 1.6. Distribution of metabolic reaction flux values (edge weights) from FBA analysis for the metabolic network of the budding yeast *S. cerevisiae* in **(a)** aerobic, glucose limited and **(b)** aerobic, acetate limited conditions.

pothesis in several conditions, but there are other meaningful objectives. See [13, 44] for a more detailed discussion of FBA.

The recent advances in whole-genome annotation have made it possible to generate high-fidelity whole-cell metabolic networks. Metabolic models of the bacteria *H. pylori* and *E. coli*, as well as the eukaryote *S. cerevisiae*, have been used to predict essential genes [21, 20, 59, 69], genetic interactions [73], and possible minimal microbial genomes [15, 56]. The fluxes from FBA measure each reaction's relative activity. In particular, the work of [4] demonstrates that similar to the degree distribution, the flux distribution of *E. coli* displays a strong overall inhomogeneity: reactions with fluxes spanning several orders of magnitude coexist in the same environment. The flux distribution for *S. cerevisiae* in Figure 1.6 is heavy-tailed, indicating that $P(\nu) \propto \nu^{-\alpha}$. The flux exponent is predicted to be $\alpha = 1.5$ by FBA methods. In a recent experiment, the strength of the various fluxes of the central metabolism of *E. coli* was measured using nuclear magnetic resonance (NMR) methods [23], revealing the power-law flux dependence $P(\nu) \propto \nu^{-1}$. This power law behavior indicates that the vast majority of reactions with small fluxes coexist with the few reactions that have large fluxes.

The FBA approach allows us to analyze a metabolism as a weighted network since each reaction is assigned a flux value. These values are node weights in the bipartite representation (R, M, E). Unfortunately,

the identity of a reaction in the other network models is opaque because each reaction is a subgraph corresponding to the metabolites of the reaction. To translate the node weights ν_r of the bipartite representation to link weights of another representation, we let

$$w_{ij} = \left| \sum_r S_{ir}\nu_r + \sum_r S_{jr}\nu_r \right|,$$

which is the aggregate rate at which metabolite i transforms into metabolite j. Generally, negative edge weights are possible and simply mean that metabolite j transforms into metabolite i.

Several measures have been introduced to study weighted networks in the context of airline transportation and the relationship between coauthors. One of these is called the "strength", s_i, of a node i, defined as $s_i = \sum_j w_{ij}a_{ij}$, which is simply a weighted node degree [7]. Figure 1.7 shows that the distribution of node strengths, $P(s)$, for the *E. coli* metabolism with glucose as the single carbon source is

$$\langle s(k) \rangle \propto k^{\beta}. \tag{1.8}$$

For networks without correlations between the node connectivity and link-weights, the weights w_{ij} are independent of i and j, and we can represent the link-weights with their average value: $w_{ij} = \langle w \rangle$, making $\beta = 1$ [7].

We continue by generalizing the clustering coefficient to weighted networks. Since c_i indicates the local density of triangles, a similar definition with link-weights should determine if large or small weights are likely to be found in clusters. One possible definition from [7] is

$$c_{a,i} = \frac{1}{s_i(k_i - 1)} \sum_{j,l} \tfrac{1}{2}(w_{ij} + w_{il})a_{ij}a_{il}a_{jl}, \tag{1.9}$$

where s_i is the node strength. The average weighted clustering is $\langle C_a \rangle = (1/N)\sum_i c_{a,i}$. If no correlations exist between weights and topology, equation (1.9) is equal to that of the unweighted clustering (see equation (1.1)). We identify two possible scenarios. If $\langle C_a \rangle > \langle C \rangle$, large weights are predominantly distributed in local clusters, whereas if $\langle C_a \rangle < \langle C \rangle$, triangles consist of low-weight links.

Other possible definitions of a weighted clustering coefficient have been proposed [41, 55, 80]. The weighted clustering coefficient expression in (1.9) only includes two weights of any triangle through node i. The following definition from [55] extends this so that all three weights are

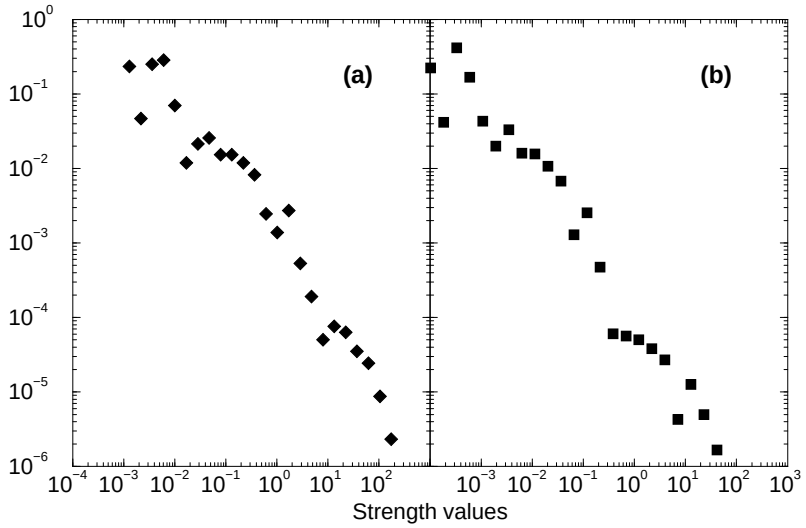

Figure 1.7. Distribution of node strength values for *S. cerevisiae* metabolism in **(a)** aerobic, glucose limited, and **(b)** aerobic, acetate limited conditions.

considered,

$$c_{b,i} = \frac{2}{(\max_{ij}\{w_{ij}\})k_i(k_i-1)} \sum_{j,l} \left(w_{ij}w_{il}w_{jl}\right)^{1/3} a_{ij}a_{il}a_{jl}. \qquad (1.10)$$

Notice that this is a geometric mean instead of an algebraic mean like (1.9). The average weighted clustering is $\langle C_b \rangle = (1/N)\sum_i c_{b,i}$. Related analysis from finance has shown that (1.9) and (1.10) can lead to significantly different interpretations [55].

5.3 Fluxes and metabolic network structure

The flux distributions of a metabolic network rely on the network topology. Some of this dependence is understood by studying the correlation between w_{ij} and k_i and k_j. The metabolic fluxes in *E. coli* scale as

$$\langle w_{ij} \rangle = \frac{\sum_{\{i,j:k_ik_j=k\}} w_{ij}}{\sum_{\{i:k_ik_j=k\}} 1} \propto (k_ik_j)^{\theta}, \qquad (1.11)$$

where $\theta \approx 0.5$ for metabolic fluxes in glucose limited conditions in *S. cerevisiae* (Figure 1.8(a)) and *E. coli* [45]. However, other values for θ are possible, as demonstrated in Figure 1.8(b), where we find $\theta \approx 0.7$ for acetate limited conditions. In the case of no correlations between

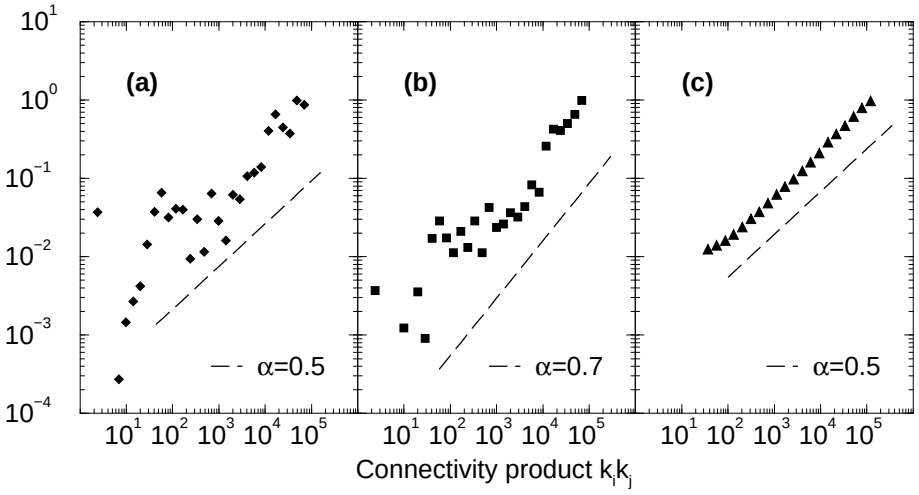

Figure 1.8. Correlation between (normalized) edge-weights and local connectivity for metabolic fluxes in *S. cerevisiae* in **(a)** glucose limited and **(b)** acetate limited conditions, as well as **(c)** betweenness-centrality for the Barabasi-Albert model [45]. The dashed lines serve as guides to the eye.

the connectivity k_i and k_j, we have from [7] that the exponent θ in equation (1.11) is related to β (equation (1.8)) in the scaling of node strength as $\beta = 1 + \theta$.

We further investigate how the flux values depend on the network topology at the single metabolite level. There are two flux structures of interest. A homogeneous local organization implies that all reactions producing (consuming) a given metabolite have comparable flux values. On the other hand, a more delocalized, or "hot backbone," is expected if the local flux organization is heterogeneous. To distinguish between these two scenarios, we define the following measure [4, 8] for each metabolite produced or consumed by k reactions, define $Y(k,i)$ by

$$Y(k,i) = \sum_{r=1}^{k} \left(\frac{S_{ir}\nu_r}{\sum_{l=1}^{k} S_{il}\nu_l} \right)^2 . \tag{1.12}$$

If all reactions producing (consuming) metabolite i have comparable values, $Y(k,i)$ scales as $1/k$. However, if a single reaction's activity dominates equation (1.12), we expect that $Y(k,i) \sim 1$. For the two cases where the *E. coli* metabolic performance is optimized with glucose and succinate as the only available carbon sources, we find that separately calculating $Y(k,i)$ for the in and out degrees follow the power law $Y(k,i) \sim k^{-0.27}$. We interpret this as intermediate behavior between

the two cases described above. However, the exponent of -0.27 indicates that the large-scale inhomogeneity observed in the overall flux distribution is increasingly valid at the level of the individual metabolites.

The local flux inhomogeneity suggests that we can identify a single reaction dominating the production or consumption of most metabolites. A simple algorithm is capable of extracting the subnetwork solely consisting of these dominate reactions, called the *high-flux backbone* (HFB) [4]. The algorithm has two steps,

1 For each metabolite, discard all incoming and outgoing links except the two dominating mass production.

2 From this set of reactions, keep only those reactions that appear as both maximal producer and maximal consumer.

The resulting HFB is specific to the particular choice of environmental conditions. Interestingly, the HFB mostly consists of reactions linked together, forming a giant component with a star-like topology that includes almost all metabolites produced in a specific growth environment. Only a few pathways are disconnected: while these pathways are members of the HFB, their end-products serve only as the second most important source for some other HFB metabolite. Connected reactions in the HFB largely agree with the traditional, biochemistry-based partitioning of cellular metabolism. For example, in *E. coli* all metabolites of the citric-acid cycle of are recovered, as well as most of the other important pathways, such as those being involved in histidine-, murein- and purine biosynthesis. While the detailed nature of the HFB depends on the particular growth conditions, the HFB captures the reactions that dominate the activity of the metabolism for this condition. As such, it offers a complementary approach to the analyses in [58, 70, 71].

Our final discussion about metabolic networks focuses on identifying the reactions that are used in varying environments, and we explore how the fluxes depend on environmental changes. Referring to Figure 1.9, we let ν_R be the collection of uptake fluxes that provide nutrients (resources, inputs, etc.) to the cell. We also let r_C be the reactions that occur within the cell (output reactions are not considered). For each ν_R, we let $r_C^*(\nu_R)$ be the point-to-set map whose image is the collection of reactions that can have a positive flux while the cell achieves optimal growth with the input fluxes fixed at ν_R. The *metabolic core* is

$$\bigcap_{\nu_R \geq 0} r_C^*(\nu_R),$$

which defines the reactions that are allowed to be active in any environment when the cell achieves optimal growth.

A stochastic procedure to calculate the metabolic core is to uniformly sample the set of input fluxes and use FBA to optimize growth for each sample. If a reaction's flux is positive, we know that this flux is in $r_C^*(\nu_R)$ for the sample. Taking the intersection of these sets over the sampled inputs yields a subset of the metabolic core. The computational results in [5] sampled $30,000$ input fluxes between 0 and 20 (20 is large enough to guarantee that a nutrient is available if needed, and hence, setting the intake fluxes to 20 assumes the cell is in an environment with unlimited resources). The metabolic core contained 138 of the 381 metabolic reactions in *H. pylori* (36.2%), 90 of 758 in *E. coli* (11.9%), and 33 of $1,172$ in *S. cerevisiae* (2.8%).

The metabolic core is partitioned into two types of reactions. The first type consists of those that are essential for biomass formation under all environmental conditions (81 out of 90 reactions in *E. coli*), while the second type of reaction is required only to assure optimal metabolic performance. In case of the inactivation of the second type, alternative suboptimal pathways can be used to ensure cellular survival. The metabolic core of *S. cerevisiae*, however, only contains reactions predicted by FBA to be indispensable for biomass formation under all growth conditions.

The analysis in [5] further suggests that optimal metabolic flows adjust to environmental changes through two distinct mechanisms. The more common mechanism is "flux plasticity," involving changes in the fluxes of already active reactions when the organism is shifted from one growth condition to another. For example, changing from glucose- to succinate-rich media alters the flux of 264 *E. coli* reactions by more than 20%. The reactions in the metabolic core always adapt to changing environmental conditions through flux plasticity. Less commonly, environmental changes induce "structural plasticity," resulting in changes to the metabolism's active wiring diagram, turning on previously zero-flux reactions and inhibiting previously active pathways. For example, when shifting *E. coli* cells from glucose- to succinate-rich media, 11 previously active reactions are turned off completely, while 9 previously inactive reactions are turned on.

A similar selection of metabolic reactions was suggested by [15]. Their "minimal reaction" contains the metabolic core as well as all reactions necessary for the sustained growth on any chosen substrate. A different definition of a minimal reaction set was proposed by [65], which consists of the 201 reactions that are always active in *E. coli* for all 136 aerobic and anaerobic single-carbon-source "minimal environments" capable of sustaining optimal growth.

A reasonable speculation is that the reactions in the metabolic core play an important role in the maintenance of crucial metabolic functions

since they are active under all environmental conditions. Consequently, the absence of individual core-reactions may lead to significant metabolic disruptions. This hypothesis is strengthened through cross-correlation with gene deletion data [27]: 74.7% of those *E. coli* enzymes that catalyze core metabolic reactions (i.e. core enzymes) are essential, compared with a 19.6% lethality fraction characterizing the noncore enzymes. A similar pattern of elevated essentiality is also observed when analyzing deletion data for *S. cerevisiae* [29], in which essential enzymes catalyze 84% of the core reactions, whereas the conditionally active enzymes have an average essentiality of only 15.6% [5]. The likelihood that the cores contain such a large concentration of essential enzymes by chance is minuscule, with p-values of $3.3e{-}23$ and $9.0e{-}13$ for *E. coli* and yeast, respectively.

Metabolic core enzymes also stand apart from the conditionally active ones when comparing their evolutionary conservation. In comparing the set of DNA sequences of core enzymes from *E. coli* with the DNA sequences for these same enzymes in a reference set of 32 bacteria, the average amount of sequence conservation is 71.1% ($P < 1e{-}6$). Similar comparisons using the set of non-core enzymes show a sequence conservation of only 47.7%. Even taking into account correlations between essentiality and evolutionary conservation, one would expect the core enzymes to have a conservation level of only 63.4% [5], thus showing that selection acts against excessive tinkering with these enzymes.

These results indicate that an organism depends largely on the continuous activity of the metabolic core, regardless of the environmental conditions. The conditionally active metabolic reactions represent the different ways in which a cell is capable of adjusting to utilize substrates from its environment. From a practical perspective, the core enzymes essential for biomass formation, both for optimal and suboptimal growth, may prove effective antibiotic targets given the cell's need to maintain the activity of these enzymes in all conditions.

6. Systems Biology and Operations Research

One of the primary research fields in Operations Research (OR) is Network Optimization, including modeling, algorithms, and analysis. The variety of problems that can be modeled via a network is staggering, and numerous OR experts have spent their careers analyzing such problems. As the previous sections demonstrate, a cell's processes can be modeled with networks that highlight the interactions within a cell. This is a powerful new tool for biologists, and the experts in OR are well positioned to help advance this important science.

The goal of this section is to highlight a few of the places where systems biology and OR overlap. This is not meant to be an exhaustive exposition, which is not possible in the confines of this chapter. We encourage interested readers to look at the cited articles to begin a more thorough investigation. No matter what the particular expertise, there is likely an important and novel application in biology.

To begin we consider the linear program that identifies the metabolic fluxes of a cell in a steady state. A simplistic but powerful depiction of the associated network is illustrated in Figure 1.9. This is a bipartite network where reactions on the left are linked to metabolites on the right. For example, if r is the reaction $A_1 + 2A_2 \rightarrow A_3 + 3A_4$, then (A_1, r), (A_2, r), (r, A_3), and (r, A_4) are arcs. The cell's inputs (resources) are modeled as reactions that transport metabolites through the cellular membrane into the cell. Similarly, the cell's outputs (products) are reactions that transport metabolites out of the cell. We let C, R and P be matrices of the form $[S_{ir}]$, where the columns are respectively indexed by reactions within the cell, reactions that add resources to the cell, and reactions that terminate in products, except growth. Growth is defined as the collection of metabolites that need to pass through the cell to achieve a unit of growth, and we let G be the column vector that expresses this relationship. As an example, suppose the metabolites used to model the cell are $A_1, A_2, \ldots, A_{10}$. Then a unit of growth being defined as $2A_3 + A_7 + 3A_8$ is the same as G being $(0, 0, 2, 0, 0, 0, 1, 3, 0, 0)^T$. We point out that the matrix $[C|R|P|G]$ is similar to the biadjacency matrix, the difference being that the nonzero components are the signed stoichiometric coefficients of the associated reaction.

Although the terms used to describe this network are new to OR, the model is not. The fluxes of the reactions control the flows across the arcs, and hence the amount of metabolites in the cell. Although researchers often discuss a metabolic flow, the fluxes are not traditional flow variables since they are associated with nodes instead of edges. In particular, a positive flux can indicate that several arcs have positive flow. We let ν_C, ν_R, ν_P and ν_G be the respective flux vectors for the reactions within the cell, the reactions that provide resources, the reactions that make products other than growth, and the amount of growth. The steady state assumption in (1.7) guarantees the conservation of metabolic flux throughout the network. This assumption essentially balances the metabolites in the cell so that they do not accrue.

Experimental results have shown that maximizing growth is a biologically relevant objective [13, 44], and the linear program that achieves

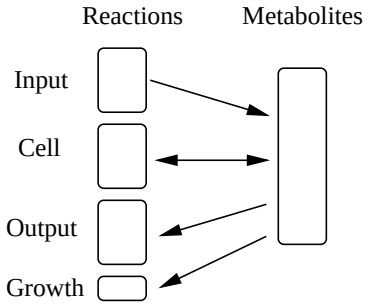

Figure 1.9. A simple bipartite representation of cellular metabolism.

this is

$$\max\{z : C\nu_C + R\nu_R + P\nu_P + G\nu_G = 0, \\ 0 \leq \nu_R \leq u, 0 \leq \nu_P, 0 \leq \nu_C \leq U\}, \tag{1.13}$$

where u limits the cell's inputs and U bounds the flux values (each bound is the maximum rate the corresponding reaction). This linear program allows us to give a mathematical definition to a few of the terms of the previous section. Let $\mathcal{P}(u)$ be the feasible region of (1.13) and make the notational convention that $\mathcal{P}(\infty)$ means the input flows are unrestricted. We also assume the $\nu = (\nu_C^T, \nu_R^T, \nu_C^T, \nu_G^T)^T$. With this notation, reaction $j \in C$ is *essential* (or *necessary*) if

$$\{\nu \in \mathcal{P}(\infty) : \nu_j = 0, \nu_G = 1\} = \emptyset. \tag{1.14}$$

So, if reaction j is turned off, then it is impossible to achieve a unit of growth no matter what resources are given to the cell.

Identifying the essential reactions can be accomplished by sequentially investigating the feasibility of (1.14) for each $j \in C$, which is possible by optimizing the zero function over the associated constraints. However, this tedious calculation has a more elegant solution. The question of partitioning the reactions into those that are necessary and those that are not is actually the problem of identifying the implied equalities in

$$C\nu_C + R\nu_R + P\nu_P = -G,\ 0 \leq \nu_C \leq U,\ 0 \leq \nu_R,\ 0 \leq \nu_P. \tag{1.15}$$

The implied equalities of this system further indicate the inputs and outputs of the cell that are necessary for growth as well as those reactions that operate at their maximum rate. Identifying implied equalities has a long and important history in OR, and we point readers to [32] and the associated bibliography. We highlight two methods, one theoretical and one more practical.

The theoretical method relies on the concept of the *optimal partition*, which is a central topic in the study of interior-point algorithms. Consider the standard form linear program

$$\min\{c^T x : Ax = b, x \geq 0\}, \tag{1.16}$$

where $A \in \mathbb{R}^{m \times n}$ and we assume that there is a strictly positive feasible element (Slater's condition). Throughout the late 1980s and 1990s interior-point algorithms were studied with regards to this problem, with the most important contribution being that these algorithms solve the problem in polynomial time. Fairly early in these investigations it was realized that the solution produced by the most popular interior-point algorithms (called path-following interior-point algorithms) differed from the solution produced by the venerable simplex algorithm. The difference is that interior-point algorithms terminate in the strict interior of the optimal set instead of at an optimal vertex. If there is a single solution, there is no difference, but in the presence of degeneracy, the solutions are different. In particular, the solution rendered by a path following interior algorithm induces the optimal partition. Let x^* be the theoretical optimal solution produced by a path-following interior-point algorithm and define

$$B = \{i : x_i^* > 0\} \text{ and } N = \{i : x_i^* = 0\}.$$

Clearly (B, N) is a two set partition of $\{1, 2, \ldots, n\}$, but this partition uniquely defines the optimal set,

$$\mathcal{P}^* = \{x : Ax = b, x \geq 0, x_N = 0\},$$

where the set subscript indicates the subvector indexed by the elements of the set. This means that a component of x^* is zero if and only if it is required to be zero to achieve optimality. A component being positive indicates that it can be positive in an optimal solution, but some optimal solutions may have a zero at this component.

The conditions identifying the optimal partition of the linear program in (1.16) are

$$Ax = b,\ A^T y + s = c,\ x \geq 0,\ s \geq 0,\ x^T s = 0,\ x + s > 0. \tag{1.17}$$

As any person in OR recognizes, these are the KKT (or Lagrange) conditions of optimality with the added condition that $x + s > 0$. Any (x, y, s) satisfying these conditions is called a *strictly complementary solution* to the linear program, and such solutions have been known to exist since 1956 [30]. Until the advent of interior-point algorithms, strictly complementary solutions held theoretical value only. If (x, y, s) satisfies all

but the last condition, i.e. $x_i + s_i = 0$ for some i, then the solution is *degenerate*. Any pair (x_i, s_i) such that $x_i + s_i = 0$ is called a *degenerate pair* and the extent of degeneracy refers to the maximum number of possible degenerate pairs. Degeneracy is a topic that is mistakingly ignored in many first courses in linear programing, a pedagogical mistake that propagates misguided analysis [42]. Understanding degeneracy provides for robust and sound analysis that appropriately explains the problem, and as we shall see momentarily, metabolic networks are highly degenerate.

A linear program that identifies the essential reactions is

$$\begin{aligned}\min\{0^T\nu_C + 0^T\nu_R + 0^T\nu_P :& \\ C\nu_C + R\nu_R + P\nu_P = -G, 0 \leq \nu_C \leq U, 0 \leq \nu_R, 0 \leq \nu_P\}& \quad (1.18)\end{aligned}$$

Adapting (1.17) to this problem, we see that we want to solve

$$\begin{aligned} C\nu_C + R\nu_R + P\nu_P &= -G,\ (\nu_C, \nu_R, \nu_P) \geq 0,\ \nu_C \leq U \\ C^T y + s^1 - \rho &= 0,\ s^1 \geq 0,\ \rho \geq 0 \\ S^T y + s^2 &= 0,\ s^2 \geq 0 \\ P^T y + s^3 &= 0,\ s^3 \geq 0, \\ \nu_C^T s^1 + \nu_R^T s^2 + \nu_P^T s^3 &= 0 \\ \rho^T (U - \nu_C) &= 0 \\ (\nu_C, \nu_R, \nu_P) + (s^1, s^2, s^3) &> 0 \\ p + (U - \nu_C) &> 0. \end{aligned}$$

In theory, solving (1.18) with a path-following interior-point algorithm should return a solution that satisfies this system. However, numerical instabilities often lead to the failure of the last two conditions —i.e. path-following interior-point algorithms regularly return a degenerate solution instead of the strictly complementary solution they theoretically should. As an example, we solved the linear program that maximizes growth for the metabolic network of yeast with two popular interior solvers, CPLEX's barrier method (with crossover turned off) and PCx. Table 1.3 indicates the difference between theory and practice appears especially wide in this metabolic network. We point out that this problem does not address the linear program in (1.18) but instead solves (1.13) over $\mathcal{P}(20e)$. From a biological perspective 20 provides sufficient resources to achieve growth, so this problem is an adequate surrogate. What is important to observe from Table 1.3 is that even if variables greater 10^{-16} are declared positive, the metabolic network is at least 33% degenerate (the true extent of degeneracy would be the maximum number of degenerate pairs). Remember that these algorithms should

	Tolerance for Zero	
	10^{-8}	10^{-16}
CPLEX	552 / 1382 (40%)	376 / 1382 (27%)
PCx	606 / 1382 (44%)	457 / 1382 (33%)

Table 1.3. CPLEX's barrier method (with crossover turned off) and PCx were used to maximize the flow into the growth node over $P(20e)$. Although the solution should be strictly complementary, both solvers terminated with highly degenerate solutions.

theoretically provide a solution that is void of degeneracy, which highlights the fact that this problem has interesting and difficult numeric properties.

As the numerical results show, the theoretical value of an interior-point algorithm can be undermined by numerical instabilities. So, we offer a recent alternative that was born out of the necessity for researchers to overcome the same problem when investigating the optimal design of radiotherapy treatments [22]. The goal of this technique is to force variables to be positive by decreasing the largest values of a solution. When this is done iteratively, the result is called the *balanced solution*. To define this solution, we let $\text{sort}(x)$ be the function that sorts the elements of x and lists them in descending order. The balanced solution is defined as the unique solution to

$$\text{lexminsort} \equiv \text{lexmin}\{sort(x) : Ax = b, x \geq 0\},$$

where lexmin is the lexicographic minimum. It is easy to show that if λe is feasible and A and b are both positive, then the solution to this problem is λe, which means that this technique correctly identified that each component of x can be positive in a feasible solution.

Adapting this idea to the metabolic network, we have

$$\begin{aligned}\text{lexminsort} \equiv \text{lexmin}\{&\text{sort}(\nu_C, \nu_R, \nu_P) : \\ &C\nu_C + R\nu_R + P\nu_P = -G,\, U \geq \nu_C \geq 0,\, \nu_R \geq 0,\, \nu_P \geq 0\}.\end{aligned}$$

This technique of identifying the implied equalities is new and has not been thoroughly tested. An interesting direction for future research is to compare the speed and results of this method to those in [5, 15, 65]. We mention that there are interpretive advantages in this approach. For example, suppose that that largest value of this calculation is $\nu_i = l$. If $i \in C$, this indicates that reaction i must have flux l to achieve a unit of growth. Similar interpretations correspond to the cases of i being in R and P.

There are only a few mathematical results regarding the calculation of lexminsort. One of these is that the solution is unique, and we let x^* be this solution. Similar to the definition of B and N, we let $\beta = \{i : x_i^* > 0\}$ and $\eta = \{i : x_i^* = 0\}$. A desirable property would be for $B = \beta$ and $N = \eta$, however there are examples for which this is not the case. This means that x^* does not generally identify the optimal partition. Preliminary numerical studies have shown that it is often the case that x^* does induce the optimal partition, and the authors suggest that it is likely for metabolic networks. The insight comes from the fact that this method 'smooths out' the flux values by reducing the maximum flux, which in turn should cause other fluxes to increase.

There are several questions left to be answered about the linear program in FBA. As mentioned earlier, the constraints of this problem require that the fluxes adhere to a steady state assumption. However, a cell's state is dynamic rather than static. A major research direction is to use this technique to understand the how the fluxes change as the cell's environment changes. The environment is currently modeled through the cell's inputs, and asking how the the fluxes change is a question in classical sensitivity analysis. Since the solutions are significantly degenerate, a more appropriate question is how does B and N change with regards to the upper bound vector u. This question was studied for general linear programs in [1, 38, 40, 49, 67], but the special properties that exist in FBA are completely open. An alternative would be the modern interpretation of robust optimization, which provides complementary information to classic sensitivity analysis.

The steady state assumption prohibits metabolite accumulation. A more realistic model would allow metabolites to accrue and then have different reactions process these metabolites. However, we do not know what objective, if any, would eliminate the extra metabolites. One simple experiment would be to replace the constraints with

$$C\nu_C + R\nu_R + P\nu_P + G\nu_G \geq 0,\ 0 \leq \nu_C \leq U,\ 0 \leq \nu_R \leq u,\ 0 \leq \nu_P,$$

which relaxes the steady state assumption and allows metabolites to accumulate. Maximizing growth with this set of constraints will likely show that some metabolites remain in the cell. This is not realistic, so a secondary (or tertiary, etc.) objective is likely governing the elimination of metabolites. This re-casts FBA into the realm of multiple objective programming, which is likely more appropriate. This is an untapped research venue.

Another area where the degeneracy in FBA has been ignored is that of calculating the HFB. This calculation depends on an optimal solution from FBA, but the high level of degeneracy implies that the dimension

of the primal and/or the dual solution spaces is significant. Categorizing the source of degeneracy as primal or dual for each pair would add insight to the problem. Moreover, it would be interesting to know the variability of the HFB is over the optimal set, see [78] for related work.

Outside of FBA, we have from earlier sections that community identification is important. The algorithms used to identify communities need to be efficient due to the size of most biological networks. Mathematical programmers are trained in algorithm design and analysis, and these skills are needed. As previously mentioned, the BC measure used for many social networks is $O(N^3)$. This polynomial bound is typically considered favorable, but the cubic growth is realized in implementation, making this attack less attractive on large networks. The alternative based on (1.6) is $O(N^2)$. These are both significantly better than clique finding, which is a classic NP-complete problem.

The recent suggestion of k-clique percolation [57] was published without complexity analysis, which is understandable since the first step is to locate a k-clique, and hence, the algorithm is NP-complete. However, k is typically smaller than the size of the maximum clique, and identifying a small clique is generally considered simple. This begs the question, What is the complexity of identifying a community from a known k-clique? A simple argument shows that the algorithm in Table 1.4 locates a community in $O(\Delta^k N^2)$, where $\Delta = \max\{\deg(v) : v \in V\}$. Since $\Delta \leq N$, we generally have the possibility of $O(N^{k+2})$, which is polynomial for fixed k but is worse than both of the other algorithms since $k \geq 2$. The numerical computations in [57] do not indicate that this bound is achieved in practice, and an interesting direction for future research would be to explain the difference between the theoretical complexity and the practical efficiency.

The traditional clustering techniques of k-means and k-medians can also be used to identify communities. Both of these problems are traditional facility location problems in OR and their application to biological networks deserves attention. Although facility location is related to community finding, it is inherently different. This is because facility location is concerned with locating positions that optimize some quality of an assignment to these positions. So these problems have the two goals of grouping entities and assigning a representative to each group, which is often (but not necessarily) a member of the group. The community idea equates nicely to grouping, but how the representative part leads to biological information is not known. We discuss the recent results of [39] to foreshadow some future applications of the k-median problem in systems biology.

1 Let $\mathcal{C}$ be the nodes of an initial k-clique.

2 Set $\mathcal{C}' = \emptyset$.

3 For each $v \in \mathcal{C}$:

  a For each $v' \in N(v) \backslash \mathcal{C}$:

    i If $|N(v') \cup N(v) \cup \mathcal{C}| \geq k-2$, add nodes $N(v') \cup N(v) \cup \mathcal{C}$ to $\mathcal{C}'$.

4 If $\mathcal{C}' \neq \emptyset$, let $\mathcal{C} = \mathcal{C} \cup \mathcal{C}'$ and go to 2.

Table 1.4. An algorithm to calculate a community from a known k-clique.

The k-median problem is one of the 4 primary questions in discrete location theory (the others being the k-means problem, the uncapacitated facility location problem, and the quadratic assignment problem). Initial investigations into the problem were undertaken by Hakimi [34], and this work spawned a substantial literature [66]. Hakimi's original intent was to locate positions from the continium of a network or graph —i.e. facilities were allowed to be positioned on an edge or vertex. This is a graph restriction of the classic Weber problem. Assuming that positions on the graph were related by a metric, Hakimi proved two significant results: 1) There is always an optimal facility location for which the facilities are located at vertices and 2) The problem of optimally locating facilities is NP-hard in N and k. An often overlooked and misunderstood property is that the problem is polynomial for a fixed k, making it fixed-parameter tractable.

The discrete k-median problem is concerned with selecting k positions on a graph from a discrete set $\mathbb{P}$ on (V, E). The positions in $\mathbb{P}$ can be located on any edge or vertex and it is assumed that $V \subseteq \mathbb{P}$. Each pair of positions is related by a nonnegative similarity score $d(p, p')$, which need not be a metric, and each node is assigned a weight $\beta(v)$. The discrete k-median problem is

$$\min \left\{ \sum_{p \in \mathbb{P}} \sum_{v \in V_p} d(p, v)\beta(v) : \mathbb{P} \subseteq (V, E),\ |\mathbb{P}| = k \right\},$$

where

$$V_{p'} = \{v \in V : d(v, p') \leq d(v, p) \ \text{ for } \ p \in \mathbb{P}\}. \tag{1.19}$$

Any collection of k positions solving this problem are called medians. The nearest neighbor condition in (1.19) assigns the vertices of the graph to the medians, but unfortunately, this definition does not uniquely define $V_{p'}$ since some nodes may be equally similar to multiple medians. However, the assumption that $|\mathbb{P}| \leq |\mathbb{N}|$ allows us to list the elements of $\mathbb{P}$, and subsequently to decide ties by assigning the vertex to the position with the least index. A result in [39] similar to Hakimi's original work shows that there is always a solution of vertices.

With regards to community location, it makes sense that $\mathbb{P} = V$. However, although the similarity measure d and the node weight β are natural in many OR applications, their interpretation in a biological framework is not clear. Indeed, the communities are defined in terms of these graph characteristics, and it is likely that they can be tailored to different biological situations, yielding a flexible model. In the discussions that follow, we assume that $\mathbb{P} = V$, $d(p, p') = \|p - p'\|_2$ and $\beta(v) = 1$. The use of the Euclidean norm implies the network is coordinitized in some meaningful way, which is awkward for biological networks. However, it is a place to start.

The main result of [39] shows that the discrete k-median problem is identical to a well studied problem in data compression that optimally designs a vector quantizer. A full discussion of vector quantization is not warranted due to space limitation, and we direct interested readers to [28]. The importance of the relationship is that it allows us to cast the graph theory problem in a way that is amenable to the efficient algorithms designed to work on the vector quantization problem. The most preeminent and significant of these techniques is the discrete Lloyd algorithm in Table 1.5. This algorithm is not an exact solution procedure since it converges to a local optimal solution. The pertinent complexity results from [39] are

- The discrete k-median problem is $O(N^{k+2})$, and
- The discrete Lloyd algorithm is $O(Nk)$.

The first result shows that the worst case complexity of the discrete k-median problem is no worse than that of the k-clique percolation's. Since $k << N$, the second result shows that the discrete Lloyd algorithm is theoretically faster than the other community finding techniques.

Using the discrete k-median problem to locate communities within a biological network is promising. The questions are numerous and include

- What similarity measure and node weight are meaningful?

1 Select an initial collection of k nodes, M.

2 Calculate the nearest neighbors V_v as in (1.19) for each $v \in M$.

3 Calculate the centroid of V_v for each $v \in M$ with each node weighted with $\beta(v)$.

4 Project each centroid onto its nearest neighbor in V forming a new collection of k nodes denoted by M'.

5 If $M = M'$, stop. Otherwise, replace M with M' and go to 2.

Table 1.5. The discrete Lloyd algorithm for $\mathbb{P} = V$.

- Can a solution to the discrete k-median problem be found as efficiently as communities can be found with k-clique percolation in practice?
- Does the discrete Lloyd algorithm outperform other community finding algorithms in practice?
- How do we initialize the discrete Lloyd algorithm so that it locates a global solution instead of a local solution?

We close by mentioning that although we have focused on the linear optimization problem associated with FBA and the community finding algorithms that identify biological structures, these are but two of the many problems in systems biology that make use of standard OR techniques. The purpose of this section was to show that the problems are plentiful, important, and natural, and we encourage the involvement of the OR community. Please contact the authors if we can be of assistance.

Acknowledgments

The authors would like to thank Trinity University's Mathematics department for their generous support. We also thank Jeremy Nolan for his discussions on the complexity of k-clique percolation.

References

[1] I. Adler and R. Monteiro. A geometric view of parametric linear programming. *Algorithmica*, 8:161–176, 1992.

[2] R. Albert and A.-L. Barabási. Statistical mechanics of complex networks. *Rev. Mod. Phys.*, 74:47, 2002.

[3] R. Albert, H. Jeong, and A.-L. Barabási. Error and attack tolerance of complex networks. *Nature*, 406:378, 2000.

[4] E Almaas, B. Kovács, T. Vicsek, Z. N. Oltvai, and A.-L. Barabási. Global organization of metabolic fluxes in the bacterium *escherichia coli*. *Nature*, 427:839, 2004.

[5] E. Almaas, Z. N. Oltvai, and A.-L. Barabási. The activity reaction core and plasticity in metabolic networks. *PLoS Comput. Biol.*, 1:e68, 2005.

[6] A.-L. Barabási and R. Albert. Emergence of scaling in random networks. *Science*, 286:509, 1999.

[7] A Barrat, M Barthelemy, R Pastor-Satorras, and A Vespignani. The architecture of complex weighted networks. *Proc. Natl. Acad. Sci USA*, 101:3747, 2004.

[8] M. Barthelemy, B. Gondran, and E. Guichard. Spatial structure of the internet traffic. *Physica A*, 319:633, 2003.

[9] N. N. Batada, L. D. Hurst, and M. Tyers. Evolutionary and physiological importance of hub proteins. *PLoS Comp. Biol.*, 2:0748, 2006.

[10] BioGrid. version 2.0.20, 2006. http://www.thebiogrid.org/.

[11] L M Blank, L Kuepfer, and U Sauer. Large-scale c-13-flux analysis reveals mechanistic principles of metabolic network robustness to null mutations in yeast. *Genome Biol.*, 6:R49, 2005.

[12] B. Bollobás. *Random Graphs*. Academic Press, New York, 2001.

[13] H. P. J. Bonarius, G. Schmid, and J. Tramper. Flux analysis of underdetermined metabolic networks: The quest for the missing constraints. *Trends Biotechnol.*, 15:308, 1997.

[14] U. Brandes. A faster algorithm for betweenness centrality. *J. Math. Soc.*, 25:163, 2001.

[15] A P Burgard, S Vaidyaraman, and C D Maranas. Minimal reaction sets for *escherichia coli* metabolism under different growth requirements and uptake environments. *Biotechnol. Progr.*, 17:791, 2001.

[16] C. Cannizzaro, B. Christensen, J. Nielsen, and U. von Stockar. Metabolic network analysis on *phaffia rhodozyma* yeast using c-13-labeled glucose and gas chromatography-mass spectrometry. *Metab. Eng.*, 6:340, 2004.

[17] F Canonaco, T A Hess, S Heri, T T Wang, T Szyperski, and U Sauer. Metabolic flux response to phosphoglucose isomerase

knock-out in *escherichia coli* and impact of overexpression of the soluble transhydrogenase Udha. *FEMS Microbiol. Lett.*, 204:247, 2001.

[18] S. Coulomb, M Bauer, D. Bernard, and M. C. Marsolier-Kergoat. Gene essentiality and the topology of protein-interaction networks. *Proc R Soc Lond Ser Biol Sci*, 272:1721, 2005.

[19] S. N. Dorogovtsev, A. V. Goltsev, and J. F. F. Mendes. Pseudofractal scale-free web. *Phys. Rev. E*, 65:066122, 2002.

[20] N C Duarte, M J Herrgard, and B O Palsson. Reconstruction and validation of *saccharomyces cerevisiae* ind750, a fully compartmentalized genome-scale metabolic model. *Genome. Res.*, 14:1298, 2004.

[21] J S Edwards and B O Palsson. The *escherichia coli* mg1655 in silico metabolic genotype: Its definition, characteristics, and capabilities. *Proc. Natl. Acad. Sci USA*, 97:5528, 2000.

[22] M. Ehrgott, A. Holder, and J. Reese. Beam selection in radiotherapy design. Technical Report 95, Trinity University Mathematics, San Antonio, TX, 2005.

[23] M Emmerling, M Dauner, A Ponti, J Fiaux, M Hochuli, T Szyperski, K Wuthrich, Bailey J E, and Sauer U. Metabolic flux responses to pyruvate kinase knockout in *escherichia coli*. *J. Bacteriol.*, 184:152, 2002.

[24] E. Fischer and U. Sauer. Metabolic flux profiling of *escherichia coli* mutants in central carbon metabolism using GC-MS. *Eur. J. Biochem.*, 270:880891, 2003.

[25] E Fischer and U Sauer. Large-scale in vivo flux analysis shows rigidity and suboptimal performance of *bacillus subtilis* metabolism. *Nat. Genet.*, 37:636, 2005.

[26] L. C. Freeman. A set of measures of centrality based upon betweenness. *Sociometry*, 40:35, 1977.

[27] S.Y. Gerdes, M.D. Scholle, J.W. Campbell, G. Balazsi, E. Ravasz, M. D. Daugherty, A. L. Somera, N. C. Kyrpides, I. Anderson, M. S. Gelfand, A. Bhattacharya, V. Kapatral, M. D'Souza, M. V. Baev, Y. Grechkin, F. Mseeh, M. Y. Fonstein, R. Overbeek, A.-L. Barabási, Z. N. Oltvai, and A. L. Osterman. Experimental determination and system level analysis of essential genes in *escherichia coli* mg1655. *J. Bact.*, 185:5673, 2003.

[28] A. Gersho and R. M. Gray. *Vector Quantization and Signal Compression*. Kluwer International Series in Engineering and Computer Science. Kluwer Academic Publishers, 1992.

[29] G. Giaever, A.M. Chu, L. Ni, C. Connelly, L. Riles, S. Véronneau, S. Dow, A. Lucau-Danila, K. Anderson, B. André, A.P. Arkin, A. Astromoff, M. El-Bakkoury, R. Bangham, R. Benito, S. Brachat, S. Campanaro, M. Curtiss, Karen Davis, Adam Deutschbauer, Karl-Dieter Entian, Patrick Flaherty, Francoise Foury, David J. Garfinkel, Mark Gerstein, Deanna Gotte, Ulrich Güldener, Johannes H. Hegemann, Svenja Hempel, Zelek Herman, Daniel F. Jaramillo, Diane E. Kelly, Steven L. Kelly, Peter Kötter, Darlene LaBonte, David C. Lamb, Ning Lan, Hong Liang, Hong Liao, Lucy Liu, Chuanyun Luo, Marc Lussier, Rong Mao, Patrice Menard, Siew Loon Ooi, Jose L. Revuelta, Christopher J. Roberts, Matthias Rose, Petra Ross-Macdonald, Bart Scherens, Greg Schimmack, Brenda Shafer, Daniel D. Shoemaker, Sharon Sookhai-Mahadeo, Reginald K. Storms, Jeffrey N. Strathern, Giorgio Valle, Marleen Voet, Guido Volckaert, Ching yun Wang, Teresa R. Ward, Julie Wilhelmy, Elizabeth A. Winzeler, Yonghong Yang, Grace Yen, Elaine Youngman, Kexin Yu, Howard Bussey, Jef D. Boeke, Michael Snyder, Peter Philippsen, Ronald W. Davis, and Mark Johnston. Functional profiling of the *saccharomyces cerevisiae* genome. *Nature*, 418:387, 2002.

[30] A. Goldman and A. Tucker. Theory of Linear Programming. In H. Kuhn and A. Tucker, editors, *Linear Inequalities and Related Systems*, volume 38, pages 53–97. Princeton University Press, Princeton, New Jersey, 1956.

[31] A K Gombert, M M dos Santos, B Christensen, and J Nielsen. Network identification and flux quantification in the central metabolism of *saccharomyces cerevisiae* under different conditions of glucose repression. *J. Bacteriol.*, 183:1441, 2001.

[32] H. Greenberg. Consistency, redundancy and implied equalities in linear systems. *Annals of Mathematics and Artificial Intelligence*, 17:37–83, 1996.

[33] H.J. Greenberg. *Mathematical Programming Glossary*. World Wide Web, `http://glossary.computing.society.informs.org/`, 1996–2006. Edited by the INFORMS Computing Society.

[34] S. L. Hakimi. Optimum distribution of switching centers in a communication network and some related graph theoretic problems. *Operations Research*, 13(3):462–475, 1965.

[35] Jing-Dong J. Han, Nicolas Bertin, Tong Hao, Debra S. Goldberg, Gabriel F. Berriz, Lan V. Zhang, Denis Dupuy, Albertha J. M. Walhout, Michael E. Cusick, Frederick P. Roth, and Marc Vidal.

Evidence for dynamically organized modularity in the yeast protein-protein interaction network. *Nature*, 430:88, 2004.

[36] L. H. Hartwell, J. J. Hopfield, S. Leibler, and A. W. Murray. From molecular to modular cell biology. *Nature*, 402:C47, 1999.

[37] Xionglei He and Jianzhi Zhang. Why do hubs tend to be essential in protein networks? *PLoS Genet.*, 2:0826, 2006.

[38] A. Holder. Simultaneous data perturbations and analytic center convergence. *SIAM J. on Optimization*, 14(3):841–868, 2004.

[39] A. Holder, G. Lim, and J. Reese. The relationship between discrete vector quantization and the p-median problem. Technical Report 102, Trinity University Mathematics, San Antonio, TX, 2006.

[40] A. Holder, J. Sturm, and S. Zhang. Marginal and parametric analysis of the central optimal solution. *Information Systems and Operational Research*, 39(4):394–415, 2001.

[41] P. Holme, S. M. Park, B. J. Kim, and C. R. Edling. Korean university life in a network perspective: Dynamics of a large affiliation network. *Physica A*, 373:821, 2007.

[42] B. Jansen, J.J. de Jong, C. Roos, and T. Terlaky. Sensitivity analysis in linear programming: Just be careful! *European Journal of Operations Research*, 101:15–28, 1997.

[43] H. Jeong, S. Mason, A.-L. Barabási, and Z. N. Oltvai. Lethality and centrality in protein networks. *Nature*, 411:41, 2001.

[44] K. J. Kauffman, P. Prakash, and J. S. Edwards. Advances in flux balance analysis. *Curr. Opin. Biotechnol.*, 14:491, 2003.

[45] P.J. Macdonald, E. Almaas, and A.-L. Barabási. Minimum spanning trees on weighted scale-free networks. *Europhys. Lett.*, 72:308, 2005.

[46] S. Maslov and K. Sneppen. Specificity and stability in topology of protein networks. *Science*, 296:910, 2002.

[47] R Milo, S Itzkovitz, N Kashtan, R Levitt, S Shen-Orr, I Ayzenshtat, M Sheffer, and U Alon. Superfamilies of evolved and designed networks. *Science*, 303:1538, 2004.

[48] R Milo, S S Shen-Orr, S Itzkovitz, N Kashtan, D Chklovskii, and U Alon. Network motifs: Simple building blocks of complex networks. *Science*, 298:824, 2002.

[49] R. Monteiro and S. Mehrotra. A General Parametric Analysis Approach and Its Implication to Sensitivity Analysis in Interior Point Methods. *Mathematical Programming*, 72:65–82, 1996.

[50] M. E. J. Newman. Scientific collaboration networks: Ii. shortest paths, weighted networks, and centrality. *Phys. Rev. E*, 64:016132, 2001.

[51] M. E. J. Newman. Assortative mixing in networks. *Phys. Rev. Lett.*, 89:208701, 2002.

[52] M. E. J. Newman. Mixing patterns in networks. *Phys. Rev. E*, 67:026126, 2003.

[53] M. E. J. Newman. The structure and function of complex networks. *SIAM Review*, 45:167, 2003.

[54] M. E. J. Newman. Fast algorithm for detecting community structure in networks. *Phys. Rev. E*, 69:066133, 2004.

[55] J.-P. Onnela, J. Saramäki, J. Kertész, and K. Kaski. Intensity and coherence of motifs in weighted complex networks. *Phys. Rev. E*, 71:065103, 2005.

[56] C Pal, B Papp, M J Lercher, P Csermely, S G Oliver, and L D Hurst. Chance and necessity in the evolution of minimal metabolic networks. *Nature*, 440:667, 2006.

[57] G. Palla, I. Derényi, I. Farkas, and T. Vicsek. Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, 435:814, 2005.

[58] J. A. Papin, J. Stelling, N. D. Price, S. Klamt, S. Schuster, and B. O. Palsson. Comparison of network-based pathway analysis methods. *Trends Biotechnol.*, 22:400, 2004.

[59] B Papp, C Pal, and L D Hurst. Metabolic network analysis of the causes and evolution of enzyme dispensability in yeast. *Nature*, 429:661, 2004.

[60] R. Pastor-Satorras, A. Vázquez, and A. Vespignani. Dynamical and correlation properties of the internet. *Phys. Rev. Lett.*, 87:258701, 2001.

[61] J.B. Pereira-Leal, A.J. Enright, and C.A. Ouzounis. Detection of functional modules from protein interaction networks. *Proteins*, 54:49, 2004.

[62] J. F. Poyatos and L. D. Hurst. How biologically relevant are interaction-based modules in protein networks? *Genome Biol.*, 5:R93, 2004.

[63] D. J. de S. Price. Networks of scientific papers. *Science*, 149:510, 1965.

[64] E. Ravasz, A. L. Somera, D. A. Mongru, Z. N. Oltvai, and A.-L. Barabási. Hierarchical organization of modularity in metabolic networks. *Science*, 297:1551, 2002.

[65] J. L. Reedi and B. O. Palsson. Genome-scale in silico models of *e. coli* have multiple equivalent phenotypic states: Assessment of

correlated reaction subsets that comprise network states. *Genome Res.*, 14:1797, 2004.

[66] J. Reese. Solution methods for the p-median problem: An annotated bibliography. *Networks*, 48(3):125–142, 2006.

[67] C. Roos, T. Terlaky, and J.-Ph. Vial. *Theory and Algorithms for Linear Optimization: An Interior Point Approach*. John Wiley & Sons, New York, NY, 1997.

[68] U Sauer, D R Lasko, J Fiaux, M Hochuli, R Glaser, T Szyperski, K Wuthrich, and J E Bailey. Metabolic flux ratio analysis of genetic and environmental modulations of *escherichia coli* central carbon metabolism. *J. Bacteriol.*, 181:6679, 1999.

[69] C. H. Schilling, Markus W. Covert, Iman Famili, George M. Church, Jeremy S. Edwards, and Bernhard O. Palsson. Genome-scale metabolic model of *helicobacter pylori* 26695. *J. Bacteriol.*, 184:4582, 2002.

[70] C. H. Schilling, D. Letscher, and B. O. Palsson. Theory for the systemic definition of metabolic pathways and their use in interpreting metabolic function from a pathway-oriented perspective. *J. Theor. Biol.*, 203:229, 2000.

[71] S. Schuster and C. Hilgetag. On elementary flux modes in biochemical reaction systems at steady state. *J. Biol. Syst.*, 2:165, 1994.

[72] P. Schwikowski B., Uetz and S. Fields. A network of protein-protein interactions in yeast. *Nat. Biotechnol.*, 18:1257, 2000.

[73] D Segre, A DeLuna, G M Church, and R Kishony. Modular epistasis in yeast metabolism. *Nat. Genet.*, 37:77, 2005.

[74] S. S. Shen-Orr, R. Milo, S Mangan, and U Alon. Network motifs in the transcriptional regulation network of *escherichia coli*. *Nat. Genet.*, 31:61, 2002.

[75] A. Vazquez, R. Dobrin, D. Sergi, J.-P. Eckman, Z. N. Oltvai, and A.-L. Barabási. The topological relationship between the large-scale attributes and local interaction patterns of complex networks. *Proc Natl Acad Sci USA*, 101:17940, 2004.

[76] S. Wasserman and K. Faust. *Social Network Analysis*. Cambridge University Press, Cambridge, 1994.

[77] D.J. Watts and S. H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393:440, 1998.

[78] S. Wiback, I. Famili, H. Greenberg, and B. Palsson. Monte carlo sampling can be used to determine the size and shape of the steady state flux space. *Journal of Theoretical Biology*, 228:437–447, 2004.

[79] S. Wuchty, Z.N. Oltvai, and A.-L. Barabási. Evolutionary conservation of motif constituents in the yeast protein interaction network. *Nat. Genet.*, 35:176, 2003.

[80] B. Zhang and S. Horvath. A general framework for weighted gene co-expression network analysis. *Stat. App. Genet. Mol. Biol.*, 4:17, 2005.