Part One
Modeling, Simulation, and Meaning of Gene Networks

Applied Statistics for Network Biology: Methods in Systems Biology, First Edition.
Edited by M. Dehmer, F. Emmert-Streib, A. Graber, and A. Salvador.
© 2011 Wiley-VCH Verlag GmbH & Co. KGaA. Published 2011 by Wiley-VCH Verlag GmbH & Co. KGaA.

1
Network Analysis to Interpret Complex Phenotypes

Hong Yu, Jialiang Huang, Wei Zhang, and Jing-Dong J. Han

1.1
Introduction

Gene network analysis is an important part of systems biology studies. Compared with traditional genotype/phenotype studies that focused on establishing the relationships between single genes and interested traits, network analysis give us a global view of how all the genes work together properly, which in turn leads to the correct biological functions [1].

Unlike the Mendelian "one gene–one phenotype" relationship, C.H. Waddington in 1957 came up with the "epigenetic landscape" to visually illustrate the multigene or network effects of genes on shaping the landscapes (various states) of cellular metabolism. Given our current knowledge, "cellular metabolism" in Waddington's landscapes model can be extended to "molecular networks," which turn steady states into network representations or snapshots. Such steady states and the transitions from one steady state to another have been computationally analyzed through simulated networks [2–4] and experimentally validated by checking gene expression profiles during proliferation/differentiation transitions, gene mutation perturbations, or environmental or physical stresses [5, 6]. The transition from one stable state to another is usually related to complex phenotypes, which could be both physiological and pathological, such as diabetes mellitus or cancerous proliferation (Figure 1.1) [7]. Gene function is not isolated, so we could not study their function separately. Not only the function of the individual gene products, but also their interaction with each other, which is increasingly more important to the success of higher organisms, determines the selective advantage of the genes and the networks they formed.

What can network analysis do? Here, we mainly talk about given a gene network, mostly validated by experiments, what information could be got from it? How could we understand the biological process with the help of a network? Basically, there are three aspects. The most traditional aspect is to identify the importance of each node in the network (e.g., which genes are more important or crucial, which genes are less

Applied Statistics for Network Biology: Methods in Systems Biology, First Edition.
Edited by M. Dehmer, F. Emmert-Streib, A. Graber, and A. Salvador.
© 2011 Wiley-VCH Verlag GmbH & Co. KGaA. Published 2011 by Wiley-VCH Verlag GmbH & Co. KGaA.

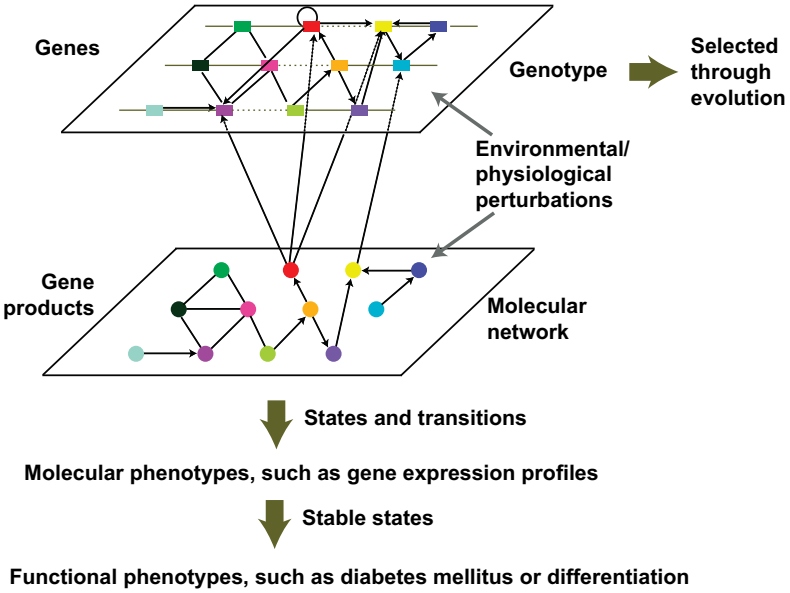


Figure 1.1 Complex phenotypes are determined by the steady state of the molecular network. A molecular network is encoded by the genetic network. The interplay of molecules in the network as well as their interactions with the environment and developmental cues determine the stable states of the network, which ultimately determines the phenotypes reflected by the system. (Adapted from [7].)

important or dispensable). Another aspect is to identify which genes are more functionally related through the whole network view, not only by measuring the direct connections, but also by considering the connections through the whole network. In this way, we could establish functional relationships between all the genes by protein–protein interaction networks or other kinds of experimentally validated networks. More recent studies have focused on identifying the paths or flows through the networks with known input and output genes. These methods could identify the unknown mediated genes and also identify which genes are more important in these processes. All these different aspects could serve well in understanding human diseases at different level and views. We will start by discussing these three aspects in detail, including some methods related to them, but not limited in pure network analysis in later sections.

Before we begin to talk about network analysis, we first explain several definitions that are very basic, but will be frequently mentioned in the following parts.

A network N consists of a set $V(N)$ of vertices (or nodes) together with a set $E(N)$ of edges (or links) that connect various pairs of vertices. Usually, nodes represent genes or proteins and edges represent interactions.

A network N is a weighted network if each of its edges has a number associated with it indicating the strength of the edge. Usually, the edge weights represent the confidences of interactions in biological experiments.

A network N is called a directed network if all of its edges are directed and a network N is called an undirected network if none of its edges is directed. Usually, signaling networks and transcriptional regulatory networks could be directed networks whose directions indicate signal transduction or transcriptional regulation.

For any network N and any particular vertex v in $V(N)$, the number of vertices v' in $V(N)$ that are directly linked to v is called the degree of v.

In particular, for any directed network N and any particular vertex v in $V(N)$, the number of vertices v' in $V(N)$ that are directly linked to v by an inward-pointing edge to v is called the in-degree of v and the number of vertices v' in $V(N)$ that are directly linked to v by an edge pointing outward from v is called the out-degree of v.

The minimum number of edges that must be traversed to travel from a vertex v to another vertex v' of a network N is called the shortest path length between v and v'. For any connected network N, the average shortest path length between any pair of vertices is called the network's "characteristic path length" (CPL).

1.2 Identification of Important Genes based on Network Topologies

Identification of important genes in biological processes is one of the most common and important aspects in all kinds of biology studies [8, 9]. The basic idea to achieve this goal in biological networks is to measure the influence or damage to the network by perturbing certain genes [10]. If removing a gene from a network leads to small changes or influences, this gene should be less important in maintaining the correct function of the biological network. In contrast, if it leads to the collapse or a large influence on the network, such as dividing the whole network into two subnetworks, this gene might play a crucial role in biological processes. This hypothesis has been increasingly supported by experimental data showing that genes with higher influences on the network were more lethal, more conserved through evolution, and basically more important in maintaining biological functions [11]. In order to evaluate genes' importance, several different measurements could be used due to different considerations.

1.2.1 Degree

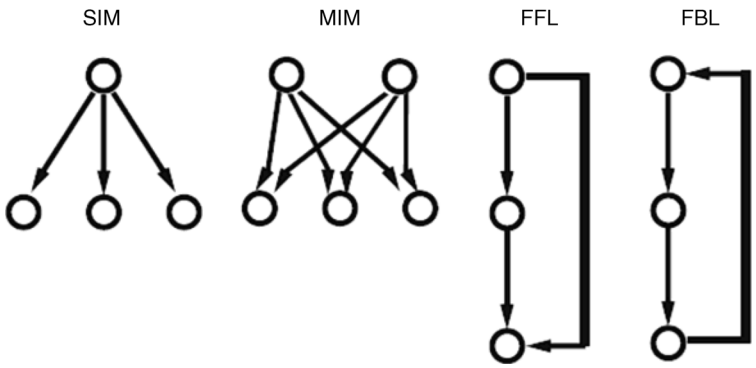
The most intuitive consideration is that the more edges are removed, the more damage is taken by the network. Thus, the genes with high degrees, known as hubs in the network, should be more important. Evidence has shown that the perturbation of hubs leads to a more dramatic increase of CPL in a biological network than random perturbations [12]. Besides, other information could be further used, such as gene expression data, to find "date hubs" and "party hubs," which indicate different biological functions [12].

1.2.2 Betweenness

The centrality or connectivity of a network can be measured by the CPL. In biological networks, the CPL indicates the speed of signal transduction or the quickness of biological response. Thus, another consideration of a gene's importance is the CPL changes when perturbing it. These changes could be measured directly by recalculating the CPL when removing each gene from the network or indirectly using the betweenness of each gene. The betweenness of a vertex v is calculated as the number of shortest paths that pass through it divided by the number of all shortest paths. Compared with betweenness, recalculating the CPL is more accurate, but more time consuming. In fact, a very high correlation exists between the CPL recalculating results and the betweenness measurements, so basically measuring the betweenness of a gene is sufficient to see its influence on the CPL. We could also easily see that a gene with high betweenness is not necessarily a hub or has a very high degree, but in view of the whole gene set, betweenness does correlate with degree.

1.2.3 Network Motifs

Compared with the former two measurements, which could be applied to any kinds of networks, network motifs are usually employed in directed networks, such as transcriptional regulatory networks or transcription factor target networks. Network motifs could be regarded as the basic blocks to form the whole network [13], and they were shown to be important in maintaining robustness, perturbation buffering, quick responses, and accurate signal transductions [13–15]. Thus, the genes that take part in multiple network motifs should be more important and counting network motifs becomes one measurement for evaluating the importance of genes. Here, we introduce several commonly used network motifs (Scheme 1.1).



Scheme 1.1 Several commonly used network motifs.

- Single-input motifs (SIM): a group of nodes regulated by a single node without any other regulation.
- Multi-input motifs (MIM): a group of nodes regulate another group of nodes together.
- Feed-forward loops (FFL): a node regulates another and then these two nodes regulate a third one together.
- Feed-back loops (FBL; also known as a multicomponent loops (MCL): an upstream node is regulated by a downstream one.

In biological networks, genes in SIMs or MIMs usually determine the bottleneck of the network, which possibly indicates that the deletion or mutation of these genes is likely to cause lethal influences. FFLs and FBLs could enable precise control or quick response, which was precisely required in biological processes and responses. Network motifs are not limited to those mentioned above, but all the motifs that have been proved to have biological meanings. By searching for different kinds of network motifs, we could find important genes for certain functions that we are interested in.

1.2.4 Hierarchical Structure

In signal transduction networks or transcriptional regulatory networks, genes can be divided into several layers and the signals flow from top to bottom (with feedback allowed). This kind of structure is called a hierarchical structure. Apart from the degree and network motifs, genes on different layers or having different offspring nodes (regulated by this gene) could provide information on understanding biological processes [16].

These network topology-based analyses have been widely used in identifying important genes in multiple studies of different species. However, some other cautions should be announced in all of these measurements besides the fact that they are based on different considerations. First, it is hard to consider the combinatorial influence of the genes, such as when removing either one of two genes with very similar connections, the network will not be badly influenced because there is a backup gene, but when removing both of them, the whole network will collapse. Backup genes exist widely in real biological processes to ensure the robustness of organisms. Currently, it is possible to detect these combinatorial effects through applying newly developed IT methods, although calculations may be very time-consuming. Another problem is that the qualities of networks negatively influence the results, especially when the edges in the networks are biased. This does happen, especially in human studies. For instance, when using literature-supported protein–protein interactions (PPIs), the "hot" genes or interesting genes are much more intensively studied than the "cold" genes and they are more likely to be hubs, because most of their interactions are discovered, while for the "cold" genes, most of their interactions are unknown.

1.3 Inferring Information from Known Networks

1.3.1 Understanding Biological Functions based on Network Modularity

The existence of modular structures (clusters of tightly connected subnetworks) has been noticed in various biological networks. In biological networks, these modules often indicate particular biological functional processes [17, 18]. The modules can be identified by various algorithms, such as the Lin Log energy model (http://www.informatik.tu-cottbus.de/~an/GD/linlog.html), the MCODE algorithm (http://baderlab.org/Software/MCODE), and the Markov Clustering algorithm (http://www.micans.org/mcl/). Then, by examining the modules' enriched Gene Ontology (GO) terms, KEGG (Kyoto Encyclopedia of Genes and Genomes) pathways, and other functional annotations, we can discover their biological functions.

1.3.2 Inferring Functional Relationships and Novel Functional Genes Through Networks

In the past few years, more and more studies have focused on identifying functional relationships between genes. These studies came from the collaborations of human association studies and gene function prediction studies. These methods aim to identify unknown disease-related genes with a candidate list derived from association studies. Usually these methods include not only PPIs, but also many other kinds of information, which could be summarized into different kinds of edges. The basic idea is that genes sharing similar functions are usually highly connected in PPI networks. Thus, in order to identify novel disease-related genes from a candidate list, we just need to find the known genes with similar phenotypes in PPI networks.

Several studies analyzed Online Mendelian Inheritance in Man (OMIM) data using PPI and description similarity between genes and phenotypes, which is the result based on human association studies over recent decades [19, 20]. With the development of new technologies, more and more association studies have been finished on large populations and specific phenotypes at high coverage and high resolution levels. These genome-wide association studies (GWAS) provided opportunities for the application of all these methods. As the integration of different kinds of networks could be seen as a whole weighted network with different weights on different edges, we would mainly introduce one method with wide applications and a good computational performance, which is based on the random walk algorithm [21].

The random walk on graphs is defined as an iterative walker's transition from its current node to all its neighbors through all weighted edges starting at given source nodes, s. Each source node could take a different weight and basically the sum value could be normalized to 1, so this value could also be considered as the probability of the information transition through the whole network. Here, compared to the traditional random walk, it added another restart process that in every step, the signal restarts at node s with a probability r. It indicated that in every step of transition,

only $(1 - r)$ of total information is continuously transitioned, with r of total restart. The goal of this method is to add a continuous input and when the stable status is achieved, all the other nodes have a stable proportion of information to be output, the sum of which is r.

Formally, the random walk with restart is defined as:

$$P^{t+1} = (1-r)*W*P^t + r*P^0$$

where W is a matrix that is based only on the network topology; basically, it is the column-normalized adjacency matrix, each none zero value represents the weight of one edge in the network. P^t is a vector in which each element holds the probability of information on a node at step t. In this application, the initial probability vector P^0 was constructed as weighted probabilities where each probability represents the influence of a source gene on the disease we are interested in, with the sum of these probabilities equal to 1. When the difference between P^t and P^{t+1} is smaller than an arbitrarily given threshold, the steady-state P^N was obtained and considered as the result. Candidate disease-related genes are then ranked according to the values in P^N.

The performance of the random walk algorithm was shown to be better than the previous algorithms. Also, this algorithm is easily applied. One obvious benefit of this method is that P^N is additive, which makes this algorithm very convenient. Take one simple example, consider the steady state P^N of only one source node A or B to be $P^N(A)$ or $P^N(B)$. When we want to consider the combinatorial effect of A and B, we can apply the weighted probabilities of the two source nodes as a and $(1 - a)$, and the steady state P^N of using both A and B as source nodes could be simply calculated as $P^N(AB) = a * P^N(A) + (1 - a) * P^N(B)$. This formula could be extended to a set s of multiple source genes. Thus, basically, for a certain network, we do not have to recalculate P^N for each set of source genes. Instead, we could calculate each source gene individually and sum the weighted results. In this algorithm, different r indicates different affinity. High r indicates more influence of input genes and less transition in the network, while low r leads to more transition steps. Empirically, the stable result could be obtained within 30–50 steps considering different r and thresholds used, and the algorithm is not very time-consuming. Thus, it is possible to calculate P^N of each gene in a network.

As mentioned above in Section 1.2, all of these algorithms are negatively influenced by the quality of networks and those "hot" genes. We were very likely to be stuck in those "hot" genes if a biased network was used.

1.3.3
Unraveling Transcriptional Regulations from Expression Data through Transcriptional Networks

Transcription factors play a crucial regulatory role in various biological processes; however, they are unlikely to be detected from expression data due to their low, and often sparse, expression. To fill this gap, Reverter *et al.* proposed a regulatory impact factor (RIF) algorithm to identify critical transcription factors from gene expression data by integrating coexpression networks [22]. RIF analysis assigns a score to each

transcription factor by considering both the correlation between the transcription factor and the differentially expressed genes and the expression level of the differentially expressed genes. In particular, for a given functional module, its potential regulators are scored by their absolute coexpression correlation averaged across all genes in the module [23].

1.3.4 Extracting the Pathway-Linked Regulators and Effectors based on Network Flows

Recently, high-throughput techniques have been widely used to detect the potential components of biological networks. So far, these high-throughput techniques cover two classes: (i) genetic screens including overexpression, deletion, or RNA interference library screens and (ii) mRNA profiling using microarray or RNA sequencing technology. By comparing the results of these two methods, Yeger-Lotem *et al*. found that genetic screens tend to identify regulators that are critical for the cell response, while the differentially expressed genes identified by mRNA profiling are likely their downstream effectors, whose changes indirectly reflect the genetic changes in the regulatory networks [24]. It is also true in diseases; using type II diabetes and hypertension as study cases [25], we found that the disease-causing genes, which have high probability to cause type II diabetes and hypertension phenotypes when perturbed, tend to be hubs in the interactome networks and enriched in signaling pathways, whereas the significantly differentially expressed genes identified by microarrays are mostly enriched in the metabolic pathways. The connection between these two gene sets is significantly tight.

To bridge the gap between the genetic screen data and the mRNA expression data using known molecular networks, Yeger-Lotem *et al*. developed an integrative approach called "Response Net" [24]. Briefly, Response Net is a flow optimization algorithm that redefines a crucial subnetwork that connects genetic hits (source) and differentially expressed genes (target) from a whole weight network, where each node or edge has been assigned a weight according to their biological importance or confidence. The cost of an edge is defined by the $-\log$ value of its weight. Thus, the goal of Response Net can be achieved by solving a linear programming optimization problem that minimizes the overall cost of the network when distributing the maximal flow from source to target. According to the solution, those edges with positive flow defined the predicted crucial subnetwork.

1.4 Conclusions

We have introduced basic methods and applications in network analysis to interpret complex phenotypes. Although these methods have many advantages, network biology still faces many challenges. Most of the methods rely on the quality of datasets, which determine the false-positives and limited coverage. Most edges in network maps are still lacking detailed attributes and directions. Post-transcriptional

modifications are hardly monitored at a large scale. Tissue- and cell-type specificities are hard to consider. However, with the development of new technologies, such as high-throughput and single-cell dynamic measurement techniques, and with increasing accuracy and coverage of high-throughput technologies, the ever-accelerating data acquisition will raise further need for data integration and modeling at the network level. More and more methods have emerged, which provide important tools for network analysis. Mastering these methods is necessary, but far from sufficient for understanding biology. More important things to do are to ask the right questions, to choose proper network analysis tools, and to validate analysis results by solid experimentation. After all, network biology is biology and the fundamental goal is the same for network biology and molecular biology – to better understand basic biological processes and the mechanisms of human diseases.

References

1 Barabasi, A.L. and Oltvai, Z.N. (2004) Network biology: understanding the cell's functional organization. *Nat. Rev. Genet.*, **5**, 101–113.

2 Bergman, A. and Siegal, M.L. (2003) Evolutionary capacitance as a general feature of complex gene networks. *Nature*, **424**, 549–552.

3 Kauffman, S.A. (1969) Metabolic stability and epigenesis in randomly constructed genetic nets. *J. Theor. Biol.*, **22**, 437–467.

4 Li, F., Long, T., Lu, Y., Ouyang, Q., and Tang, C. (2004) The yeast cell-cycle network is robustly designed. *Proc. Natl. Acad. Sci. USA*, **101**, 4781–4786.

5 Chen, J.F., Mandel, E.M., Thomson, J.M., Wu, Q., Callis, T.E., Hammond, S.M., Conlon, F.L., and Wang, D.Z. (2006) The role of microRNA-1 and microRNA-133 in skeletal muscle proliferation and differentiation. *Nat. Genet.*, **38**, 228–233.

6 Huang, S., Eichler, G., Bar-Yam, Y., and Ingber, D.E. (2005) Cell fates as high-dimensional attractor states of a complex gene regulatory network. *Phys. Rev. Lett.*, **94**, 128701.

7 Han, J.D. (2008) Understanding biological functions through molecular networks. *Cell Res.*, **18**, 224–237.

8 Jeong, H., Mason, S.P., Barabasi, A.L., and Oltvai, Z.N. (2001) Lethality and centrality in protein networks. *Nature*, **411**, 41–42.

9 Tew, K.L., Li, X.L., and Tan, S.H. (2007) Functional centrality: detecting lethality of proteins in protein interaction networks. *Genome Inform.*, **19**, 166–177.

10 Albert, R., Jeong, H., and Barabasi, A.L. (2000) Error and attack tolerance of complex networks. *Nature*, **406**, 378–382.

11 He, X. and Zhang, J. (2006) Why do hubs tend to be essential in protein networks? *PLoS Genet.*, **2**, e88.

12 Han, J.D., Bertin, N., Hao, T., Goldberg, D.S., Berriz, G.F., Zhang, L.V., Dupuy, D., Walhout, A.J., Cusick, M.E., Roth, F.P. *et al.* (2004) Evidence for dynamically organized modularity in the yeast protein–protein interaction network. *Nature*, **430**, 88–93.

13 Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D., and Alon, U. (2002) Network motifs: simple building blocks of complex networks. *Science*, **298**, 824–827.

14 Milo, R., Itzkovitz, S., Kashtan, N., Levitt, R., Shen-Orr, S., Ayzenshtat, I., Sheffer, M., and Alon, U. (2004) Superfamilies of evolved and designed networks. *Science*, **303**, 1538–1542.

15 Wuchty, S., Oltvai, Z.N., and Barabasi, A.L. (2003) Evolutionary conservation of motif constituents in the yeast protein interaction network. *Nat. Genet.*, **35**, 176–179.

16 Yu, H. and Gerstein, M. (2006) Genomic analysis of the hierarchical structure of regulatory networks. *Proc. Natl. Acad. Sci. USA*, **103**, 14724–14731.

17 Bader, G.D. and Hogue, C.W. (2003) An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinformatics*, **4**, 2.

18 Eisen, M.B., Spellman, P.T., Brown, P.O., and Botstein, D. (1998) Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. USA*, **95**, 14863–14868.

19 Lage, K., Karlberg, E.O., Storling, Z.M., Olason, P.I., Pedersen, A.G., Rigina, O., Hinsby, A.M., Tumer, Z., Pociot, F., Tommerup, N. *et al.* (2007) A human phenome–interactome network of protein complexes implicated in genetic disorders. *Nat. Biotechnol.*, **25**, 309–316.

20 Wu, X., Jiang, R., Zhang, M.Q., and Li, S. (2008) Network-based global inference of human disease genes. *Mol. Syst. Biol.*, **4**, 189.

21 Kohler, S., Bauer, S., Horn, D., and Robinson, P.N. (2008) Walking the interactome for prioritization of candidate disease genes. *Am. J. Hum. Genet.*, **82**, 949–958.

22 Reverter, A., Hudson, N.J., Nagaraj, S.H., Perez-Enciso, M., and Dalrymple, B.P. (2010) Regulatory impact factors: unraveling the transcriptional regulation of complex traits from expression data. *Bioinformatics*, **26**, 896–904.

23 Hudson, N.J., Reverter, A., Wang, Y., Greenwood, P.L., and Dalrymple, B.P. (2009) Inferring the transcriptional landscape of bovine skeletal muscle by integrating co-expression networks. *PLoS ONE*, **4**, e7249.

24 Yeger-Lotem, E., Riva, L., Su, L.J., Gitler, A.D., Cashikar, A.G., King, O.D., Auluck, P.K., Geddie, M.L., Valastyan, J.S., Karger, D.R. *et al.* (2009) Bridging high-throughput genetic and transcriptional data reveals cellular responses to alpha-synuclein toxicity. *Nat. Genet.*, **41**, 316–323.

25 Yu, H., Huang, J., Qiao, N., Green, C.D., and Han, J.D. (2010) Evaluating diabetes and hypertension disease causality using mouse phenotypes. *BMC Syst. Biol.*, **4**, 97.