Graph Theory and Analysis of Biological Data in Computational Biology

Shih-Yi Chao
Ching Yun University
229,Chien-Hsin R.d., Jung-Li, Taiwan

1. Introduction

The theory of complex networks plays an important role in a wide variety of disciplines, ranging from communications to molecular and population biology. The focus of this article is on graph theory methods for computational biology. We'll survey methods and approaches in graph theory, along with current applications in biomedical informatics. Within the fields of Biology and Medicine, potential applications of network analysis by using graph theory include identifying drug targets, determining the role of proteins or genes of unknown function. There are several biological domains where graph theory techniques are applied for knowledge extraction from data. We have classified these problems into several different domains, which are described as follows.

1. Modeling of bio-molecular networks. It presents modeling methods of bio-molecular networks, such as protein interaction networks, metabolic networks, as well as transcriptional regulatory networks.
2. Measurement of centrality and importance in bio-molecular networks. To identify the most important nodes in a large complex network is of fundamental importance in computational biology. We'll introduce several researches that applied centrality measures to identify structurally important genes or proteins in interaction networks and investigated the biological significance of the genes or proteins identified in this way.
3. Identifying motifs or functional modules in biological networks. Most important biological processes such as signal transduction, cell-fate regulation, transcription, and translation involve more than four but much fewer than hundreds of proteins or genes. Most relevant processes in biological networks correspond to the motifs or functional modules. This suggests that certain functional modules occur with very high frequency in biological networks and be used to categories them.
4. Mining novel pathways from bio-molecular networks. Biological pathways provide significant insights on the interaction mechanisms of molecules. Experimental validation of identification of pathways in different organisms in a wet-lab environment requires monumental amounts of time and effort. Thus, there is a need for graph theory tools that help scientists predict pathways in bio-molecular networks.

Our primary goal in the present article is to provide as broad a survey as possible of the major advances made in this field. Moreover, we also highlight what has been achieved as well as some of the most significant open issues that need to be addressed. Finally, we hope that this chapter will serve as a useful introduction to the field for those unfamiliar with the literature.

2. Definitions and mathematical preliminaries

2.1 The concept of a graph

The concept of a graph is fundamental to the material to be discussed in this chapter. A graph G consists of a set of vertices $V(G)$ and a set of edges $E(G)$. In a simple graph, two of the vertices in G are linked if there exists an edge $(v_i, v_j)\in E(G)$ connecting the vertices v_i and v_j in graph G such that $v_i\in V(G)$ and $v_j\in V(G)$. The number of vertices will be denoted by $|V(G)|$, and the set of vertices adjacent to a vertex v_i is referred to as the neighbors of v_i, $N(v_i)$. The degree of a vertex v_i is the number of edges with which it is incident, symbolized by $d(v_i)$. Two graphs, G_1 and G_2, are said to be isomorphic ($G_1 \cong G_2$) if a one-to-one transformation of V_1 onto V_2 effects a one-to-one transformation of E_1 onto E_2. A subgraph G' of a graph G is a graph whose set of vertices and set of edges satisfy the relations: $V(G')\subseteq V(G)$ and $E(G') \subseteq E(G)$, and if G' is a subgraph of G, then G is said to be a supergraph of G'. The line graph $L(G)$ of an undirected graph G is a graph such that each vertex in $L(G)$ indicates an edge in G and any pairs of vertices of $L(G)$ are adjacent if and only if their corresponding edges share a common endpoint in G.

2.2 Directed and undirected graphs

A graph may be *undirected*, meaning that there is no distinction between the two vertices associated with each edge, or its edges may be *directed* from one vertex to another. Formally, a finite directed graph, G, consists of a set of vertices or nodes, $V(G) = \{v_1, . . . ,v_n\}$, together with an edge set, $E(G) \subseteq V(G) \times V(G)$. Intuitively, each edge $(u, v) \in E(G)$ can be thought of as connecting the starting node u to the terminal node v. An undirected graph, G, also consists of a vertex set, $V(G)$, and an edge set $E(G)$. However, there is no direction associated with the edges in this case. Hence, the elements of $E(G)$ are simply two element subsets of $V(G)$, rather than ordered pairs as directed graphs. As with directed graphs, we shall use the notation uv (or vu as direction is unimportant) to denote the edge $\{u, v\}$ in an undirected graph. For two vertices, u, v, of an undirected graph, uv is an edge if and only if vu is also an edge. We are not dealing with multi-graphs, so there can be at most one edge between any pair of vertices in an undirected graph. That is, we are discussing the simple graph. A simple graph is an undirected graph that has no loops and no more than one edge between any two different vertices. In a simple graph the edges of the graph form a set and each edge is a pair of *distinct* vertices. The number of vertices n in a directed or undirected graph is the size or order of the graph.

2.3 Node-degree and the adjacency matrix

For an undirected graph G, we shall write $d(u)$ for the degree of a node u in $V(G)$. This is simply the total number of edges at u. For the graphs we shall consider, this is equal to the number of neighbors of u, $d(u) = |N(u)|$. In a directed graph G, the *in-degree* , $d^+(u)$ (outdegree, $d^{-}(u)$) of a vertex u is given by the number of edges that terminate (or start) at u. Suppose that the vertices of a graph (directed or undirected) G are ordered as $v_1, \ldots, v_n$. Then the adjacency matrix, A, of G is given by

$$a_{ij} = \begin{cases} 1 & if\ v_i v_j \in E(G) \\ 0 & if\ v_i v_j \notin E(G) \end{cases}$$

Thus, the adjacency matrix of an undirected graph is symmetric while this need not be the case for a directed graph.

2.4 Path, path length and connected graph

Let u, v be two vertices in a graph G. Then a sequence of vertices $u = v_1, v_2, \ldots, v_k = v$, such that for $i = 1, \ldots, k$-1, is said to be a path of length k-1 from u to v. The geodesic distance, or simply distance, $d(u, v)$, from u to v is the length of the shortest path from u to v in G. If no such path exists, then we set $d(u, v) = 1$. If for every pair of vertices, (u, v), in graph G, there is some path from u to v, then we say that G is connected.

3. Modeling of bio-molecular networks

3.1 Introduction

Several classes of bio-molecular networks have been studied: Transcriptional regulatory networks, protein interaction network, and metabolic networks. In Biology, transcriptional regulatory networks and metabolic networks would usually be modeled as directed graphs. For instance, in a transcriptional regulatory network, nodes would represent genes with edges denoting the transcriptional relationships between them. This would be a directed graph because, if gene A regulates gene B, then there is a natural direction associated with the edge between the corresponding nodes, starting at A and terminating at B. In recent years, attentions have been focused on the protein-protein interaction networks of various simple organisms (Itzkovitz & Alon, 2005). These networks describe the direct physical interactions between the proteins in an organism's proteome and there is no direction associated with the interactions in such networks. Hence, PPI networks are typically modeled as undirected graphs, in which nodes represent proteins and edges represent interactions. In next sections, we individually introduce these bio-molecular networks.

3.2 Transcriptional regulatory networks

Transcriptional regulatory networks describe the regulatory interactions between genes. Here, nodes correspond to individual genes and a directed edge is drawn from gene A to gene B if A positively or negatively regulates gene B. Networks have been constructed for the transcriptional regulatory networks of *E. coli* and *S. cerevisiae* (Salgado et al., 2006; Lee et al., 2002; Salgado et al., 2006; Keseler et al., 2005) and are maintained in databases such as RegulonDB (Salgado et al., 2006) and EcoCyc (Keseler et al., 2005). Such networks are usually constructed through a combination of high-throughput genome location experiments and literature searches. Many types of gene transcriptional regulatory related approaches have been reported in the past. Their nature and composition are categorized by several factors: considering gene expression values (Keedwell & Narayanan, 2005; Shmulevich et al., 2002), the causal relationship between genes, e.g. with Bayesian analysis or Dynamic Bayesian Networks (Zou & Conzen, 2005; Husmeier, 2003), and the time domain e.g. discrete or continuous time (Li et al., 2006; He & Zeng, 2006; Filkov et al., 2002; Qian et al., 2001). One of the limitations of graph theory applications in analyzing biochemical networks is the static quality of graphs. Biochemical networks are dynamical, and the abstraction to graphs can mask temporal aspects of information flow. The nodes and links of biochemical networks change with time. Static graph representation of a system is, however, a prerequisite for building detailed dynamical models (Zou & Conzen, 2005). Most dynamical modeling approaches can be used to simulate network dynamics while using the graph representation as the skeleton of the model. Modeling the dynamics of biochemical networks provides closer to reality recapitulation of the system's behavior *in silico*, which can be useful for developing more quantitative hypotheses.

3.3 Protein interaction networks

Understanding protein interactions is one of the important problems of computational biology. These protein-protein interactions (PPIs) networks are commonly represented by undirected graph format, with nodes corresponding to proteins and edges corresponding to protein-protein interactions. The volume of experimental data on protein-protein interactions is rapidly increasing by high-throughput techniques improvements which are able to produce large batches of PPIs. For example, yeast contains over 6,000 proteins, and currently over 78,000 PPIs have been identified between the yeast proteins, with hundreds of labs around the world adding to this list constantly. Humans are expected to have around 120000 proteins and around 10^6 PPIs. The relationships between the structure of a PPI network and a cellular function are waited to be explored. Large-scale PPI networks (Rain et al., 2001; Giot et al., 2003; Li et al., 2004; Von Mering et al., 2004; Mewes et al., 2002) have been constructed recently using high-throughput approaches such as yeast-2-hybrid screens (Ito et al., 2001) or mass spectrometry techniques (Gavin et al., 2002) to identify protein interactions.

Vast amounts of PPI related data that are constantly being generated around the world are being deposited in numerous databases. Data on protein interactions are also stored in databases such as the database of interacting proteins (DIP) (Xenarios et al., 2000). We briefly mention the main databases, including nucleotide sequence, protein sequence, and PPI databases. The largest nucleotide sequence databases are EMBL (Stoesser et al., 2002), DDBJ (Tateno et al., 2002), and GenBank (Benson et al., 2002). They contain sequences from the literature as well as those submitted directly by individual laboratories. These databases store information in a general manner for all organisms. Organism specific databases exist for many organisms. For example, the complete genome of yeast and related yeast strains can be found in *Saccharomyces* Genome Database (SGD) (Dwight et al., 2002). FlyBase (Ashburner, 1993) contains the complete genome of the fruit fly *Drosophila melanogaster*. It is one of the earliest model organism databases. Ensembl (Hubbard et al., 2002) contains the draft human genome sequence along with its gene prediction and large scale annotation. SwissProt (Bairoch & Apweiler, 2000) and Protein Information Resource (PIR) (McGarvey et al., 2000) are two major protein sequence databases. SwissProt maintains a high level of annotations for each protein including its function, domain structure, and post-translational modification information.

Understanding interactions between proteins in a cell may benefit from a model of a PPIs network. A full description of protein interaction networks requires a complex model that would encompass the undirected physical protein-protein interactions, other types of interactions, interaction confidence level, or method and multiplicity of an interaction, directional pathway information, temporal information on the presence or absence of PPIs, and information on the strength of the interactions. This may be achieved by designing a scoring function and assigning weights to nodes and edges of a PPIs network.

3.4 Metabolic networks

Metabolic networks describe the bio-chemical interactions within a cell through which substrates are transformed into products through reactions catalysed by enzymes. Metabolic networks generally require more complex representations, such as hyper-graphs, as reactions in metabolic networks generally convert multiple inputs into and multiple outputs with the help of other components. An alternative is a weighted bipartite graph to reduce representation for a metabolic network. In such graphs, two types of nodes are used to represent reactions and compounds, respectively. The edges in a weighted bipartite graph connect nodes of different types, representing either substrate or product relationships. These networks can represent the complete set of metabolic and physical processes that determine the physiological and biochemical properties of a cell. Metabolic networks are complex. There are many kinds of nodes (proteins, particles, molecules) and many connections (interactions) in such networks. Even if one can define sub-networks that can be meaningfully described in relative isolation, there are always connections from it to other networks. As with protein interaction networks, genome-scale metabolic networks have been constructed for a variety of simple organisms including *S. cerevisiae* and *E. coli* (Jeong et al., 2000; Overbeek et al., 2000; Karp et al., 2002; Edwards et al., 2000), and are stored in databases such as the KEGG (Kanehisa & Goto, 2000) or BioCyc (Karp et al., 2005) databases. A common approach to the construction of such networks is to first use the annotated genome of an organism to identify the enzymes in the network and then to combine bio-chemical and genetic information to obtain their associated reactions (Kauffman et al., 2000; Edwards et al., 2001). While efforts have been made to automate certain aspects of this process, there is still a need to validate the networks generated automatically manually against experimental biochemical results (Segre et al., 2003). For metabolic networks, significant advances have also been made in modelling the reactions that take place on such networks. The overall structure of a network can be described by several different parameters. For example, the average number of connections a node has in a network, or the probability that a node has a given number of connections. Theoretical work has shown that different models for how a network has been created will give different values for these parameters. The classical random network theory (Erdös & Renyi, 1960) states that given a set of nodes, the connections are made randomly between the nodes. This gives a network where most nodes have the same number of connections. Recent research has shown that this model does not fit the structure found in several important networks. Instead, these complex networks are better described by a so-called scale-free model where most nodes have only a few connections, but a few nodes (called hubs) have a very large number of connections. Recent work indicates that metabolic networks are examples of such scale-free networks (Jeong *et al.*, 2000). This result is important, and will probably lead to new insights into the function of metabolic and signaling networks, and into the evolutionary history of the networks. Robustness is another important property of metabolic networks. This is the ability of the network to produce essentially the same behavior even when the various parameters controlling its components vary within considerable ranges. For example, recent work indicates the segment polarity network in the *Drosophila* embryo can function satisfactorily with a surprisingly large number of randomly chosen parameter sets (von Dassow *et a.l*, 2000). The parameters do not have to be carefully tuned or optimized. This makes biological sense, which means a metabolic network should be tolerant with respect to mutations or large environmental changes.

Another important emerging research topic is to understand metabolic networks in term of their function in the organism and in relation to the data we already have. This requires combining information from a large number of sources, such as classical biochemistry, genomics, functional genomics, microarray experiments, network analysis, and simulation. A theory of the cell must combine the descriptions of the structures in it with a theoretical and computational description of the dynamics of the life processes. One of the most important challenges in the future is how to make all this information comprehensible in biological terms. This is necessary in order facilitate the use of the information for predictive purposes to predict what will happen after given some specific set of circumstances. This kind of predictive power will only be reached if the complexity of biological processes can be handled computationally.

4. Measurement of centrality and importance in bio-molecular networks

Biological function is an extremely complicated consequence of the action of a large number of different molecules that interact in many different ways. Genomic associations between genes reflect functional associations between their products (proteins) (Huynen et al., 2000; Yanai et al., 2001). Furthermore, the strength of the genomic associations correlates with the strength of the functional associations. Genes that frequently co-occur in the same operon in a diverse set of species are more likely to physically interact than genes that occur together in an operon in only two species ((Huynen et al., 2000), and proteins linked by gene fusion or conservation of gene order are more likely to be subunits of a complex than are proteins that are merely encoded in the same genomes (Enright et al., 1999). Other types of associations have been used for network studies, but these focus on certain specific types of functional interactions, like subsequent enzymatic steps in metabolic pathways, or physical interactions. Elucidating the contribution of each molecule to a particular function would seem hopeless, had evolution not shaped the interaction of molecules in such a way that they participate in functional units, or building blocks, of the organism's function (Callebaut et al., 2005). These building blocks can be called modules, whose interactions, interconnections, and fault-tolerance can be investigated from a higher-level point of view, thus allowing for a synthetic rather than analytic view of biological systems (Sprinzak et al., 2005). The recognition of modules as discrete entities whose function is separable from those of other modules (Hartwell et al., 1999) introduces a critical level of biological organization that enables *in silico* studies.

Intuitively, modularity must be a consequence of the evolutionary process. Modularity implies the possibility of change with minimal disruption of function, a feature that is directly selected for (Wilke et al., 2003). However, if a module is essential, its independence from other modules is irrelevant unless, when disrupted, its function can be restored either by a redundant gene or by an alternative pathway or module. Furthermore, modularity must affect the evolutionary mechanisms themselves, therefore both robustness and evolvability can be optimized simultaneously (Lenski et al., 2006). The analysis of these concepts requires both understanding of what constitutes a module in biological systems and tools to recognize modules among groups of genes. In particular, a systems view of biological function requires the development of a vocabulary that not only classifies modules according to the role they play within a network of modules and motifs, but also how these modules and their interconnections are changed by evolution, for example, how they constitute units of evolution targeted directly by the selection process (Schlosser et al., 2004). The identification of biological modules is usually based either on functional or topological criteria. For example, genes that are co-expressed or coregulated can be classified into modules by identifying their common transcription factors (Segal et al., 2004), while genes that are highly connected by edges in a network form clusters that are only weakly connected to other clusters (Rives et al., 2003). From viewpoint of evolutionary, genes that are inherited together but not with others often form modules (Snel et al., 2004; Slonim et al., 2006). However, the concept of modularity is not at all well defined. For example, the fraction of proteins that constitutes the core of a module and that is inherited together is small (Snel et al., 2004), implying that modules are fuzzy but also flexible so that they can be rewired quickly, allowing an organism to adapt to novel circumstances (Campillos et al., 2006).

A set of data is provided by genetic interactions (Reguly et al., 2006), such as synthetic lethal pairs of genes or dosage rescue pairs, in which a knockout or mutation of a gene is suppressed by over-expressing another gene. Such pairs are interesting because they provide a window on cellular robustness and modularity brought about by the conditional expression of genes. Indeed, the interaction between genes epistasis (Wolf et al., 2000) has been used to successfully identify modules in yeast metabolic genes (Segre et al., 2005). However, often interacting pairs of genes lie in alternate pathways rather than cluster in functional modules. These genes do not interact directly and thus are expected to straddle modules more often than lie within one (Jeong et al., 2000).

In silico evolution is a powerful tool, if complex networks can be generated that share the pervasive characteristics of biological networks, such as error tolerance, small-world connectivity, and scale-free degree distribution (Jeong et al., 2000). If furthermore each node in the network represents a simulated chemical or a protein catalyzing reactions involving these molecules, then it is possible to conduct a detailed functional analysis of the network by simulating knockdown or over-expression experiments. This functional datum can then be combined with evolutionary and topological information to arrive at a more sharpened concept of modularity that can be tested in vitro when more genetic data become available. Previous work on the in silico evolution of metabolic (Pfeiffer et al., 2005), signaling (Soyer & Bonhoeffer, 2006; Soyer et al., 2006), biochemical (Francois et al., 2004; Paladugu et al., 2006), regulatory (Ciliberti et al., 2007), as well as Boolean (Ma'ayan et a., 2006), electronic (Kashtan et al., 2005), and neural (Hampton et al., 2004) networks has begun to reveal how network properties such as hubness, scaling, mutational robustness as well as short pathway length can emerge in a purely Darwinian setting. In particular, *in silico* experiments testing the evolution of modularity both in abstract (Lipson et al., 2002) and in simulated electronic networks suggest that environmental variation is key to a modular organization of function. These networks are complex, topologically interesting (Adami, 2002), and function within simulated environments with different variability that can be arbitrarily controlled.

5. Identifying motifs or functional modules in biological networks

Biological systems viewed as networks can readily be compared with engineering systems, which are traditionally described by networks such as flow charts. Remarkably, when such a comparison is made, biological networks and engineered networks are seen to share structural principles such as modularity and recurrence of circuit elements (Alon, 2003). Both biological systems function and engineering are organized with modularity. Engineering systems can be decomposed into functional modules at different levels (Hansen et al., 1999), subroutines in software (Myers, 2003) and replaceable parts in machines. In the case of biological networks, although there is no consensus on the precise groups of genes and interactions that form modules, it is clear that they possess a modular structure (Babu et al., 2004). Alon proposed a working definition of a module based on comparison with engineering. A *module* in a network is a set of nodes that have strong interactions and a common function (Alon, 2003). A module has defined input nodes and output nodes that control the interactions with the rest of the network.

Various basic functional modules are frequently reused in engineering and biological systems. For example, a digital circuit may include many occurrences of basic functional modules such as multiplexers and so on (Hansen et al., 1999). Biology displays the same principle, using key wiring patterns again and again throughout a network. For instance, metabolic networks use regulatory circuits such as feedback inhibition in many different pathways (Alon, 2003). Besides basic functional modules, recently a small set of recurring circuit elements termed *motifs* have been discovered in a wide range of biological and engineering networks (Milo et al., 2002). Motifs are small (about 3 or 4 nodes) sub-graphs that occur significantly more frequently in real networks than expected by chance alone, and are detected purely by topological analysis. This discover kindled a lot of interest on organization and function of motifs, and many related papers were published in recent years. The observed over-representation of motifs has been interpreted as a manifestation of functional constraints and design principles that have shaped network architecture at the local level (Milo et al., 2002). Some researchers believe that motifs are basic building blocks that may have specific functions as elementary computational circuits (Milo et al., 2002). Although motifs seem closely related to conventional building blocks, their relation lacks adequate and precise analysis, and their method of integration into full networks has not been fully examined. Further, it is not clear what determines the particular frequencies of all possible network motifs in a specific network.

6. Mining novel pathways from bio-molecular networks

In the studying organisms at a systems level, biologists recently mentioned (Kelley et al. 2003) the following questions: (1) Is there a minimal set of pathways that are required by all organisms? (2) To what extent are the genomic pathways conserved among different species? (3) How are organisms related in terms of the distance between pathways rather than at the level of DNA sequence similarity? At the core of such questions lies the identification of pathways in different organisms. However, experimental validation of an enormous number of possible candidates in a wet-lab environment requires monumental amounts of time and effort. Thus, there is a need for comparative genomics tools that help scientists predict pathways in an organism's biological network. Due to the complex and incomplete nature of biological data, at the present time, fully automated computational pathway prediction is excessively ambitious. A metabolic pathway is a set of biological reactions where each reaction consumes a set of metabolites, called substrates, and produces another set of metabolites, called products. A reaction is catalyzed by an enzyme (or a protein) or a set of enzymes. There are many web resources that provide access to curated as well as predicted collections of pathways, e.g., KEGG (Kanehisa et al. 2004), EcoCyc (Keseler et al. 2005), Reactome (Joshi-Tope et al. 2005), and PathCase (Ozsoyoglu et al 2006). Work to date on discovering biological networks can be organized under two main titles: (i) Pathway Inference (Yamanishi et al., 2007; Shlomi et al., 2006), and (ii) Whole-Network Detection (Tu et al., 2006; Yamanishi et al. 2005). Even with the availability genomic blueprint for a living system and functional annotations for its putative genes, the experimental elucidation of its biochemical processes is still a daunting task. Though it is possible to organize genes by broad functional roles, piecing them together manually into consistent biochemical pathways quickly becomes intractable. A number of metabolic pathway reconstruction tools have been developed since the availability of the first microbial genome, Haemophilus influenza (Fleischmann et al., 1995). These include PathoLogic (Karp & Riley, 1994), MAGPIE (Gaasterland & Sensen, 1996) and WIT (Overbeek et al., 2000) and PathFinder (Goesmann et al., 2002). The goal of most pathway inference methods has generally been to match putatively identified enzymes with known or reference pathways. Although reconstruction is an important starting point for elucidating the metabolic capabilities of an organism based upon prior pathway knowledge, reconstructed pathways often have many missing enzymes, even in essential pathways. The issue of redefining microbial biochemical pathways based on missing proteins is important since there are many examples of alternatives to standard pathways in a variety of organisms (Cordwell, 1999). Moreover, engineering a new pathway into an organism through heterologous enzymes also requires the ability to infer new biochemical routes. With more genomic sequencing projects underway and confident functional characterizations absent for many of the genes, automated strategies for predicting biochemical pathways can aid biologists inunraveling the complex processes in living systems. At the same time, pathway inference approaches can also help in designing synthetic processes using the repertoire biocatalysts available in nature.

7. Conclusion

The large-scale data on bio-molecular interactions that is becoming available at an increasing rate enables a glimpse into complex cellular networks. Mathematical graph theory is a straightforward way to represent this information, and graph-based models can exploit global and local characteristics of these networks relevant to cell biology. Moreover, the need for a more systematic approach to the analysis of living organisms, alongside the availability of unprecedented amounts of data, has led to a considerable growth of activity in the theory and analysis of complex biological networks in recent years. Networks are ubiquitous in Biology, occurring at all levels from biochemical reactions within the cell up to the complex webs of social and sexual interactions that govern the dynamics of disease spread through human populations. Network graphs have the advantage that they are very simple to reason about, and correspond by and large to the information that is globally available today on the network level. However, while binary relation information does represent a critical aspect of interaction networks, many biological processes appear to require more detailed models. A comprehensive understanding of these networks is needed to develop more sophisticated and effective treatment strategies for diseases such as Cancer. This may eventually prove mathematical models of large-scale data sets valuable in medical problems, such as identifying the key players and their relationships responsible for multi-factorial behavior in human disease networks. In conclusion, it can be said of biological network analysis is needed in Bioinformatics research field, and the challenges are exciting. It is hoped that this chapter will be of assistance to researchers by highlighting recent advances in this field.

8. References

Adami, C (2002). What is complexity. *BioEssays*, Vol. 24, pp. 1085–1094.

Alon,U. (2003). Biological networks: the tinkerer as an engineer. *Science*, Vol. 301, No. 5641.

Ashburner, M. (1993). FlyBase. Genome News, Vol. 13, pp. 19–20.

Babu, M. M. (2004). Structure and evolution of transcriptional regulatory networks. *Curr Opin Struct Biol*, Vol. 14, No. 3, pp.283-291.

Bairoch, A. & Apweiler, R. (2000). The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Research*, Vol. 28, pp. 45–48.

Benson, D. A. (2002). GenBank. *Nucleic Acids Research*, Vol. 30, pp. 17–20.

Callebaut W & Rasskin-Gutman D (2005). Modularity: understanding the development and evolution of complex systems. Mueller GB, Wagner GB, Callebaut W, editors. Cambridge (Massachusetts): MIT Press.

Campillos, M. et al. (2006). Identification and analysis of evolutionarily cohesive functional modules in protein networks. *Genome Res.*, Vol. 16, pp. 374–382.

Ciliberti, S. et al. (2007). Robustness can evolve gradually in complex regulatory gene networks with varying topology. *PLoS Comput Biol.*, Vol. 3.

Cordwell, S. (1999). Microbial genomes and missing enzymes: redefining biochemical pathways. *Arch. Microbiol.*, Vol. 172, pp. 269–279.

Dwight, S. S. (2002). Saccharomyces genome database (SGD) provides secondary gene annotation using the gene ontology (GO). *Nucleic Acids Research*, Vol. 30, pp. 69–72.

Edwards, J. & Palsson, B. (2000). The Escherichia coli MG1655 in silico metabolic genotype: Its definition, characteristics and capabilities. *Proceedings of the National Academy of Sciences*, Vol. 97, No. 10, pp. 5528–5533.

Edwards, J. (2001). In silico predictions of Escherichia coli metabolic capabilities are consistent with experimental data. *Nature Biotechnology*, Vol. 19, pp. 125–130.

Erdős, P.& Rényi, A. (1960). The Evolution of Random Graphs. *Magyar Tud. Akad. Mat. Kutató Int. Közl.*, Vol. 5, pp. 17–61.

Enright, A. J. et al. (1999). Protein interaction maps for complete genomes based on gene fusion events. *Natur*e, Vol. 402, pp. 86–90.

Filkov, V. et al., (2002). Analysis techniques for microarray time-series data. *J Comput Boil*, Vol. 9, pp. 317-331.

Fleischmann, R. et al., (1995). Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. *Science*, Vol. 269, pp. 469–512.

Francois, P. & Hakim, V. (2004). Design of genetic networks with specified functions by evolution in silico. *Proc Natl Acad Sci U S A*, Vol. 101, pp. 580–585.

Gaasterland,T.& Sensen,C. (1996). MAGPIE: automated genome interpretation.Trends *Genet.*, Vol. 12, pp. 76–78.

Gavin, A. et al. (2002). Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature*, Vol. 415, pp. 141–147.

Giot, L. et al. (2003). A protein interaction map of Drosophila Melanogaster. *Science*, Vol. 302, pp.1727–1736.

Goesmann, A., et al. (2002). PathFinder: reconstruction and dynamic visualization of metabolic pathways. *Bioinformatics*, Vol. 18, pp. 124–129.

Hampton, A.N. & Adami, C. (2004). Evolution of robust developmental neural networks. Pollack JB, Bedau MA, Husbands P, Ikegami T, Watson R, editors. Boston: MIT Press. pp. 438–443.

Hansen, M. C. et al. (1999). Unveiling the iscas-85 benchmarks: A case study in reverse engineering. *IEEE Des. Test*, Vol. 16, No. 3, pp. 72-80.

Hartwell, L.H., et al. (1999). From molecular to modular cell biology. *Nature*, Vol. 402, pp. C47–C52.

He, F. & Zeng, A.P. (2006). In search of functional association from time-series microarray data based on the change trend and level of gene expression. *BMC Bioinformatics,* Vol. 7, pp. 69-84.

Hubbard, T. (2002). The ensembl genome database project. *Nucleic Acids Research*, Vol. 30, pp. 38–41.

Husmeier, D. (2003). Sensitivity and Specificity of Inferring Genetic Regulatory Interactions from Microarray Experiments with Dynamic Bayesian Networks. *Bioinformatics*, Vol. 19, pp. 2271-2282.

Huynen, M. (2000). The identification of functional modules from the genomic association of genes. *Genome Res.*, Vol. 10, pp. 1204–1210.

Ito, T. et al. (2001).A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proceedings of the National Academy of Sciences*, Vol. 98, No. 8, pp. 4569–4574.

Itzkovitz, S. & Alon, U. (2005). Subgraphs and network motifs in geometric networks. *Physical Review E*, Vol. 71, pp. 026117-1-0261179, ISSN 1539-3755.

Jeong, H. et al. (2000). The large-scale organization of metabolic networks. *Nature*, Vol. 407, pp. 651–654.

Jeong, H. B. (2000). The large-scale organization of metabolic networks. *Nature*, Vol. 407, pp. 651-654.

Joshi-Tope G. et al. (2005) Reactome: a knowledgebase of biological pathways. *Nucleic Acids Res. Database Issue*, Vol. 33, pp. D428-32

Kanehisa, M. & Goto, S. (2000). KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, Vol. 28, No. 1, pp. 27–30.

Karp, P. et al. (2002). The EcoCyc Database. *Nucleic Acids Research*, Vol. 30, No. 1, pp. 56–58.

Karp, P. et al. (2005). Expansion of the BioCyc collection of pathway/genome databases to 160 genomes. *Nucleic Acids Research*, Vol. 33, No. 19, pp.6083–6089.

Kauffman, K. et al. (2003). Advances in flux balance analysis. *Current Opinion in Biotechnology*, Vol. 14, pp. 491–496.

Kashtan, N. & Alon, U. (2005). Spontaneous evolution of modularity and network motifs. P*roc Natl Acad Sci USA,* Vol. 102, pp. 13773–13778.

Kanehisa, M. et al. (2004) The KEGG resource for deciphering the genome. *Nucleic Acids Res*., Vol. 32, pp. D277–280.

Karp,P. & Riley,M. (1994). Representations of metabolic knowledge: pathways. In Altman,R., Brutlag,D., Karp,P., Lathrop,R. and Searls,D. (ed.) *Second International Conference on Intelligent Systems for Molecular Biology.* AAAI Press, Menlo Park, CA.

Keedwell, E. & Narayanan, A. (2005). Discovering Gene Networks with a Neural-Genetic Hybrid. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*. Vol. 2, pp. 231-242.

Kelley, P et al. (2003). Conserved pathways within bacteria and yeast as revealed by global protein network alignment. *Proc. of National Academy of Sciences USA,* Vol. 100, No. 20, pp. 11394-11395.

Keseler, I. et al. (2005). EcoCyc: a comprehensive database resource for Escherichia coli. *Nucleic Acids Research*, Vol. 33, No. 1.

Lee, T. et al. (2002). Transcriptional regulatory networks in Saccharomyces cerevisiae. *Science*, Vol. 298, pp. 799–804.

Lenski, R.E. et al. (2006). Balancing robustness and evolvability. *PLoS Biol*., Vol. 4.

Li, S. et al. (2004). A map of the interactome network of the metazoan C. elegans. *Science*, Vol. 303, pp. 540–543.

Li, X. et al. (2006). Wand QK: Discovery of time-delayed gene regulatory networks based on temporal gene expression profiling. *BMC Bioinformatics*, Vol. 7, pp.26-46.

Lipson, H. et al. (2002). On the origin of modular variation. *Evolution*, Vol. 56, pp. 1549–1556.

Ma'ayan, A. et al. (2006). Topology of resultant networks shaped by evolutionary pressure. *Phys Rev E*, Vol. 73, pp. 061912.

McGarvey, P. B. (2000). PIR: a new resource for bioinformatics. *Bioinformatics*, Vol. 16, pp. 290–291.

Mewes, H. et al. (2002). MIPS: a database for genomes and protein sequences. *Nucleic Acids Research*, Vol. 30, No. 1, pp. 31–34.

Milo, R. et al. (2002). Net- work motifs: simple building blocks of complex networks. *Science*, Vol. 298, No. 5594, pp. 824-827.

Myers, C. R.(2003). Software systems as complex networks: structure, function, and evolvability of software collaboration graphs. *Physical Review E*, Vol. 68.

Overbeek, R. et al. (2000). WIT: integrated system for high-throughput genome sequence analysis and metabolic reconstruction. *Nucleic Acids Research*, Vol. 28, No. 1, pp. 123–125.

Ozsoyoglu, M. et al. (2006). Genomic Pathways Database and Biological Data Management. *Animal Genetics*, Vol. 37, pp. 41-47.

Paladugu, S.R. et al. (2006). In silico evolution of functional modules in biochemical networks. *IEE Proc Syst Biol.*, Vol. 153, pp. 223–235.

Pfeiffer, T. et al. (2005). The evolution of connectivity in metabolic networks. *PLoS Biol*., Vol. 3.

Qian, J. et al. (2001). Beyond synexpression relationships: local clustering of time-shifted and inverted gene expression profiles identifies new, biologically relevant interactions. *J Mol Biol*, Vol. 314, pp. 1053-1066.

Rain, J. et al. (2001). The protein-protein interaction map of Heliobacter Pylori. *Nature*, Vol. 409, pp. 211–215.

Reguly, T. et al. (2006). Comprehensive curation and analysis of global interaction networks in Saccharomyces cerevisiae. *J Biol.*, Vol. 5, No. 11.

Rives, A.W. & Galitski, T. (2003). Modular organization of cellular networks. *Proc Natl Acad Sci U S A*, Vol. 100, pp. 1128–1133.

Salgado, H. et al. (2006). The comprehensive updated regulatory network of Escherichia coli K-12. *BMC Bioinformatics*,Vol. 7, No. 5.

Salgado, H. et al. (2006). RegulonDB (version 5.0): Escherichia coli K-12 transcriptional regulatory network, operon organization, and growth conditions. *Nucleic Acids Research*, Vol. 34, No. 1.

Schlosser, G. & Wagner, G.P. (2004). Modularity in development and evolution. Chicago: University of Chicago Press.

Segal, E. et al. (2004). A module map showing conditional activity of expression modules in cancer. *Nat Genet*, Vol. 36, pp. 1090–1098.

Segre, D. et al. (2003). From annotated genomes to metabolic flux models and kinetic parameter fitting. *Omics*, Vol. 7, No. 3, pp. 301–316.

Segre, D. et al. (2005). Modular epistasis in yeast metabolism. *Nat Genet*, Vol. 37, pp. 77–83.

Shlomi, T. et al. (2006). QPath: a method for querying pathways in a protein-protein interaction network. *BMC Bioinformatics*, Vol. 7, No. 199.

Shmulevich, I. et al. (2002). From Boolean to Probabilistic Boolean Networks as Models of Genetic Regulatory Networks. *Proceedings of the IEEE*, Vol. 90, pp.1778-1790.

Slonim, N. et al. (2006). Ab initio genotype-phenotype association reveals intrinsic modularity in genetic networks. *Mol Syst Biol.*, Vol. 2.

Snel, B & Huynen, M.A. (2004). Quantifying modularity in the evolution of biomolecular systems. *Genome Res*, Vol. 14, pp. 391–397.

Sprinzak, D. & Elowitz, M.B. (2005). Reconstruction of genetic circuits. *Nature*, Vol. 438, pp. 443–448.

Stoesser, G. et al. (2002). The EMBL nucleotide sequence database. *Nucleic Acids Research*, Vol. 30, pp. 21–26.

Soyer, O.S. & Bonhoeffer, S. (2006). Evolution of complexity in signaling pathways. *Proc Natl Acad Sci U S A*, Vol.103, pp. 16337–16342.

Soyer, O.S. et al. (2006). Simulating the evolution of signal transduction pathways. *J Theor Biol.*, Vol. 241, pp. 223–232.

Tateno, Y. (2002). DAN data bank of japan (DDBJ). *Nucleic Acids Research*, Vol. 30, pp. 27–30.

Tu, Z. et al. (2006). An integrative approach for causal gene identification and gene regulatory pathway inference. *Bioinformatics*, Vol. 22, No. 14, pp. e489-96.

Von Dassow, G. (2000). The segment polarity network is a robust developmental module. *Nature*, Vol. 406, No. 6792, pp.188-192.

Von Mering, C. et al. (2002). Comparative assessment of large-scale data sets of protein-protein interactions. *Nature*, Vol. 417, pp. 399–403.

Wilke, C. O. & Adami, C. (2003). Evolution of mutational robustness. *Mutat Res.*, Vol. 522, pp. 3–11.

Xenarios, I. et al. (2000). DIP: the database of interacting proteins. *Nucleic Acids Research*, Vol. 28, No. 1, pp.289–291.

Wolf, J.B. (2000). Epistasis and the evolutionary process. Oxford: Oxford University Press.

Yanai, I., et al. (2001). Genes linked by fusion events are generally of the same functional category: a systematic analysis of 30 microbial genomes. *Proc. Natl. Acad. Sci. USA*, Vol. 98, pp. 7940–7945.

Yamanishi, Y. et al. (2005). Supervised enzyme network inference from the integration of genomic data and chemical information. *ISMB (Supplement of Bioinformatics)*, pp. 468-477.

Yamanishi, Y. et al. (2007). Prediction of missing enzyme genes in a bacterial metabolic network. *FEBS J.*, Vol. 274, No. 9, pp. 2262-73.

Zou, M. & Conzen, SD (2005). A new dynamic Bayesian network (DBN) approach for identifying gene regulatory networks from time course microarray data. *Bioinformatics*, Vol. 21, pp. 71-79.

Advanced Technologies
Edited by Kankesu Jayanthakumaran
INTECH