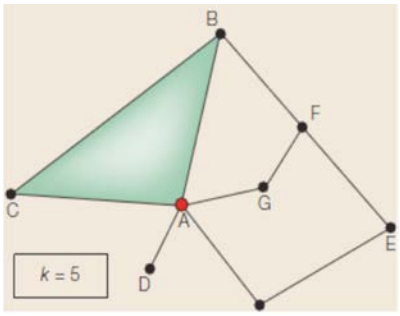
Chapter 2 Basics of Biological Networks

2.1Overview

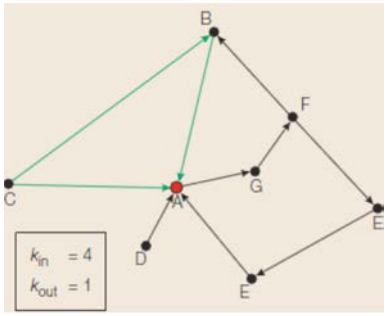
Protein-protein interaction (PPI) networks, biochemical networks, transcriptional regulation networks, signal transduction or metabolic networks are often sharing some characteristics and properties. Network analysis can be employed to reveal the underlying network topological structures and provides useful network measures for ranking network nodes. Therefore, identification of drug target, determining protein function, and designing effective strategies for treating various diseases become possible. This chapter will introduce the basic knowledge for performing analysis of complex biological networks.

2.2 Graph Theory and Definitions

A graph G can be defined as a pair (V, E) where V is a set of vertices representing the nodes and E is a set of edges $E = \{(i, j) \mid i, j \in V\}$ representing the connections between the nodes. An edge between nodes i and j can also be associated with it a weight function $w : E \to R$, where R denotes a real number. If an edge (i, j) exists between nodes i and j, we say that the vertex i is *adjacent* to the vertex j.

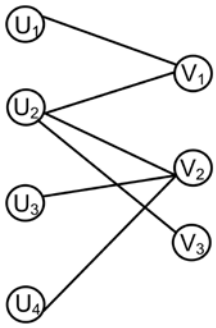


A directed graph is defined as an ordered triple $G = (V, E, f)$, where f is a function that maps each edge element in E to an *ordered* pair of vertices in V. Thus, an edge $E = (i, j)$ is considered to have direction from i to j to reflect the flow of information throughout the network.



Bipartite graph is an undirected graph $G = (V, E)$ in which V can be partitioned into 2 sets U and V such that $(u, v) \in E$ implies either $u \in U$ and $v \in V$ or

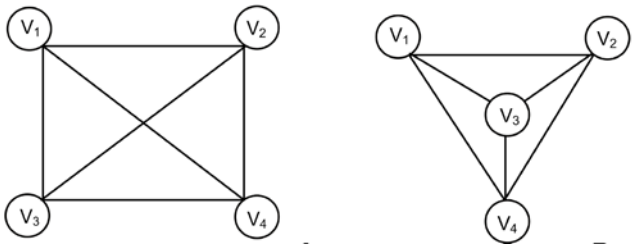
$v \in U$ and $u \in V$. Applications of this type of graph to visualization or modeling of biological networks range from representation of enzyme-reaction linked metabolic networks to ontologies.



The degree of a node in an undirected graph is the number of connections (or edges) the node has to other nodes and is defined as deg(i) = k(i) = |N(i)| where N(i) is the number of the nodes adjacent to node i. If a network is directed, then each node has two different degrees, the in-degree degin(i) which is the number of incoming edges to node i, and the out-degree degout(i) which is the number of outgoing edges from node i.

Graph Isomorphism

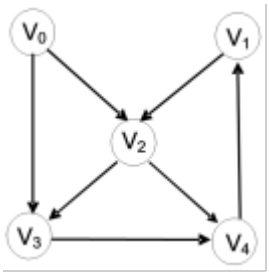
Let G_1= (V_1, E_1) and G_2= (V_2, E_2) be two undirected graphs. $f : V_1 \rightarrow V_2$ is called isomorphism if f is an *edge-preserving* mapping; that is for all $a, b \in V_1$, $(a,b) \in E_1$ if and only if $(f(a), f(b)) \in E_2$.



The two graphs have different topology but they are isomorphs.

Adjacency Matrix

The most common data structures that are used to make these networks computer readable are adjacency matrices. Given a graph G = (V, E) the adjacency matrix consists of a $|V| \times |V| = n \times n$ matrix $A = [a_{ij}]$ such that $a_{ij} = 1$ (or $= w_{ij}$ for a weighted edge) if $(i, j) \in E$ or $a_{ij} = 0$ otherwise.



$$A=\begin{bmatrix}0&0&1&1&0\\0&0&1&0&0\\0&0&0&1&1\\0&0&0&0&1\\0&1&1&0&0\end{bmatrix}$$

2.3 Network Measures

Looking at different network properties can provide valuable insight into the internal organization of a biological network.

The graph density *g* defined as density $g=2|E|/[|V|(|V|-1)]$ shows how dense or sparse a graph is according to the number of connections per node set. A sparse graph is a graph where $|E|=O(|V|^{\gamma})$ with $1<\gamma<2$. It has been argued that biological networks are generally sparsely connected, as this can confer an evolutionary advantage for preserving robustness. A *complete* graph is a graph in which every pair of nodes is *adjacent*. A clique is a complete subgraph of an undirected graph *G*.

A *walk* is a pass through a specific sequence of nodes $(v_1,v_2,...,v_L)$, which are connected with edges $\{(v_1,v_2),(v_2,v_3),...,(v_{L-1},v_L)\}\subseteq E$. A *simple path* is a *walk* with no repeated nodes. A *trail* is a path where no edge can be repeated.

The distance $\delta(i,j)$ from *i* to *j* is the length of the shortest path from *i* to *j* in *G*. If no such path exists, then we set $\delta(i,j)=\infty$. The average path length of a graph *G* is defined to be the average of shortest path lengths $\delta(i,j)$ over all pairs of distinct nodes $i,j\in V(G)$: $\bar{\delta}=\frac{2}{N(N-1)}\sum_{i=1}^{N}\sum_{j=1}^{N}\delta(i,j)$, where $\delta(i,j)$ is the minimum distance between nodes *i* and *j*. The diameter of a graph *G* is the maximum of the shortest path from *i* to *j* in *G*, defined as $D=\max_{i,j}[\delta(i,j)]$.

Local Network Measures

(1) The Degree Distribution $P(k)$

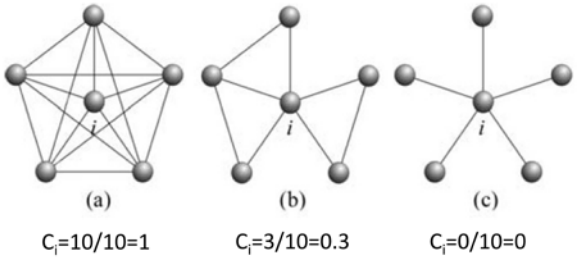
It is the distribution of percentages of nodes that have degree *k* (*i.e.*, the number of edges it has). Node degree in PPI networks correlates with gene essentiality, conservation rate, and disease causing likelihood. Networks that have a power-law degree distribution are called 'scale free' networks because the shape of degree distribution does not change with the size of the network.

(2) The Distribution of Clustering Coefficient

It is a measure that shows the tendency of a graph to be divided into clusters. A

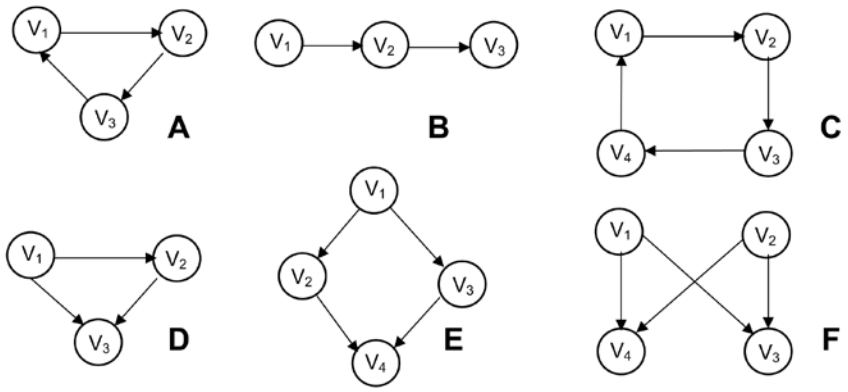
cluster is a subset of vertices that contains a lot of edges connecting these vertices to each other.

Assuming that i is a vertex with degree $\deg(i) = k(i)$ in an undirected graph G and that there are $e(i)$ edges between the k neighbors of i, then the local clustering coefficient of i in G is given by $C_i = \frac{2e(i)}{k(i)[k(i)-1]}$. The closer the local clustering coefficient is to 1, the more likely it is for the network to form clusters.



(3) Network Motifs

Network motifs are small subgraphs that occur in a network statistically more often than in randomized networks. Motifs can define classes of networks, because networks of similar types usually share many motifs, whereas networks of different types do not. However, focusing only on overrepresented patterns in a network can lead to losing valuable biological information about patterns that are functionally significant but not over-represented. Signal transduction and gene regulatory networks tend to be described by various motifs.



Some network motifs: A) Three-node feedback. B) Three-node chain. C) Four-node feedback. D) Feed-forward loop. E) Biparallel. F) Bi-Fan.

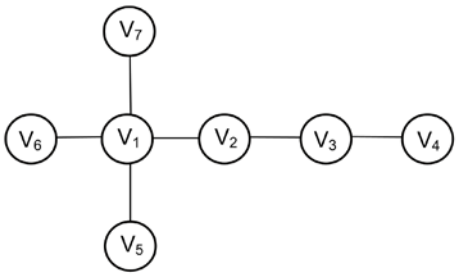
(4) Node Ranking with Network Centralities

In biological networks, it is important to detect central nodes or intermediate nodes that affect the topology of the network.

Degree Centrality $C_d(i)$ can reflect an important node involved in a large number of interactions. For a node i, the degree centrality is calculated as $C_d(i) = \deg(i)$. Nodes

with very high degree centrality are called *hubs* since they are connected to many neighbors. The removal of such central nodes has great impact on the topology of the network. Biological networks are robust against random perturbations, but disruption of hubs often leads to system failure.

Closeness Centrality $C_{clo}(i)$ indicates important nodes that can communicate quickly with other nodes of the network $C_{clo}(i) = N_{acc}(i) \Big/ \sum_{j \in V}^{N} \delta(i, j)$ with $N_{acc}(i)$ being the number of nodes accessible by node *i*.



For the network shown above, the node V_1 can access 4 nodes (V_2, V_5, V_6, V_7) with step 1, 1 node (V_3) with step 2, and 1 node (V_4) with step 3, the sum of the shortest path lengths is $d_1 = \sum_j \delta(1, j) = 4 \times 1 + 1 \times 2 + 1 \times 3 = 9$. There are 6 nodes accessible from node V_1, so $C_{clo}(1) = 6/9$.

For the node V_2, 6 nodes can also be accessed by V_2. The sum of the shortest path lengths is $d_2 = \sum_j \delta(2, j) = 2 \times 1 + 4 \times 2 = 10$, so $C_{clo}(2) = 6/10$. As a result, V_1 is more central than node V_2.

Betweenness Centrality $C_b(i)$ shows that nodes which are intermediate between neighbors rank higher. High betweenness centrality reflects important nodes that lie on a high proportion of paths between other nodes in the network and become "bottlenecks", for their role as key connector with essential functional and dynamic properties.

For the same network shown above, $N_p(1) = 12$ because there are 12 shortest paths that pass through node V_1. These paths from the starting to the ending node are {V_2-V_5, V_2-V_6, V_2-V_7, V_3-V_5, V_3-V_6, V_3-V_7, V_4-V_5, V_4-V_6, V_4-V_7, V_5-V_6, V_5-V_7, V_6-V_7}.

$N_p(2) = 8$ because 8 there are shortest paths passing through node V_2. These paths are {V_1-V_3, V_1-V_4, V_3-V_5, V_3-V_6, V_3-V_7, V_4-V_5, V_4-V_6, V_4-V_7}.

$N_p(3) = 5$ with {V_1-V_4, V_2-V_4, V_4-V_5, V_4-V_6, V_4-V_7}.

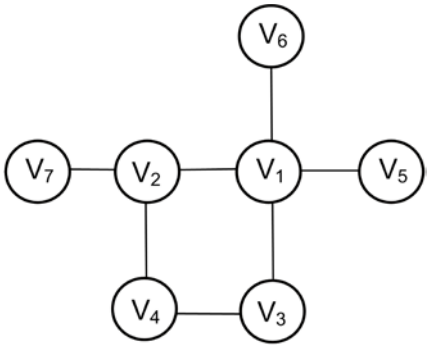
$N_p(4) = N_p(5) = N_p(6) = N_p(7) = 0$. The total sum of shortest paths that pass through the nodes was calculated to be $N_p = \sum_j N_p(j) = 25$. Thus the centralities are

$C_b(1) = 12/25 = 0.48$, $C_b(2) = 8/25 = 0.32$, $C_b(3) = 5/25 = 0.20$,
$C_b(4) = C_b(5) = C_b(6) = C_b(7) = 0$, indicating node V_1 to be more central.

Eigenvector Centrality ranks higher the nodes that are connected to important neighbors.

Eccentricity Centrality is the measure that shows how easily accessible a node is from other nodes. Let $G = (V, E)$ be an undirected graph. The eccentricity centrality is defined as $C_{ecc}(i) = 1/\max_j[\delta(i, j)]$, where $\delta(i, j)$ is the shortest path between nodes i and j. The eccentricity $C_{ecc}(i)$ Cecc of a node i is the greatest distance between i and any other node.

With the following tool network as an example,



V_1 can access 4 nodes (V_2, V_3, V_5, V_6) with step 1 and 2 nodes (V_4, V_7) with step 2. The maximum shortest path $\max_j[\delta(1, j)] = 2$.

V_2 can access 3 nodes (V_4, V_7, V_1) with step 1 and 3 nodes (V_3, V_5, V_6) with step 2. The maximum shortest path $\max_j[\delta(2, j] = 2$.

V_3 accesses 2 nodes (V_1, V_4) with step 1, 3 nodes (V_2, V_5, V_6) and one node (V_7) with step 3, leading to $\max_j[\delta(3, j] = 3$;

V_4: 2 × 1, 2 × 2, 2 × 3; The maximum shortest path $\max_j[\delta(4, j] = 3$;

V_5: 1 × 1, 3 × 2, 2 × 3; The maximum shortest path $\max_j[\delta(5, j] = 3$;

V_6: 1 × 1, 3 × 2, 2 × 3; The maximum shortest path $\max_j[\delta(6, j] = 3$;

V_7: 1 × 1, 2 × 2, 3 × 3; The maximum shortest path $\max_j[\delta(7, j] = 3$.

As a result, the ordering of the nodes according to $C_{ecc}(i)$ is $\{V_1,V_2\},\{V_3,V_4,V_5,V_6,V_7\}$. In biological networks, proteins with high eccentricity are easily reachable by other components of the network, and thus can readily perceive changes in concentration of other proteins they are linked to. In contrast, those proteins that have lower eccentricities will often play a marginal functional role in the system.

Subgraph Centrality is the measure that ranks nodes according to the number of subgraphs of the overall network in which the node participates, with more weight given to small subgraphs.

Global Network Measures

Some useful global network measures include (1) average degree of a network: $d = \sum_k kP(k)$; (2)Number of edges in a network: $Nd/2$; (3)Average network diameter $D = \max_{i,j}[\delta_{\min}(i, j)]$; and (4)The average clustering coefficient of the whole network given by $C_{ave} = \frac{1}{N}\sum_{i=1}^{N}\frac{2E_i}{k_i(k_i-1)}$, where $N=|V|$ is the number of vertices.

It was noted that biological networks have a significantly higher average clustering coefficient compared to random networks, which proves their modular nature. But these properties alone are not descriptive enough to capture complex topological characteristics of biological networks. Local network properties represent more constraining measures of network structure than global network properties and provide additional means for describing networks.

2.4 Network Models

To visually represent the properties of the network we usually rank the vertices according to their degree and then plot the degree versus the rank of each vertex. Another representation is to create a histogram by plotting the vertices of the graph sorted according to their degree using a logarithmic scale. A third and very popular

representation is to plot the degrees of the nodes sorted versus either their degree distribution $P(k)$.

Erdös-Rényi model for random networks

This model was mainly introduced to describe the properties of a random network. The simple model involves taking a number of vertices N and connecting nodes by selecting edges from the $N(N\text{-}1)/2$ possible edges randomly. The probability of obtaining a random graph G with N nodes and n edges is given by $P(G) = p^n(1-p)^{N(N-1)/2-n}$. Thus, the probability of a vertex to have degree k becomes $P(k) = e^{-<k>} < k >^k / k!$, where $< k >$ is the average connectivity of the network. The average degree has a value of $< k >= 2n/N = p(N-1) \simeq pN$ does have a critical value that is independent of the system size. The probability that two of the neighbors in a random network are connected is equal to the probability that two randomly selected nodes are connected. Consequently the clustering coefficient of a random graph is $C_{rand} = p = < k > / N$.

Watts and Strogatz model for small-world networks

This model was introduced to describe networks with the small world topology. This type of topology characterizes many biological networks, like metabolic networks where paths of few (3-4) reactions link most metabolites. As a consequence, local changes in metabolite concentration will propagate throughout the entire network. In this model, the degree distribution follows a power-law equation $P(k) = k^{-\gamma}$, in which most nodes are connected with small proportion of other nodes and a small proportion of nodes are highly connected. Thus each vertex is connected to $N/2$ nearest neighbors.

Barabasi-Albert model for scale-free networks

The networks are built to mimic gene duplication events, such that they expand continuously by addition of new nodes and the new nodes attach preferentially to sites that are already well connected.

We start with small number of nodes m_0. At each step, a new node $m < m_0$ is added and gets linked to the existing network. The probability that a new node is connected to node i is $P(k_i) = k_i / \sum_j k_j$, where k_i is the degree of node i. The rate of connecting new nodes to node i is $\partial k_i / \partial t = \Delta k [k_i / \sum_j k_j] = m k_i / (2mt) = k_i / (2t)$. The connection is time-dependent so $k_i(t) = m\sqrt{t/t_i}$, where t_i is the time point when node i enters network. The probability that a node has degree smaller than k is $t_i > m^2 t / k^2$. So the

probability density of the network is $P(k) = \partial p(k_i(t) < k) / \partial k \sim k^{-3}$, a power law distribution of $\gamma \sim 3$.

2.5 Integration of Networks and Data

To understand a living cell, one needs to study all of its components as an interconnected system rather than a collection of individual parts. Current high-throughput technologies do not capture the details of spatial and time heterogeneity of interactions. To understand complex biological phenomena, we should try to combine and use all biological data that are available. Biological networks are commonly combined with microarray data, proteomics data, metabolomics data, genomic data, isotope labeling experiments and biomarkers. At present, proteomics technologies are still in development and many proteins, especially those with low abundance, are difficult to quantify. When using transcriptomics in place of proteomics, one should be aware that some factors, e.g., posttranscriptional and posttranslational regulations, may lead to poor correlation between RNA abundance and protein levels.

Metabolic networks include all biochemical reactions inferred from genome annotations. These networks in their stoichiometric form can be used in diverse applications, ranging from estimating the flux distribution of an organism under specific conditions to understanding of the robustness and evolution of metabolism. Recent improvements on the accuracy of metabolomics have enabled estimating parameters of kinetic models of large-scale metabolic network to convert the '-omics' data into model parameters.

Unlike metabolic networks, which represent flow of mass, gene regulatory networks (GRNs) represent information flow. Hence, algorithms for integration of high-throughput data with GRNs require different computational approaches. Parts of the GRN that are activated under specific time points and conditions, called activated subnetworks, can be identified by integrating GRN with gene expression data. A cell might use different activated subnetwork to respond to different environmental stresses. It is easier to find activated subnetworks if the GRN has a modular structure. Topological analysis of the GRN can identify the isolated modules. Microarray data from different environmental stimuli can be combined with GRN modules by calculating *covariance* between the input stimuli and its transcriptional response to identify subnetworks activated by the stimuli.

Compared to GRN, PPI networks are larger, more suitable for applying omics approaches and can be used to predict protein function. Network-based approaches for predicting protein function or involvement in a disease can be divided into two types: ‘direct annotation’ approaches, which infer the characteristic of a protein based on its connections in the network; and ‘module-assisted’ approaches, which first identify modules of related proteins, and then annotate each module based on characteristics of its annotated members.

Network-based approaches were used to study drug resistance, which is one of the major challenges in current treatment strategies, from infectious disease to cancer. When cells are exposed to hostile environment, such as drug treatment, cells tend to rewire themselves to survive in the new environment. It is a multifactorial process and can involve several proteins. Therefore, a systems level approach is required to understand the mechanism of drug resistance and to identify combinatorial therapy.