

# Chapter 7

## Complex Networks and Systems Biology

Ushasi Roy, Rajdeep Kaur Grewal and Soumen Roy

**Abstract** Modern biology has decisively moved in a direction where we scrutinise systems holistically rather than looking at entities in different levels discretely or in isolation. Unlike previous reductionist approaches; in this new approach called Systems Biology, networks play a crucial role in arriving at and summing up the holistic picture and in understanding the emergent properties of the system. In this chapter, we give an overview of how network approaches are useful at various levels in biology. After a conceptual introduction to networks and various network metrics used to quantify networks; we discuss various concepts like network motifs and random networks. We then examine at length about how networks shed insight at virtually every layer of life like gene regulatory networks, networks involving proteins and metabolic networks. We end the chapter with a discussion of the application of networks to epidemiology.

**Keywords** Network · Directed networks · Weighted networks · Degree · Degree distribution · Assortativity · Shortest path length · Connectedness · Eccentricity · Diameter · Closeness centrality · Betweenness centrality · Clustering coefficient · Cliques · Community structure · Modularity · k-core decomposition · Erdos-Renyi graphs · Small-world · Scale-free · Motifs · Feed forward loops · Gene Regulatory Network (GRN) · Protein Structure Network (PSN) · Protein Energy Network (PEN) · Allosteric · Protein Protein Interaction network (PPI networks) · Protein folding network · Metabolic networks · Epidemiology · Susceptible Infectious Recovered (SIR) · Susceptible Infectious Susceptible (SIS)

### 7.1 Introduction

#### 7.1.1 Systems Biology

The study of biological systems has historically been a largely phenomenological or observational science. However, in the last quarter of the twentieth century; in-depth

---

S. Roy (✉) · R. K. Grewal · U. Roy  
Bose Institute, 93/1 Acharya Prafulla Chandra Roy Road,  
Kolkata 700 009, India  
e-mail: soumen@jcbosc.ac.in

quantitative studies of various biological phenomena started gaining momentum. Over the course of the last decade and half, the advent of high-throughput technologies have only made the application of quantitative techniques imperative to biology. They also inculcated the realisation that biological systems are far too complex to be solved by classic reductionist approaches. It was becoming increasingly apparent that the study of biological systems need an integrated, multidisciplinary approach whose essence is underscored by an effective cycle of modelling and experimentation. “Systems” approaches are definitely poised to occupy mainstream biology over the course of the next decade or so. These approaches examine the structure and dynamics of cellular and organismal function, contrary to the study of isolated parts of cell or organism (Kitano 2002). Thus, “Systems Biology” is a new branch of science which integrates techniques from Mathematics, Physics, Chemistry, Computer Science, Engineering and Information theory to model various biological phenomena from a holistic point of view.

Intrinsic to this development, is the concept of “emergent properties” which refer to holistic properties at the system level, since the behaviour of the system as a whole will not merely be an agglomeration of the properties of its segregated constituents. For studying this composite system, consolidation of the diverse interactions among various components of the system is required. The theory of networks which is based on a well established graph-theoretic approach; enables us to do so efficiently (Albert et al. 2002; Newman 2010).

## 7.1.2 Networks

From the perspective of Graph Theory, a network can be represented by a graph. A graph is defined as  $G = \{V, E\}$  where  $V$  is the set of nodes (or vertices or simply points) and  $E$  denotes the set of edges (or links or arcs or simply lines), which establishes an interconnection among the nodes. A real complex system can be mapped onto a network structure where one needs to identify the major components of the system as the nodes and the interactions among them as the edges. This concept has been illustrated below by two simple graphs. In Fig. 7.1a, the set of nodes  $V = \{a, b, c, d\}$  and the set of edges is given by  $E = \{(e_1 = (a, b)), (e_2 = (b, c)), (e_3 = (a, c)), (e_4 = (c, d))\}$ . Similarly,  $V = \{v_1, v_2, v_3, v_4, v_5\}$  and  $E = \{(e_1 = (v_1, v_5)), (e_2 = (v_2, v_5)), (e_3 = (v_2, v_3)), (e_4 = (v_3, v_5)), (e_5 = (v_4, v_5))\}$  correspond to the set of nodes and edges in Fig. 7.1b.

### 7.1.2.1 Subgraph

A subgraph  $G' = \{v', e'\}$ , having  $v'$  vertices and  $e'$  edges is defined to be a subgraph of  $G = \{V, E\}$  if  $v'$  is a subset of  $V$  and  $e'$  is a subset of  $E$ .

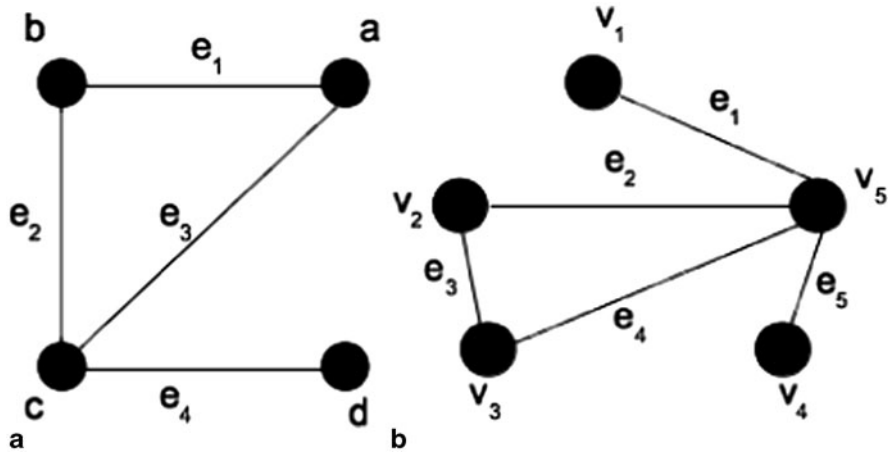


Fig. 7.1 (a) and (b): Simple graphs

### 7.1.2.2 Directed Networks

In a directed network, the edges have a direction, i.e., identification of the “source” and “sink” nodes for a particular connection is important. Thus, a particular node will have both incoming and outgoing edges and will have different in and out degree distributions. Many important networks, viz., World Wide Web (WWW) and metabolic networks are directed in nature.

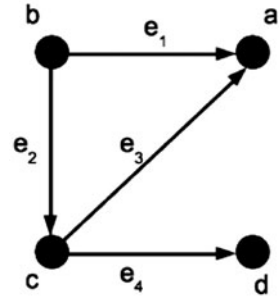
### 7.1.2.3 Weighted Networks

Generally we construct binary networks with the edge weights having two possible values, 0 and 1; representing absence and presence of connections respectively. In contrast, many real networks are weighted in nature. In these networks, in addition to the binary values, edge weights can have any fractional values in between 0 and 1, depending on the strength of interactions. Here all the edges are not equally important and the edge with higher edge weight will have a higher significance in the network. Examples are social networks, internet and cellular networks as they are characterized by the level of acquaintance between individuals, band widths and reaction rates which may have different values (Fig. 7.3).

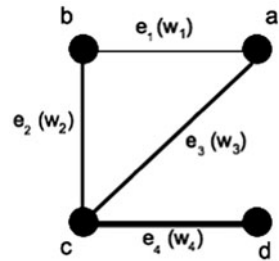
## 7.2 Network Metrics

Network metrics help in the characterisation of a given network—both quantitatively and qualitatively. Their significance lies in analysing both the local property, i.e., the individual behaviour of nodes or edges, as well as the global property of the whole

**Fig. 7.2** A simple directed graph with node set  $V = \{a, b, c, d\}$  and the set of directed edges  $E = \{(e_1 = (b, a)), (e_2 = (b, c)), (e_3 = (c, a)), (e_4 = (c, d))\}$  where the first node in the edge set denotes origin while the second one represents the end of an edge



**Fig. 7.3** A simple undirected weighted graph with the set of nodes defined as  $V = \{a, b, c, d\}$  and the set of edges  $E = \{e_1, e_2, e_3, e_4\}$  having edge weight  $W = \{w_1, w_2, w_3, w_4\}$



network. These structural network metrics may also serve as a great tool for exploring the unified behaviour of the network.

### 7.2.1 Degree

A degree of a node is defined as the number of edges incident on that node. It signifies the number of connections made by a node  $i$  with the remaining nodes in the network, termed as neighbours of node  $i$ . The nodes which have comparatively much higher degree than that of the other nodes in a network correspond to the *hub*.

For directed networks, degree of a node is specified using two distinct centrality measures *in-degree* and *out-degree*. In a directed network, the number of edges directing outward from the particular node is its out-degree and the number of nodes directing towards it correspond to the in-degree of that node in a network. In Fig. 7.1a, the degree of each of the nodes  $\{a, b, c, d\}$  in the graph  $G$  are  $\{2, 2, 3, 1\}$  respectively. For the directed graph  $H$  in Fig. 7.2 the in-degree and out-degree of the nodes  $\{a, b, c, d\}$  are  $\{2, 0, 1, 1\}$  and  $\{0, 2, 2, 0\}$  respectively.

### 7.2.2 Degree Distribution

The degree distribution  $P(k)$ , the probability that a randomly chosen node has degree  $k$  or fraction of nodes in the network having degree  $k$ , of a network provides one

of the basic topological characterisation of a network. Various types of networks can sometimes be distinguished by their degree distribution. For instance, scale free networks have a power law degree distribution,

$$P(k) \approx k^{-\gamma} \quad (7.1)$$

and it has been claimed that when  $2 \leq \gamma \leq 3$ ; the hubs play a significant role in the network (Barabasi and Oltvai 2004). In contrast, small random networks follow Binomial distribution which in the limit of large  $N$  approaches the Poisson distribution

$$P(k) \approx e^{-\langle k \rangle} \frac{\langle k \rangle^k}{k!} \quad (7.2)$$

where  $\langle k \rangle$  denotes the average degree of the graph. For directed networks, there might be different distributions of in-degree, out-degree and total degree of the nodes in the network.

### 7.2.3 Assortativity

Assortativity refers to the affinity of nodes in a network to become linked to other nodes having similar degree distribution. This tendency of correlation among nodes of similar degree is also sometimes called as assortative mixing. In contrast, sometimes high degree nodes are somewhat inclined towards low degree nodes. This kind of dissimilar preferential attachment gives rise to a disassortative network. Most biological and technological networks exhibit disassortative mixing while social networks belongs to the former class, i.e., they are assortative in nature. Mathematically, *assortativity* of a complex network can be expressed as

$$r = \frac{\langle k_1 k_2 \rangle - \langle k_1 \rangle \langle k_2 \rangle}{\sigma_k^2} \quad (7.3)$$

where the averages are taken over all edges and  $\sigma_k^2$  is the variance of the node-degree  $k$ . For all practical purposes, calculating assortativity of real world networks, the above equation can be modified as (Newman 2002)

$$r = \frac{E^{-1} \sum_e j_e k_e - [E^{-1} \sum_e \frac{1}{2}(j_e + k_e)]^2}{E^{-1} \sum_e \frac{1}{2}(j_e^2 + k_e^2) - [E^{-1} \sum_e \frac{1}{2}(j_e + k_e)]^2} \quad (7.4)$$

where  $j_e, k_e$  are the degrees of the nodes at the ends of the  $e$ th edge, with  $e = 1, 2, \dots, E$ .

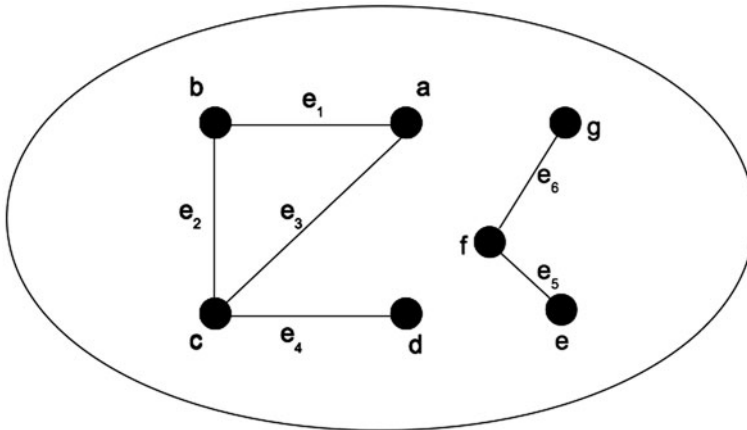


Fig. 7.4 A graph  $G$  with two disconnected components

### 7.2.4 Shortest Path Length

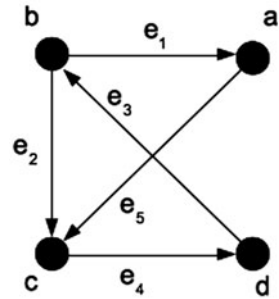
A *path* is an alternate sequence of nodes and edges, starting and ending with a node, such that each edge in the sequence is incident on the node preceding and following it. There is no repetitions of nodes and edges in a path. In Fig. 7.1a  $\{a, e_1, b, e_2, c, e_4, d\}$  represents a path connecting the nodes  $a$  and  $d$ . *Shortest path* between a pair of vertices  $(i, j)$ , where  $i, j \in V$ , in a graph is the geodesic distance ( $d_{ij}$ ) between them i.e the minimum number of edges traversed while moving from node  $i$  to node  $j$ .

### 7.2.5 Connectedness

A graph is said to be connected if there exists at least a path between any pair of nodes constituting the graph. It may so happen that there exists a pair of nodes in a graph having no path connecting them. Such graphs are known as disconnected graphs. For a disconnected graphs, each connected component is termed as a *cluster*. *Giant cluster* in a network refers to the largest connected component of the network (Fig. 7.4).

Directed graphs, in terms of connectedness, are defined to be strongly or weakly connected graphs. If each pair of nodes in the directed graph has at least one directed path (each edge in the sequence is incident out- and in- on the node preceding and following it, respectively) between them, the graph is said to be strongly connected. If the underlying undirected graph (graph obtained from the directed graph by removing the directions of edges from it) of the directed graph is connected, we call it as weakly connected graph. It is quite obvious that a strongly connected graph will definitely be a weakly connected graph (Fig. 7.5).

**Fig. 7.5** An example of a strongly connected graph: Say, for example, the set of directed paths  $P$  from node  $a$  to the other three nodes is given by  
 $P = \{(a, e_5, c, e_4, d, e_3, b), (a, e_5, c), (a, e_5, c, e_4, d)\}$



The connection between a pair of nodes in a network is often represented by *adjacency matrix* or connection matrix. The adjacency matrix of the graph in Fig. 7.1 a of  $N$  nodes and no parallel edges is an  $N$  by  $N$  symmetric binary matrix  $A = [a_{ij}]$ , where

$$x_{ij} = 1, \text{ if there is an edge between node } i \text{ and } j \\ = 0, \text{ if there is no edge between them}$$

$$A = \begin{matrix} & \begin{matrix} a & b & c & d \end{matrix} \\ \begin{matrix} a \\ b \\ c \\ d \end{matrix} & \begin{pmatrix} 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{pmatrix} \end{matrix} \quad (7.5)$$

### 7.2.6 Average Shortest Path Length

*Average Shortest Path Length* ( $l$ ) or the characteristic path length of a network is the sum of all the shortest path lengths between each pair of nodes in a graph averaged over all possible edges in a network.

$$L = \frac{1}{N(N-1)} \sum_{i,j \in V, i \neq j} d_{ij} \quad (7.6)$$

The above definition, however, fails in case the network has more than one connected component. One way of dealing with it is to restrict the sum over the nodes belonging to the largest connected component of the network. Another approach is to assign infinite distance between the pair of disconnected nodes or the pair of nodes having no connected path, and then take the harmonic mean of the shortest path between

the pair of nodes in the network. The latter gives a quantitative measure, called the Efficiency of the network, which is defined as follows

$$E = \frac{1}{N(N-1)} \sum_{i,j \in V, i \neq j} \frac{1}{d_{ij}} \quad (7.7)$$

### 7.2.7 Eccentricity

Eccentricity  $E(i)$  of a node  $i$  in a graph  $G$  is the maximum value of all the geodesic distances calculated from that particular node  $i$  to all other nodes  $j$  in the network.

$$E(i) = \max_{j \in V} d(i, j) \quad (7.8)$$

The eccentricity of a node  $i$  represents how close or distant is  $i$  from the farthest node of the network. The node with minimum eccentricity in graph  $G$  is called the *centre* of  $G$ .

### 7.2.8 Diameter

The diameter of a graph refers to the maximal distance between any pair of its nodes. The diameter of a disconnected network, composed of more than one isolated components or clusters, is infinite. So, for practical purposes, in such cases, it may be defined as the maximum diameter of its components.

### 7.2.9 Closeness Centrality

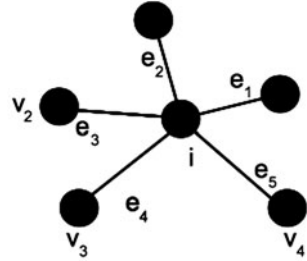
The closeness centrality  $C$  of a node  $n_i$  is the inverse of the sum of its distances to all other nodes,  $n_j$ . Mathematically, it is defined as

$$C(n_i) = \frac{N-1}{\sum_{j=1}^g d(n_i, n_j)} \quad (7.9)$$

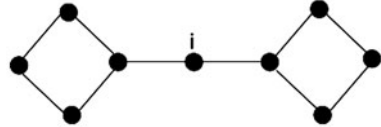
Closeness of a node signifies the efficiency of a node to convey information within the network. For example, consider a star graph as shown in Fig. 7.6. In this graph the node  $i$  is the most centrally located node in the graph. Thus, it spreads information much faster than any other node in the network can.



**Fig. 7.6** Star graph



**Fig. 7.7** In this figure, the node having highest betweenness is  $i$



### 7.2.10 Betweenness Centrality

The betweenness centrality of a node measures the node’s involvement in the communication paths of other nodes in the network.

$$B(v) = \sum_{s \neq v \neq t} \frac{\sigma_{st}(v)}{\sigma_{st}} \tag{7.10}$$

where  $\sigma_{st}$  is the total number of shortest paths from node  $s$  to node  $t$  and  $\sigma_{st}(v)$  is the number of those paths that pass through  $v$  (Freeman 1977).

For better understanding of this centrality, consider the graph shown in Fig. 7.7. Here, nodes in the graph can be divided into two groups. These two group of nodes are connected by a single node  $i$ . Hence the betweenness centrality value of node  $i$  is the highest among others. If one wants to travel from one node lying in one cluster to another in the other cluster, then the path passing through node  $i$  is the only way. Another important realisation of this centrality can be gained while analysing this graph. If the node  $i$  from the network is removed (along with the edges incident on it), the graph becomes disconnected, with two connected components. Removal of the high betweenness nodes will result in either of the two following consequences. In one case, the communication among different clusters may get completely lost, as in the above mentioned example. In the other one, the cost of traveling may get enhanced since the path will comprise of more edges than before. These high betweenness nodes are often called as bottlenecks of the network.

### 7.2.11 Clustering Coefficient

It is a measure which accounts for the tendency of a node in a network to cluster together. This behaviour is commonly observed in most real world networks, in

particular social networks. Clustering can either be global or local, depending on the overall clustering of nodes in the whole network or the property of the single node. The definition of *Global Clustering Coefficient* (GCC) is based on the concept of triples of nodes. A triple consists of three nodes which remain connected by either three (closed triple) or two (open triple) undirected edges. GCC is the ratio of the number of triangles to the number of connected triples.

$$C' = \frac{3 \times \text{Number of triangles}}{\text{Number of connected triples}} \quad (7.11)$$

The *Local Clustering Coefficient* (LCC) of a node in a graph gives a quantification of the proximity of its neighbours from becoming a completely connected graph. It can be defined in the following way. A node,  $i$  with  $k_i$  neighbours, can have, at most,  ${}^{k_i}C_2 = \frac{k_i(k_i-1)}{2}$  number of possible edges in its neighbourhood. Suppose, the neighbours of node  $i$  are connected by  $e_i$  edges, then the LCC of that node is defined as

$$c_i = \frac{2e_i}{k_i(k_i - 1)} \quad (7.12)$$

Therefore the Clustering Coefficient of the whole graph can be obtained by taking average of  $c_i$  over all the nodes in G:

$$C = \langle c \rangle = \frac{1}{N} \sum_{i \in N} c_i \quad (7.13)$$

### 7.2.12 Cliques and Community Structure

In a complex network having large number of nodes and edges, a  $k$ -*Clique* is defined as a completely connected subgraph having a set of  $k$  nodes in which each node is connected to every other node by an edge in that subgraph. Two  $k$ -cliques will belong to the same *community* when they share  $k - 1$  nodes.

### 7.2.13 Modularity

A relatively independent unit, called modules (also called groups, clusters or communities), is often present in a complex network. Modularity is a quantitative measure which describes the extent to which a system is divided into modules. A network with high modularity value will be endowed with intense connections among nodes within a module but sparse or minimal links to other modules in the network. Mathematically, modularity is defined as

$$M = \sum_{i=1}^m \left[ \frac{e_i}{E} - \left( \frac{d_i}{2E} \right)^2 \right] \quad (7.14)$$

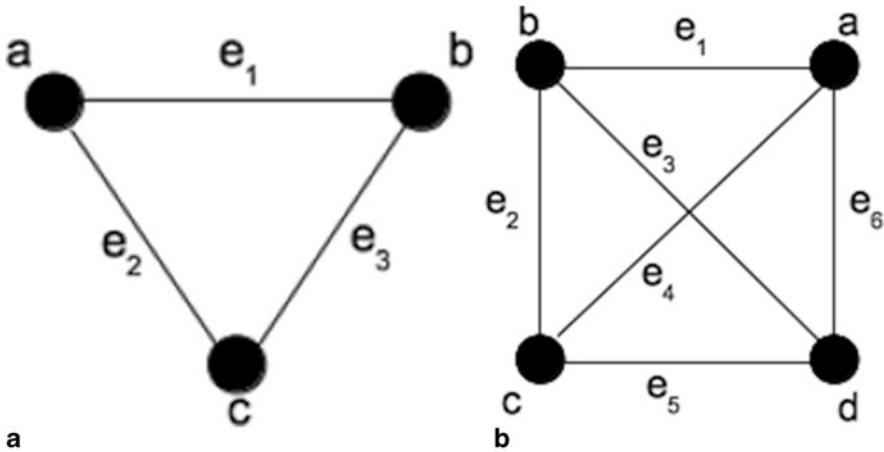


Fig. 7.8 a 3-clique and b 4-clique

where  $E$  is the total number of edges in the network,  $e_i$  is the number of edges within module  $i$ ,  $d_i$  is the sum of degrees of all the nodes of module  $i$ , and the summation runs over total number of modules  $m$  in the network (Fig. 7.8).

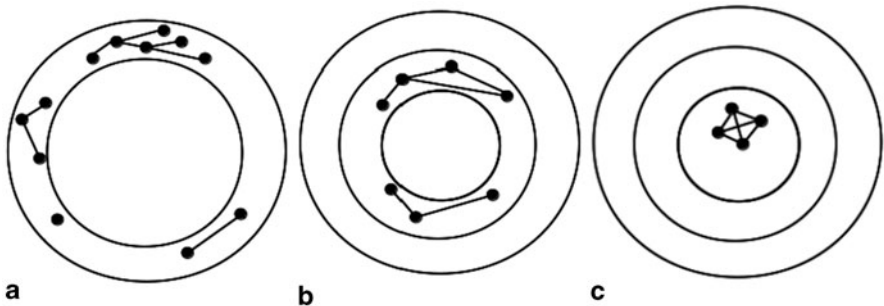
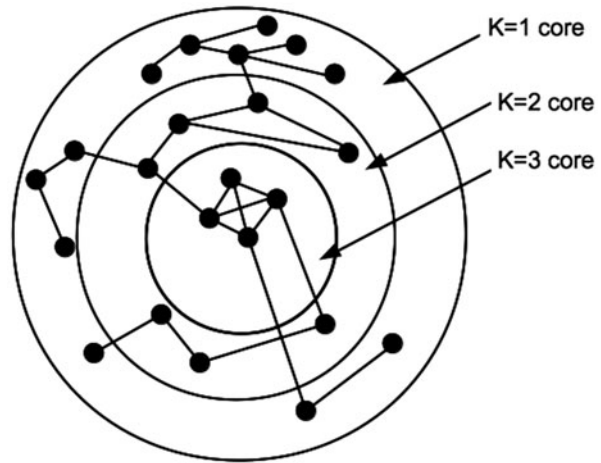
### 7.2.14 *k*-Core (or *k*-Shell) Decomposition

*K*-core decomposition method provides us a hierarchical representation of the network. A  $k$ -core of a graph  $G$  is a maximal subgraph of  $G$  in which each node is connected to at least  $k$  other nodes in the subgraph. A node  $i$  belongs to a  $k$ -shell if and only if it belongs to the  $k$ th-core but not to the  $k + 1^{\text{th}}$ -core.

The  $k$ -core decomposition is based on sequential removal of nodes along with its edges. Let us consider a connected graph  $G$ . At first, all nodes with degree  $d = 1$  are removed from the graph  $G$ . After their removal, new nodes with degree  $d = 1$  may appear in  $G$ . The pruning process is continued until all the nodes with degree  $d = 1$  are removed. These nodes together with their incident edges forms the  $k_s = 1$  shell. In a similar fashion the higher degree nodes are removed to obtain the  $k_s = 2$  shell and so on. The process is repeated until all the nodes from the graph  $G$  have been removed.

The network topology plays a significant role in portraying the interactions within the nodes. Such decomposition have been used by many researchers to analyse the real world networks (Wuellner et al. 2010). The  $k$ -core decomposition of PPI network of yeast has revealed that the proteins belonging to the innermost core have higher probability of being both essential and evolutionary conserved (Wuchty et al. 2005). Judicious introduction of new parameters like synthetic accessibility have demonstrated sufficient promise in predicting the viability of knockout strains with accuracy comparable to approaches using biochemical parameters (like FBA etc.)

**Fig. 7.9** *k*-core decomposition of a simple graph



**Fig. 7.10** *k*-shells ( $k_s$ ) of the graph  $G$  in Fig. 7.9. **a**  $k_s = 1$ , **b**  $k_s = 2$  and **c**  $k_s = 3$

on large, unbiased mutant data sets (Wunderlich et al. 2006). Another recent topic where network metrics are thought to play a significant role is the controllability of biological networks (Banerjee et al. 2012; Fig. 7.10).

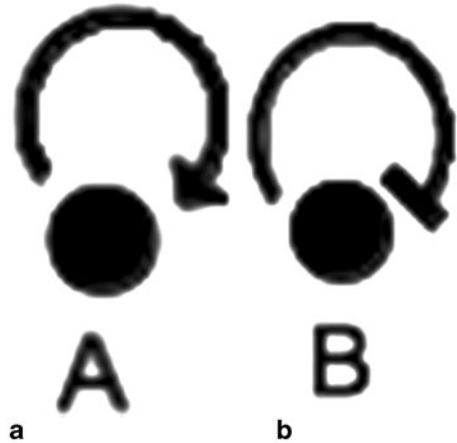
In this section we have hopefully presented an elaborate introduction to network metrics. Recent research has however conclusively shown that instead of looking at just one or two metrics, it is imperative that we look at multiple metrics in parallel to get the most informative picture (Filkov et al. 2009; Roy 2012, 2014).

## 7.3 Random Graph Theory

### 7.3.1 Erdos-Renyi Graphs (ER Graphs)

Erdos-Renyi Graphs are random graphs where edges are constructed between all pairs of nodes with some equal probability (say  $p$ ), independent of one another. The

**Fig. 7.11** **a** Positive autoregulation: activation of gene *A* by its own product, **b** Negative autoregulation: deactivation/inhibition of gene *B* by its own product



degree distribution profile of ER graphs shows Poisson distribution. The ER Graphs have low clustering coefficients and the average path length are found to be smaller compared to the real world networks.

### 7.3.2 *Small World Networks*

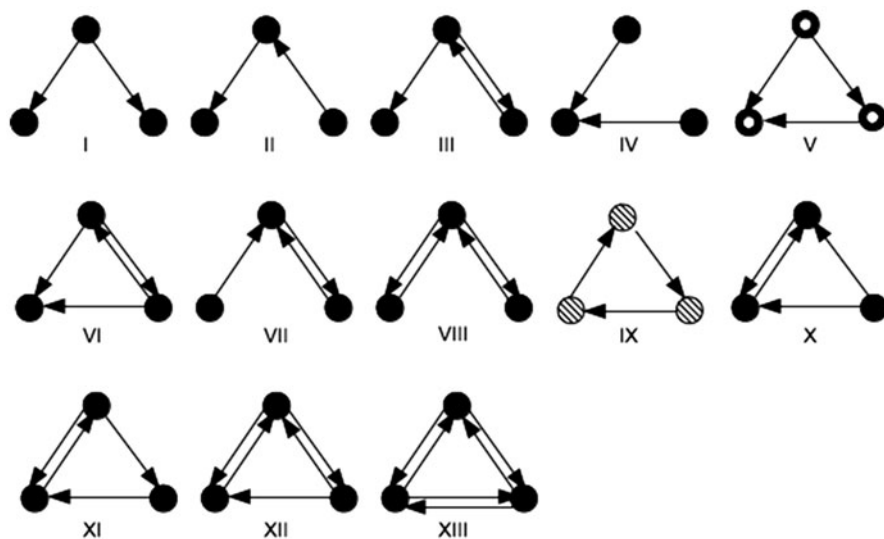
Networks having smaller average path length comparable to the ER graphs of similar size and order but larger clustering coefficient than ER graphs are termed as small world networks. The average shortest path length of the small world networks scale as logarithm of the number of nodes in the network i.e.

$$L \propto \log N \quad (7.15)$$

Most of the real networks exhibit small-wold property. The small world feature is thus common to most biological networks such as neural network of *C. elegans* and Food web.

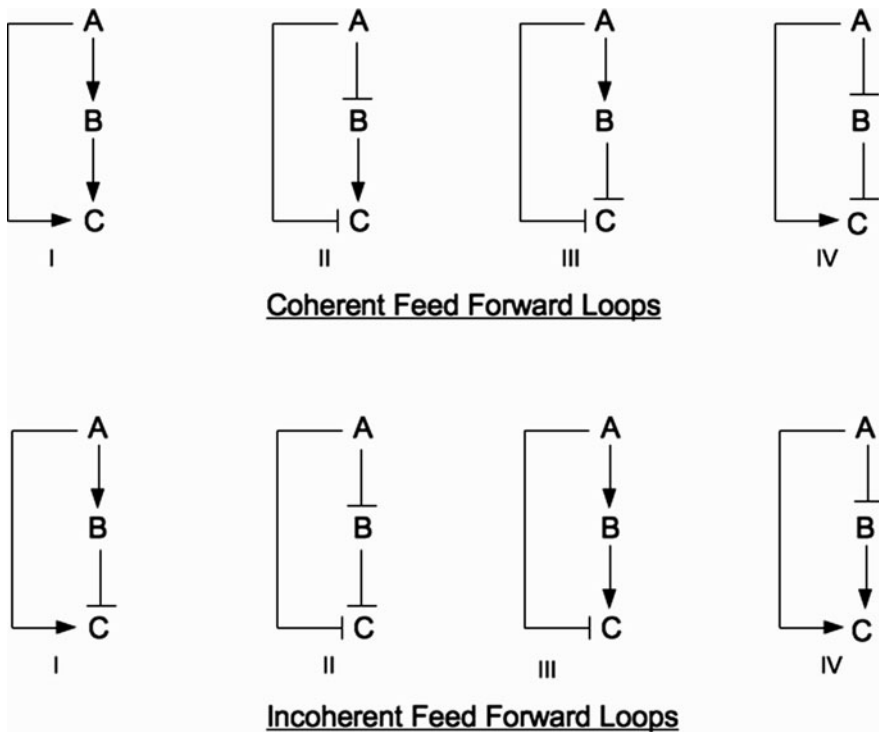
## 7.4 Motifs in Network

Motifs in a network refer to a particular pattern of subgraphs that appear more commonly than what is expected to occur in a random graph. Motifs are much more abundantly present in biological networks than other type of networks. Self loops, i.e., the edges which originate and terminate in the same node, can be thought of as the simplest network motif. This will refer to autoregulation, or autogeneous control, e.g., regulation of a gene by its own gene product, in a transcription network (Fig. 7.11).



**Fig. 7.12** The 13 possible three-node directed subgraphs. Subgraph V, having annular nodes, is the Feed Forward Loop (FFL), while subgraph IX, with striped nodes, is the Feed Back Loop (FBL)

Autoregulatory network may be positive or negative. For instance, in the former case, the genes activate their own transcription, while in the latter, the genes act as repressors. Negative autoregulation has many advantages. It speeds up the response time of gene circuits. Also, it promotes robustness of the steady-state expression level to fluctuations in production rate. In contrast, positive autoregulation slows down responses. In addition, the system exhibits bistability when the rate of positive autoregulation is strong compared to the degradation/dilution rate. The next interesting step will be to look at three-node patterns. There are 13 such patterns, as shown in Fig. 7.12. Out of these thirteen patterns, the only significant one is the Feed Forward Loop (FFL), Fig. 7.12 (V), as found in the sensory transcription network of *E. coli* and yeast (Lee et al. 2002; Milo et al. 2002). It is a strong network motif which appears more often than its randomised version. A straight forward description of a FFL would be as follows. It is composed of a transcription factor, say X, which regulates a second transcription factor, Y, and both X and Y regulate gene Z. It has two parallel paths of regulation, a direct path that goes from X to Z, consisting of a single edge, and another indirect one via Y, having a cascade of two edges. A plus sign or a minus sign is assigned to each of the edges corresponding to activation and repression respectively. So there are  $2^3 = 8$  possibilities, out of which four are coherent FFL and the rest four are incoherent. This grouping is based on the comparison between the signs of the direct and the indirect paths. If both comes out to be the same, then we get coherent FFLs, and incoherent ones have opposite signs. Incoherent FFLs have an odd number of minus signs and the two paths possess an antagonistic effect. Among all the eight different types, Coherent Type-I, followed by the Incoherent Type-I, are the two most abundant FFLs present across various biological networks. Feedback Loops (FBL) (Fig. 7.13).



**Fig. 7.13** The eight possible Feed Forward Loops (FFLs). The *upper four* are the coherent FFLs, while *lower four* are incoherent FFLs.  $\downarrow$  denotes the activation (+ sign) and  $\perp$  denotes inhibition (- sign)

## 7.5 Gene Regulatory Network (GRN)

Genes are fragments of DNA molecules which carry the genetic code in the form of a sequence constituting four nucleotides, viz., adenine (A), thymine (T), guanine (G) and cytosine (C). Each individual gene has its own characteristic genetic code and genes are collectively responsible for various functions in a living organism. The two step process in which at first the information encoded in the nucleotide sequence of a DNA gets decoded to messenger RNA (mRNA) and then proteins are synthesised to perform all the essential biochemical functions is called gene expression. The former step is called *Transcription* while the latter is the *Translation*. A number of genes act together to perform a definite biological function. To depict this, we can think of an interactive network of fragments of DNA or mRNA (nodes) which governs the rate of gene expression, i.e., the rate of protein synthesis, which is known as a Gene Regulatory Network or GRN.

## 7.6 Networks of Proteins

Protein, the most important biological macromolecule, which performs almost all the essential functions in a living organism; is a polypeptide chain formed from 20 possible amino acids. To accomplish various biological functions, the protein folds to attain a well defined three dimensional spatial conformation (often called as the native state). This native state correspond to the global minima of the energy landscape. The protein folding is driven by a number of non covalent interactions, viz., hydrogen bonding, van der Waals force, ionic and hydrophobic interactions, among its constituent amino acids. To visualise this interaction, one may take recourse to networks. Proteins can be modelled into a network containing amino acid residues as nodes and two of the residues are linked together if they interact.

### 7.6.1 Protein Structure Network (PSN)

*Protein Structure Networks (PSN)* are based on the geometrical distance between different amino acids. Geometrical considerations provide deep insights to protein folding. PSN's identify the  $C\alpha$  atoms of the amino acid residues as nodes. Two residues are said to interact with each other if the geodesic distance between their  $C\alpha$  atoms is less than a fixed cut-off value like  $8.5 \text{ \AA}$  (Vendruscolo et al. 2002). Such a representation mainly emphasises the backbone chain interactions of the proteins. A few selected nodes (often called *key residues*), from these networks which have high betweenness centrality; correspond to the previously known nucleation centres for protein folding. The residues identified by such graphical properties are sometimes investigated further for their role in providing unique structure to the protein native structure. However such a formalism of PSN disregards the side chain interactions of the amino acids within the polypeptide chain. Side chain interactions are essential for maintaining the 3D structure of the protein. To encapsulate these interactions, a different mechanism for designing PSN has been proposed. Instead of considering the  $C\alpha$  atoms only, connections were established for any two atoms of the amino acid residues whose distance falls within the fixed cut-off. Many such PSNs with varying cut-off distances to probe the long-range and short-range interactions within a protein have been explored (Greene et al. 2003). The short-range interactions networks show small world property while single-scale behaviour in degree distribution was observed for long-range interactions networks. The latter was thought to confer robustness in the overall topology of the protein structure against random mutations. An alternative study incorporated only the non-covalent side chain interactions of the amino acid residues (Kannan et al. 1999). The interactions were defined on the basis of specific minimum *interaction strength*. The cluster profile and hubs in these networks were identified to play a significant role in secondary structural integration in a tertiary structure of proteins. The hubs also play a crucial role in enhancing the thermal stability of the thermophilic proteins when compared to their mesophilic counterparts (Brinda et al. 2005).



### 7.6.2 Protein Energy Network (PEN)

Thus, we have seen that PSNs can capture the atomic interactions of proteins at geometric level very well. Though they overlook the the basic chemistry of bonded and non-bonded interactions. The energies of these interactions result from various types of interactions, e.g., hydrogen bonding, hydrophobic interactions, cation-pi interactions etc. taking place within a protein. The networks, which account only non-bonded interaction energies, viz., van der Waals interaction (vdW) and electrostatic interaction energy of the side chain atoms of the amino acid residues, are termed as Protein Energy Networks (PEN) (Vijayabaskar et al. 2010).

The various amino acid residues are the nodes of the network. Edges are defined between the residues  $i$  and  $j$ , if the non-bonded interaction energy,  $E_{ij}$ ; is less than a cut-off energy  $e$ . Since interaction energies between different pairs vary, the resulting PEN is an undirected weighted network. Vijayabaskar et al. had explored PEN for six different proteins. The interaction energies were calculated from equilibrium ensembles obtained by performing Molecular Dynamics (MD) simulations. They observed that the networks are densely connected i.e they have more number of interactions for small energy cut-off  $e$  (less negative,  $\sim -5 \text{ kJ/mol}$ ). As the cut-off interaction energy is increased to high negative values ( $\sim -25 \text{ kJ/mol}$ ) the network becomes more sparsely connected i.e it has low number of interactions or edges connecting the nodes. The fractional contribution of vdW and electrostatic energy to the total energy was also analysed. The vdW interaction energy dominates the region of low interaction energy (less negative values) and its value falls off to zero for  $e \sim -35 \text{ kJ/mol}$  while reverse is the case for electrostatic interaction energy which dominates high interaction energy region (high negative values). Another important observation was that the PEN breaks down into small independent clusters within a small window of  $e$ . For less negative values of  $e$ , a large cluster percolates within the network which can be quantified by the tethering together of small independent clusters within the PEN by weak vdW interactions; as the value of  $e$  is made to have less negative values. This provides an evidence for weak interactions (rather than strong interactions) holding together the 3D structure of a protein. The cluster profile of the network helps in understanding the structural integrity of the proteins.

### 7.6.3 Allostery and Protein Energy Network

Recently allosteric mechanism has drawn much attention in the field of research. Allostery can be defined as the control of protein structure, function and/or flexibility induced by the binding of a ligand or another protein, which is called an effector, at a site away from the active site (allosteric site) (Goodey et al. 2008)

Loosely speaking, allostery is a regulation between two distant sites of a protein caused by binding of ligands. PEN serves as a useful tool to explore this mechanism of communication within the proteins. The communication paths between the two functional sites of a protein can be elucidated by tracking the shortest path in the

weighted PEN (Bhattacharya et al. 2011). The shortest paths between a pair of residues in these networks, from energy point of view, will be the ones which are less costly or energetically more favourable. To achieve this weights assigned to the edges have values proportional to the reciprocal of the interaction energy among the pair of residues. The suboptimal paths of the network with reduced efficiency were also explored by deleting all the edges incident on any one of the residues belonging to the optimal paths (the shortest path). An interesting observation was the presence of these suboptimal paths as the optimal paths in less frequently accessed conformations during MD simulations and thus effectively act as alternate paths of communication adapted due to mutation/ligand induced perturbations. Such insights gained by analysing PENs support theoretical as well as experimental observations of the concept of transmission of allosteric signals through multiple, preexisting pathways (de sol et al. 2009).

#### **7.6.4 Protein Protein Interaction Network (PPI Networks)**

Most fundamental biological processes are carried out by proteins and their interactions. Proteins usually execute their functions through interactions with other biomolecular units, rather than acting in isolation. In this type of networks, proteins are nodes and if there is an experimental verification regarding binding between two proteins, then an edge is drawn between the two. Previous studies have discussed whether PPI networks are scale-free in nature. Such a study of a PPI network for yeast shows that its degree distribution follows a power law with an exponential cut-off (Jeong et al. 2001). In scale-free protein networks, most proteins participate in very few interactions, while few hubs are involved in most of the interactions. Another characteristic property is that small-world effect is also present in PPI networks which indicates that any two proteins are connected by a short path of very few links. These networks are disassortative in nature, i.e., highly connected nodes are seldom connected among themselves. The elimination of a protein often causes functional disruption of a module in a PPI network. Such proteins are termed as *lethal*. Thus lethality of a protein is the decisive factor characterising the biological indispensability of a protein.

#### **7.6.5 Protein Folding Network**

During folding, a protein takes up consecutive conformations. Distinct conformational states are represented by nodes in the network and two of them are linked by an edge if one can be obtained from another by an elementary move. It has been studied that the network formed by the various conformations of a 2D lattice polymer has small world properties (Scala et al. 2001). The degree distribution has been found to be consistent with a Gaussian (Amaral et al. 2000)

## 7.7 Metabolic Networks

Metabolism, a set of biochemical reactions essential for sustaining life, is one of the various life processes taking place within an organism. The metabolism of a compound involves a sequence of reactions, termed metabolic pathway, in which the initial compound is transformed into various other intermediary compounds to get the product by the action of enzymes. The intermediaries and the products of such chain reactions are termed as metabolites. It may happen that the product of one pathway is served to initiate some other pathway.

In metabolic networks, the nodes correspond to the substrates (ADP, ATP, H<sub>2</sub>O) and the edges represent the predominantly directed chemical reactions among these substrates. For 43 organisms, these networks have been studied (Jeong et al. 2001) and for all of them; the degree distribution of the incoming and outgoing links have been claimed to follow a power law, with the exponent value in the range 2.0–2.4. There have also been alternate representation of these networks: ATP, ADP, NADH are included as nodes only if they directly take part in the reaction (Ma et al. 2003). Such metabolites are called current metabolites and are ignored while measuring the average path length of the network during their indirect participation in the reaction.

It was found that the path lengths of the metabolic networks in eukaryotes are longer than that of bacteria. Small world property was found in *E. coli* by representing metabolic networks as two complementary networks—substrate graph and reaction graph. It was hypothesised that since metabolic networks respond to perturbations (like changes in concentration of the metabolite or the enzyme), their function could be optimised by the small-world behaviour of the network (Wagner et al. 2001).

## 7.8 Networks and Epidemiology

We can get deep insights into the dynamics of disease spreading in an interacting population of species by applying network theory. Here, we briefly describe two well known spreading models on networks and recent developments about influential spreaders in networks.

### 7.8.1 *Susceptible Infectious Recovered (SIR)*

In a network of  $N$  nodes, initially we assume one node is in the infectious state (I) and the rest in the susceptible state (S). This node, denoted by  $I$ , is the origin of infection. The infection gets propagated in successive time steps. In each time step, nodes of type  $I$  infects neighbours, which are susceptible to infection, with some probability  $\beta$ . They then enter the recovered state (R), where they cannot be infected again, i.e., they achieve immunity against infection.

### 7.8.2 *Susceptible Infectious Susceptible (SIS)*

Here the immunised or recovered state of the origin, just after infecting the neighbours, is absent. Infected individuals still possess the capability of infecting their neighbours with probability  $\beta$ . However, they may subsequently return to the susceptible state with probability  $\lambda$ ; thus remaining infectious with probability  $(1-\lambda)$ .

### 7.8.3 *Influential Spreaders in Networks*

A common belief related to infection or disease spreading is that the best (efficient) spreaders will correspond to a highly connected nodes (high degree) or to the most central nodes (having high betweenness value). It has been argued that the network topology should naturally play an important role in infection spreading or information spread. The position of a node in the network serves as a deciding factor for it to be the most influential spreader. The k-shell decomposition method was performed on a set of eight real social networks and both SIS and SIR model were studied (Kitsak et. al. 2010). The nodes in the innermost k-shell were claimed to be the most efficient spreaders.

## 7.9 Conclusion

In this chapter, we have hopefully given an overview of how complex networks are important at every level in biology. In Sect. 7.1, we mention how biology has shifted from a reductionist approach to holistic approach. Hence deriving a network picture is of immeasurable value because complex networks understandably play an integral part in this new approach. We went on to introduce the very basics of a network or graphical representation; namely nodes, edges, weighted networks etc. In Sect. 7.2, we dwell in-depth on common network metrics like degree, shortest path length, connectedness, giant clusters, cliques and community structure, eccentricity, diameter, closeness and betweenness centralities, clustering coefficient, assortativity, k-core and modularity. In the next section, we briefly discuss about small-world properties and random networks which serve as a good reference points in networks. We then discuss the concepts regarding motifs and their importance in biological networks. In Sect. 7.5, we discuss about interactive Gene Regulatory Networks of fragments of DNA or mRNA (nodes) which governs the rate of gene expression, i.e., the rate of protein synthesis. In Sect. 7.6, we discuss about networks of proteins: protein structure networks, protein energy networks and protein-protein interaction networks and protein folding networks. In Sect. 7.7, we discuss about metabolic networks. Finally, in Sect. 7.8 we end this chapter with a discussion of concepts and models which deal with spread of infection on networks. Thus, we have hopefully been able to portray the importance of complex networks to understand processes at virtually every level of life.

## References

- Albert R, Barabasi A-L (2002) Statistical mechanics of complex networks. *Rev Mod Phys* 74:47–97
- Albert R, Jeong H, Barabasi A-L (2000) Error and attack tolerance of complex networks. *Nature* 406:378–382
- Amaral LAN, Scala A, Barthelemy M, Stanley HE (2000) Classes of small-world networks. *Proc Natl Acad Sci U S A* 97:11149–11152
- Banerjee SJ, Roy S (2012) Key to network controllability arxiv:1209.3737
- Barabasi AL, Oltvai ZN (2004) Network biology: understanding the cells functional organization. *Nat Rev: Genet* 5:101–113
- Bhattacharyya M, Vishveshwara S (2011) Probing the allosteric mechanism in Pyrrolysyl-tRNA synthetase using energy-weighted network formalism. *Biochem* 50:6225–6236
- Brinda KV, Vishveshwara S (2005) A network representation of protein structures: implications for protein stability. *Biophys J* 89:4159–4170
- Filkov V, Saul ZM, Roy S, D’Souza RM, Devanbu PT (2009) Modeling and verifying a broad array of network properties. *EPL (Europhys Lett)* 86:28003
- Freeman LC (1977) A set of measures of centrality based on betweenness. *Sociometry* 40:35–41
- Goodey NM, Benkovic SJ (2008) Allosteric regulation and catalysis emerge via a common route. *Nat Chem Biol* 4:474–482
- Greene LH, Higman VA (2003) Uncovering network systems within protein structures. *J Mol Biol* 334:781–791
- Hongwu M, An-Ping Z (2003) Reconstruction of metabolic networks from genome data and analysis of their global structure for various organisms. *Bioinformatics* 19:270–277
- Jeong H, Tombor B, Albert R, Oltvai ZN, Barabasi AL (2000) The large-scale organization of metabolic networks. *Nature* 407:651–654
- Jeong H, Mason SP, Barabasi A-L, Oltvai ZN (2001) Lethality and centrality in protein networks. *Nature* 411:41–42
- Kannan N, Vishveshwara S (1999) Identification of side-chain clusters in protein structures by a graph spectral method. *J Mol Biol* 292:441–464
- Kitano H (2002) Systems biology: a brief overview. *Science* 295(5560):1662–1664
- Kitsak M, Gallos L, Havlin S, Liljeros F, Muchnik L, Stanley HE, Makse HA (2010) Identification of influential spreaders in complex networks. *Nat Phys* 6:888–893
- Lee TI et al (2002) Transcriptional regulatory networks in *Saccharomyces cerevisiae*. *Science* 298(5594):799–804
- Ma H et al (2003) Reconstruction of metabolic networks from genome data and analysis of their global structure for various organisms. *Bioinformatics* 19:270–277
- Milo R et al (2002) Network motifs: simple building blocks of complex networks. *Science* 298(5594):824–827
- Newman MEJ (2002) Assortative mixing in networks. *Phys Rev Lett* 89:208701
- Newman MEJ (2010) *Networks: an introduction*. Oxford University Press, Oxford
- Roy S (2012) Systems biology beyond degree, hubs and scale-free networks. *Syst Synth Biol* 6:31–34. doi:10.1007/s11693-012-9094-y
- Roy S (2014) Networks, metrics and systems biology. In Kulkarni V, Stan G-B, Raman K (eds) *A systems theoretic approach to systems and synthetic biology I: models and system characterizations*. Springer, Heidelberg. DOI:[http://dx.doi.org/10.1007/978-94-017-9041-3\\_8](http://dx.doi.org/10.1007/978-94-017-9041-3_8)
- Roy S, Filkov V (2009) Strong associations between microbe phenotypes and their network architecture. *Phys Rev E* 80:040902 (R)
- Scala A et al (2001) Small-world networks and the conformation space of a short lattice polymer chain. *Europhys Lett* 55:594
- Vendruscolo M, Dokholyan NV, Paci E, Karplus M (2002) Small-world view of the amino acids that play a key role in protein folding. *Phys Rev E* 65:061910
- Vijayabaskar MS, Vishveshwara S (2010) Interaction energy based protein structure networks. *Biophys J* 99:3704–3715

- Wagner A, Fell D (2001) The small world inside large metabolic networks. *Proc Roy Soc London Series B* 268:1803–1810
- Wuchty S, Almaas E (2005) Peeling the yeast protein network. *Proteomics* 5:444–449
- Wuellner DR, Roy S, D'Souza RM (2010) Resilience and rewiring of the passenger airline networks in the United States. *Phys Rev E* 82:056101
- Wunderlich Z, Mirny LA (2006) Using the topology of metabolic networks to predict viability of mutant strains. *Biophys J* 91:2304–2311