Structure of biological networks

Atanas Kamburov

04.06.2007

Abstract

Complex physiological functions of the cell are often carried out by myriads of molecules that interact with each other in different ways. These interactions form complex interaction networks whose structure is far from random. Instead, most of these networks possess certain properties like a scale-free and small-world nature and are often marked by the existence of subgraphs that are overrepresented in comparison with other complex networks. Here, we are going to look at the common global topological properties of molecular interaction networks, as well as at local architectural features in terms of characteristic network motifs.

1 Introduction

For a long time biologists, while analysing complex physiological processes taking place in the cell like apoptosis, transcription and translation, were concentrated on the meaning of single molecules for the whole physiological process. Today it is however clear that such processes can hardly be linked to single molecules, but are rather determined by whole myriads of molecules and the different functional interactions between them [1]. These complex interactions form molecular interaction networks, or interaction graphs, where nodes of the networks represent molecules (genes, proteins, RNA, metabolites) and edges (links, arcs) link molecules that interact. The simplest example are maybe protein-protein interaction networks, where nodes stand for proteins and links between couples of nodes exist if both proteins interact physically. However, it is also possible to represent more complex functional interactions as edges of a molecular interaction graph: for example, a metabolic pathway can be represented graphically by putting links between metabolites if there exists a biochemical reaction that transforms one metabolite into the other. Gene regulatory networks are another class of biological networks, where nodes represent genes and the existence of a link from gene A to gene B means that the product of gene A serves as a transcription factor for gene B.

Speaking of networks, different people often imagine different objects. Physicists and electricians would probably imagine networks as regular objects, like in Figure 1. a), b). Economists probably imagine networks as objects with a hierarchical nature (see Figure 1. c). Others may imagine networks as completely random objects like in Figure 1. d). Biological networks are often very different than any of the networks shown in Figure 1. a) - d) but still share certain properties with all of them. Like in the case of other complex real networks

(the WWW, social networks, road or flight maps), the architecture of biological networks obeys certain principles. It becomes more and more clear that studying these principles will lead to better understanding of complex cellular processes. The terms "Network biology" and "Systems biology", despite being relatively young, already occur frequently in modern scientific literature and evidence that more and more biologists are looking away from the single-molecule reductionism of the past century and towards analysis of molecular interaction networks that drive major cellular functions.

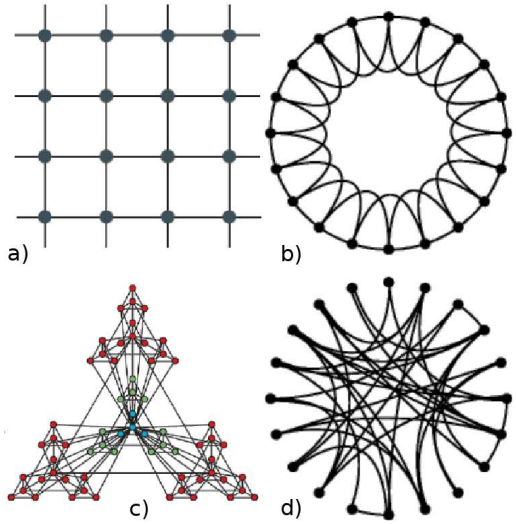

Figure 1: Various graph shapes: a), b) are completely regular graphs where all nodes have the same number of interactors; c) is a hierarchical model of molecular interaction networks as described in Ref [2]; d) is a randomized version of b) (see Ref [3]).

To analyse the structure of biological networks, one can start from their global architectural properties and move towards modules and molecules, or one can analyse their local properties by finding frequent patterns of interactions and move towards pattern clusters and modules. Here, we are going to discuss both the 'top-down' and the 'bottom-up' approaches.

2 Global architectural features of biological networks

2.1 The scale-free nature of biological networks

In order to characterize the global topological properties of cellular networks, we need to introduce some network notions. To start with something simple: networks can be directed or not, depending on the nature of the interactions between their nodes. Protein-protein interaction maps are undirected because the links represent mutual binding relationships. Other interaction networks

like gene regulatory and metabolic networks are directed: edges here represent a flow of information or mass and most edges do not have counterparts running in the opposite direction (e.g., if gene A regulates gene B, this does not mean that gene B also regulates gene A). The *degree (connectivity)* k of a node A denotes the number of links between A and other nodes. In the directed case, one differs between in-degree (the number of edges ending in A) and out-degree (the number of edges from A to other nodes). See Figure 2. for examples. Important global properties of undirected molecular interaction graphs are the average degree $\langle k \rangle$ (the average of the degrees of all nodes) and the degree distribution $P(k)$. $P(k)$ gives the probability that a randomly picked node has k links. In the directed case, there are separate measures for the out-degrees and in-degrees of nodes.

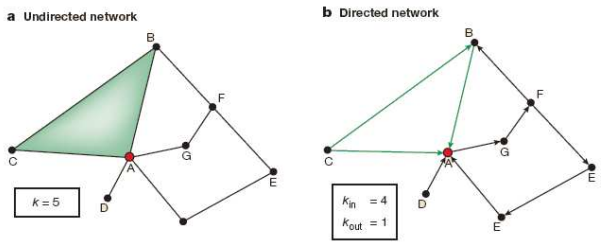

Figure 2: Examples of a) a directed and b) of a non-directed network. In the undirected case, the degree of node A is the number of links k between A and other nodes; in the directed case, one differs between in-degree k_{in} and out-degree k_{out}

For decades, molecular interaction networks were considered either completely regular or completely random. However, the obvious existence of molecules with a very high number of interactions is a fact that cannot be explained by either of the two models. In regular networks, all nodes have the same connectivity. In random graphs, the connectivity of nodes follows a Poisson distribution, which means that the existence of nodes with an extraordinarily high number of links is very improbable. Recent studies have shown that in many biological networks, the degree distribution follows a power law, that is $P(k) \sim k^{-\gamma}$ with parameter γ being often between 2 and 3. In such networks, most nodes have a small number of links, but a small number of nodes, called hubs, exist that have many links. Because in such networks no 'typical node' (typical 'scale') exists, the networks are called **scale-free**. Historically, the scale-free property was first shown for metabolic networks – in fact, most metabolites participate in few biochemical reactions, whereas some metabolites like coenzyme A, ATP and pyruvate participate in a large number of reactions. More recent studies (for example, see Refs [4], [5]) show that most protein-protein interaction networks and gene regulatory networks also have a scale-free nature. This is apparent from Figure 3. that illustrates a protein-protein interaction network of *S. cerevisiae* – most proteins interact with a small number of other proteins, but hubs with many interactions also exist. It is important to note however, that not every biological network has a scale-free nature – for example, the transcriptional regulatory networks of *S. cerevisiae* and *E. coli* were shown to possess mixed

scale-free and exponential properties [6].

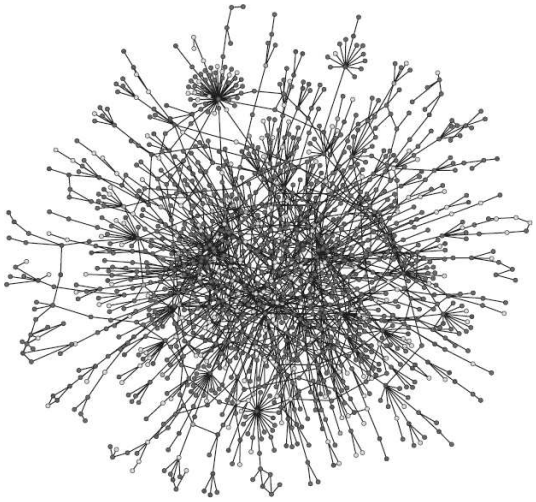

Figure 3: A protein-protein interaction map of *S. cerevisiae*, see Ref [4]

Scale-free networks are very robust – they are extraordinarily resilient to random component failures. Even after a high number of nodes are removed, the rest are still held together by the hubs so that the network often does not become disintegrated and can still fulfill its function. As the number of hubs is relatively very small compared to the number of nodes with few links, the chance that a randomly removed node is a hub is small. The intentional removal of hubs, on the other hand, is often critical to the network's integrity and proper function – that is, scale-free networks have a high hub vulnerability. In *S. cerevisiae* for example, only about 10% of the proteins with less than 5 links are essential, whereas if proteins with more than 15 links are deleted, this has a deteriorating effect on the yeast's viability in more than 60% of the cases.

The scale-free nature of complex networks is often a result of two fundamental processes – network growth and preferential attachment. Network growth is the process where new nodes join the network over a long time period, and preferential attachment means that new nodes prefer to link to nodes that already have a high number of edges. Both processes together are probably responsible for the scale-free nature of most complex networks. In protein interaction networks, they are likely to have a common origin which is rooted in gene duplication – genes that are duplicated have identical products that interact with the same partners, so each protein that interacts with a duplicated gene's product gains a new link after the duplication. If the duplication probabilities for all genes are approximately equal, then proteins with many interaction partners are statistically more likely to gain new links than proteins with few interactions (Figure 4.). Gene duplication might not be the only mechanism accounting for the scale-free nature of protein interaction networks. However, with appropriate adjustments, it is able to explain the power law degree distribution in gene regulatory and metabolic networks.

Biological networks are not the only class of networks that has mostly scale-

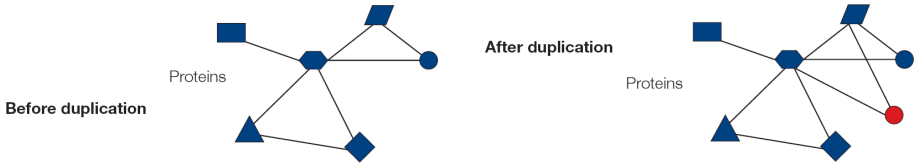

Figure 4: Schematic representation of the outcome of gene duplication. In this case, the gene encoding the protein represented as a circle is duplicated and the product of the new gene has the same interaction partners as the gene that has been duplicated. Note that the hub (the central protein shown as a hexagon) gains a link if any of the other proteins is duplicated, but the rectangle – the protein with a single link – gains a new edge only if the hub is duplicated.

free members – in fact, many complex real-word networks have a power-law node degree distribution. Examples include the world wide web, electric power grids and airplane flight webs.

2.2 The small-world effect

Like connectivity measures, path length measures offer a possibility to quantify and compare topologies of complex networks. The path length l_{AB} between nodes A and B is defined as the number of edges on the shortest path from A to B. In the directed graph case, l_{AB} is often different than l_{BA}. A measure of a network's navigability is the average over all shortest distances between nodes – the average path length $\langle l \rangle$.

Typically, regular networks have relatively long average path lengths. This is generally not true for random networks, where a relatively small number of edges connect couples of nodes, even if the network has a very large number of nodes. This phenomenon is observed also in certain classes of irregular but non-random networks – in fact the phenomenon was first observed in social networks back in the sixties of the 20^{th} century. Social studies show that in our social world of over 6 billion people, only 6 friendship relations ("handshakes") on average are sufficient to link every two people, an interesting observation known as "six degrees of separation". Because these short part lengths make large networks look small, the phenomenon was called "small-world" effect. Many biological networks have even shorter average path lengths than random networks, which is why they are often referred to as **"ultra-small-world"** networks. In metabolic networks, for example, 3-4 reactions often suffice to convert one metabolite into another. This brings a benefit to the organism, because local changes in metabolites' concentrations reach the whole network very quickly so the system can quickly respond to them. For this reason, the ultra-small-world property of metabolic networks most probably underlies selective pressure, which is why we observe the same average path lengths in the metabolic networks of a parasitic bacterium and a large multicellular organism [7].

2.3 Assortativity

In social networks, "famous" people (i.e. people with many connections) often know each other, which makes social networks assortative. In most molecular interaction networks, on the other hand, hubs seem to avoid linking to each other but prefer to link to nodes with few interaction partners. This property is called **disassortativity**. The origin of disassortativity in biological networks remains unexplained.

2.4 Modularity

Cellular functions are often carried out in highly **modular** manner. This means that functional modules of highly interconnected molecules control complex cellular functions. A measure of the modularity of a network is the average clustering coefficient $\langle C \rangle$, which can be explained as follows: if nodes A and B are the only ones connected to a node C, and if there is a link between A and B (forming a triangle of links between A, B and C), then C has a clustering coefficient $C_C = 1$ because all its neighbors are linked to each other. If no neighbours of a node are linked to each other (that is, if there are no triangles a node is part of), then the clustering coefficient of the node is 0. Each node in a network has a clustering coefficient between 0 and 1 which depends on the fraction of couples of *linked* neighbours of the node. The average clustering coefficient $\langle C \rangle$ of a network is the mean of the clustering coefficients of all nodes. Biological networks often have a significantly high $\langle C \rangle$ compared to randomized networks, which indicates their modular nature [2]. The observation is consistent with the fact that in the cell, certain molecules interact to govern certain processes and thus form interaction subgraphs (modules) that are relatively but yet not completely independant of other subgraphs governing other processes. For example, the glycolysis pathway is the one that converts glucose to pyruvate and pyruvate is used in the TCA cycle to obtain energy, but both pathways are linked via ATP and pyruvate. This modularity is the reason why complete cellular interaction networks (like the complete metabolism of an organism or the complete map of protein-protein interactions) are often viewed as "networks of networks".

2.5 Network models

As mentioned earlier, the earliest models of biological networks were completely random or completely regular. Network models with a power law degree distribution are nearer to reality as they reflect the scale-free and small-world properties of biological networks. Subtly constructed hierarchical models, however, are even closer to real molecular interaction networks as they are able to reflect their scale-free, small-world and modular properties (see [2]).

3 Local architectural features of biological networks

Biological networks can be approached either by describing their global topological properties, as we saw above, or by describing the local properties in terms

of subgraph analysis. Subgraphs are subsets of nodes and the according interactions between them. Those subgraphs that appear in a network at significantly higher numbers than in randomized versions of the same network are called network motifs and are central to the bottom-up approach to biological network analysis because of their role as elementary units determining network function. As Figure 5. shows, different classes of complex networks are signed by the presence of different motifs. For example, the feed-forward loop motif is significantly overrepresented in gene regulatory networks and neuronal nets. In gene regulatory networks, this means that some gene X regulates two other genes Y and Z and that gene Y also regulates gene Z. The functional role of feed-forward loops in gene regulatory networks is still not completely understood; however, there is evidence that feed-forward loops function as sign-sensitive delay elements. A motif that is typical for gene regulatory networks apart from the feed-forward loop is the bi-fan motif, where two genes have two common regulators.

An interesting empirical observation is that motifs usually do not appear isolated but tend to form motif clusters, that is, different occurrences of a motif often share edges and / or nodes.

Network	Nodes	Edges	N_{real}	$N_{rand} \pm SD$	Z score	N_{real}	$N_{rand} \pm SD$	Z score	N_{real}	$N_{rand} \pm SD$	Z score
Gene regulation (transcription)					**Feed-forward loop**			**Bi-fan**			
E. coli	424	519	40	7 ± 3	10	203	47 ± 12	13			
*S. cerevisiae**	685	1,052	70	11 ± 4	14	1812	300 ± 40	41			
Neurons					**Feed-forward loop**			**Bi-fan**			**Bi-parallel**
C. elegans†	252	509	125	90 ± 10	3.7	127	55 ± 13	5.3	227	35 ± 10	20
Food webs					**Three chain**			**Bi-parallel**			
Little Rock	92	984	3219	3120 ± 50	2.1	7295	2220 ± 210	25			
Ythan	83	391	1182	1020 ± 20	7.2	1357	230 ± 50	23			
St. Martin	42	205	469	450 ± 10	NS	382	130 ± 20	12			
Chesapeake	31	67	80	82 ± 4	NS	26	5 ± 2	8			
Coachella	29	243	279	235 ± 12	3.6	181	80 ± 20	5			
Skipwith	25	189	184	150 ± 7	5.5	397	80 ± 25	13			
B. Brook	25	104	181	130 ± 7	7.4	267	30 ± 7	32			
Electronic circuits (digital fractional multipliers)					**Three-node feedback loop**			**Bi-fan**			**Four-node feedback loop**
s208	122	189	10	1 ± 1	9	4	1 ± 1	3.8	5	1 ± 1	5
s420	252	399	20	1 ± 1	18	10	1 ± 1	10	11	1 ± 1	11
s838‡	512	819	40	1 ± 1	38	22	1 ± 1	20	23	1 ± 1	25
World Wide Web					**Feedback with two mutual dyads**			**Fully connected triad**			**Uplinked mutual dyad**
nd.edu§	325,729	1.46e6	1.1e5	2e3 ± 1e2	800	6.8e6	5e4±4e2	15,000	1.2e6	1e4 ± 2e2	5000

Figure 5: Network motifs from several biological and technological networks [8]. N_{real} is the observed number of appearances of a motif in the real network and N_{rand} is the number of appearances of the same motif in randomized versions of the same network.

4 Discussion

Network biology is a relatively new field of study concerned with the topological properties of biological networks. Network biologists approach the cell from the top down, starting from the networks' global properties such as their scale-free, small-world, hierarchical nature and moving towards modules and molecules, or from the bottom up, starting from interaction motifs and moving to motif clusters and modules. Results of both approaches support the thesis that network topology is far from random, and deeply linked with network function. Although network biology has witnessed remarkable progress in the recent years, it is still in its infancy, and is likely to involve enormously in the next few decades. However, we have to mention here that biological systems can hardly be understood by analysing interaction network topologies alone. Instead, one should also consider the system's dynamics. For example, not all metabolic reactions in the cell take place at the same time and at the same rate – a fact that can perfectly be expressed in terms of network dynamics but not in terms of topology alone. Integrating our knowledge of topology and dynamics of molecular interaction networks can definitely help us to better understand the cell – the unit of life on Earth.

References

[1] Barabási A and Oltvai Z. Network biology: understanding the cell's functional organization. *Nat. Rev. Genet*, 5:101–113, 2004.

[2] Ravasz E, Somera A, Mongru D, Oltvai Z, and Barabási A. Hierarchical organization of modularity in metabolic networks. *Science*, 297:1551–1555, 2002.

[3] Watts D and Strogatz S. Collective dynamics of 'small-world' networks. *Nature*, 393:440–442, 1998.

[4] Jeong H, Mason S, Barabási A, and Oltvai Z. Lethality and centrality in protein networks. *Nature*, 298:41–42, 2001.

[5] Featherstone D and Broadie K. Wrestling with pleiotropy: genomic and topological analysis of the yeast gene expression network. *Bioassays*, 24:267–274, 2002.

[6] Shen-Orr S, Milo R, Mangan S, and Alon U. Network motifs in the transcriptional regulation network of escherichia coli. *Nat. Genet.*, 31:64–68, 2002.

[7] Jeong H, Tombor B, Albert R, Oltvai Z, and Barabási A. The large-scale organization of metabolic networks. *Nature*, 407:651–654, 2000.

[8] Milo R, Shen-Orr S, Itzkovitz S, Kashtan N, and Alon U. Network motifs: simple building blocks of complex networks. *Science*, 298:824–827, 2002.