

# Protein Networks: Generation, Structural Analysis and Exploitation

Enrico M. Bucci<sup>1,2</sup>, Massimo Natale<sup>1</sup> and Alice Poli<sup>1</sup>

<sup>1</sup>*Biodigitalvalley Srl – Pont Saint Martin- AO-*

<sup>2</sup>*CNR - IBB – Naples  
Italy*

## 1. Introduction

The scientific community is well aware of the fact that the very presence of internet profoundly affects the way research is done. Internet is the major infrastructure through which results and data are communicated and shared, computing is parallelized, collaborations are started and enlarged, papers are published, and a myriad of other transactions are performed, so to involve nearly all aspects of everyday researcher's life.

Perhaps surprisingly, scientists paid little attention to the theoretical study of internet structure (i.e. topology) and of its dynamical behavior, until quite recently.

Even more surprisingly, however, when they did it, it quickly emerged some very rarely found scientific truth, of such a general kind, to directly reverberate from Internet studies into the field of molecular and cellular biology, with very small (if any) changes. Consequences manifested immediately, so that, at the turning of the millennium, in a now classical Nature paper (Vogelstein B. et al., 2000), Vogelstein, Lane and Levine, the discoverers of p53 and of its role as tumor-suppressor, wrote that "The cell, like the Internet, appears to be a 'scale-free network'." To let the reader fully appreciate this revolution, it is useful to recall that in 1999, just one year before the appearance of this particular paper, more than 15.000 independent articles have been already published on p53 and its role in cancer biology, making this protein one of the most studied topics ever. Yet, after 20 years of research, despite the enormous amount of available data, some aspects of p53 biology were still missing, and were so crucial to understand why this protein is found mutated in about 50% of cancer patients, to let its discoverer write "One way to understand the p53 network is to compare it to the Internet" (Vogelstein B. et al., 2000).

What had it happened? To understand this, we must look only two years before. In 1998, Zoltan Oltvai, a molecular pathologist, and Lazlo Barabási, a physicist studying Internet topology, were both working at the Northeastern University of Chicago. They both were Hungarians –actually Barabási was born Romanian, but lived and studied in Hungary - have small kids, and were home neighbors; thus it is hardly surprising that, as recalled by Barabási himself (Barabási AL, 2002), they usually met for dinner. By that time, Barabási had already found out that the Internet structure is a peculiar one: he had collected evidence that it is far from random. This was a non trivial result, given the fact that all large networks were modeled at the time as random. To understand this point, we have to think of

networks made of a huge number of nodes, like human social networks, communication networks and alike. To study how they work, i.e. how information spreads through the network, or whether the network is sensitive to external attacks, or how to find the best pathway from one node to another, it is not possible to perform a direct experiment, given the size of the object under study; instead, one has to model the network, then find out a proper set of equations, and simulate the behavior of the network varying the equation parameters. Since the 60s, the model of choice for large networks was that of Erdős and Rényi, which assumes that each node in a network is randomly connected to a fixed, average number of other nodes. As Barabási explains (Barabási AL, 2002), Erdős and Rényi “acknowledged for the first time that real graphs, from social networks to phone lines, are not nice and regular. They are hopelessly complicated. Humbled by their complexity, [they] assumed that these networks are random.” In a random network, each node is equivalent to every else. Were the real network random, this would have several practical implications. For example, removing one server from Internet, was it a random network, would have on average the same effect of removing every other node, so that to protect Internet from hackers one should only care that, on average, a sufficiently high number of servers is shielded, wherever they are.

However, you could have already guessed that removing 100 servers from the Google facilities would have a larger impact on Internet than shutting down 100 servers in rural China (at least presently). This is due to the fact that Google machines are Internet *hubs*, i.e. they are continuously connected to an enormous number of other machines, and mediate a big amount of Internet data exchange. Barabási and its group were the first to notice the presence of hubs in the web (for example, the New York Times web site has an immense number of links, whereas an obscure blogger may have none), and recognized that the classical network theory of Erdős and Rényi was totally unable to deal with them.

To see whether this was a peculiarity of Internet or a general finding, they began to map the topology of other networks as well. It turned that 1) most real networks are different from both regular lattices and random structures and 2) they all exhibit a common underlying organization, based on few hubs and many poorly connected nodes. This last point is evident if one plots the number of nodes having a defined amount of connections, which is called node *degree* or *connectivity*, versus the degree itself. One gets a curve (the *degree distribution*), which smoothly descends from a maximum (many nodes with very low connectivity) to a minimum (few nodes with very high connectivity); since this curve is exponential, the obtained degree distribution obeys a power-law and is said to be *scale-free*.

Having already obtained the first evidence for the generality of its finding on network structure, Barabási met Oltvai, who, like the majority of the biologists, was very well aware of the intricacy of metabolic connections between the molecular constituents of a living cell. Indeed, the complicate diagrams on the walls of biochemistry labs represent complex networks, where the nodes are biomolecules of any sort, and the links are biological interactions (let us keep this description vague for the moment). The two researchers wondered if these biological networks were also scale free as those made by man. By 2002, Barabási and Oltvai had published their results obtained from 43 different organisms (Ravasz E et al., 2002): the metabolic networks connecting the main metabolites have essentially the same large-scale structure of complex, non-biological networks. They are all scale-free, with hubs and poorly connected nodes, despite significant differences in the particular biochemical pathways included, so that each cell of every examined organism resembles a tiny Internet and can be studied in pretty much the same way.

With a perfect timing and a great deal of intuition, Vogelstein, Lane and Levine realized that, when looking to cancer, p53 is indeed a crucial hub, sitting in the center of a complex protein network, and, very much like internet hacking, cellular hijacking by cancer proceeds by attacking hubs. This is why, *ex post*, one finds p53 mutated so many times: touching a cellular hub causes a great deal of effects, while mutating less prominent “client” proteins passes nearly unnoticed. This is also what brought the three researchers to publish the paper which changed p53 science forever.

Since these early observations, a lot of progresses have been made, to the point that protein networks are useful tools in the hands of molecular biologists. The rest of this chapter is devoted to a simple introduction to their structure and properties, in a way that purportedly simplify mathematical descriptions, keeping an eye on the biological meaning of protein networks.

## 2. The structure of protein networks: Scale freeness

Before entering in some details about protein networks, we want to point out some special characteristics of this type of networks.

First of all, one should keep in mind that a protein network is an abstract representation of the real world, instead of a physical entity like the Internet infrastructure or a phone line web. In particular, while for the latter the links between the nodes are physical (cables in both cases), in the former case the links represent only a *potential* interaction between two proteins. With the notable exception of macromolecular complexes, which can be thought of as networks of interacting proteins, one would never be able to visualize a protein web under a microscope.

Secondly, with the same exception mentioned before, molecular biologists do usually deal with a special type of protein network, one where the nodes represent all of the protein copies coded by a single gene, instead of all the individual proteins which are floating around in a cell. The protein network we will refer to in the following, thus, is a graph which resumes all the known interactions (the links) occurring between the product of every gene out of some list (the nodes); in this respect, such a network is more like a map of our current knowledge about some specific ensemble of proteins than a representation of a real molecular web.

As a third point, a simplification is usually made, by considering only one type of interaction between the proteins composing a given network. The links connecting the nodes thus correspond to one out of a number of possible biological interactions, ranging from very specific types -such as a network where a link between two proteins represents a physical interaction in a molecular complex- to broader concepts -such as a network where a link between two proteins occurs if they are found co-expressed in a given condition. Correspondingly, different protein networks can be obtained, joining the nodes according to different types of interaction: protein-protein interaction networks, transcription networks, enzymatic networks, signaling networks, co-expression networks and so on. However, one should not consider this simplification as an absolute constraint: software does exist, for example, which is able to color the links of a network according to the type of interaction, filtering them as wanted.

We can now start examining an example. Let us consider the human protein-protein interaction network, which can be downloaded from the Reactome organization institutional site (<http://www.reactome.org/download/index.html>). At the time when this



*robustness*. In the case of protein networks, robustness has some interesting biological consequences. First of all, it implies that low-frequency, random events such as protein mutations will really affect the cell – which relies for its functioning on several different types of proteins and biochemical networks – with an exceedingly low probability. Thus, we can say that, even before our immune system takes action against malfunctioning cells, we are protected by potentially dangerous random insults – which could give rise to serious diseases such as cancer – by the very architecture of the cell protein networks. On the opposite side, scale-free determined robustness means that the cell has some true Achilles' heels – the few highly connected proteins – which can be exploited by selective attacks. Once again, we can refer to Internet for an useful comparison: selective, non random attacks to central routers are preferred by hackers, which purportedly aim to take control over the attacked networks. At cellular level, viruses can be considered hackers, which divert the protein network operations toward an illegitimate scope. Indeed, it has been found by several independent groups that viruses selectively target central proteins, causing large effects on the host protein network (de Chassey B. et al., 2008; Navratil V. et al., 2011; Zou X. et al., 2010).

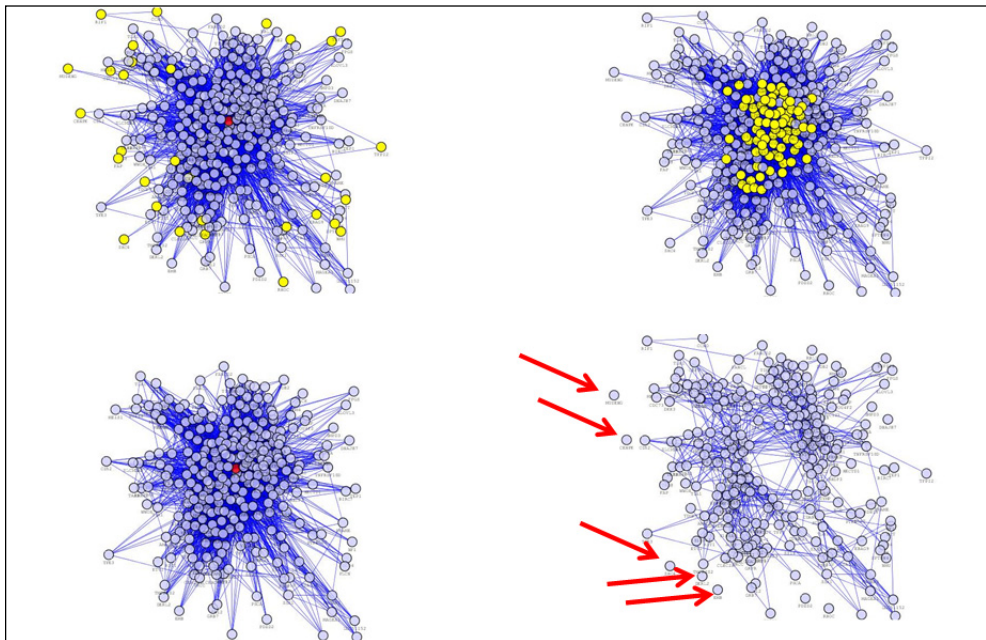


Fig. 2. The effects of removing nodes from a protein network: random removal (left) versus hub removal (right)

Beside robustness, scale-free protein networks show another interesting feature. They exhibit a *small-world* behavior: hopping from one node to a neighbor, any node can be reached from any other in few steps. The distance  $L$  between any couple of nodes, in particular, grows roughly proportional to the logarithm of the number of nodes  $N$  which are included in the network:

$$L \sim \text{Log}(N) \quad (2)$$

For protein-protein interaction networks,  $L$  grows even slower with  $N$ , and the average distance between two nodes is so small that they have been defined ultrasmall networks (Cohen R & Havlin S. 2003). The small-world property, shared by many complex networks, is particularly relevant to biology when the considered network is made of proteins which can influence the neighbors through the links. This is typically the case of transcriptional networks, where two proteins are connected if one influences the expression of the other, signaling network, where proteins are coupled through phosphorylation and other post-translational modifications, and in general holds true for any protein network where the activity of one protein affects its connected partners. In these cases, the small-world structure implies that whatever stimulus changes the status or activity of a protein, its effects will rapidly propagate to the entire network, since on average only few proteins will separate the starting node from every other in the net. This in turn has the consequence that a cell, once its protein network has been stimulated at a single, peripheral node, may quickly change the status of a large number of proteins in response, so that the original signal propagates to a vast number of different proteins, synchronizing their status to the variation of the external stimulus. In other words, small-world protein networks, like any other small-world, display enhanced signal-propagation speed, computational power, and synchronizability (Watts DJ & Strogatz SH. 1998).

In the case of scale-free networks, the observed robustness and the small-world effect are mutually connected. In particular, since protein hubs are linked to the vast majority of all other nodes, most of the pathways connecting any couple of nodes pass through hubs, so that the average distance between any two nodes in the network does not change much if nodes are removed randomly: this is a formulation of network robustness equivalent to the one we mentioned before. At the same time, the more a hub is prominent, i.e. it is connected to a higher amount of protein nodes, the more the distance between any two uncoupled nodes will tend to a single hop through the hub. Hubs are thus key features of scale-free network, mediating both robustness and small-world properties of protein networks; we will dedicate the next paragraph to examine their properties.

### 3. The structure of protein networks: Hubs

Since the early times of protein network studies, the few, always present hubs attracted a lot of attention: quite naturally, it was thought that since highly connected proteins have a lot of different molecular partners, they should also be implied in the majority of the cellular processes. In case of a protein interaction network like the one depicted in figure 2, this reasoning goes as follows: proteins found in many different macromolecular complexes, represented as hubs in the interaction network, should be either core components of a single molecular complex, or elements conserved in many different molecular complexes, which works as switches and are used by the cell to coordinate the activation or repression of different molecular machineries. As a consequence, any alteration of the hubs of a protein-protein interaction network is predicted to have large effects on the cell biology. This assumption, which has been dubbed as “*centrality-lethality rule*”, has been extensively explored by experimentally knocking-down protein interaction hubs and quantitatively assessing the effects in different models (Jeong H. et al., 2011). Going a step further in the reasoning, it has been hypothesized that mutations affecting these proteins should be

particularly related to the insurgence of diseases. Some experimental validation of this prediction has been indeed obtained: Rambaldi and his group (Rambaldi D. et al., 2008) provided evidence that virtually all proteins having a degree higher than 80 in the human protein-protein interaction network are target of known cancer-related mutations. Similarly, Ortutay and Vihinen (Ortutay C. & Vihinen M. et al., 2009), after building an interaction network comprising all human proteins involved in immune response, found that the network hubs include known disease-causing genes as well as 26 new genes related to primary immunodeficiency. In a further example, Chang and colleagues (Chang W. et al., 2009), found new gastric cancer candidate markers by looking to hubs in a protein-protein interaction network build from genes differentially expressed in the patient tissues.

So far, we looked to hubs in protein-protein interaction networks. However, hubs are a common characteristic of any complex web, albeit their biological meaning and relevance change according to the particular type of protein network considered. To understand this point, let us compare the three different human protein networks reported in figure 3.

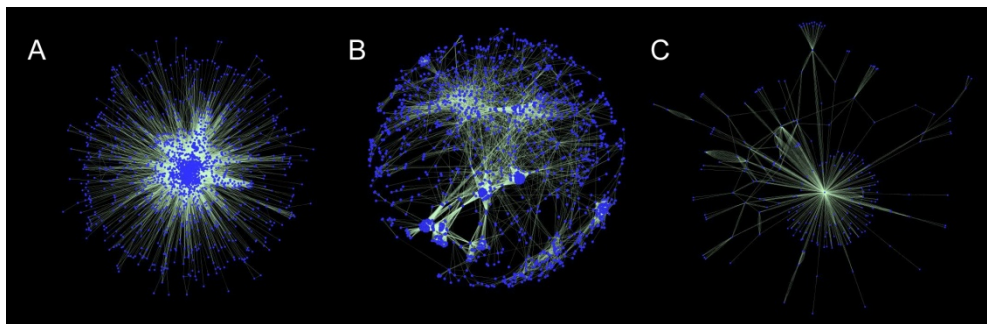


Fig. 3. A) Co-expression network; B) Biochemical/metabolic network; C) Process specific network

The first network on the left (network A) is obtained by considering all proteins differentially expressed in breast cancer patients. Proteins are connected if they are found co-expressed in at least 2 different and independent experiments, and the resulting network is a *co-expression network*. The central network (network B) is build by considering all the proteins which are linked to the p53 human protein by some known biochemical pathway, and is then a *biochemical/metabolic network*. The last network on the right (network C) is a network reporting all those proteins controlled by p53 or controlling it during the unfolding of the apoptotic process, and can be seen as a *process specific network*. You may have already noticed that hubs (and scale-freeness) are present in all three webs, despite the fact that the network size decreases from left to right. However, the biological relevance of hubs is very different in the three network.

In network A, hubs are cancer markers which are found co-expressed with nearly any other cancer protein. Hubs in this network are not granted to be very relevant for the pathogenesis of breast cancer: they can be proteins which are deregulated by the inflammation accompanying cancer, as well as cytoskeletal proteins altered due to the hyperproliferation of cancer cells or other type of very abundant proteins, with no specific role in cancer progression and insurgence. While these proteins are indeed dysregulated in breast cancer, and thus are useful for diagnosing it, their status of hubs do not privilege them with respect

to other, less connected nodes, since their co-occurrence with many partners does not imply that their expression level changes more than that of any other node in the network. Indeed, the list of hubs of network A includes useful diagnostic proteins, such as ERBB2, ESR1 and BCRA1, as well as proteins with little meaning for cancer diagnosis, such as complement proteins. Thus, for co-expression networks like the one depicted in figure 3A, being an hub is of no particular merit for a protein.

In network B, since by construction each neighbor of a specific hub is connected to p53 by some biochemical chain, hubs are at the crossroad of several cellular pathway involving p53. In this respect, hubs of this network are in a prominent position to act as checkpoints for controlling the (very redundant) flow of biochemical information from and toward p53, and thus we can expect them to be important controllers and mediators of p53 activity. For example, we found in network B that the 15 hubs with the highest degree are all different subunits of all the three mammalian RNA polymerase, but two, which are important transcription factors (TF2A and TF2B); this is hardly surprising, since p53 in the very end exerts its prominent and multiple actions regulating the transcriptional process, so that all p53 pathways converge into the regulation of the RNA polymerase machinery. By considering hubs with a lower degree, we find the mitosis controlling kinase NEK-2, the nuclear cap-binding protein 1 and 2, and several other proteins which have prominent roles in regulating the cellular status. As a general rule, although there are exceptions, the lower is the degree, the more specific is the position of the protein in the p53 network (or the lesser is known about it). For example, among the proteins having  $k=1$ , we find the liprin alpha 4, a protein which binds to the intracellular membrane-distal phosphatase domain of tyrosine phosphatase LAR, and appears to localize LAR for regulating the disassembly of focal adhesion and orchestrating cell-matrix interactions; or E2F-3, a transcription factor which binds specifically to the protein RB1, in a cell-cycle dependent manner; or MDB4, the Methyl-CpG-binding domain protein 4, which is a mismatch-specific DNA N-glycosylase involved in DNA repair, specific for G:T mismatches within CpG sites. Thus, for biochemical/metabolic networks like the one depicted in figure 3B, hubs are checkpoints for most of the pathways considered in building the network (in the presented case, p53-related pathways), acting as crucial mediators of biological activity and behaving like switches for several biochemical pathways. On the opposite site, if interested to specific, less studied biochemical players, one should concentrate on low-degree nodes of the network, a group which is enriched in proteins involved in few, specific metabolic modules.

In network C, starting from p53, nodes are attached if they co-occur in at least one biochemical pathway and are involved in apoptosis. The fact that two proteins co-occur in more than one pathway is represented by multiple links. This network can be considered as extracted from network B, by filtering out those proteins not involved in apoptosis. As for network B, hubs of this network are to be considered prominent biochemical regulators; however, since we are restricted to a single, specific biological process, there is no special role for low-degree proteins, which are simply peripheral players in a specific apoptotic pathway, among the many redundant possibilities. Hubs are thus the only targets for the analysis of network C: they are important mediators of p53-related apoptosis, controlling most of the network, and their knocking out can be expected to perturb largely the apoptotic control of the cell. As a matter of fact, ordering by degree the nodes of this network, after p53, which is trivially an hub, we find MDM2, possibly the most important regulator of p53 mediated apoptosis, and the apoptosis-stimulating protein of p53 ASPP2, which influences the apoptotic response of cells without affecting p53-induced cell cycle arrest. On the



opposite side, we find *cdc42*, an important cellular protein, which nonetheless mediates only one of the apoptotic pathways controlled by p53 (Thomas A. et al., 2000). Thus, for networks like that depicted in figure 3C, hubs may be considered the most relevant proteins to be found involved in the selected biological process, and they can be safely assumed as targets for further analysis.

After the preceding discussion, it should be clear at this point that protein hubs are extremely variable in their relevance, and that before considering the degree of a node as a topological guide to prioritize protein lists, one must carefully select the type of network to be used, i.e. the rule to generate links between nodes. However, even having the best network may be not enough. To understand why, let us first make a general consideration and then go on with an example.

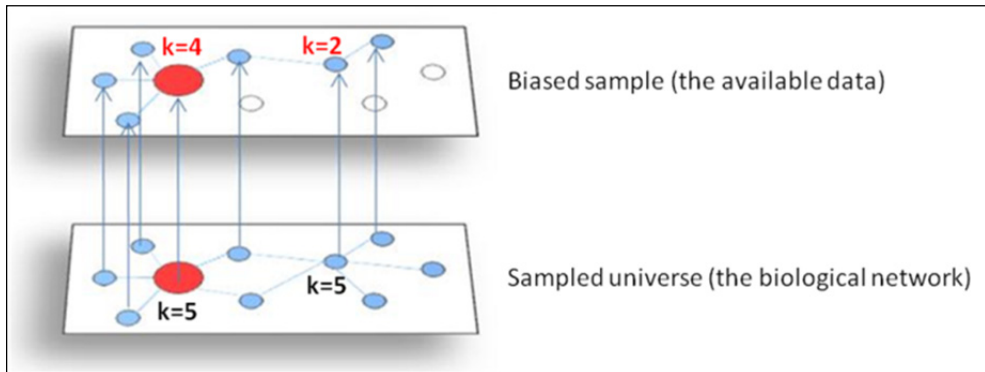


Fig. 4. Sampling bias in biological networks

As shown in figure 4, we must face a sampling problem. “Sampling”, in this context, means to accumulate knowledge on a specific node or part of the network, useful to define its connectivity. In facts, whatever biological network we are exploring, we are only getting an incomplete representation of the real thing, one which was produced by a finite number of experiments interrogating a biological entity. If sampling of the real network (sitting on the lower plane in figure 4) is non-random, i.e. it is concentrated around some “hot” protein (represented in red), then we get a skewed representation (sitting on the upper plane in figure 4) where nodes originally having the same degree are represented as very different in the reconstructed network. Besides being biased, the representation we have can also be error-prone. For example, in figure 4 many links are missing on the upper plane as compared to the lower one (negative error); the opposite situation, where extra links are erroneously added to the representation – for example due to non-specific binding in protein interaction experiments – is also common. Both errors and biases obviously affect the definition of hubs in a network. However, while simple tests exist to check whether an identified hub is a genuine one in an error-prone web (Vallabhajosyula RR. et al., 2009), bias may have subtler effects, much more difficult to deal with. To see this last point, let us consider a further example.

On the left of figure 5, there is a co-expression network which includes all proteins studied in breast cancer. Two proteins are connected if, by any experimental method, they were found to be co-expressed in a breast cancer human sample, whatever the stage or the

provenance of the sample. Proteins which are known targets for drug currently used to treat breast cancer or under development are highlighted in red. On the right of the same figure, there is a box-plot which shows the degree distribution for nodes which have been never entered the drug development process (first box on the left), are in preclinical development (second box), are in clinical development phases (phase I, II and III corresponding to the third, fourth and fifth box respectively) and are already on the market (last box on the right). A clear trend may be seen, with the degree regularly increasing as the clinical development of a target proceeds. Is this a genuine trend to be used for drug target identification, i.e. is it true that the more a protein is an hub, the better is to target it from a pharmacological perspective? Quite the opposite. If we consider the same network in a temporal perspective, we will see why. Have a look at figure 6. For the sake of simplicity, we will focus on three exemplary pharmaceutical targets in breast cancer (the vascular endothelial growth factor VEGF, the thymidylate synthase TYMS and the clusterin CLU).

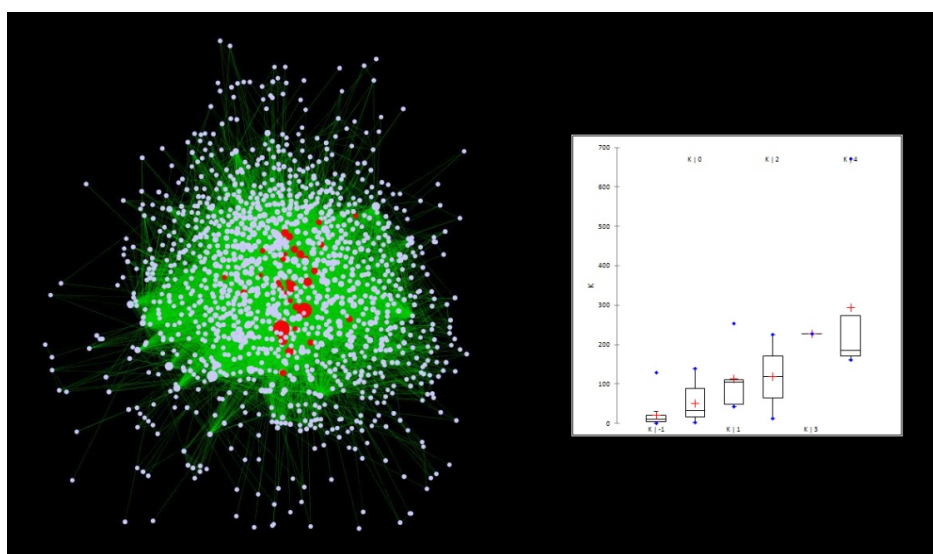


Fig. 5. Degree distribution for pharmacological targets in a breast cancer co-expression network

We want to study their position as hub during time, to see whether it is constant or changes, as new experiments are performed and new network nodes are added. Since the network grows in time, instead of the degree we will consider the ratio between the degree and the total number of nodes; this is the fraction of network nodes connected to the considered protein, and we will refer to this quantity as to “net occupancy”. You may have already noticed that this quantity varies in an unpredictable manner. Something connected to about 6.5% of all network nodes in 1993 (TYMS), a true hub for the network, became connected to less than 1% in 2002, to go back to about 3% in 2009. VEGF, which was an important hub in 2009, was barely connected to the network before 1997, and was certainly not an hub by that time. From the graph, we can notice three temporal points associated to an abrupt trend change for all the selected proteins: 2002 for TYMS, 1997 for VEGF and 2005 for CLU. What happened at the

time? In 2002, pemetrexed, a drug targeting TYMS, was introduced for breast cancer therapy; in 1997, the antiangiogenic therapy was hypothesized as an option to treat breast cancer; in 2002, the experimental drug OTX-111, targeting CLU, was shifted to prostate cancer, due to mixed results in breast cancer trials. We can thus directly observe that, in the selected cases, the industrial interest immediately precedes a topological change of a protein in the network, promoting to hubs those proteins which are under industrial development, and downsizing those proteins which were not up to the standard in clinical trials. Such kind of an effect may also be caused by interests different from the industrial ones. For example, it is probable that strong academic groups tend to produce a lot of data on their “pet” proteins; moreover, most studied proteins tend understandably to be of human origin, well soluble, stable and easily detected. Large scale “unbiased” experiments, such those using microarrays, two-yeast hybrid or proteomic techniques, produce data which are also biased toward detectable proteins (Ivanic J. et al., 2009), and are still very often affected by the interest of the experimenter (think to the study of knock-out models). Some possible solutions which have the potential to mitigate biases as well as errors in reconstructing protein networks have been recently proposed. These approaches make use of network alignment between different organisms (Tan CS. Et al., 2009). In particular, evidence has been recently produced demonstrating that even though the present protein network data are strongly biased by the experimental methods used to produce them, they still exhibit species-specific similarity and reproducibility (Fernandes LP. Et al., 2010). While intra-species conservation approaches tend to contribute “core” networks, i.e. networks made of conserved proteins and conserved topologies which do not account for inter-species variability, they have the indubitable advantage to average biases (because the networks used for the alignment come from different scientific communities, and are less vexed by pharmaceutical industry interests) and errors (because more large scale experiments are taken into accounts). Moreover, hubs conserved among different species are likely to be very relevant for the basic biology of the cell, as shown by the fact that they tend to be duplicated so to increase the mutational robustness of the corresponding biological network (Kafri R. et al., 2008)

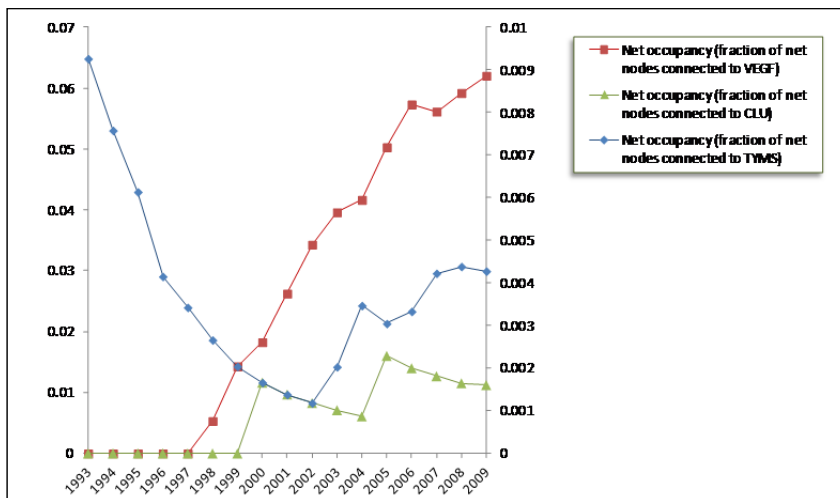


Fig. 6. VEGF, TYMS, and CLU network occupancy

We want to conclude this paragraph by the following message: when properly taking into account biases and errors, the topological prominence of hubs is indeed informative and useful for protein/gene prioritization; however, the real biological meaning of “hubbiness” is strictly dependent on the linking rule applied for building a specific network, as shown in this paragraph for co-expression networks, biochemical/metabolic networks and process specific networks.

#### 4. The structure of protein networks: Neighborhoods

In a graph, the neighbors of a given node consist in all those other nodes that are connected to it up to a certain distance. Distance in this context is intended as the minimal number of steps connecting the source node to any other. In other words, for a particular protein  $x$  in a network (which we will call the seed), we define the neighborhood of  $x$ ,  $N(x)$ , to be the subgraph of the network whose vertex set consists of all of  $x$ 's interaction neighbors and the edges between them, up to a preselected distance  $D$ .

According to the type of graph, neighborhoods can be used to derive useful biological information.

We will try to illustrate this by showing how:

1. in a network, the biological roles of the neighbors can be used to infer the unknown functions of a seed;
2. in protein-protein interaction networks, a group of highly interconnected neighbors sharing a given biological function likely coincides with a macromolecular complex or part of it.

As for the first example, it is useful to remember that traditionally the function of a protein is inferred from its sequence and/or structure by homology modeling. Unsurprisingly, this approach performs poorly for those proteins which have unusual sequences and unknown structures. In this particular circumstances, an analysis of the biological functions of the network neighbors of the protein can be decisive. In particular, it has been proved that in protein networks the probability that a certain biological function is shared between two proteins is higher if the 2 considered proteins are proximal neighbors, and then decreases as the distance  $D$  increases (Shamir R. et al., 2007). This is true in many different network types, such as protein-protein interaction networks, metabolic/biochemical protein networks, genetic interaction networks etc. Moreover, if a given protein with an unknown function is at short distance (usually  $D=1$ ) from several proteins sharing a given function, the probability that it too shares that particular function is obviously even higher. On this basis, a neighborhood-guided labeling strategy is possible to assign biological functions to virtually any protein in a network, providing that at least a fraction of the nodes in its neighborhood has a known biological role. The process is exemplified in figure 7, where functional annotation is symbolized by node coloring.

As can be intuitively understood by looking at figure 7, the functional annotation of a given node is guided by several factors, including distance and number of neighbors with a given biological function, their own connectivity and their heterogeneity (which led to the lack of propagation for the red and the blue colors in the example). Mathematical modeling of the labeling procedure basically consists in weighting all these factors in a single probability function, so to obtain a score for the assignment of a given biological role to all the network nodes. While the details of the proposed methods are out of the scope of this introductory text, we would like to stress here that the procedure depends always on the local topology,

which affects the label propagation by determining the number of neighbors a given node communicate with, and on the type of network considered, which limits the distance and the direction of propagation of a label along the edges.

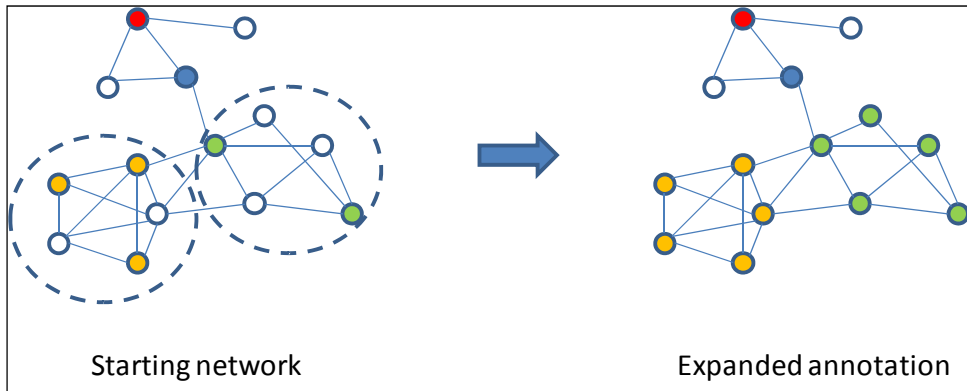


Fig. 7. Functional annotation of a given node

As for the second example, we will refer to a recent work of Fox et al. (Fox AD et al., 2011) on protein-protein interaction networks. Consider in particular the two alternative situations illustrated in figure 8.

In A, the neighborhood for  $D=1$  of the selected seed (shown in blue) is made of two groups of nodes, which are not directly connected; on the opposite, in B the neighborhood is highly interconnected in a single cluster.

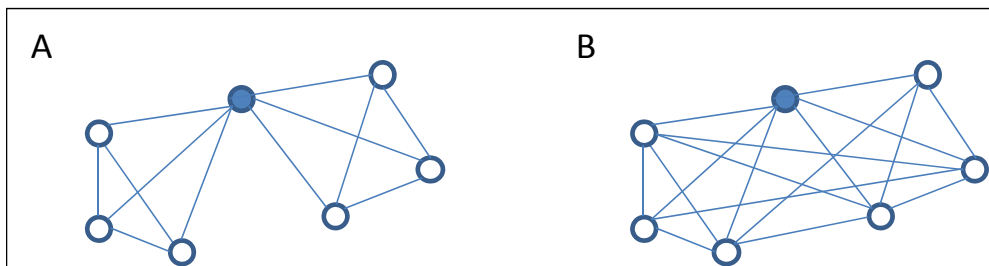


Fig. 8. A) Two disconnected neighborhood; B) Highly interconnected neighborhood

As reported by the authors, the structure observed in A suggests the possibility that the two groups of neighbors might be active under different conditions, as opposite to B. Indeed, it was found that single-component neighborhoods like the one represented in B are enriched in protein sharing similar functions and participating to molecular complexes, and are thus more likely to represent a single, defined protein complex, while multiple-components neighborhoods like the one represented in A tend to represent different molecular complexes, sharing a single component. Interestingly, we found that this concept can be extended beside protein-protein interaction networks. Let us consider, for example, all those proteins, which are reported as changed in expression by at least two different papers on Parkinson's Disease. We will consider two proteins connected, if they co-occur at least 2

times, i.e. if they are reported together by at least two papers. The obtained co-expression network is shown in figure 9A.

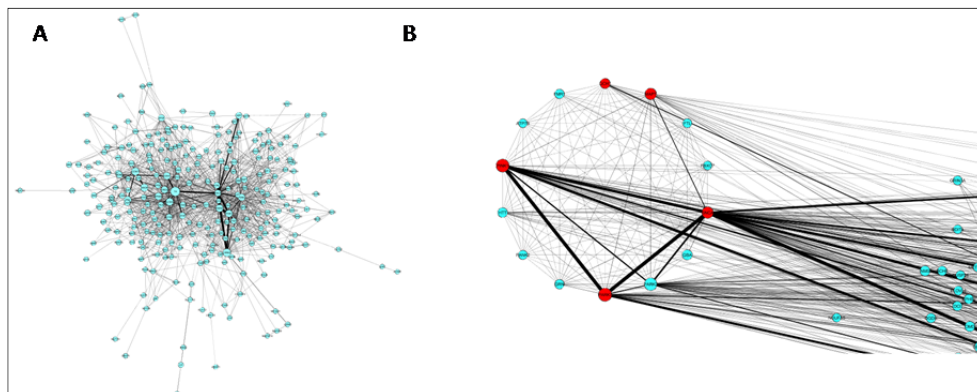


Fig. 9. A) Parkinson's Disease co-expression network; B) A clique from the same network (red nodes have many connections outside the clique).

On the right, in figure 9B, a neighborhood of 14 proteins is extracted from the network, which are all fully interconnected (meaning that each protein is connected to any other). Among these 14 proteins, the red ones are those which have at least as many bonds outside the neighborhood as they have inside it (i.e. at least 13 bonds outside the network). These 14 proteins are arranged in a way similar to that exemplified in figure 8B: a single cluster of highly connected nodes. Much in the same way predicted for protein-protein interaction networks, the cluster is enriched in proteins sharing some functional aspect: in particular, it turns out that 13 out of the 14 components are found in inclusion bodies, a hallmark of neurodegeneration in Parkinson's Disease. Intriguingly, in a sense they represent once again a macromolecular complex – albeit a non-specific one, being a structurally random aggregate, which may vary in its particular composition from case to case. Thus, while the starting network is a co-expression network, where edges do not represent physical interactions among proteins, also in this case proteins in well connected neighborhoods tend to share biological functions and to be involved in the formation of complexes.

## 5. The structure of protein networks: Graphlet degree signatures

Until now, we have examined pretty simple topological features of the nodes in a protein network. Recently, however, more complex metrics have been introduced, which have several advantages over the older ones. In particular, many of these sophisticated parameters are useful because they recapitulate a larger amount of information with respect to simpler ones. One of such parameter is the “graphlet degree signature” of a node, first introduced by Milenković T. & Przulj N. (2008). To understand what is it, let us consider figure 10.

Imagine that we want to study the local topology around the two colored nodes shown in figure 10A. A possible way would be to count all the graphlets of a certain type which pass through the nodes. Graphlets are small connected network subgraphs with a pre-

determined number of nodes. In figure 10B, we reported all the possible graphlets with 2 nodes and 3 nodes, with the designation  $G_0$ ,  $G_1$  and  $G_2$  originally introduced by Pržulj. As evident by figure 10C, node 1 is touched by 3, 3 and 1  $G_0$ ,  $G_1$  and  $G_2$  graphlets respectively. Node 2 is touched by 5, 5 and 2  $G_0$ ,  $G_1$  and  $G_2$  graphlets respectively. You can check the number of  $G_0$  and  $G_1$  graphlets on the left part of figure 10C, and the number of  $G_2$  graphlets (triangles) on the right; these numbers are called  $G_0$ ,  $G_1$  and  $G_2$  graphlet degree of a node. Thus, with respect to two- and three nodes graphlets, it is possible to define an ordered vector of the type  $\langle g_0, g_1, g_2 \rangle$ , which will describe for each node how many graphlets of any possible type actually pass through the node. For node 1 and node 2, this vector assumes the values of  $\langle 3, 3, 1 \rangle$  and  $\langle 5, 5, 2 \rangle$  respectively. The vector obtained considering all the 29 possible graphlets having from 2 to 5 nodes has been originally dubbed “graphlet degree signature” or simply “signature” of a node.

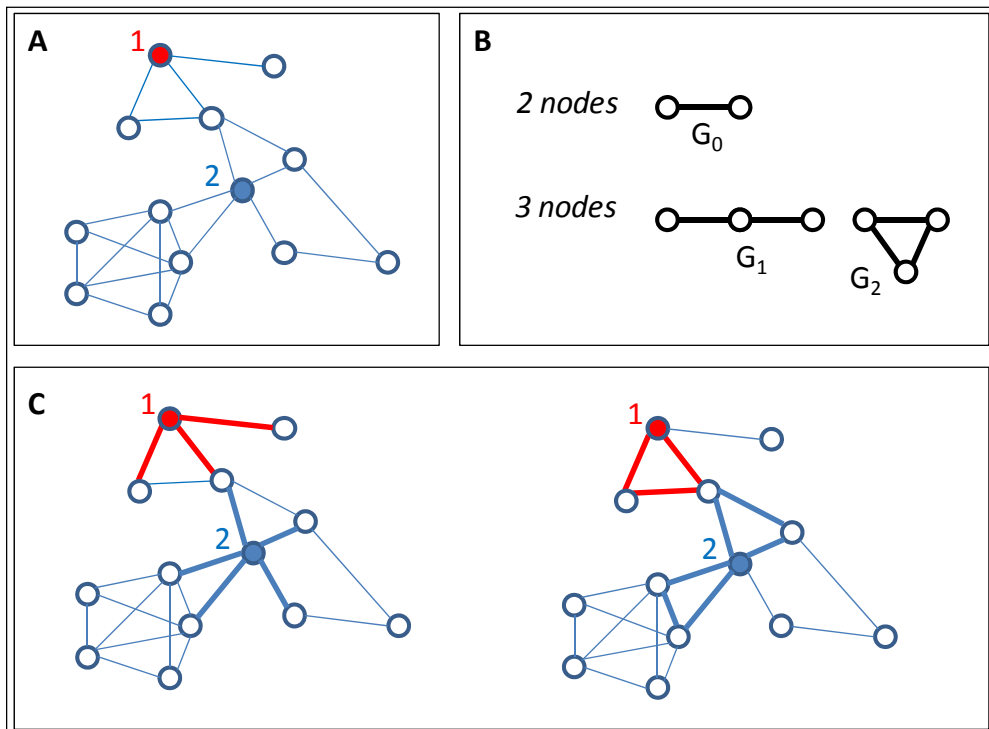


Fig. 10. A) Two nodes with a distinct topology; B) All the possible connection arrangement (graphlets) for groups of 2 or 3 nodes; C) Left, prevalence of  $G_0$  and  $G_1$  graphlets passing by nodes 1 and 2; Right, prevalence of  $G_2$  graphlets passing by nodes 1 and 2.

Before going further, it is important to note the following:

1. the  $G_0$  degree is equivalent to the node degree we saw in the preceding paragraphs; in this respect, the graphlet degree signature can be seen as a generalization of the node degree, which is not limited to the count of only a single type of graphlet;
2. By considering the type and number of graphlet connected to a node, the graphlet degree signature captures in a single metric both the degree of the node, the neighborhood abundance and its topology, recapitulating in a single measure the complementary aspects we introduced in the previous paragraphs.

Once defined in the way we have seen, the graphlet degree signature can be used to cluster all the nodes of a network according to their similarity. We will not enter into the details of the method, which is fully described elsewhere (Milenković T. & Przulj N. 2008); to our purposes, it is enough to understand that nodes with a graphlet degree signature similar above a certain threshold (which implies a similar centrality and a topologically equivalent neighborhood) can be grouped together. The resulting groups, however, may contain nodes which are quite far in the original network, so that nodes in the same cluster are in general scattered all over the network. Mostly relevant to the biologist, it has been shown how these clusters contain proteins of similar biological role and functioning (Milenković T. & Przulj N. 2008). This means that, at least in principle, if one selects a node with known biological features, it is possible to calculate its graphlet degree signature, search for nodes with a similar signature and transpose the biological features to what has been found, without considering the distance of the newly identified nodes from the starting point, as opposite to what we have seen in the previous paragraph.

## 6. A brief review of successful applications

We will see at this point how the topological analysis of biological networks has been already applied to achieve interesting results. Due to the limited space, we will restrict ourselves to few examples; however, the literature describing successful applications of network analysis in biology is growing at exponential rate, as evident by comparing the papers produced yearly and indexed by Pubmed for “network biology” in 2000 (372) to the corresponding figure for 2010 (2324).

A first, obvious application of topological network analysis consists in illuminating new aspects of the cell biology, which are evident only when looking to the full puzzle represented by a molecular net, instead then to the single pieces of it. The instruments used for such an analysis are many; however, even the simplest topological descriptors we have introduced in this chapter, such as the cliques, may be very useful.

To illustrate this point, let us refer initially to the classic work of Spirin and Mirny on the yeast protein-protein interaction network (Spirin V. & Mirny LA. 2003). These two authors were the first to describe the presence of densely connected modules in protein-protein interaction network, i.e. neighborhoods whose internal connectivity is very high compared to the average network connectivity. As we already know, in extreme cases – i.e. in case all the neighborhood's components are fully connected – these protein groups are cliques. As discovered by the aforementioned authors, cliques and very connected neighborhoods represent molecular complexes and/or functional modules. Thanks to this fact, the authors were able to identify a full wealth of new functional modules, including several previously



unknown molecular machineries, such as an eight-member module of cyclin-dependent kinases, cyclins and their inhibitors regulating the cell cycle, a six-member module of proteins involved in bud emergence and polarity establishment and a six-member module of CDCs, septins, and Ser/Thr protein kinases involved in mitotic control. From their starting seminal work, which for the first time shifted the network analysis from single node centrality to community of nodes, a deluge of research followed. This trend culminated in several complex applications of clique analysis, such as a recent work which nicely illustrated how the mitotic spindle functioning is regulated by a cascade of events which involves cliques (i.e. molecular complexes) instead of single proteins (Chen TC. Et al., 2009). With regard to more complex topological parameters, such as the graphlet degree signature introduced in the previous paragraph, there are obviously fewer examples, given the fact that they have been introduced much later. However, being refined instruments, the results obtained by their systematic application are somehow superior in generality, and uncover the real potency of the topological approach in molecular network analysis. To understand this, is sufficient to read a recent paper by Milenkovic T. et al., 2010. The authors describe how in a human protein-protein interaction network oncogenes do have a very similar graphlet degree signature, which is different from that of genes unrelated to cancer, at a point that they are able to use this signature to identify new oncogenes. If this finding will be confirmed by others, we will be forced to admit that the detailed topology around a node in a global protein-protein interaction map is important in determining the function of the corresponding protein at least at the same level as its sequence and three-dimensional structure – a somehow unexpected result, given the fact that protein-protein interaction networks are only a very abstract map of all the interactions which have been observed, without spatial and temporal resolution, and do not corresponds to any physical entity. However, we want to conclude this paragraph by stressing the fact that, albeit this and similar fundamental problems rest to be solved, and are matter of current and future research in the field, we are seeing already the first applications of network analysis in human therapy. In particular, although network science is still in its infancy, it is currently shifting from a better understanding of why a given drug works or not to the identification of new therapeutic interventions. As an example, consider the case of multi-drug therapy, which is a very active field of research and experimental work, due to its high potential in overcoming several obstacle to the effective pharmacological treatment of different conditions. As opposed to the classical “magic bullet” pharmacological paradigm, aiming to the ultra-specific targeting of a single protein, a new kind of approach to the design of a therapy is emerging, which relies on simultaneously targeting several molecular processes. The topological analysis of the molecular network underlying a specific disease is the only way to rational implement such an approach, allowing the quest for modulators acting on different network areas, so to attack different cellular pathways. This way to proceed was recently validated by some groups, which could identify the right combination of drugs to be used in a number of oncological conditions, such as incurable pancreatic adenocarcinoma (Azmi AS. Et al., 2010), as well as head and neck chemoresistant cancer (Ratushny V. et al., 2009).

From the point of view of the network topology, the approaches described in these papers can be seen as the targeting of control hubs within neighborhoods with quite distinct compositions and cellular functions (i.e. separated neighborhoods enriched in proteins with different functional annotation), a practical strategy which relies on the concepts discussed previously in this chapter and which wait to be extended to several other cases.

## 7. Conclusion: A concept-map for the analysis of network topology

Having listed some few examples, we would like to recall to the attention of the reader those elements which allow a successful analysis: a proper selection of the data set to start with, a correct identification of the rules used to build the network (i.e. the type of network to be analyzed), few general assumptions on the relationships between the topology and the biological properties of the proteins to be found, and a correctly chosen null-hypothesis for the minimization of false positives (which, if possible, should also take into account bias and errors).

Let us discuss briefly the first point, i.e. the selection of a proper data set to derive the nodes of the network. This step is crucially influenced by the scope of the network. For example, if the aim is to find potential drug targets for a given condition, a literature-derived dataset, including all the proteins known to be related to a certain disease – irrespectively of the type of relation they have with the studied condition – might be useful. A protein expression data set, containing data on differential protein expression, would be equally useful. On the contrary, taking into account a complete protein-protein interaction data set may be both misleading – given the fact that there is no guarantee that the proteins contained in it are expressed in the selected condition – and useless, because this type of database lacks information on those proteins which have strong activity and expression in the selected condition, but do not have any identified molecular partner.

As for the second point, usually people select the type of network (and thus the node linking rule) they want to build at the very first step – i.e., they use protein interaction databases to build protein-protein interaction networks, expression databases for co-expression networks and so on. However, there are certain cases where this passage is not automatic. For example, if the data source for the node list is the scientific literature, instead of building a literature co-occurrence network one can derive the linking rule from a different source, like a microarray experiment database. By combining a literature-derived list of nodes with microarray information for linking them, one would obtain a network, whose nodes are selected on the basis of a specific scientific topic, and are bound by co-expression, without the need to perform an actual experiment in the condition of interest.

As for the third point, it is true that, in general, the topology of a node is correlated to the relevance of the role that the corresponding protein plays in the particular condition the network refers to. However, one has to recall that:

1. The meaning of “topologically relevant proteins” varies with the type of network – for example, hubs in co-expression networks are usually housekeeping proteins, while in protein-protein interaction networks they may be core constituent of molecular complexes;
2. the specific meaning varies also with the network dimension – so that in a network including the full yeast proteome, topologically prominent proteins are heterogeneous in function, while in a network made of proteins involved in apoptosis the hubs are key apoptosis regulators;
3. obviously protein prioritization is affected by the particular topological quantity one is measuring – a protein may be an hub, yet may have no clique including it;
4. the relevance of a protein for the cell may be in gross contrast with what is perceived as relevant by the investigator – housekeeping proteins are very relevant to the functioning of the cell, but not so to someone wanting to find new drug targets.

Finally, coming to the forth point, we want to stress here that control models to be used for underpinning significant topological properties should vary, depending on the topological quantity under study. Thus, to get a control network for testing the relevance of some topological characteristic of a certain group of node, one may compare the results obtained on the actual network with those obtained in:

- a random network, i.e. a network made of the same number of nodes and edges, with fully random connection between the nodes- this is enough to test for the global distribution of topological quantities, such as the degree distribution or the existence of statistically relevant neighborhoods;
- a degree-preserving random network, i.e. a network made of the same number of nodes and edges, with the degree of each node preserved, but a completely different wiring – this is the proper control, when one want to test the association between some topological parameter and a specific biological attribute, which depends on the particular nodes considered;
- a set of random network (degree preserving type or not) – this is the proper control, when one want to test the probability of the emergence of the observed topology in a network

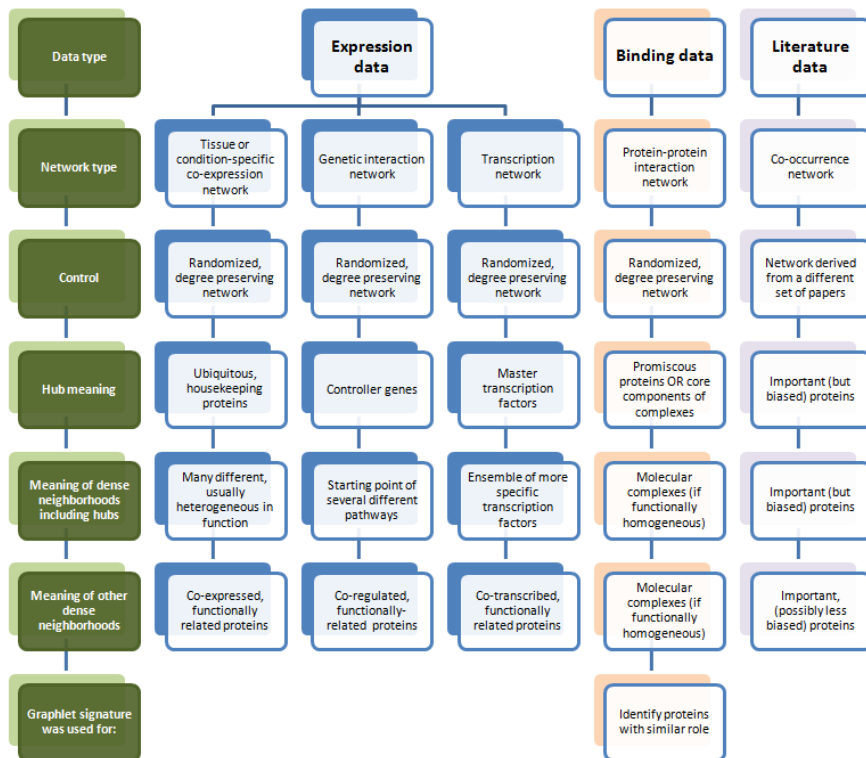


Fig. 11. Concept map for network topology analysis

After these considerations, we will conclude this chapter by outlining a general concept-map, reported in figure 11, which we feel can be useful in analyzing the topology of protein networks. This map should be regarded as a contribution to avoid common misinterpretations of the meaning of topological parameters in different contexts, not as an all-inclusive description of the possible applications and types of protein networks.

## 8. Acknowledgment

We would like to thank all the BioDigitalValley team involved in developing ProteinQuesttm, the tool which we used to explore the wonderful world of biological networks and to outline the concepts described in this chapter.

## 9. References

- Azmi AS, Wang Z, Philip PA, Mohammad RM. & Sarkar FH. (2010). Proof of concept: network and systems biology approaches aid in the discovery of potent anticancer drug combinations. *Mol Cancer Ther.*, Vol. 9, No. 12, (December 2010), pp. 3137-44
- Barabási AL. (2002). *Linked: How Everything Is Connected to Everything Else and What it Means for Business, Science, and Everyday Life*. ISBN 0-452-28439-2
- Chang W, Ma L, Lin L, Gu L, Liu X, Cai H, Yu Y, Tan X, Zhai Y, Xu X, Zhang M, Wu L, Zhang H, Hou J, Wang H. & Cao G. (2009). Identification of novel hub genes associated with liver metastasis of gastric cancer. *Int J Cancer*, Vol. 125, No.12, (December 2009), pp. 2844-53
- Chen TC, Lee SA, Chan CH, Juang YL, Hong YR, Huang YH, Lai JM, Kao CY. & Huang CY. (2009). Cliques in mitotic spindle network bring kinetochore-associated complexes to form dependence pathway. *Proteomics*, Vol. 9, No. 16, (August 2009), pp. 4048-62.
- Cohen R & Havlin S. (2003). Scale-free networks are ultrasmall. *Phys Rev Lett.*, Vol. 90, No. 5, (February 2003), 058701
- de Chassey B, Navratil V, Tafforeau L, Hiet MS, Aublin-Gex A, Agaagué S, Meiffren G, Pradezynski F, Faria BF, Chantier T, Le Breton M, Pellet J, Davoust N, Mangeot PE, Chaboud A, Penin F, Jacob Y, Vidalain PO, Vidal M, André P, Rabourdin-Combe C. & Lotteau V. (2008). Hepatitis C virus infection protein network. *Molecular Systems Biology*, Vol. 4, No. 230, (November 2008)
- Fernandes LP, Annibale A, Kleinjung J, Coolen AC. & Fraternali F. (2010). Protein networks reveal detection bias and species consistency when analysed by information-theoretic methods. *PLoS One*, Vol. 5, No. 8, (August 2010), e12083
- Fox AD, Hescott BJ, Blumer AC. & Slonim DK. (2011). Connectedness of PPI network neighborhoods identifies regulatory hub proteins. *Bioinformatics*, Vol. 27, No. 8, (Apr 2011), pp.1135-42

- Ivanic J, Yu X, Wallqvist A. & Reifman J. (2009). Influence of protein abundance on high-throughput protein-protein interaction detection. *PLoS One*, Vol. 4, No. 6, (June 2009), e5815
- Jeong H, Mason SP, Barabási AL, Oltvai ZN. (2001). Lethality and centrality in protein networks. *Nature*, Vol. 411, No. 6833, (May 2001), pp.41-2
- Kafri R, Dahan O, Levy J. & Pilpel Y. (2008). Preferential protection of protein interaction network hubs in yeast: evolved functionality of genetic redundancy. *Proc Natl Acad Sci U S A*, Vol. 105, No. 4, (January 2008), pp. 1243-8
- Milenkovic T, Memisevic V, Ganesan AK. & Przulj N. (2010). Systems-level cancer gene identification from protein interaction network topology applied to melanogenesis-related functional genomics data. *J R Soc Interface*, Vol. 7, No. 44, (March 2010), pp. 423-37
- Milenković T. & Przulj N. (2008). Uncovering biological network function via graphlet degree signatures. *Cancer Inform*, Vol. 6, (Apr 2008), pp. 257-73
- Navratil V, de Chassey B, Combe CR. & Lotteau V. (2011). When the human viral infectome and diseasome networks collide: towards a systems biology platform for the aetiology of human diseases. *BMC Syst Biol*, Vol. 21, (January 2011), pp. 5-13
- Ortutay C. & Vihinen M. (2009). Identification of candidate disease genes by integrating Gene Ontologies and protein-interaction networks: case study of primary immunodeficiencies. *Nucleic Acids Res.*, Vol. 37, No. 2, (February 2009), pp. 622-8
- Rambaldi D, Giorgi FM, Capuani F, Ciliberto A. & Ciccarelli FD. (2008). Low duplicability and network fragility of cancer genes. *Trends Genet.*, Vol. 24, No. 9, (September 2008), pp. 427-30
- Ratushny V, Astsaturov I, Burtress BA, Golemis EA. & Silverman JS. (2009). Targeting EGFR resistance networks in head and neck cancer. *Cell Signal*, Vol. 21, No. 8, (August 2009), pp. 1255-68
- Ravasz E, Somera AL, Mongru DA, Oltvai ZN. & Barabási AL. (2002). Hierarchical organization of modularity in metabolic networks. *Science*, Vol. 297, No. 5586, (August 2002), pp. 1551-5
- Sharan R, Ulitsky I. & Shamir R. (2007). Network-based prediction of protein function. *Mol Syst Biol.*, Vol. 3, No. 88, (March 2007)
- Spirin V. & Mirny LA. (2003). Protein complexes and functional modules in molecular networks. *Proc Natl Acad Sci U S A*, Vol. 100, No. 21, (October 2003), pp.12123-8
- Tan CS, Bodenmiller B, Pasculescu A, Jovanovic M, Hengartner MO, Jørgensen C, Bader GD, Aebersold R, Pawson T. & Linding R. (2009). Comparative analysis reveals conserved protein phosphorylation networks implicated in multiple diseases. *Sci Signal.*, Vol. 2, No. 81, (July 2009), ra39
- Thomas A, Giesler T. & White E. (2000). p53 mediates bcl-2 phosphorylation and apoptosis via activation of the Cdc42/JNK1 pathway. *Oncogene*, Vol. 19, No. 46, (November 2000), pp. 5259-69

- Vallabhajosyula RR, Chakravarti D, Lutfali S, Ray A. & Raval A. (2009). Identifying hubs in protein interaction networks. *PLoS One*, Vol. 4, No. 4, (April 2009), e5344
- Vogelstein B, Lane D, Levine AJ. (2000) Surfing the p53 network. *Nature*, Vol. 408, No. 6810, (November 2000), pp. 307-10.
- Watts DJ & Strogatz SH. (1998). Collective dynamics of 'small-world' networks. *Nature*, Vol. 393, No. 6684, (June 1998), pp. 440-2
- Zou X. (2010). The Topological Properties of Virus-Human Protein Interaction Networks, *The Fourth International Conference on Computational Systems Biology*, ISB 2010, Suzhou, China, September, 2010