Studying protein–protein interaction networks: a systems view on diseases

Ferenc Jordán, Thanh-Phuong Nguyen and Wei-chung Liu

Advance Access publication date 20 August 2012

Abstract

In order to better understand several cellular processes, it is helpful to study how various components make up the system. This systems perspective is supported by several modelling tools including network analysis. Networks of protein–protein interactions (PPI networks) offer a way to depict, visualize and quantify the functioning and relative importance of particular proteins in cell function. The toolkit of network analysis ranges from the local indices describing individual proteins (as network nodes) to global indicators of system architecture, describing the total interaction system (as the whole network). We briefly introduce some of these network indices and present a case study where the connectedness and potential functional relationships between certain disease proteins are inferred. We argue that network analysis can be used, in general, to improve databases, to infer novel functions, to quantify positional importance and to support predictions in pathogenesis studies. The systems perspective and network analysis can be of particular importance in studying diseases with complex molecular processes.

Keywords: PPI network; network analysis; centrality; disease

INTRODUCTION

Modern systems biology was born only recently, in the era of modern computing technology, but systems thinking in biology has a long tradition. Whereas classical works focused more on conceptual developments (e.g. cybernetics [1]), increased computing power and large databases make it possible to extend this research avenue towards a detailed analysis of systems architecture [2, 3], managing massive amounts of data [4] and allowing also for the dynamical simulations of complex systems [5]. In general, systems models provide a methodological and conceptual framework for the analysis of part-to-whole problems: how connected parts make up the whole system and how the architecture of the system constrains the behaviour of its parts. This approach makes it possible to better understand the relative importance of system components and make quantitative predictions for system-level understanding and management. This integrative view is more important as the studied biological functions are the results of more complicated background processes (e.g. several genes are expressed, multiple signalling pathways are involved). Major applications include drug design [6] and systems medicine [7]. Interactome networks have recently become one of the most appealing research topics in systems biology. In a previous study [8], the three interactome networks of interest are reviewed: metabolic networks, protein–protein interaction (PPI) networks and gene regulatory networks. Among them, PPI networks have taken much interest of lot of ongoing research as interactions among proteins are intrinsic to almost all cellular functions and biological processes.

Discovering PPIs has been a key problem in molecular biology and bioinformatics. Some good surveys about PPI research have been available [9]. Generally, there are experimental and computational methods for prediction of protein interactions. The experimental methods are divided into two groups, the traditional and the high-throughput ones. Traditional experimental methods typically include
co-immuno-precipitation and synthetic lethal screening. Although the high-throughput experimental detection methods for PPI (typically, yeast two-hybrid [10, 11], phage display [12], affinity purification and mass spectrometry [13] and protein micro-arrays) present many advantages over traditional experimental methods, they are still tedious, labour-intensive and usually have high false positive and high false negative rates.

Because the work itself involves quantitative tasks, computer science came to the scene offering computational approaches. Depending on the source of information used, computational approaches can be categorized into three groups: structure-based, sequence-based and genome-based, as shown in [14], respectively. Besides the methods based on a single data source, many bioinformaticians use multiple data sources, including Bayesian network approach [15], kernels method [16], probabilistic decision tree approach [17] and the hybrid approach [18].

The number of methods both experimentally and computationally has produced a huge amount of PPI data. Some of the well-known PPI database are MPPI (MIPS) (http://mips.helmholtz-muenchen.de/proj/ppi/), BioGRID (http://thebiogrid.org/), Human Protein Reference Database (http://www.hprd.org/), STRING (http://string.embl.de/), MINT (http://mint.bio.uniroma2.it/mint/Welcome.do) and the I2D (http://ophid.utoronto.ca/ophidv2.201/index.jsp).

With recent blooming of PPI data, there has been a shift from attempting to understand the molecular networks of other species to understanding the networks that underlie human diseases [19–23].

**NETWORK ANALYSIS IN SYSTEMS BIOLOGY**

One important toolkit of systems biology is network analysis. Network models are useful representations of several biological systems, ranging from metabolic pathways [24–26] to ecosystems [27]. Network analysis has its roots in sociology where sociologists were (and still are) interested in the patterns of interactions between people in groups. Various methodologies of network analysing tools were invented to describe network structures at the microscopic and the macroscopic levels [28]. At the level of nodes, one might want to identify important nodes basing on their network positions. For this, one can simply count how many direct neighbours a node has and this is often called the degree centrality \( D \); or one might employ more complicated methods by quantifying, for instance, the shortest distance \( d \) between a focal node and one other node, and then sum those distances up for all others as a proxy to its network position (a distance here refers to the number of links the focal node needs to traverse in order to reach the target node) with all others in the same social network. At the level of a network, network properties can be quantified by collecting information from the nodal level; this ranges from simple measures like link density, which is the number of observed links divided by the total amount possible, to more complicated one such as the averaged shortest distance between any node pairs. Several disciplines have borrowed the concept of network analysis from the sociologists in the last decade. For instance, with the advent of advanced techniques in molecular biology and bioinformatics, huge amount of data were generated on the interaction between various molecular species (e.g. metabolites, genes, proteins, enzymes); and by using network analysing tools, a picture starts to emerge regarding the organization of these interactions. An example of this is the gene regulation networks and a great deal is known about their structural organization [29–31]. Those networks have the ‘scale-free’ degree distribution, meaning the presence of a few nodes with huge number of connections, whereas the vast majority of others only have a few; and they also have the ‘small-world’-like structure where a node’s direct neighbours are densely connected among themselves, although the distance between any two nodes are typically small. Furthermore, nodes in gene regulation networks tend to cluster into modules, with smaller modules forming the parts of much larger modules in a hierarchical manner.

In PPI networks, graph nodes represent proteins, and links represent their interactions. Interactions can often be of two types. In the simplest case, an unsigned and undirected link exists between two proteins if they form a (part of a) protein complex pertaining to certain cellular functions. In the other case, a directed and signed link from one protein to another one exists if the former regulates (positively or negatively) the latter one (e.g. see [32]). The first type of interaction is more common in PPI network studies, whereas the second type is more adequate for modelling signal transduction networks. It is important to note that the particular research problem and
the available database largely determine exactly which network analytical tools can be used for analysis. Like the gene regulation network, PPI networks have also been extensively studied and their structures are also well known. For instance, the PPI network of yeast also possesses the scale-free degree distribution, and it has been demonstrated empirically that the removal of highly connected nodes is detrimental to the survival of yeasts [33]; and in human PPI networks, proteins involved in housekeeping tasks are found to form a core network, whereas other clusters of proteins performing tissue-specific duties tend to occupy more peripheral positions in the network [34].

Any network can be quantified by its global and local properties. Global properties describe the general architecture of the network, characterizing it by a single number, in a macroscopic way. Some of these metrics are quite simple and straightforward but provide a basic characterization of the network. These include the number of nodes \( N \), the number of links \( L \), the average distance in the network \( \langle d \rangle \), the diameter of the network (the maximum of \( d \)), the number of components, the average degree \( \langle d \rangle \left\langle D \right\rangle \) and the clustering coefficient \( CC \). The distance \( d \) here is defined as above, whereas the clustering coefficient measures the extent of interconnection between a focal node’s direct neighbours [35]. In some cases, certain combinations of these indices provide a meaningful metric. For example, the combination of the average distance \( \langle \omega d \rangle \) and the CC quantifies the ‘small world character’ \( SW = CC/\langle \omega d \rangle \) of a network. Beyond these simple indices, more complicated ones may help in special cases. For example, the degree-based measure of network centralization \( (NCI_D) \) takes a value that ranges from 0% (totally non-centralized, like a lattice) to 100% (totally centralized, like a star); it is defined as the sum of deviations of individual node degrees from the maximum degree in the observed network, and then normalized by what this sum could have been if the network assumes the most centralized structure (i.e. a star graph, [28]). NCI_D can be helpful, for instance, if network vulnerability is to be assessed. Using these macroscopic indices is particularly helpful when different networks need to be compared.

Local properties of networks describe the positions of their individual graph nodes. The simplest measure is node degree which is the number of neighbours of a given node. If degree is determined for all nodes in the network, its distribution can offer a simple and also visually appealing characterization of the architecture of the whole network. In many networks, we can observe that there are few nodes with many connections (hubs) and many nodes with only a few connections. If this scale-free property characterizes a network, it may have functional implications (e.g. error and attack tolerance, [36]).

Beyond studying the node degree and focusing on hubs in complex networks [34, 37], novel network analytical tools offer also more sophisticated characterization of network nodes (e.g. the communicability property [38]), the groups of nodes (e.g. clique, and gene set analysis [39], where a clique is a set of nodes where all of them are connected to each other, i.e. a complete subgraph), network modules (well-defined building blocks of large networks, such as a feed-forward loop, e.g. [40]) or nodes in particular positions (e.g. bottlenecks, i.e. nodes or small sets of nodes that are responsible for connecting large parts of the network, [41]) in large molecular networks. All of these approaches try to link structure to functionality, but as they suggest, understanding function is not simply related to the high connectedness of hub proteins. For example, gene set analysis is often used to identify what molecular species are involved in what regulatory pathways [39] by using methods akin to hypothesis testing in statistics; and in contrast, network analysis provides a way to identify nodes in bottleneck or important positions that are essential to maintain the integrity of the whole network together [41].

Recent network indices make it also possible to compare indirect relationships between network nodes. Topological importance \( (TI^n) \) is a general measure of centrality, focusing on how effects originated from a focal node can spread throughout the network: in short, the focal node can influence its immediate neighbours directly, and through those neighbours, it can then affect other more distant neighbours via indirect routes up to \( n \) steps [42]. In an additive and multiplicative way, this index provides a full quantification of all effects between nodes \( i \) and \( j \) up to \( n \) steps. Because \( n \) can be changed, indirect effects of different length can be compared with each other. This may provide information about the range of non-local interactions spreading in a molecular network. These indirect effects may not seem to be of highest importance intuitively, but mapping these relationships may help in quantifying the network structure behind pleiotropy,
interconnected pathways and probably also weak interactions (cf. stress response proteins [43]).

Certain proteins may be of high importance not because they are central in the network, but because of their unique position (their function may be harder to replace, as there is less structural redundancy in the network). The topological overlap (TO

i

) measure is derived from the previous index (TI

i

). With a threshold level defined (t), we can determine which are the stronger (over the threshold) and which are the weaker (below the threshold) interactors of a particular network node i. Then, for a pair of nodes i and j, we determine the overlap between the two sets of strong interactors. This is the overlap between the effective interaction fields of nodes i and j. For a given node i, we can sum all of these overlaps with all other nodes, and this sum provides the topological overlap value for node i (as a function of n and t). Large TO values show highly overlapping interaction fields, indicating structural redundancy, whereas low TO values indicate unique network positions [44].

Non-local, multi-node network indices provide novel, sophisticated tool for network analysis. They can help in quantifying indirect relationships [45], provide a structural characterization to linking diverse functions [46, F. Vaggi et al., submitted for publication], help to better understand evolution [47] and development [48], to map the background of pleiotropic effects [49–51] or making predictions based on integrated data and information [4].

A CASE STUDY: DISEASE-RELATED PROTEINS IN A WIDER CONTEXT

Large proteomic databases make it possible to take a systems view on proteins, to visualize and quantify their relationships and hopefully better understand their functions. Figure 1 shows the human PPI network of proteins assigned to five different diseases (cancer, diabetes, obesity, heart diseases and autism; see colour code in figure legend) and their interactive partners (in white). Data are obtained from the i2d [52] and OMIM [53] databases [45]. Cancer proteins are likely to play crucial roles in the network because, they are found to be well connected with proteins of the other four diseases. In spite of several documented connections between diseases, certain pairs of diseases (i.e. groups of disease-related proteins) are unconnected. For example, there is no link between diabetes-related and heart disease-related proteins. This structural isolation may imply the functional independence of these diseases. However, network analysis may offer alternative views on this as it can identify the proteins linking them indirectly, in several steps of interactions, and quantify their mediator effects (e.g. the strength of the indirect interactions between the group of diabetes proteins and the group of heart disease-related proteins). The analysis of gene ontology terms of these mediators shows a strong correlation between the TI measure and functional importance of these proteins [45].

Beyond quantifying network structure, according to the currently available databases, some particular findings may infer connections not yet described or assignments not yet added to databases.

Figure 2 shows the neighbourhood of the 14–3–3 protein zeta/delta protein (UniprotID: P63104 corresponding to the YWHAZ gene). It is the ego-network of P63104, showing its direct partners and the interactions between these partners (thus, the graph in Figure 2 is a subgraph of the one in Figure 1). Network analysis reveals that even though it is not richly connected to others, its centrality is quite high in the PPI network. It is not assigned to any disease in OMIM database but its neighbourhood makes it one of the most important mediators between different groups of disease proteins. Literature also suggests its hidden importance as a chaperon [54] with several weak links to other proteins involved in various functions such as cell growth and carcinogenesis [55], breast cancer re-occurrence after chemotherapy resistance [56], luteal sensitivity to PGF [57] and cell proliferation [58]. In this case, the position of this protein is informative only in a wider network context, considering indirect relationships and non-local neighbourhood.

Figure 3 shows the neighbourhood of a richly connected protein assigned to cancer, proto- oncogene tyrosine–protein kinase Src (UniprotID: P12931) in the OMIM database. Similar to Figure 2, here we present another ego–network, being a subgraph of the original network. In contrast to P63104, this protein is well linked in the human PPI network. Being mostly involved in colorectal cancer [59], P12931 plays a role in various signal transduction pathways in the cell. These include epidermal growth factor receptor and nerve growth factor signalling, as well as signalling in the immune system. Beyond being involved in all these functions, indirect indices of network analysis
also reveals that according to the i2d database, P12931 is the only protein that links otherwise un-connected groups of diabetes-related and heart disease-related proteins (note that it also links the heart disease-related proteins to the obesity group). Also in this case, quantifying the structure of the system of PPIs may reveal or infer biologically plausible functions.

PPI network analysis of disease proteins suggests that putting them in a wider network context provides insightful information. Network analytical techniques that identify key mediators suggest that these are not necessarily network hubs (e.g. P63104): relatively poorly connected network nodes may also play crucial role in linking different parts of large and complex networks (see also [41, F. Vaggi et al., submitted for publication]). Mediators between diseases, and in general, the central nodes of PPI networks, are likely to have pleiotropic functions due to their involvement in connecting various pathways. This role cannot be revealed by local approaches, and only by adopting a systems view and a network perspective can elucidate these functions.

DISCUSSION

Network analysis provides local and global quantification of network structure, characterizing also sets of nodes, indirect pathways and unique positions. These recently suggested, sophisticated tools offer a rich toolkit to better understand the functioning of the cell (in particular, the systems-level basis of diseases, see [19, 22, 60, 61]).
Xu and Li [62] studied the clustering of disease proteins and have found major differences between various PPI networks based on different databases. This calls for methodological standardization, even before performing network analysis. They have identified 178 disease proteins by considering their indirect neighbourhood: the topological features they studied include average distance from known disease proteins. In some cases, network analysis is used in combination with other tools [63]. For example, Dezso et al. [64] use a scoring statistics based on topology and functional information to detect disease pathways.

Goh et al. [65] make a clear difference between essential and disease genes. They have found that proteins corresponding to essential genes are frequently central in PPI networks but disease proteins may well be at the periphery. They provide an evolutionary explanation for this, suggesting that selection pressure may be responsible for the peripheral position of disease proteins in the network.

Beyond the academic interest, the applicative domain is also expanding. A key emerging area is

Figure 2: The neighbourhood of the P63104 protein (14-3-3 protein zeta/delta, corresponding to the YWHAZ gene). P63104 is in central position and its white colour indicates that it is not assigned to any disease in the database. Note that it is not linked to many proteins (medium degree) but its neighbours are of different colours, so it connects proteins involved in different diseases.

Figure 3: The neighbourhood of the P12931 protein (proto-oncogene tyrosine-protein kinase Src). P12931 is in central position and its blue colour indicates its assignment as a cancer-related protein in the database. Note that it has both pink and red partners, so it mediates effects between heart disease-related and diabetes-related proteins.
systems medicine, providing lots of examples for hopeful applications. Gon˜i et al. [66] have compared two neurodegenerative diseases in two tissues. Their PPI network analysis shows that proteins related to Alzheimer’s disease are more central in the brain, whereas proteins related to multiple sclerosis are more central in blood tissues. They used betweenness centrality, a non-local index. In other cases, networks are used rather for hypothesis generation and visualization [7]. It is suggested that some type of statistical analysis is needed to supplement the network analysis when identifying parts of the relevant network or study particular proteins of interest.

Finding key proteins (and the corresponding key genes and pathways) may shed new light on the structure-to-function relationship for these problems. Moreover, identifying key pathways, nodes linking different pathways or key sets of nodes are ways that can make network analysis even more systems-focused. These non-local, multi-node techniques may be of special interest in the real integration of large databases. In the case of complex diseases (such as cancer), instead of finding individual genes or proteins as potential candidates for medicine, the systems view is extremely important due to the fact that pleiotropic effects, indirect relationships and non-additive effects can often complicate our understanding of the system and making predictions of its behaviour.

**Key Points**

- Network analytical tools quantify network structure at local, meso-scale and global level.
- Meso-scale network indices are the least explored ones so far but they are crucial for understanding the role of central and unique nodes.
- The network approach can help disease research, as reductionist approaches are often limited by pleiotropy, indirect and non-additive effects.

**Acknowledgements**

We are grateful to two anonymous reviewers for their useful comments.

**References**


