Biotechnology Advances 31 (2013) 274–286

Contents lists available at SciVerse ScienceDirect

Biotechnology Advances

journal homepage: www.elsevier.com/locate/biotechadv

Research review paper

On the functional and structural characterization of hubs in protein–protein interaction networks

Paola Bertolazzi [a], Mary Ellen Bock [b], Concettina Guerra [c,*]

[a] IASI-CNR, viale Manzoni 30, Roma, Italy
[b] Department of Statistics, Purdue University, West-Lafayette, IN, USA
[c] College of Computing, Georgia Institute of Technology, Atlanta, GA, USA

ARTICLE INFO

Article history:
Received 12 October 2012
Received in revised form 13 November 2012
Accepted 1 December 2012
Available online 8 December 2012

Keywords:
Protein–Protein interaction networks
Network topology
Hubs
Protein 3D structure
Evolution
Lethality
Structural disorder

ABSTRACT

A number of interesting issues have been addressed on biological networks about their global and local properties. The connection between the topological properties of proteins in Protein–Protein Interaction (PPI) networks and their biological relevance has been investigated focusing on hubs, i.e. proteins with a large number of interacting partners. We will survey the literature trying to answer the following questions: Do hub proteins have special biological properties? Do they tend to be more essential than non-hub proteins? Are they more evolutionarily conserved? Do they play a central role in modular organization of the protein interaction network? Are there structural properties that characterize hub proteins?

Contents

* Corresponding author at: School of Interactive Computing, Tech Square Research Building, 85 Fifth Street NW, Atlanta, GA 30308, USA. Tel.: +1 4044256907.
E-mail addresses: paola.bertolazzi@iasi.cnr.it (P. Bertolazzi), mbock@purdue.edu (M.E. Bock), guerra@cc.gatech.edu (C. Guerra).

0734-9750/$ – see front matter
http://dx.doi.org/10.1016/j.biotechadv.2012.12.002

1. Introduction

This paper deals with Protein–Protein Interaction (PPI) networks and focuses on hubs, i.e. proteins in the network with a high number of interacting partners. A PPI network is represented by a graph, a mathematical entity $G(V,E)$, where V is a set of vertices (or nodes) and E is a set of edges, i.e. pairs of nodes with the meaning that the two nodes have some relation. In a PPI network nodes are proteins and the relation is an interaction between two proteins.

Proteins in PPI networks have a wide range of degrees, i.e. numbers of interacting proteins. It is not well understood why some proteins interact with hundreds of proteins and others interact with only a few or even only one (Gunasekaran et al., 2003). However, it seems intuitive that proteins interacting with multiple partners may have a major role in the functional and modular architecture of the interactomes. For instance, it seems quite reasonable to assume that hubs are more indispensable or essential for life, in that their knock-out could be more disastrous than that of the other proteins. Similarly, one would expect hubs to be more conserved throughout evolution. From a structural viewpoint, an interesting question is whether the hub proteins exhibit features, either geometric or physico-chemical, that can explain their ability to bind to different partners (Gursoy et al, 2008). These intuitive observations stimulated a lot of studies trying to show a link between topological properties, structural properties and biological function. However these investigations did not seem to reach definite conclusions, mostly because of the concerns raised on the quality of data examined. Sometimes robust correlations between those properties were detected in some organisms; often however the interaction data lacked any significant correlation between the examined features. Finding the reasons for such correlations, when detected, also raised an interesting debate.

In this paper we review the literature on hub proteins and their functional and structural characterization. The structure of the paper is the following. First, we briefly introduce PPI networks, present distinct types of protein interactions, provide reference to the main data bases and discuss some issues related to data accuracy. Then we introduce the concept of hub, using definitions and notations from graph theory. At this point we are ready to review the literature on hubs through three coordinates: topological, structural and conservation characteristics of such proteins.

2. PPI networks: notation, definitions and topological properties

A protein–protein interaction network for an organism is a list of proteins and their interactions. An interaction is defined to be physical contact of the two proteins (see De Las Rivas and Fontanillo, 2010 for more detail). In network science terminology, the PPI network is an undirected graph with each protein as a node. A graph $G(V, E)$ consists of a set of nodes V and a set E of pairs (u,v), $u, v \in V$, called edges. If the pairs are unordered then the graph is said to be undirected. If there is an edge between the nodes v and u the two nodes are said to be adjacent. The degree of a node v is the number of its adjacent nodes. Each edge connecting v to its adjacent nodes is incident to v.

In the PPI network two nodes have an edge between them if an interaction has been observed between the two proteins. The number of interacting partner proteins is the degree of the protein.

Two nodes (first and last) are "connected" in the terminology used in network science and graph theory if there is a path from the first node to the last node, i.e. there is a sequence of nodes each of whom has an edge with the next one in the sequence. A node in the sequence is only required to have an interaction with the node preceding it in the sequence and the one following it in the sequence. Thus, the first and last nodes in the sequence may not actually have an edge between them.

2.1. Distribution of degree

It has been observed in PPI networks that proteins with high degree are rare but proteins with low degree are quite common. We describe the empirical distribution of degree in a PPI network by defining the probability $P[k]$ of degree k to be the fraction of proteins in the PPI network with degree k. It has been observed (Jeong et al., 2001) for this empirical distribution that when $logP[k]$ is plotted on the vertical axis against $logk$ on the horizontal axis, then the points of the plot appear to form (approximately) a downward sloping line. The fact that the slope of the line is constant over various ranges of k is referred to as the "scale-free" property (Barabasi, 1999). The downward sloping line is the signature of a power law distribution, i.e. one for which $P[k]$ is proportional to k^{-A} where A is the slope of the line and A is a positive value. Thus, the distribution of degree is often modeled as following a power law. Typically, $2<A<3$ for PPI networks.

2.2. Complexes in PPI networks

Protein complexes are groups of proteins performing similar function or involved in the same biological process. They are the building blocks of molecular organization. As we will describe later in this survey, hubs play an important role in interconnecting such complexes. An extensive map of the complexes of the yeast PPI network was derived by large-scale experimental studies which integrated information from different sources (Gavin et al., 2006; Krogan et al., 2006).

Computational approaches to detect protein complexes in PPI networks have been designed based on the observation that complexes tend to correspond to highly interacting sets of proteins. In graph terminology, they correspond to dense subgraphs in a PPI network. Protein complexes are often evolutionary conserved, as they can be found in several organisms with an identical or similar interaction pattern. This observation is supported by computational studies on local alignment of two or multiple PPI networks that identified a large number of complexes common to yeast and fly and to human and fly, among others (Ciriello et al, 2012).

3. PPI databases and accuracy of interaction data

Interaction information is obtained by a combination of low-throughput and high-throughput experiments and computational techniques (Ito et al., 2001; Uetz et al., 2000). Two of the most common large scale methods for inferring the interactions are TAP-MS and Yeast two-hybrid (Y2H). Large databases documenting protein interactions are publicly available for several organisms, such as *Homo sapiens* (human), *Saccharomyces cerevisiae* (yeast), *Rattus norvegicus* (rat), *Mus musculus* (mouse), *Drosophila melanogaster* (fly), and *Caenorhabditis elegans* (worm). The databases include DIP (Xenarios et al., 2002), HIPPIE (Schaefer et al., 2012), MIPS (Pagel et al., 2005), MINT (Chatraryamontri et al., 2007), Biogrid (Stark et al., 2006), and HPRD (HPRD).