ISTITUTO DI ANALISI DEI SISTEMI ED INFORMATICA
"Antonio Ruberti"
CONSIGLIO NAZIONALE DELLE RICERCHE

P. Bertolazzi, M. Bock, C. Guerra

ON THE FUNCTIONAL AND STRUCTURAL CHARACTERIZATION OF HUBS IN PROTEIN-PROTEIN INTERACTION NETWORKS

R. 18, 2012

Paola Bertolazzi – IASI-CNR, viale Manzoni 30, Roma, Italy; paola.bertolazzi@iasi.cnr.it.

Mary Ellen Bock – Department of Statistics, Purdue University, West-Lafayette, IN, USA; mbock@purdue.edu.

Concettina Guerra – College of Computing, Georgia Institute of Technology, Atlanta, GA, USA; guerra@cc.gatech.edu.

ISSN: 1128–3378

Collana dei Rapporti dell'Istituto di Analisi dei Sistemi ed Informatica "Antonio Ruberti", CNR

viale Manzoni 30, 00185 ROMA, Italy

tel. ++39-06-77161
fax ++39-06-7716461
email: iasi@iasi.cnr.it
URL: http://www.iasi.cnr.it

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?

Key words: Protein-Protein interaction networks; network topology; hubs; protein 3D structure; evolution; lethality; structural disorder.

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; paola.bertolazzi@iasi.cnr.it
b. Department of Statistics, Purdue University, West-Lafayette, IN, USA; mbock@purdue.edu
c. College of Computing, Georgia Institute of Technology, Atlanta, GA, USA. guerra@cc.gatech.edu
∗ E-mail Corresponding Author: guerra@cc.gatech.edu
Address: Concettina Guerra, School of Interactive Computing, Tech Square Research Building, 85 Fifth Street NW, Atlanta, GA 30308

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?

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 node 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 $\log P[k]$ is plotted on the vertical axis against $\log k$ 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, 2003], HIPPIE [Schaefer et al., 2012], MIPS [Pagel et al., 2005], MINT [Chatraryamontri et al., 2007], Biogrid [Start et al., 2006], and HPRD [HPRD].

Accuracy for the presence of protein interactions obtained by high-throughput experiments suffers from high rates of false positives, especially in the unedited TAP-MS data for pairs of proteins that are in the same complex but not in direct physical contact. Furthermore, there is a bias in databases for the presence of interactions with well-studied proteins because the many experiments conducted with these proteins offer more opportunity to observe their interaction with other proteins. Other less well studied proteins may have a large number of interactions but they are not yet observed. False negatives for the observation of interactions between two proteins are believed to be numerous not only because the proteins have not been studied sufficiently but also because they were not investigated in the appropriate cells or under the appropriate conditions in the organism.

In addition, some interactions are temporal and may even be reversible, i.e. the two proteins disassociate and are no longer in physical contact. Only rarely are the protein interaction databases annotated to provide the spatial, temporal or context circumstances under which the interactions were observed. Most of the PPI databases are static rather than dynamic, i.e they do not show changes in interaction status between two proteins over time. (An exception is DYNSIN [Bhardwaj et al., 2011].)

In these existing databases, subsets of more reliable interactions may be selected based on biological knowledge available in the scientific literature or on the amount of experimental evidence, for instance when an interaction has been duplicated several times. The more reliable interactions are often referred to as "core data" [Ito et al., 2001] when trying to validate results.

4 Types of Interactions

Classification of interaction by type can provide important annotation to the PPI. Temporal quality of an interaction is denoted by classifying it as transient or permanent. A transient interaction occurs for a limited time and is reversible; the interface involved in a transient interaction may be used by multiple partners at different times. A permanent protein-protein interaction is strong and irreversible, and the interface is used by one partner only. An example of a transient interaction is that of a kinase binding a substrate within a particular signaling pathway. In general, the interactions involved in post-translational modifications are transient in nature. In the human proteome a large fraction of interactions are transient [Brown and Jurisica, 2007]. Examples of permanent interactions are those among proteins participating in a stable network complex that performs some specific function, for instance the 20S Proteasome.

A further characterization of a pair of proteins before an interaction adds to the temporal description of an interaction. An obligate interaction occurs between two proteins when they form a stable structure but neither of them has a stable structure on its own in vivo. A non-obligate interaction occurs between two proteins each of which has a stable structure on its own. Some interactions may not fall distinctly

in either category since stability can be described on a continuum. However, most obligate interactions are permanent [Nooren et al., 2003, Przytycka et al., 2010].

5 Hub Definitions

In network science terminology a hub is a node with high degree. We will use this definition for a hub protein. While not universal, most of the definitions of a hub protein follow this convention, i.e. a hub protein is a protein with many interacting protein partners. A hub protein is often referred to as "highly connected" in the sense that it has high degree [Barabasi, 1999]. The list of hub proteins for a PPI network in an organism is defined by choosing a minimum threshold for the degree, say k. Then all proteins with at least k interacting protein partners are hub proteins. The choice of this threshold parameter varies but it is important to note that even when the choice of k is pinned down, the list of hub proteins will still be highly dependent on the interaction database being used or the subset of it being used. As mentioned previously, the existing databases often have different degrees specified for a given protein because some databases include only very reliable interactions while others list the interactions obtained with all possible experimental and computational techniques. Furthermore the databases may change with time as more interactions are discovered.

Some definitions for hubs directly specify a numerical value for the threshold k (typically 5, or a larger number, most often 10). Other definitions give an indirect way of choosing the threshold k. Some rank the proteins of a PPI by degree and designate the hub proteins as those whose degrees fall in the top r percent (where r might be 10 or 5 or less) for the PPI. This ultimately results in a choice of threshold k. Another method uses the relative connectivity of the potential list of hub proteins to pick a threshold (see below). Based on a set of thresholds, some propose a classification of proteins into highly connected proteins, intermediately connected proteins, and non-hubs [Ekman et al., 2006].

Among hub nodes, there are some selected on the basis of additional topological properties that have biologically interesting properties more pronounced than the remaining hubs.

In [Vallabhajosyula et al., 2009] a methodology is proposed that identifies hubs based on the observation [Maslov et al., 2002, Maslov et al., 2004] that high degree proteins are often found to have lower connectivity among themselves than non-hub proteins. Therefore, one way to define the list of hub proteins involves identifying the set of high-degree proteins that has significantly lower mutual connectivity than proteins that do not lie in this set. This is done by introducing a simple topological measure of a graph, the relative connectivity, which is the relative size of its largest connected component, i.e. the number of nodes in the largest connected component divided by the total number of nodes in the graph. Hubs are selected as the top degree nodes whose corresponding subgraph has a small (according to some criterion) relative connectivity. Starting from the list of nodes ranked by their degree in decreasing order, subgraphs of increasing size are iteratively built by adding one node at time, each time computing the measure of relative connectivity of the resulting subgraph. This measure is small as long as high degree nodes are added, due to the property mentioned above. Then it should begin to increase as more low degree nodes are added. The experiments conducted in [Vallabhajosyula et al., 2009] show a sharp increase at a certain point during this iterative process. Thus hubs are identified as the subset of nodes (proteins) where such increase in the measure is detected.

6 Hub Classification

Hubs in PPI networks have been classified into categories along different directions. First, they can be classified into single and multi interface hubs, party and date, static and dynamic hubs, based on biological properties of interacting proteins such as co-expression or structural and temporal properties of their interfaces. Another type of classification takes into account topological properties of the PPI network

such as centrality measures and stresses the role played by hubs in maintaining the overall connectivity of the network. It has to be noted that these categories are often overlapping and in some cases only the terms used to denote them are different. In the following we review these categories and their relations.

Single and multi interface hubs

Single (or singlish)-interface hubs have only a few interaction interfaces (two at most) that may be used by multiple partners at different times since these interactions are mutually exclusive. By contrast, multi-interface hubs allow simultaneous interactions. Thus, the interactions of a single interface hub are generally transient, whereas those of a multi-interface hub are more likely to be permanent. Singlish-interface hubs tend to be enriched in signaling proteins, whereas multi-interface hubs are often present in protein complexes [Kim et al., 2006]. Statistically significant differences exist for these two classes of hubs in terms of important biological features, as we will discuss later.

Party and Date

This classification was introduced in [Han et al., 2004] and since then considered and tested by various authors raising some controversy. It is based on the correlation of mRNA expression of hubs with their interaction partners; the average Pearson correlation coefficient of hubs over all partners appears to follow a bimodal distribution allowing a clear separation of the two types of hubs: "party" hubs, which are highly correlated in their mRNA expression with their partners while "date" hubs show lesser correlation.

The authors suggest that the date hubs are more likely to be global regulators linking lower-level functional modules comprised of party hubs and their neighbors.

Date hubs tend to have transient interactions due to their low average co-expression correlation with their interaction partners, while party hubs were designated as permanent because of their high mRNA co-expression correlation.

It was suggested that party and date hubs play an important role in the modular organization of networks, i.e. party hubs have high connectivity to the members in a module, whereas date hubs are higher-level connectors between modules that perform varying functions.

Removing date hubs seemed to lead to very rapid disintegration into multiple components, whereas removal of party hubs had much less effect on global connectivity [Bertin et al., 2007, Han et al., 2004].

An appealing aspect of this separation is that it introduced some temporal and spatial characteristics into an otherwise static set of data. Party hubs are believed to interact with most of their partners at the same time while date hubs interact with their partners at different times and/or locations.

A somewhat different classification of hub proteins of the human interactome based on co-expression profiles quantifies the extent to which a hub and its interacting partners were co-expressed in the same tissues [Taylor et al., 2009]. Based on this, the authors identified intermodular hub proteins that are co-expressed with their interacting partners in a tissue-restricted manner and intramodular hub proteins that are co-expressed with their interacting partners in all or most tissues.

Since its appearance, the bimodality of the distribution at the basis of the classification of party and date hubs has been questioned in several papers [Agarwal et al., 2010, Batada et al., 2007, Yu et al., 2008, Wilkins and Kummerfeld., 2008] where little or no evidence for such separation was found. Furthermore, in [Agarwal et al., 2010] the role of date hubs in interconnecting separate modules of a PPI network was rejected by showing that a betweenness centrality measure is not a generic property of date hubs but instead of only a small subset of all date hubs. More generally, they observed that topological properties of hubs do not in general correlate with co-expression, reaching the conclusion that the dichotomy date/party for hubs in protein interaction networks is not meaningful.

Despite these observations, the classification in date and party hubs is widely used and cited in the

literature, and is supported by many studies, as we will see later in this survey.

Bottlenecks and Betweenness

Centrality is an important property of nodes in biological networks. While the degree itself may be considered a measure of centrality, it is only a local measure since it considers only the immediate neighbors of a node. More global definitions of centrality have been introduced that take into consideration the entire network and paths within the network to assign a value to each node that relates to the importance of a protein in the overall communication. A comprehensive overview of different centrality measures is published in [Koschtzki, 2005]. Here we focus on the shortest-path betweenness and its relationships with degree. Nodes with high value of shortest-path betweenness are called bottlenecks [Yu et al., 2007].

The shortest-path betweenness centrality assigns to each node a value given by the fraction of shortest paths that pass through it. It is defined as:

$$Bu = \sum_{i,j} \frac{\sigma(i,u,j)}{\sigma(i,j)}$$

where $\sigma(i,u,j)$ is the number of shortest paths between nodes i and j that pass through vertex or edge u, $\sigma(i,j)$ is the total number of shortest paths between i and j, and the sum is over all pairs i,j of distinct nodes.

Betweenness centrality can also be viewed as a measure of network resilience; it tells us how many shortest paths will get longer when a vertex is removed from the network.

In general, it has been shown that degree and betweenness are highly correlated quantities in biological networks. However, only a weak correlation could be detected for the yeast PPI network in [Wuchty and Stadler, 2003]. Although many bottlenecks tend to be hubs, a wide range of degrees was observed [Yu et al., 2007] for nodes of yeast interactome with high value of betweenness leading to the classification of hub nodes into $hub-nonbottlenecks$, $nonhub-bottlenecks$, $nonhub-nonbottlenecks$ and $hub-bottlenecks$. Among the non hubs, the bottlenecks appear to have an important role as connectors of modules in the network [Joy et al., 2005]. Because protein bottlenecks in the interaction network connect different functional modules, it is conceivable that bottlenecks with high degrees should have a higher tendency to be date hubs.

Globally central versus locally central

This classification stresses the importance of the hubs placement in the network and their role in linking the network modules [Wuchty, 2004, Wuchty and Almaas, 2005]. It arises from the observation that the degree alone may not be sufficient to characterize the proteins but more important is the participation of the proteins into subgraphs termed *cores*. A *k-core* is a subgraph obtained from the original graph through recursive removal of all nodes of degree less than k, with k varying in some given range [Seidman, 1983]. Nested layers of the network are identified with the innermost cores (corresponding to larger values of k) containing proteins which are not necessarily the highest degree nodes in the network.

Proteins in the innermost k-cores are defined to be globally central while highly connected proteins which are members of the outer k-cores are defined to be locally central.

7 Structural Properties

As more and more interaction data and protein structures became available, the natural question to ask was whether there exist structural properties that characterize hub proteins and explain their ability to recognize multiple interfaces in their interacting partners. We show an example of a node of the PPI

network of the Kaposi herpes virus and the 3D interaction of the corresponding protein with one of its incident nodes in figure 1

There seems to be consensus on the initial observations that hub proteins have higher propensity to possess unstructured or disordered regions that render them more flexible. It is in fact the flexibility, especially if operating at the global level, that allows a protein to bind to several partners by adopting different three-dimensional conformations.

Although the amount of structural data has significantly increased over the last decades, the number of complexes (the 3D structures of two or more proteins bound together) is still relatively small and in most cases limited to proteins binding to small fragments of other molecules. In particular, not many structures are available of the same protein bound to different targets. For this reason, most of the studies in this area have used disorder data produced by means of computational prediction tools, although more recently studies have analyzed dynamic conformational changes of a set of crystallized protein complexes present in the PDB [Bhardwaj et al., 2011]. Thus, the structural characterization supported by the current studies is valid as long as the data (both predicted and crystallized) do not vary significantly.

In this section we review the literature that linked structural properties to highly connected proteins in interaction networks concentrating on disorder and other geometrical properties. We do not discuss the biochemical properties such as residue composition and charge of hub proteins and of their interfaces; for a review on the subject see [Aloy and Russell, 2006, Patil et al., 2011].

7.1 Intrinsic Disorder

It was observed in the past that some proteins contain long extended regions that are unstructured in that their residues do not have a rigid 3D structure in physiological conditions [Huber and Bennett, 1983, Gunasekaran et al., 2003, Tompa, 2002, Uversky et al., 2000, Wright and Dyson, 1999]. However, only in the last decade a systematic study of the structural and functional role of the so-called disorder was conducted leading to a widespread recognition of its importance in many biological processes [Bellay et al., 2011, Chouard, 2011, Mittag et al., 2010, Uversky and Dunker, 2010]. This brought a relevant paradigm shift in structural bioinformatics where the unquestioned dogma per decades has been that structure dictates function. The well-known lock-and-key metaphor that describes the recognition process crucial to binding appears to be not always valid, as there is experimental evidence of cases in which a disordered region of a protein binds to a partner and in doing so assumes a specific three-dimensional conformation [Sugase et al., 2006].

Although there are several different definitions of a disorder region, there are ways of identifying such regions for the different definitions. One way is to resort to the 3D description of a protein in terms of the bond angles along the backbone or the side chains. In this definition of disordered proteins, such angles vary for different structures of the same protein available in the PDB; typically these differences are observed in long stretches of consecutive amino acids of the polypeptide chain. One case arises with highly flexible proteins that appear in remarkably different global conformations in bound and unbound states. This flexibility is often obtained by a small region or loop connecting well structured domains; it is this loop that, by changing its conformation, allows the domains to move with respect to one another. Likely evidence for a disordered region may also be present where the X-ray experiments fail to detect a 3D structure and therefore appears as missing in the PDB. The importance of disordered regions in ubiquitous processes such as phosphorylation is well documented [Iakoucheva et al., 2004], as typically the binding region of a kinase with its substrate is located at an extended linear region.

These observations stimulated a lot of research in the characterization of unstructured regions in terms of biochemical composition and sequence motifs that would allow the design of computational methods for their predictions. We do not survey such work here; for a related literature see [Uversky and Dunker, 2010]. We only mention that the available prediction tools are shown to achieve a relatively good success rate of about 80% in assigning any individual residue of a protein to either an ordered or a disordered part.

Based on predicted data, it was possible to estimate the distribution of disorder in a number of organisms. It was found that disorder is more present in complex organisms: in the human proteome about one third of the proteins residues are in disordered regions as apposed to only a few percent in other less complex species [Dunker et al., 2009, Dyson and Wright, 2005, Ward et al., 2004], in archaea and bacteria about 2-4%. This is attributed to the increased need in eukaryotes for cell signaling and regulation. Among the eukaryotes, the human proteins exhibit the highest percentage of disordered residues with fly, yeast, and worm following in this order.

The link between the number of interacting partners and disorder came under scrutiny in the work by [Dunker et al., 2005] where several structures of hub proteins binding to a large number of partners were analyzed displaying different degree of disorder, from completely to partly disordered to completely ordered. A notable example reported in the literature of a hub which features many disorder regions but also a structured part is that of the tumour suppressor p53, a protein with hundreds of links implicated in multiple signalling pathways in connection with human cancer. It consists of a large structured globular part surrounded by long highly flexible regions that seem to assume diverse conformations from order to disorder depending on their binding partners [Oldfield et al., 2008]. Interestingly, for some hubs, as 14-3-3, the binding regions of their partner proteins were found to be intrinsically disordered [Radivojac et al., 2006]. However, due to the limited dataset a statistical difference between the disorder of interacting partners of hub and non-hub proteins could not be established.

Although it seemed intuitive that the tendency to bind to many other proteins requires some versatility in its conformation, the confirmation came from systematic studies of the available topological and predicted disorder data [Dosztanyi et al., 2006, Ekman et al., 2006, Haynes et al., 2006, Kim et al., 2008, Patil et al., 2006, Singh et al., 2007]. The disorder parameters used in such analyses included the ratio of disorder residues over the total number of residues of a protein and the number of continuous predicted disorder regions of length greater than a certain threshold (typically > 30) or its proportion. These parameters were found to be significantly higher in hub proteins than in non-hub.

The disorder propensities of proteins with various numbers of interacting partners from four eukaryotic organisms (Caenorhabditis elegans, Saccharomyces cerevisiae, Drosophila melanogaster, and Homo sapiens) were investigated [Haynes et al., 2006].The analysis was carried out on the predicted structural disorder on four datasets obtained using PONDR VL-XT [Li et al., 1999, Romero et al., 2001]. A systematic analysis of the hubs in Saccharomyces cerevisiae proteome was conducted in [Kim et al., 2008] on disordered data predicted using the software DISOPRED [Ward et al., 2004].

The work in [Miyamoto-Sato et al., 2010] generated PPI data for 50 human transcription factors (TFs) and included the sequences involved in the interactions (i.e., the interacting regions, IRs). Analysis of the IR data set revealed the existence of regions that interact with multiple partners and are preferentially associated with intrinsic disorder. The results of all these works clearly indicate that structural disorder is a distinctive characteristic of hub proteins in eukaryotes.

The next question was whether disordered hub proteins have a tendency to interact among themselves. It was determined that the occurrence of interactions in the human proteome between disordered proteins was significantly frequent, and that between a disordered protein and a structured protein was significantly infrequent; furthermore this propensity for interaction was much stronger between non-hub proteins [Shimizu et al., 2009].

The analysis then focused on different classes of hubs, for instance, single and multi-interface hubs. It turns out that single interface hubs are enriched for disorder while multi-interface hubs are not [Kim et al., 2008]. Although single-interface hubs have a higher propensity for disorder, their interfaces appear to be structured; in fact the fraction of disordered residues at their interface is about the same as of multi-interface hubs. The promiscuity in the binding for single interface hubs, despite their somewhat structured interface, can be explained by: (a) the higher disorder of the interacting partners, as already observed in [Dunker et al., 2005], or (b) the geometric similarity in the surfaces patches of the multiple partners. Examples of both cases are available in the literature but not enough structural data

are available for meaningful statistics.

Other studies have explored the role of intrinsic disorder in party and date hubs (whether or not a hub has high mRNA co-expression with its partners) offering support for distinctive characteristics of these two types of hubs [Ekman et al., 2006]. Fewer of the party hubs contain long disordered regions compared to date hubs, indicating that these regions are important for flexible binding but less so for static interactions. Furthermore, party hubs interact to a large extent with each other, which is consistent with the idea that party hubs are at the core of highly connected functional modules [Ekman et al., 2006]. The work by [Singh et al., 2007] confirms that intrinsic disorder is significantly enriched in date hub proteins when compared with party hub proteins. Intrinsic disorder has been largely implicated in transient binding interactions. The disorder to order transition, which occurs during binding interactions in disordered regions, renders the interaction highly reversible while maintaining the high specificity [Yura et al., 2009]. The enrichment of intrinsic disorder in date hubs may facilitate transient interactions, which might be required for date hubs to interact with different partners at different times.

7.2 Conformational Rearrangements

Interactions are typically accompanied by conformational changes, which may involve the interface alone or may affect the entire structure. Thus proteins in their bound conformation exhibit structural changes with respect to their unbound conformation. The 'induced fit' model describes the binding process in which proteins achieve shape complementarity at their interface after a structural rearrangement.

An approach to the analysis of the structural rearrangements of hubs was taken in [Bhardwaj et al., 2011] where alternate conformations of proteins structures in human and yeast were extracted from the PDB databank and mapped into the DynSIN network to determine the dynamic conformational changes occurring at their interfaces with different partners. Thus, unlike other studies mostly based on predicted disorder and flexibility, they analyze actual protein structures: this has the advantage of not introducing errors in predicting the data, on the other hand renders the statistics less reliable due to the limited amount of structural data available (about 10% of the human proteins are present in the PDB with more than one conformation [Bhardwaj et al., 2011]). Alternate conformations of proteins are superimposed with a structural alignment heuristic that, rather than trying to maximize the overall RMSD of the aligned structures, focuses on the interface regions and tries to capture their changes.

The results of their analysis, in agreement with previous studies, show that hubs in human and yeast exhibit higher conformational flexibility than non-hubs. Furthermore, multi-interface hubs display a greater degree of conformational change than do single-interface ones; this is perhaps the feature that enables them to utilize more interfaces for interactions. They also find that transient associations involve smaller conformational changes than permanent ones, a fact that can be explained because proteins involved in transient interactions often interact with domains that are similar to each other and so do not require drastic structural changes for their activity [Bhardwaj et al., 2011].

7.3 Domain Composition

Protein domains are the building blocks of the protein modular architecture and play an important role in protein interactions. Most proteins are composed of two or more domains each of which is a substructure capable of folding independently. A domain may be present in diverse proteins performing different functions. It typically acts as the binding partner in various complexes. Domains may interface multiple proteins despite their often rigid structure using multiple surface patches. Examples of reusable and promiscuous domains are SH2 and SH3 which are important components of the signal transduction pathways.

The domain composition of proteins has been highly investigated to determine which laws govern their combinations and how the function of a protein relates to the presence of certain domains. Studies have been conducted to determine the number of co-occurring domain sets in nature versus the number of

putative combinations. It was shown that the sets of domains in yeast occurring significantly more often than by chance consist of ancient domains conserved from bacteria or archea [Cohen-Gihon et al., 2012].

The identification of domain families from sequences is a mature field in bioinformatics. Web tools are available to find domains on the basis of sequence similarity, protein order/disorder prediction. Domain databases such as Pfam [Punta et al., 2012] are also available.

It is conceivable that multiple domains potentially allow multiple molecular recognition sites and therefore multiple interactions thus suggesting a connection between number of constituent domains and degree of a protein. Although multiple interfaces exist even in small hubs with only a single domain [Humphris and Kortemme, 2007], an over representation of multi-domain proteins among the hubs was in fact detected in the yeast proteins [Ekman et al., 2006, Patil et al., 2010, Schuster-Bockler et al., 2007]. In fact a large fraction of protein interactions can be attributed to a small number of domains and often such domain interactions are conserved in different species. Interestingly, single-domain hubs have a greater fraction of disorder than multi-domain hubs [Patil et al., 2010].

Exploring the relation between degree and domains, another feature of the domain composition was considered besides their number, i.e. domain coverage. Domain coverage refers to the percentage of the residues in a protein that belong to a domain, thus excluding residues in loops or other structures outside a domain. It turns out that domain coverage has a better correlation with degree than the number of constituent domains [Xia et al., 2008]. Interestingly, the PPI domain coverage per protein appears to increase with the complexity of the organisms so that proteins in complex organisms contain more domains and perform more specialized function. This was observed for a set of several hundreds protein domains that are involved in PPI in 19 different organisms ranging from Kluyveromyces lactis to Homo sapiens, as well as two plants, Oryza sativa and Arabidopsis thaliana [Xia et al., 2008]. The over-representation of domains is one possible way of explaining the higher connectivity of the human network [Koonin, 2005] that is richer in interaction patterns, despite a similar number of nodes in lower organisms.

7.4 Geometric Surface Properties

Are there geometric properties that characterize the hub surfaces and their interfaces? Do hub interactions tend to occur in pockets, as in the case of protein-ligand binding? These questions have been addressed by many researchers working in the area of protein classification, molecular recognition and docking, to predict whether two proteins are likely to interact and if so what would be their interfaces. The work on surface patches classification has focused on particular classes of hubs. The analysis of the 3-D structure at the interfaces for permanent and transient (non-obligate) interactions did not reveal any pattern that could clearly discriminate between obligate/non-obligate and transient/permanent interactions. No single structural parameter of an interface either geometric, such as size of contact area or planarity, or biochemical, such as polarity or hydrophobicity, could identify the type of interaction [Nooren et al., 2003, Perkins et al., 2010]. However, some geometric properties appear to be more pronounced in some cases. For instance, the size of the interface areas of intrinsically disordered proteins is in general much larger per residue relative to ordered proteins, which supports the finding that disorder occurs preferentially in hub proteins. Interactions occurring in large surface pockets can be observed in some instances of dynamic interactions, although the majority take place in less constrained binding surfaces, such as in the instance of kinase-substrate interactions [Batada et al., 2006].

More recently, a classification of all domain interfaces of known structures was attempted, resulting in nearly 6000 distinct types of interfaces. Distinctiveness of interfaces was measured computing geometric features related to angles and overlap of the aligned interfaces followed by a hierarchical clustering to group the interfaces. A subset of known hubs was also examined showing very distinct surface regions at the interface with different partners [Kim et al., 2006].

8 Evolutionary Conservation

Like other characteristics of hubs, the connection between evolutionary conservation and high degree of nodes in PPI networks has been controversial [Jancura et al., 2011]. This is testified by the numerous papers which in turn demonstrated, rejected and reconfirmed this connection. That rigorous statistical analyses by different authors could lead to conflicting results should not surprise; in fact, as remarked from the very start in [Bloom et al., 2003, Fraser et al., 2003], the findings depend on the accuracy of the protein interaction and orthology data. Almost all works analyzed the yeast proteins and their conservation relative to a variety of eukaryotes at different evolutionary distances and of prokaryotes. The results were produced over a period of almost a decade during which the yeast interaction data varied significantly; furthermore, different databases were used and sometimes only core data were selected, i.e. subsets of very reliable interactions. Table 1 summarizes PPI data and results in several references.

In this review we first describe ways of defining evolutionary conservation and then survey the various approaches that try to establish a connection between protein interactions and evolution.

Central to the concept of evolutionary conservation is that of orthology. Orthologs are proteins that have originated from the same ancestor protein via speciation but exist in different species. Although this definition does not explicitly refer to the protein's function, in practice there exists a strong relation between orthology and biological function since orthologuos proteins typically perform an equivalent function in two species [Koonin, 2005]. Functionally related proteins from different species are characterized by small differences in the amino acid sequences, due to amino acids substitutions, the predominant type of change during evolution of such proteins, but also to insertions or deletions of one or more amino acids. Thus, sequence similarity provides evidence of functional conservation as well as of evolutionary relationships between the proteins.

Sequence similarity in orthologus proteins is measured by *evolutionary distance D* given by the formula $q = \frac{[ln(1+2D)]}{2D}$, where q is the proportion of identical residues in a sequence alignment of the protein pair [Grishin, 1995]. Orthology is a many-to-many relation, thus to compute the evolutionary distance a specific protein is selected among the putative orthologs according to some criterion; for instance a well-conserved ortholog characterized by a sequence identity above a certain threshold [Fraser et al., 2002].

For pairs of orthologous proteins, a correlation is found between D, the evolutionary distance between the pair, and the degree of the first member of the pair Several papers [Batada et al., 2006, Fraser et al., 2002, Fraser et al., 2003, Han et al., 2004, Jordan et al., 2003, Saeed et al., 2006] follow this approach and compute either the Pearson' correlation coefficient, denoted by r, or Spearman's rank, denoted by ρ.

Table 1 lists the characteristics of the data and methodologies utilized in several references, which study the evolutionary conservation of yeast proteins with respect to other organisms. Few results on the evolutionary preferences in organisms different from yeast were also produced (see [Lemos et al., 2005] for results on drosophila), but they are not listed in the table.

Perhaps, the first experimental study showing the preferential evolutionary conservation of proteins with higher degree, already hypothesized in [Hurst and Smith, 1999], was in [Fraser et al., 2002] where it was shown that the proteins of the yeast Saccharomyces cerevisiae with a high number of interacting partners have a smaller evolutionary distance to their orthologs in Caenorhabditis elegans.

Conflicting results were shown in a later study [Jordan et al., 2003] that analyzed two closely related species and led to the conclusion that clear correlations between the connectivity and evolutionary conservation of proteins could not be detected even in the presence of more accurate information about the orthologs between these two species. Their claim was that the previous results were due to a few highly interacting proteins that evolved more slowly and could not be attributed to all hubs. The experiments in [Wuchty, 2004] involving several higher eukaryotes, such as Homo sapiens, Mus musculus, Caenorhabditis elegans, Drosophila melanogaster, and Arabidopsis thaliana and in [Batada et al., 2006] also failed to indicate convincing correlations. However, a large scale study over multiple data sets demonstrated once more that the major reason for conflict between previous studies is the use of different datasets

Table 1: **Relation between the degree of the yeast protein and its evolutionary distance relative to its ortholog in a reference organism. The network data used to analyze the relation are described in column 2-4 and the orthology data in column 6-7. Column 8 gives the Pearson' correlation coefficient r, the Spearman's rank ρ, the orthologous excess retention ERk and the retention rate ERρ, when applicable.**

Reference	*Database*	*Proteins*	*PPI*	*Reference Organism*	*Orthology method*	*Orthology Pairs*	*Correlation*
[Fraser et al., 2002]	core data	2,445	3,541	C.elegans	Seq. ident.	309	r = -0.24
[Jordan et al., 2003]	MIPS	-	-	S. pombe	SEALS	1,004	ρ = -0.029
[Fraser et al., 2003]	MIPS et al	3,575	13,925	C. albicans	Max.Likel.	3,727	ρ = -0.25
				S. pombe	Max.Likel.	2988	ρ = -0.24
[Wuchty, 2004]	DIP	3,677	11,249	H. sapiens	InParanoid	1,997	r = -0.04; $ERk = 0.92$
				D. melanogaster	InParanoid	1,757	$r = -0.02; ERk = 0.82$
				C. elegans	InParanoid	1,489	r =-0.03; $ERk = 0.85$
				M. musculus	InParanoid	1,754	r = -0.06; $ERk = 0.96$
				A. thaliana	InParanoid	1,898	r = -0.06; $ERk = 0.89$
[Saeed et al., 2006] *	DIP	4,773	15,481	Mus musculus	InParanoid	2,354	ρ =-0.121
	MIPS	4,154	7,458	Mus musculus	InParanoid	2,064	ρ =-0.017
	INTACT	1,577	3,618	Mus musculus	InParanoid	1,036	ρ =-0.227
[Brown and Jurisica, 2007]	OPHID	5,652	95,104	H. sapiens	RBH		$ER\rho = 0.52$
				R.norvegicus	RBH		$ER\rho = 0.58$
				Mus musculus	RBH		$ER\rho = 0.58$
				D. melanogaster	RBH		$ER\rho = 0.62$
				C. elegans	RBH		$ER\rho = 0.55$

$*$ Other datasets, BIND, GRID, and MINT are used but are not reported here

[Saeed et al., 2006]. In general when no correlation was found, it was because the dataset had a large number of interactions derived from experimentally inaccurate methods. Datasets derived from robust experimental methods showed a better relationship.

An alternative approach to establishing a connection between number of interacting partners and evolutionary conservation is based on a different measure of conservation that refers to an entire proteome rather than to individual proteins: it is estimated by the presence of orthologs in related proteomes, more precisely by the fraction of proteins of a given organism that have an ortholog in another organism. This measure was analyzed in [Wuchty, 2004] as a function of node degree and was referred to as *orthologous excess retention (ERk)*. To compute ERk, all proteins of a given species are binned according to their degree. The binning of protein's degree is logarithmic to account for the scale-free property of the PPI networks. In bin k the fraction of the proteins that have an ortholog in the other species is computed. The orthologous excess retention for the bin is the ratio of this observed fraction and the fraction obtained from randomly distributed orthologous proteins. In all tests the values of ERk [Wuchty, 2004] showed a stronger relationship between conservation and degree than could be detected when using the evolutionary rate as a measure.

In [Brown and Jurisica, 2007] a large scale analysis of the relation of evolutionary conservation and degree was conducted with the goal of showing its effect on the transfer of interactions from pairs of proteins of one organisms to their orthologs in another organism. A variant of ERk, based on linear rather than logarithmic binning, denoted here as ERρ, was computed for the yeast proteins and their orthologs in several other species including D. melanogaster, Caenorhabditis elegans, M. musculus and A. thaliana confirming the results in [Wuchty, 2004].

More recently, the attention has shifted towards the analysis of the relation of evolution and number of interactions in conjunction with other topological and biological properties. In other words, the question was whether a restricted set of hubs possessing some additional property could be more highly conserved than the rest of the hubs. It appears that a variety of factors, in addition to degree, including expression

level, gene essentiality, may affect evolutionary rates in yeast [Plotkin and Fraser., 2007].

In [Wuchty and Almaas, 2005] the ERk measure was computed for a restricted set of hubs, the globally central hubs (introduced in section 6). In the yeast proteome they appear to be more evolutionary conserved than other hubs, suggesting they serve as the evolutionary backbone of the proteome.

The difference in evolution of party and date hubs was also investigated. It was shown that party hubs evolve more slowly than date hubs [Batada et al., 2007, Bertin et al., 2007, Fraser, 2005]. Nor surprisingly, multi-interface hubs, that are more likely to correspond to party hubs, were found to be more evolutionary conserved [Kim et al., 2006]. Related to these works are those that focused on the evolutionary conservation of hubs based on their participation in highly interconnected complexes, such as 26S Proteasome, reaching the conclusion that hubs present in stable complexes are more conserved evolutionarily than those participating in transient interactions [Amoutzias and Van de Peer, 2010, Brown and Jurisica, 2007, Wuchty et al., 2003]. This finding is consistent with the ones on party hubs that are believed to be mostly present in complexes. While a significant difference in the average evolutionary rate exists between hub and non-hub proteins which are not present in protein complexes [Coulomb et al., 2005], in [Manna et al., 2009] it was revealed that there exists no significant difference of hub and non-hub proteins present in stable complexes.

The study on protein evolutionary rate in [Pang et al., 2010] examined a different yeast PPI dataset which integrated protein interaction and gene co-expression data to derive a co-expressed protein-protein interaction degree (ePPID) measure, which reflects the number of partners with which a protein can permanently interact. Using this dataset they were able to detect a stronger correlation between degree of a node and its (protein) essentiality than when using PPI data alone.

Along a different direction, in [Fox et al., 2006] the existence of a positive correlation was determined between the degree of a protein and the conservation of its interaction partners; in addition, they show that a protein is more likely to be a hub if it has a high-degree ortholog.

In conclusion, it seems reasonable to say that, although in general no strong statement can be made about the preferential conservation of hubs, a weak correlation between evolutionary distance and degree could be detected in some organisms, especially when the data came from accurate experimental procedures. Furthermore, the orthologous excess retention ERk obtained much higher correlation values in support of the existence of a relation between conservation and degree. However, some of the improvement in correlation is due to the fact that binning is used. Finally, the correlation was stronger when other properties of proteins in addition to degree were incorporated in the analysis.

9 Functional Properties

A link between degree and function of a protein was always hypothesized [Kunin et al, 2004]. Here we review ways of showing such a link based on the impact of degree on essentiality and on protein functional classification.

9.1 Essentiality - Lethality

An informal widely used way of defining essential or lethal proteins/genes is that they are indispensable for the survival of an organism. In other words, they are the ones that, when knocked out, render the cell unviable [Giaever et al., 2002]. Since essential genes may cause the death of an organism if they are not properly expressed or malfunction, their identification is of paramount interest in biology.

Essential genes are frequently identified experimentally through deletion experiments (by the analysis of haploid deletion mutant strain growth rates). This process is particularly useful because it does not require knowledge of the function of genes. A systematic gene deletion screen [Giaever et al., 2002], for instance, determined 1,105 yeast genes essential for growth on rich glucose media. Other databases of essential genes are available for some prokaryotes and eukaryotes [Chen et al., 2010, Zang and Lin, 2009].

However, these experimental approaches are time-consuming, thus not very suitable for large-scale analysis. As a result, the essentiality profiles for a large fraction of genes are still unknown. Recently, computational approaches have been proposed to complement the experimental ones in predicting essentiality. One computational problem that has been addressed is that of identifying a minimal gene set needed to sustain a life form [Koonin, 2000]. Also studies have been conducted to predict the biological significance of a protein from its topological relevance; they basically use machine learning techniques to exploit the known topological properties of essential proteins, as well as other features such as gene expression, cellular localization and biological process information, in the prediction of putative essential proteins [Acencio and Lemke, 2009, Estrada, 2006, Li et al., 2012]. A complication arises when non-essential genes are discovered to cause cell death: it happens when, for instance, a pair of non-essential genes is deleted simultaneously. In this case we can call the interaction lethal. These phenomena are mainly studied in systems biology, that deals with the behavior of sets of genes instead of single genes.

The earliest work establishing the connection between high connectivity and essentiality was in [Jeong et al., 2001] where it was shown that in the yeast network high-degree nodes are three times more likely to be essential than nodes having few interaction partners. In the next few years, several studies reconfirmed this relationships [Batada et al., 2006, He et al., 2006, Yu et al., 2004, Yu et al., 2007, Wuchty, 2002, Zotenko et al., 2008]. However there have also been some contradictory results that rejected this hypothesis [Coulomb et al., 2005, Yu et al., 2008]. The tests in support of the different and sometimes contradictory claims were performed on various organisms, including yeast, Caenorhabditis elegans, Drosophila melanogaster, and human, and different datasets. The reason for testing several networks is to try to avoid the bias introduced by the concentration of experimental studies on essential proteins with the result that the corresponding nodes in the network have higher degree.

The statistical significance of the positive correlation initially found in the yeast was confirmed by successive studies performed on variants of the network of the yeast [Batada et al., 2006, Zotenko et al., 2008]. However, the authors of [Yu et al., 2008] disputed the correlation using a compilation of yeast high quality binary interaction data. No significant correlation was observed in any of three proteome-wide high-throughput binary data sets (table 2), as well as in Caenorhabditis elegans and human interactome maps. The discrepancy with the results of previous papers was attributed to the biases in the old data sets. In table 2 a summary of the network data and results is presented.

Essentiality was also examined for different kinds of hubs in the SIN network, a PPI network of the yeast that incorporates structural information [Kim et al., 2006]. Multi-interface hubs are found to be twice as likely to be essential as singlish-interface ones, which, in turn, are no more likely to be essential than the average protein.

Summarizing the above studies and results, it appears, as the intuition suggests, that essentiality is correlated with degree although the statistical significance is questionable. There is consensus however on the fact that when the interaction data are restricted to those obtained by Y2H experiments, the yeast network in its various instances exhibits no correlation or only a weak one between degree and essentiality [Batada et al., 2006, Yu et al., 2008, Zotenko et al., 2008]. It has to be noted that there is a difference in judging the quality of the H2Y data: they are considered not reliable in [Batada et al., 2006] while the H2Y binary dataset assembled in [Yu et al., 2008] is shown to be of high quality.

The next step was to provide an explanation for the connection between high degree and essentiality, assuming it really exists, but this has been a matter of debate among the researchers. In [Jeong et al., 2001] it was suggested that the over-representation of essential proteins among high-degree nodes in a protein interaction network was due to the important role of essential proteins in guaranteeing network integrity by interconnecting low-degree nodes. This hypothesis has been challenged with different arguments; in [He et al., 2006] it is suggested that the majority of proteins are essential due to their involvement in one or more essential protein−protein interactions that are distributed uniformly at random along the network edges. Under this hypothesis, hubs are proposed to be predominantly essential because they are involved in more interactions and thus are more likely to be involved in one

Table 2: **The relation between the degree and essentiality of the yeast proteins. The network data used to analyze the relation are described in column 2-4. Column 5 gives the Kendall's tau τ or R^2.**

Reference	*Database*	*Proteins*	*PPI*	*Correlation*
[Zotenko et al., 2008]	DIP CORE [Xenarios et al, 2003]	2,316	5,569	$\tau = 0.22$
	LC	3,224	11,291	$\tau = 0.32$
	HC [Batada et al., 2006]	2,752	9,097	$\tau = 0.32$
	Y2H	400	491	$\tau = 0.09$
[Yu et al., 2008]	Y2H-union	2018	2930	$R^2 = 0.001$

In the table, LC refers to a Literature Curated network; HC is a High Confidence network that combines small-scale data with high-throughput data that were independently reported at least twice. The Y2H network is obtained solely from high-throughput data that were experimentally detected at least three times. Y2H-union is the union of three reliable datasets Uetz-screen, Ito-core, and CCSBYI1. The datasets used in the experiments are rather different, as the correlation between the degrees in the sets is generally weak.

which is essential. In [Zotenko et al., 2008] the idea of essentiality being a function of a global network structure was rejected by showing that essential hubs are not more important in maintaining the network connectivity than non−essential hubs. This was done by studying the effect of the removal of nodes from a network in terms of the size of the largest connected component in the original and reduced network. The result was that the removal of essential nodes is not more disruptive than the removal of an equivalent number of random nonessential nodes that have the same degree distribution. Rather, according to [Zotenko et al., 2008], the majority of hubs are essential due to their participation in modules or complexes that consist of densely connected proteins with shared biological function that are enriched in essential proteins. The last surveyed results support a systems biology vision since they show that in many cases the biological function is not performed by a single protein but by two or more proteins together

Furthermore, [Kafri et al., 2008] showed that highly connected essential proteins tend to have duplicates which can compensate their deletion thus decreasing the disruptive impact of their removal. In both Saccharomyces cerevisiae and Caenorhabditis elegans duplicate genes evolve more slowly than singletons, indicating that some essential proteins are more likely to be redundant.

9.2 Correlation of Hub Properties with Essentiality

Because of the discrepancies of the above findings, the analysis on essential proteins was extended to their characterization in terms of other topological properties, to determine, for example, whether higher order centrality could be effective in predicting gene lethality. The studies revealed that a correlation with essentiality exists [Hahn et al., 2005, Park et al., 2009] for the betweenness measure in the three protein-protein interaction networks of yeast, worm, and fly , i.e. the centrality value for essential proteins is significantly higher than the centrality value of non-essential proteins. In contrast, the results in [Zotenko et al., 2008] obtained on several variants of the yeast network show that essentiality is no better correlated with other centrality measures than with the node degree. Hubs appear to be better predictors of essentiality than bottlenecks in PPI networks, although nonhub-bottlenecks are significantly more essential than nonhub-nonbottlenecks [Yu et al., 2007] .

A further characterization of hubs based on centrality measures was introduced in [Cho and Zhang, 2010] where a hierarchy of hubs based on the *path strength model* is identified in which highly scored hubs correspond to proteins located in critical positions of the network. The model assigns a value to a protein based not only on the paths traversing it, as the betweeness measures does, but also on the probabilities

of such paths given the weights of the edges incident to each node of the paths. The experimental results in the yeast protein interactome network demonstrate that the hubs in top positions in the obtained hierarchy are essential proteins for performing functions.

Another class of hubs identified based on the notion of relative connectivity (defined in section 6) are also found to be more enriched for essentiality [Vallabhajosyula et al., 2009]. Similarly, the globally central proteins that play a role in interconnecting network modules (defined in section 6) have higher probabilities to be essential to the survival of the organism as well as to be evolutionary conserved. In [Wuchty and Almaas, 2005] it was observed that essentiality and degree together correlated well with ERk, a measure of protein conservation (see section 8). In fact, when only essential hubs were selected better values of ERk were reported.

Essentiality was also studied in relation with cluster coefficient of a node in the network. The cluster coefficient is measure of local connectivity; for a node u it is defined as the number of edges connecting the nodes adjacent to u divided by $k \times (k-1)$, where k is the degree of node u. In [Coulomb et al., 2005, Yu et al., 2008] the degree of a node and the clustering coefficient of all the neighboring nodes were examined in relation with essentiality; however, no significant correlation with essentiality was observed in the yeast network. By contrast, when only a subset of reliable interactions of the yeast network was considered, a correlation was detected for the essentiality and the clustering coefficient of nodes [Yu et al., 2008].

From all the above studies, it appears that the addition of higher order topological properties to the degree often makes the correlation with essentiality stronger.

9.3 Functional Classification

One way to establish a link between degree and function is to resort to a functional classification of proteins and analyze the annotations of interacting pairs of proteins according to such classification.

The relationships between interacting pairs and Gene Ontology (GO) [GO] attributes was investigated in [Yu et al., 2008] using binary interaction yeast datasets of high quality. In general, significant enrichment (relative to random data) is detected for functionally similar pairs in the three categories, biological process, cellular component, and molecular function of GO ontology. But do hubs have higher propensity to be in one of those categories?

The analysis in [Ekman et al., 2006] uses KOG [Tatusov et al., 2003] functional classification that consists of four main functional categories: metabolism; information storage and processing; cellular processes and signaling; and poorly characterized. It shows that high degree is often associated with proteins involved in 'information storage and processing' (transcription in particular) and 'cellular processes and signaling'. Among the non-hubs, on the other hand, there are many proteins that participate in metabolism. As expected, proteins with no or few interacting partners are typically poorly characterized in terms of functions. This functional characterization of hubs is maintained when restricted classes of hubs are examined. The analysis on party and date hubs of the yeast proteome derived from the DIP database shows no significant difference in the percentage of proteins belonging to the four functional classes for the two types of hubs.

The analysis in [Alberghina et al., 2012] of a sub-network of the yeast PPI network, corresponding to the cell growth and cell cycle, examined the functional role of hubs, in particular of the 20 hubs of highest degree, and determined that they are involved in biological processes such as ribosome biosynthesis and maturation, protein synthesis, folding and glycosylation and in microtubule formation. Contrary to the belief that hubs may be major regulatory molecules, they found that only 2 out of the top 20 molecules play in fact a regulatory role.

A somewhat different approach was taken in the study of [Kim et al., 2006] that examined if the two different kinds of interactions, "simultaneously possible" and "mutually exclusive", could be distinguished in terms of functional properties of the linked proteins. Using the Gene Ontology classification into cellular component, molecular function, and biological process designations, they showed that proteins

connected by simultaneously possible interactions are more likely to share the same function than are those connected by mutually exclusive ones.

The fact that hub proteins tend to share certain functional features that enable them to participate in multiple protein interactions was utilized for the theoretical identification of such hub proteins without prior knowledge of the corresponding PPIs [Hsing et al., 2008]. Using machine learning methods, they developed a hub classifier that could predict highly-connected proteins, even in organisms that lack protein interaction data. The classifier is based on Gene Ontology data, which provide functional annotations for individual proteins in hundreds of species.

10 Discussion and Conclusions

The roles of hubs within PPI networks has been examined from different perspectives taking into account functional properties and biological importance. Despite the large body of literature on hub characterization, for some of the above questions there is not much consensus on the findings as detailed in the survey. A complicated and somewhat confusing picture emerged from all the works conducted on different organisms and different datasets with different levels of reliability. Although correlation, albeit weak, could be detected between degree and some functional and structural properties of proteins, the main concerns were experimental artifacts and other biases present in the networks that favor some proteins believed to be important. Other important factors that raised much concern, besides the current limited size of the interaction datasets, were the interaction reliability, and the lack of annotations of interactions. Restricted sets of reliable and high quality interactions seemed to resolve the issues in some cases, however some concerns still remain about the validity of conclusions drawn in the presence of ever changing data sets.

Since the results have been susceptible to the various datasets used, it is evident that there is still need for systematic studies on extended datasets of improved quality which include interaction annotations (transient, permanent, regulatory, kinase-substrate, etc.). The proteome-wide studies rarely take into account the variability of hubs in terms of degree and interaction patterns; hubs are defined as nodes with degree higher than 10 (in some cases 5) and, for instance, in the human the highest degree protein MYC has about 1000 partners and there many proteins with more than 200 partners. On the other hand, hub proteins of important functional classes, as for example kinases, have much smaller degrees and some distinctive structural interaction patterns when binding to their substrate. Due to the wide range of the number of interaction partners and to the variety of types of interactions, the significance of the possible correlation of biological and functional properties with degree is often difficult to establish.

Even when the analysis focused on restricted classes of hubs or restricted set of interactions, the variability within the sets is perhaps still too high to allow to draw some general conclusions. Transient and permanent interactions have been the focus of several investigations and although some properties are more frequently observed for one of the two types of interaction, the sets of proteins within each class show a wide range of different characteristics from a functional perspective. As observed in [Valente et al., 2009], transient interactions include those associated with a protein/complex performing a standard function on many target proteins/complexes as well as those occurring when two proteins complexes come together in a more delimited functional context. Examples of the first type are the transient interactions between a chaperone and its hundreds of targets, while examples of the second type are those for kinase-substrate within a particular signaling pathway.

At the proteome-wide level, the incorporation of 3D structural data and of genetic information, in addition to consideration of protein dynamics, will help gaining deeper insights into the basic biological functionality underlying networks. The more recent connection of hubs with intrinsic disorder is adding to the picture.

In this survey we have not discussed the link between global and local network properties and human diseases. It seems intuitive that hubs should preferentially encode disease-related genes. In fact, several

well-known and extensively studied proteins that are implicated in diseases are hubs. Examples include p53, p21, p27, BRCA1, kalirin, ubiquitin, calmodulin and many others which play central roles in various cellular mechanisms. Cancer-related proteins have, on average, twice as many interaction partners as non-cancer proteins in protein-protein interaction networks [Jonsson and Bates, 2006]. However, as observed already, this fact may be attributed to the the more extensive study of cancer-related proteins that led to their higher connectivity. The analysis of network properties with a systems biology approach that studies the combined effect of sets of proteins and relations between essentiality and degree may provide the rationale for combinatorial drugs that target less prominent nodes to increase drug efficacy and create fewer side effects [Hase et al., 2009, Keskin et al., 2007]. We will certainly see many research endeavors in this area in the next years.

Acknowledgements
The work of Paola Bertolazzi is partially supported by Project Sysbio of CNR.

References

Acencio and Lemke, 2009. Acencio, M., Lemke, N. (2009) Towards the prediction of essential genes by integration of network topology, cellular localization and biological process information. BMC Bioinformatics, 10, 290.

Agarwal et al., 2010. Agarwal, S., Deane, C.M., Porter, M.A., Jones2, N. S. (2010) Revisiting Date and Party Hubs: Novel Approaches to Role Assignment in Protein Interaction Networks, PLoS Comput. Biol., 6, 6.

Alberghina et al., 2012. Alberghina, L., Mavellib, G., Drovandi, G., Palumbo, P., Pessina, S., Tripodia, F., Coccettia, P., Vanonia,M. (2012) Cell growth and cell cycle in Saccharomyces cerevisiae: Basic regulatory design and protein-protein interaction network Biotechnology Advances, 30, 1, 5272

Aloy and Russell, 2006. Aloy, P., Russell, R.B. (2006) Structural systems biology: modelling protein interactions. Nat. Rev. Mol. Cell. Biol., 7, 188-197

Amoutzias and Van de Peer, 2010. Amoutzias, G., Van de Peer, Y. (2010) Single-gene and whole-genome duplications and the evolution of protein-protein interaction networks. Book Chapter in the book Evolutionary Genomics and Systems Biology (,Ed. Caetano-Anolls, G.), 413-429.

Barabasi, 1999. Barabasi, A.L., (1999) Emergence of scaling in random networks. Science, 286, 509-512.

Batada et al., 2006. Batada, N.N., Hurst, L.D., Tyers, M. (2006) Evolutionary and physiological importance of hub proteins. PLoS Comput. Biol., 2, 7, e88.

Batada et al., 2007. Batada, N.N., Reguly, T., Breitkreutz, A., Boucher, L., Breitkreutz, B.J., Hurst, L.D., Tyers, M(2007) Still stratus not altocumulus: Further evidence against the date/party hub distinction. PLoS Biol., 5, e154.

Bellay et al., 2011. Bellay, J., Han, S., Michaut, M., Kim, T., Costanzo, M. Andrews B.J., Boone, C., Bader G.D., Chad, L., Myers, C.L, and Kim, P.M. (2011) Bringing order to protein disorder through comparative genomics and genetic interactions Genome Biology, 12:R14.

Bertin et al., 2007. Bertin, N., Simonis, N., Dupuy, D., Cusick, M.E., Han, J.D., Fraser, H.B., Roth, F.P., Vidal, M. (2007) Confirmation of organized modularity in the yeast interactome. PLoS Biol., 5, e153.

Bhardwaj et al., 2011. Bhardwaj, N., Abyzov, A.,Clarke, D., Shou, C., Gerstein, M. B. (2011) Integration of protein motions with molecular networks reveals different mechanisms for permanent and transient interactions. Protein Science, 20, 10, 1745-1754.

Bloom et al., 2003. Bloom, J., Adami, C. (2003) Apparent dependence of protein evolutionary rate on number of interactions is linked to biases in protein-protein interactions data sets, BMC Evolutionary Biology, 3, 1, 21.

Brown and Jurisica, 2005. Brown, K.R., Jurisica, I., (2005) Online predicted human interaction database. Bioinformatics, 21, 2076-2082.

Brown and Jurisica, 2007. Brown, K. R. and Jurisica, J. (2007) Unequal evolutionary conservation of human protein interactions in interologous networks. Genome Biology, 8:R95.

Chatraryamontri et al., 2007. Chatraryamontri, A., Ceol, A., Montecchi Palazzi, L., Nardelli, G., Schneider, M.V., Castagnoli, L., Cesareni, G. (2007) MINT: the Molecular INTeraction database. Nucleic Acids Res., 35(Database issue): D572D574.

Chen et al., 2010. Chen, W.H. Minguez, P., Lercher, M.J., and Bork, P. (2012) OGEE: an online gene essentiality database. Nucleic Acids Res., 40, D1, D901-D906.

Cho and Zhang, 2010. Cho, Y., Zhang, A., (2010) Identification of functional hubs and modules by converting interactome networks into hierarchical ordering of proteins. BMC Bioinformatics, 11(Suppl 3):S3.

Chouard, 2011. Chouard, T., (2011) Breaking the protein rules, Nature, 471,151-153.

Ciriello et al, 2012. Ciriello, G., Mina, M., Guzzi, P.H., Cannataro, M., Guerra, C. (2012) AlignNemo: A Local Network Alignment Method to Integrate Homology and Topology, PlosOne,

Cohen-Gihon et al., 2012. Cohen-Gihon, I., Nussinov, R., Sharan, R. (2007) Comprehensive analysis of co-occurring domain sets in yeast proteins. BMC Genomics, 8,161, doi:10.1186/1471-2164-8-161.

Coulomb et al., 2005. Coulomb, S., Bauer, M., Bernard, D., Marsolier-Kergoat, M.-C. (2005) Gene essentiality and the topology of protein interaction networks, Proceedings of the Royal Society B: Biological Sciences. 272, 1573, 1721-1725.

De Las Rivas and Fontanillo, 2010. De Las Rivas, J., Fontanillo, C. (2010) Protein-Protein interactions essentials: key concepts to building and analyzing interactome networks. PLoS Comput. Biol., 6, 6, e1000807.

Dosztanyi et al., 2006. Dosztanyi, Z., Chen, J., Dunker, A.K., Simon, I., Tompa, P. (2006) Disorder and sequence repeats in hub proteins and their implications for network evolution. J. Proteome Res., 5, 2985-2995.

Dunker et al., 2005. Dunker, A.K., Cortese, M.S., Romero, P., Iakoucheva, L.M., Uversky, V.N. (2005) Flexible nets. The roles of intrinsic disorder in protein interaction networks. FEBS J. 272, 5129-5148.

Dunker et al., 2009. Dunker, A.K., Oldfield, C.J., Meng, J., Romero, P., Yang, J.Y., Chen, J.W., Vacic, V., Obradovic, Z., Uversky, V.N. (2008) The unfoldomics decade: an update on intrinsically disordered proteins. BMC Genomics, 16, 9.

Dyson and Wright, 2005. Dyson, H.J., Wright, P.E. (2005) Intrinsically unstructured proteins and their functions. Nat. Rev. Mol. Cell Biol., 6, 197-208.

Ekman et al., 2006. Ekman, D., Light, S., Bjrklund A.S. and Elofsson, A. (2006) What properties characterize the hub proteins of the protein-protein interaction network of Saccharomyces cerevisiae? Genome Biology, 7, R45.

Estrada, 2006. Estrada, E. (2006) Virtual identification of essential proteins within the protein interaction network of yeast. Proteomics, 6, 35-40.

Fox et al., 2006. Fox, A., Taylor, D., Slonim, D.K. (2009) High throughput interaction data reveals degree conservation of hub proteins. Pac Symp Biocomput., 391-402.

Fraser, 2005. Fraser, H.B. (2005) Modularity and evolutionary constraint on proteins. Nat. Genet., 37, 351-352.

Fraser et al., 2002. Fraser, H., Hirsh, A., Steinmetz, L., Scharfe, C., and Feldman, M. (2002) Evolutionary rate in the protein interaction network. Science, 296, 750-752.

Fraser et al., 2003. Fraser, H., Wall, D., and Hirsh, A. (2003) A simple dependence between protein evolution rate and the number of protein-protein interactions. BMC Evolutionary Biology, 3, 11.

Gavin et al., 2006. Gavin, A.C., et al. (2006) Proteome survey reveals modularity of the yeast cell machinery. Nature, 440,7084, 631-6.

Giaever et al., 2002. Giaever, G., Chu, A.M., Ni, L., Connelly, C., Riles, L., et al. (2002) Functional profiling of the Saccharomyces cerevisiae genome. Nature, 418, 6896, 387-391.

Grishin, 1995. Grishin, N. (1995) Estimation of the number of amino acid susbtitutions per site when the subsitution rate varies among sites. J. of Mol. Evol., 41, 5, 675-679.

Gunasekaran et al., 2003. Gunasekaran, K., Tsai, C.J., Kumar, S., Zanuy, D., Nussinov,, R. (2003) Extended disordered proteins: targeting function with less scaffold. Trends Biochem. Sci., 28, 81-85.

Gursoy et al, 2008. Gursoy, A., Keskin, O., Nussinov, R. (2008) Topological properties of protein interaction networks from a structural perspective. Biochem Soc Trans., 36, 1398-403;

Hahn et al., 2005. Hahn, M.W. and Kern, A.D. (2005). Comparative genomics of centrality and essentiality in three eukaryotic protein-interaction networks. Mol. Biol. Evol., 22, 4, 803-806.

Han et al., 2004. Han, J.-D.J., Bertin, N., Hao, T., Goldberg, D.S., Berriz, G.F., Zhang, L.V., Dupuy, D., Walhout, A.J.M., Cusick, M.E., Roth, F.P., Vidal, M. (2004) Evidence for dynamically organized modularity in the yeast protein-protein interaction network. Nature, 430, 88-93.

Hase et al., 2009. Hase, T., Tanaka. H., Suzuk. Y., Nakagawa, S., Kitano, H. (2009) Structure of Protein Interaction Networks and Their Implications on Drug Design. PLoS Comput Biol., 5, 10, e1000550. doi:10.1371/journal.pcbi.1000550

Haynes et al., 2006. Haynes, C., Oldfield, C.J., Ji, F., Klitgord, N., Cusick, M.E., Radivojac, P., Uversky, V.N., Vidal, M., Iakoucheva, L.M. (2006) Intrinsic disorder is a common feature of hub proteins from four eukaryotic interactomes. PLoS Comput. Biol., 2, e100.

He et al., 2006. He, X., Zhang, J. (2006) Why do hubs tend to be essential in protein networks? PLoS Genetics, 2, 6, e88doi:10.1371.

Hsing et al., 2008. Hsing, M., Byler, K.G., Cherkasov, A. (2008) The use of Gene Ontology terms for predicting highly-connected 'hub' nodes in protein-protein interaction networks. BMC Systems Biology, 2, 80.

Huber and Bennett, 1983. Huber, R., Bennett Jr., W.S. (1983) Functional significance of flexibility in proteins. Biopolymers, 22, 261-279.

Humphris and Kortemme, 2007. Humphris, E.L., Kortemme, T. (2007) Design of Multi-Specificity in Protein Interfaces. PLoS Comput. Biol., 3, e164.

Hurst and Smith, 1999. Hurst, L. and Smith, N. 1999. Do essential genes evolve slowly? Curr. Biol., 9, 747-750.

Iakoucheva et al., 2004. Iakoucheva, L.M., Radivojac, P., Brown, C.J., O'Connor, T.R., Sikes, J.G., Obradovic, Z., Dunker, A.K. (2004) The importance of intrinsic disorder for protein phosphorylation. Nucl. Acids Res., 32, 1037-1049.

Janin et al., 2008.

Ito et al., 2001. Ito, T., Chiba. T., Ozawa, R., Yoshida, M., Hattori, M., Sakaki, Y. (2001) A comprehensive two-hybrid analysis to explore the yeast protein interactome. Proc. Natl. Acad. Sci. USA, 98, 4569-4574.

Jancura et al., 2011. Jancura, P., Marchiori, E. (2011) A survey on evolutionary analysis in PPI networks. Protein Interaction / Book 2. InTech.

Janin, J., Bahadur, R.P., Chakrabarti, P. (2008) Protein-protein interaction and quaternary structure. Q. Rev. Biophys., 41, 133-180.

Jeong et al., 2001. Jeong, H.,Mason, S.P., Barabasi, A.L., Oltvai, Z.N. (2001) Lethality and centrality in protein networks. Nature, 411, 41-42.

Jin et al., 2001. Jin, G., Zhang, S., Zhang, X., Chen, L. (2001) Hubs with Network Motifs Organize Modularity Dynamically in the Protein-Protein Interaction Network of Yeast. PLoS ONE 2, 11, e1207. .

Jonsson and Bates, 2006. Jonsson P.F., Bates, P.A. (2006) Global topological features of cancer proteins in the human interactome. Bioinformatics, 22, 18, 2291-2297.

Jordan et al., 2003. Jordan, I.K., Wolf, Y.I., Koonin, E.V. (2003) No simple dependence between protein evolution rate and the number of protein-protein interactions: only the most prolific interactors tend to evolve slowly. BMC Evolutionary Biology, 3, 1.

Joy et al., 2005. Joy, M.P., Brock, A., Ingber, D.E. and Huang, S. (2005) High-betweenness proteins in the yeast protein interaction network. J. Biomed. Biotechnol., 2, 96-103.

Kafri et al., 2008. Kafri, R., Dahan, O., Levy, J., Pilpe, Y. (2008) Preferential protection of protein interaction network hubs in yeast: Evolved functionality of genetic redundancy. PNAS, 105, 4, 1243-1248.

Keskin et al., 2007. Keskin, O., Gursoy, A., Ma, B., Nussinov, R. (2007) Towards drugs targeting multiple proteins in a systems biology approach. Curr Top Med Chem., 7, 10, 943-51.

Kiel et al., 2008. Kiel, C., Beltrao, P., Serrano, L. (2008) Analyzing protein interaction networks using structural information. Annu. Rev. Biochem., 77, 415-441.

Kim et al., 2006. Kim, W.K., Henschel, A., Winter, C., Schroeder, M. (2006) The many faces of protein-protein interactions: a compendium of interface geometry. PLOS Comp. Biol., 2, 9, e124.

Kim et al., 2006. Kim, P.M., Lu, L.J., Xia, Y., Gerstein, M.B. (2006) Relating three-dimensional structures to protein networks provides evolutionary insights. Science, 314, 1938-1941.

Kim et al., 2008. Kim, P.M., Sboner, A., Xia, Y. Gerstein, M. (2008) The role of disorder in interaction networks: a structural analysis. Molecular Systems Biology, 4, 179.

Komurov, 2007. Komurov, K., White, M. (2007) Revealing static and dynamic modular architecture of the eukaryotic protein interaction network. Molecular System Biology, 3, 110.

Koonin, 2000. Koonin, E. (2000) How Many Genes Can Make a Cell: The Minimal-Gene-Set Concept. Annu. Rev. Genomics Hum. Genet. 2000. 01:99 116.

Koonin, 2005. Koonin, E. (2005) Orthologs, paralogs and evolutionary genomics. Annu. Rev. Genet, 39, 309-38.

Koschtzki, 2005. Koschtzki, D., Lehmann, K.A., Peeters, L., Richter, S., Tenfelde-Podehl, D., Zlotowski, O. (2005) Centrality Indices, in Brandes and Erlebach Network Analysis: Methodological Foundations, 3418 LNCSl, Springer.

Krogan et al., 2006. Krogan, N.J., et al (2006) Global landscape of protein complexes in the yeast Saccharomyces cerevisiae. Nature, 440, 637-643.

Kunin et al, 2004. Kunin, V., Pereira-Leal, J.B., Ouzounis, C.A. (2004) Functional evolution of the yeast protein interaction network. Mol Biol Evol., 21, 1171-1176.

Lemos et al., 2005. Lemos, B., Bettencourt, B. R., Meiklejohn, C. D., Hartl, D. L. (2005) Evolution of proteins and gene expression levels are coupled in drosophila and are independently associated with mrna abundance, protein length, and number of protein-protein interactions, Molecular Biology and Evolution, 22, 5, 1345-1354.

Li et al., 1999. Li, X., Romero, P., Ran,i M., Dunker, A.K., Obradovic, Z. (1999) Predicting protein disorder for N-, C-, and internal regions. Genome Inform. Ser. Workshop, 10, 3040.

Li et al., 2012. Li, M., Zhang, H., Wang, J., Pan, Y. (2012) A new essential protein discovery method based on the integration of protein-protein interaction and gene expression data. BMC Systems Biology, 6, 15. .

Manna et al., 2009. Manna, B., Bhattacharya, T., Kahali, B., Ghosh, T. C. (2009) Evolutionary constraints on hub and non-hub proteins in human protein interaction network: Insight from protein connectivity and intrinsic disorder. Gene, 434, 1, 50-55.

Maslov et al., 2002. Maslov, S., Sneppen, K. (2002) Specificity and stability in topology of protein networks. Science. 296, 910-913.

Maslov et al., 2004. Maslov, S., Sneppen, K. (2004) Protein interaction networks beyond artifacts. FEBS Letters. 530, 255-256.

Mittag et al., 2010. Mittag, T., Kay, L.E., Forman-Kay, J.D. (2010) Protein dynamics and conformational disorder in molecular recognition. J. Mol. Recognit., 23, 2, 105-16.

Miyamoto-Sato et al., 2010. Miyamoto-Sato, E., Fujimori, S., Ishizaka, M., Hirai, N., Masuoka, K., et al. (2010) A comprehensive resource of interacting protein regions for refining human transcription factor networks. PLoS ONE 5(2): e9289.

Ning et al., 2010. Ning, K., Ng, H.K., Srihari, S., Leong, H.W., Nesvizhskii, A.I. (2010) Examination of the relationship between essential genes in PPI network and hub proteins in reverse nearest neighbor topology. BMC Bioinformatics, 11-505.

Nooren et al., 2003. Nooren, I. M.A. , Thornton, J. M. (2003) Diversity of protein-protein interactions EMBO J., 15, 22, 14, 34863492.

O'Brien et al., 2005. O'Brien, K.P., Remm, M., Sonnhammer, E.L. (2005) Inparanoid: a comprehensive database of eukaryotic orthologs. Nucleic Acids Res., 33, D476-D480.

Oldfield et al., 2008. Oldfield, C.J., Meng, J., Yang, J., Yang, M.Q., Uversky, V.N., Dunker, A.K. (2008) BMC Genomics, 9 (Suppl. 1), S1.

Pagel et al., 2005. Pagel, P., Kovac, S., Oesterheld, M., Brauner, B., Dunger-Kaltenbach, I., Frishman, G., Montrone, C., Mark, P., Stmpflen, V., Mewes, HW., Ruepp, A., Frishman, D. (2005) The MIPS mammalian protein-protein interaction database. Bioinformatics, 21, 6, 832-834.

Pang et al., 2010. Pang, K., Cheng, C., Xuan, Z., Sheng, H., Ma, X. (2010) Understanding protein evolutionary rate by integrating gene co-expression with protein interactions. BMC Systems Biology, 4, 179 .

Park et al., 2009. Park, K., Kim, D. (2009) Localized network centrality and essentiality in the yeast-protein interaction network. Proteomics, 9, 22, 5143-5144.

Patil et al., 2010. Patil, A., Kinoshita, K., Nakamura, H. (2010) Domain distribution and intrinsic disorder in hubs in the human protein-protein interaction network. Protein Science, 19, 8, 1461. doi:10.1002/pro.425

Patil et al., 2011. Patil, A., Kinoshita, K., Nakamura, H. (2010) Hub promiscuity in protein-protein interaction networks. Int. J. Mol. Sci., 11, 1930-1943

Patil et al., 2006. Patil, A., Nakamura, H. (2006) Disordered domains and high surface charge confer hubs with the ability to interact with multiple proteins in interaction networks. FEBS Lett., 580, 2041-2045.

Perkins et al., 2010. Perkins, J.R., Diboun,I., Dessailly, B.H., Lees, J.G., Orengo, C. (2010) Transient protein-protein interactions: structural, functional, and network properties. Structure, 18, 13.

Plotkin and Fraser., 2007. Plotkin, J.B., Fraser, H.B. (2007) Assessing the determinants of evolutionary rates in the presence of noise. Mol. Biol. Evol., 24, 1113-1121.

Przytycka et al., 2010. Przytycka, T.M., Singh, M., Slonim, D.K (2010) Toward the dynamic interactome: it's about time. Briefings in Bioinformatics, 11, 1, 15. doi:10.1093/bib/bbp057.

Punta et al., 2012. Punta, M., Coggill, P.C., Eberhardt, R.Y., Mistry, J., Tate, J., Boursnell, C., Pang, N., Forslund, K., Ceric, G., Clements, J., Heger, A., Holm, L., Sonnhammer, E.L.L., Eddy, S.R., Bateman,A., Finn, R.D. (2012) The Pfam protein families database. Nucleic Acids Research, Database Issue 40:D290-D301.

Radivojac et al., 2006. Radivojac, P., Vucetic, S., O'Connor, T.R, Uversky, V.N., Obradovic, Z., Dunker, A.K. (2006) Calmodulin signaling: analysis and prediction of a disorder-dependent molecular recognition. Proteins, 63, 398-410.

Romero et al., 2001. Romero, P., Obradovic, Z., Li, X., Garner, E.C., Brown, C.J. (2001) Sequence complexity of disordered protein. Proteins, 42, 38-48.

Saeed et al., 2006. Saeed, R., Deane, C.M. (2006) Protein-protein interactions, evolutionary rate, abundance and age. BMC Bioinformatics, 7, 128.

Schaefer et al., 2012. Schaefer, M.H., Fontaine, J.F., Vinayagam,A., Porras, P., Wanker, E.E., Andrade-Navarro. M.A. (2012) HIPPIE: integrating protein interaction networks with experiment based quality scores. PLoS One. 7, 2, :e31826.

Schuster-Bockler et al., 2007. Schuster-Bockler, B., Bateman, A., (2007) Reuse of structural domain-domain interactions in protein networks. BMC Bioinformatics, 8, 259.

Seidman, 1983. Seidman, S. (1983) Network structure and minimum degree, Social Network, 5, 269-287.

Shimizu et al., 2009. Shimizu, K., Toh, H. (2009) Interaction between intrinsically disordered proteins frequently occurs in a human protein-protein interaction network. J Mol Biol., 392, 5, 1253-65. .

Singh et al., 2007. Singh, G.P., Ganapathi, M., Dash, D. (2007) Role of intrinsic disorder in transient interactions of hub proteins. Proteins, 66, 4, 761-5.

Start et al., 2006. Stark, C., Breitkreutz, B.J., Reguly, T., Boucher, L., Breitkreutz,A., Mike Tyers, M. (2006) BioGRID: a general repository for interaction datasets. Nucleic Acids Res. 34(Database issue): D535D539.

Sugase et al., 2006. Sugase, K., Dyson, H. J.,Wright, P. E. (2007) Mechanism of coupled folding and binding of an intrinsically disordered proteis. Nature, 447, 1021-1025.

Tatusov et al., 2003. Tatusov R, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B, Koonin EV, Krylov DM, Mazumder R, Mekhedov SL, Nikolskaya AN, Rao BS, et al.: (2003) The COG database: an updated version includes eukaryotes. BMC Bioinformatics, 4, 41.

Taylor et al., 2009. Taylor, I.W., Linding, R., Warde-Farley, D., Liu, Y. Pesquita C, et al. (2009) Dynamic modularity in protein interaction networks predicts breast cancer outcome. Nature Biotechnology 27, 199-204.

Tompa, 2002. Tompa, P. (2002) Intrinsically unstructured proteins. Trends Biochem. Sci., 27, 527-533.

Tzakos et al., 2012. Tzakos, A.G. (2012) Intrinsic protein disorder as a drug target in oncology: designing drugs targeting plasticity. Biochem & Pharmacol 1:e107.

Tsai et al., 2009. Tsai, C.J., Ma, B., Nussinov, R. (2009) Protein-protein interaction networks: how can a hub protein bind so many different partners? Trends in Biochemical Sciences, 34, 12, pp. 594-600.

Tun et al., 2008. Tun, K., Rao, R.K., Samavedham, L., Tanaka, H., Dhar, P.K. (2008) Rich can get poor: conversion of hub to non-hub proteins. Syst Synth Biol. 2,, 34, 7582.

Tuncbag et al., 2009. Tuncbag, N., Kar, G., Gursoy, A., Keskin, O, Nussinov, R. (2009) Toward inferring time dimensionality in protein-protein networks by integrating structures: the p53 example. Mol Biosyst., 5. 12, 1770-1775.

Uetz et al., 2000. Uetz, P., Giot, L., Cagney, T.A., Mansfield, G. (2000) A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. Nature, 403, 623-627.

Uversky and Dunker, 2010. Uversky, V. N., Dunker, A. K. (2010) Understanding protein non-folding. Biochim. Biophys. Acta, 1804, 1231-1264.

Uversky et al., 2000. Uversky, V. N., Gillespie, J. R., Fink, A. L. (2000) Why are natively unfolded proteins unstructured under physiologic conditions? Proteins, 41, 415-427

Valente et al., 2009. Valente, A.X.C.N., Roberts, S.B., Buck, G.A., Gao, Y. (2009) Functional organization of the yeast proteome by a yeast interactome map. PNAS, 106, 5, 1490-1495.

Vallabhajosyula et al., 2009. Vallabhajosyula, R.R., Chakravarti, D., Lutfeali, S., Ray, A., Raval, A. (2009) Identifying Hubs in Protein Interaction Networks. PLoS ONE, 4, 5344.

von-Mering et al., 2002. von-Mering, C., Krause, R., Snel, B., Cornell, M., Oliver, S.G., Fields, S.., Bork, P. (2002) Comparative assessment of large-scale data sets of protein-protein interactions. Nature, 417, 399-403

Xia et al., 2008. Xia, K., Fu, Z., Hou,L., J. Han, J-D J. (2008) Impacts of protein-protein interaction domains on organism and network complexity. Genome Res. 2008. 18: 1500-1508.

Yu et al., 2004. Yu, H., Greenbaum, D., Xin Lu, H., Zhu, X., Gerstein, M. (2004) Genomic analysis of essentiality within protein networks. Trends Genet., 20, 6, 227-231.

Yu et al., 2007. Yu, H., Kim, P.M., Sprecher, E., Trifonov, V., Gerstein, M. (2007) The importance of bottlenecks in protein networks: correlation with gene essentiality and expression dynamics. PLoS Comput. Biol., 3, 4, e59. doi:10.1371/journal.pcbi.0030059.

Yu et al., 2008. Yu, H., Braun, P., Yildirim, M.A., Lemmens, I., Venkatesan, K., et al. (2008) High-quality binary protein interaction map of the yeast interactome network. Science 322, 104-110.

Yura et al., 2009. Yura, K., Hayward, S. (2009) The interwinding nature of protein-protein interfaces and its implication for protein complex formation. Bioinformatics, 25, 3108-3113.

Walker et al., 1997. Walker, D.R., Koonin, E.V. (1997) SEALS: a system for easy analysis of lots of sequences. Proc Int Conf. Intell. Syst. Mol. Biol., 5, 333-339.

Wall et al., 2003. Wall, D.P., Fraser, H.B., Hirsh, A.E. (2003) An improved method for detecting putative orthologs. Bioinformatics, 19, 13, 1710-1.

Ward et al., 2004. Ward, J.J., Sodhi, J.S., McGuffin, L.J., Buxton, B.F., Jones, D.T. (2004) Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. J. Mol. Biol., 337, 635-645.

Wilkins and Kummerfeld., 2008. Wilkins, M.R., Kummerfeld, S.K. (2008) Sticking together? Falling apart? Exploring the dynamics of the interactome. Trends in Biochemical Sciences, 33, 195-200.

Wright and Dyson, 1999. Wright, P. E., Dyson, H. J. J. (1999) Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. Mol. Biol., 293, 321-331.

Wuchty, 2002. Wuchty, S. (2002) Interaction and domain networks of yeast. Proteomics, 2, 1715-1723.

Wuchty, 2004. Wuchty, S., (2004) Evolution and Topology in the Yeast Protein Interaction Network. Genome Res., 14, 1310-1314.

Wuchty and Almaas, 2005. Wuchty, S., Almaas, E. (2005) Peeling the yeast protein network. Proteomics , 5, 444-449.

Wuchty and Stadler, 2003. Wuchty, S., Stadler, P.F. (2003) Centers of complex networks. J. Theor. Biol., 223, 1, 45-53.

Wuchty et al., 2003. Wuchty, S., Oltvai, Z.N., Barabasi, A.L. (2003) Evolutionary conservation of motif constituents in the yeast protein interaction network. Nat. Genet., 35, 176-179.

Xenarios et al, 2003. Xenarios, I., Salwnski, L., Duan X.J., Higney, P., Kim, S.M., Eisenberg, D. (2002) DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions. Nucl. Acids Res., 30, 1, 303-305.

Zang and Lin, 2009. Zhang, R., Lin, Y. (2009) DEG 5.0, a database of essential genes in both prokaryotes and eukaryotes. Nucl. Acids Res., 37 (suppl 1): D455-D458. doi: 10.1093/nar/gkn858.

Zotenko et al., 2008. Zotenko, E., Mestre, J., OLeary, D.P., Przytycka, T.M. (2008) Why Do Hubs in the Yeast Protein Interaction Network Tend To Be Essential: Reexamining the Connection between the Network Topology and Essentiality. PLoS Comput. Biol., 4, 8, e1000140. doi:10.1371/journal.pcbi.1000140

HPRD. Human Protein Reference Database, http://www.hprd.org/

GO. Gene Ontology database, http://www.geneontology.org/

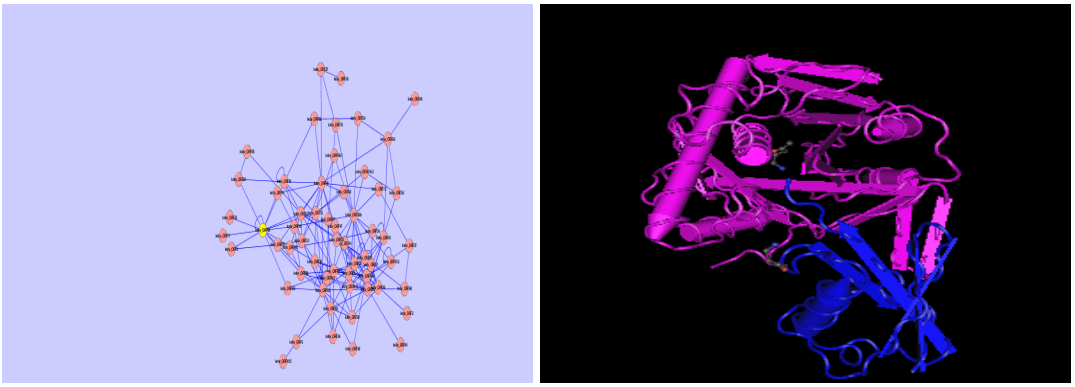

Figure 1: **The PPI network of the Kaposi herpes virus (left); the structure of the yellow protein of the network in complex with one of its interacting partners (pdb entry: 2j7q) (right).**