Protein-Protein Interaction Networks

William Andreopoulos
Center for Computational Biology and Bioinformatics
Department of Electrical Engineering
Columbia University, New York, USA

Dirk Labudde
Bioinformatics group Mittweida (bigM)
University of Applied Sciences Mittweida, Germany

1 Introduction

Biological networks are now the starting point of many studies for understanding and curing human diseases. A protein–protein interaction (PPI) involves two or more proteins binding together, often to carry out their biological function. Many of the most important molecular processes in the cell, such as DNA replication, are carried out by molecular machines that are built from a large number of protein components organised by their PPIs. A set of PPIs occurring at the same time and location of the cell is known as a complex of proteins. A protein-protein interaction network (PPIN) is a collection of PPIs, often deposited in online databases. PPINs may complement other datasets, such as protein structural information, which may lead to understanding the different subparts that contribute to the function of a whole biological system (Bapat *et al*., 2010). Known molecular pathways can be embedded in PPINs to derive new knowledge, such as on spatio-temporal dynamics (Schmid & McMahon, 2007; Srinivasan *et al*., 2007). Understanding PPINs allows one to target them with drugs (Pujol *et al*., 2009). PPINs are also playing a role in the emerging field of synthetic biology that promises to create biological systems with new properties, and modules of networks are available (http://parts.mit.edu) (Russell & Aloy, 2008; Aloy, 2007). A major issue in using PPINs in practice involves dealing with errors in the form of missing interactions and false signals.

2 Experimental Methods for Extracting PPINs and Applications

The yeast two-hybrid (Y2H) system and affinity purification followed by mass spectrometry (AP-MS) are two commonly used experimental methods for detecting interacting proteins. The Y2H system identifies direct interactions between pairs of proteins that physically interact with each other, represented as binary relationships. Large-scale experiments involve making yeast colony arrays; each yeast colony expresses a defined pair of bait and prey proteins that are scored for activity in an automated manner, where high activity indicates interaction. AP-MS detects presence of a protein in a complex, but may not identify the direct interactions between proteins within a complex (Yu *et al.*, 2008). In an AP-MS experiment, a tagged protein is expressed in yeast and then pulled from a cell extract, along with any proteins associated with it. The pull-down is done by co-immunoprecipitation or by tandem affinity purification and the set of pulled-down proteins is identified by MS. Similar to Y2H, the tagged protein in AP-MS may be called the bait and the pulled-down proteins the prey. Figure 1 compares the information on protein complexes given by Y2H and AP-MS experiments (Gentleman & Huber, 2007). Other methods exist for detecting both binary interactions and presence in the same complex, but an issue with all technologies is how to routinely scale them up for high-throughput assays (Tarassov *et al.*, 2008; Venkatesan *et al.*, 2009; Cusick *et al.*, 2009; Gentleman & Huber, 2007).

These procedures have been systematically applied to large sets of yeast proteins. Tong *et al.* combined Y2H experiments to generate two networks. Then they did a computational analysis to identify highly likely PPIs common to both networks. They predicted yeast PPIs mediated by a specific domain, and the interactions were validated in vivo (Tong *et al.*, 2002). Giot *et al.* (2003) produced a PPIN of the fly D. melanogaster. First they produced a draft map of 7048 proteins and 20,405 interactions. A computational method of rating two-hybrid interaction confidence was developed to refine the draft to a higher confidence map of 4679 proteins and 4780 interactions. More recently, the groups of Krogan *et al.* and Gavin *et al.* both derived high-throughput PPI networks on yeast (Krogan *et al.*, 2006; Gavin *et al.*,

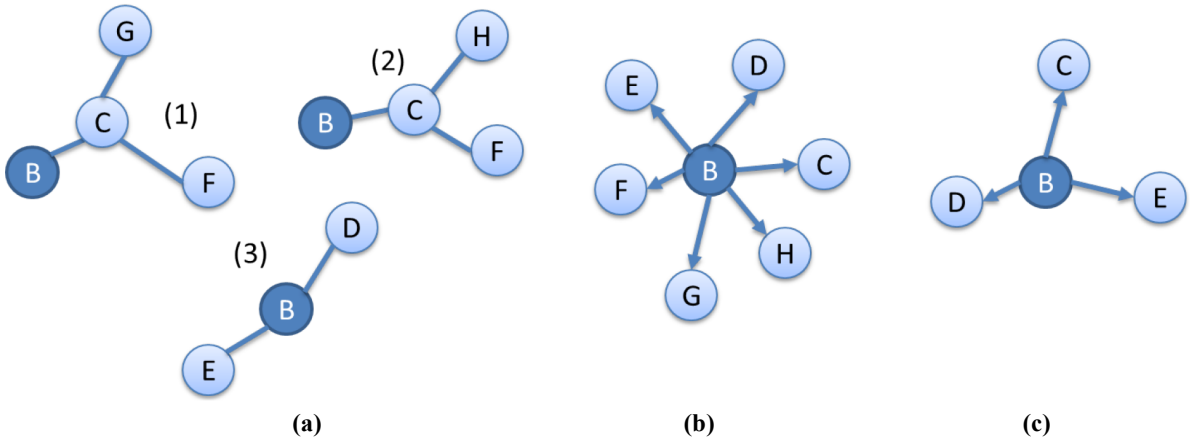


Figure 1: The representation of protein complexes in Y2H and AP-MS data. (a) The actual involvement of protein B in three different multiprotein complexes, numbered 1, 2, 3. In each complex B interacts with different proteins. (b) Protein B is used as a bait in a hypothetical AP-MS experiment. If there are no false positives and false negatives, B will appear to have a connection to all proteins, but the individual complexes will be indistinguishable. (c) Protein B is used as a bait in a hypothetical Y2H experiment with a genome-wide set of prey proteins. If there are no false positives and false negatives, B will be pulled out with its direct interaction partners only (adapted from Gentleman and Huber 2007). (Gentleman & Huber, 2007)

2006), using AP-MS and Y2H respectively. There was low overlap between the two networks. Simonis *et al.* presented a C. elegans PPIN, by testing a matrix of proteins, with a dimension of ~10,000x~10,000, using a high throughput Y2H system. The worm (C. elegans) has ~10,000 open reading frames (ORFs). This interactome consists of 1,816 PPIs between 1,496 proteins. By integrating with previous worm PPIs they estimated the size of the worm interactome at ~116,000 PPIs (Simonis *et al*., 2009).

A major problem in dealing with PPINs is the high noise rate in high-throughput experiments. False positives in a network are erroneous interactions, while false negatives are missing interactions, which may limit the reliability of a PPI network. To evaluate the correctness and reliability of high throughput PPI networks, we used known complexes from the MIPS database. We compared the overlaps of two high-throughput PPI networks mentioned above, by Gavin *et al.* and Krogan *et al.*, with the MIPS protein complexes dataset (Mewes *et al.*, 2006). Table 1 shows full results for the overlaps of Gavin and Krogan networks to the MIPS complexes. For protein pairs that appear in both PPINs and MIPS complexes, we evaluated the number of overlapping edges PPIN ∩ *complexes*. We found *Gavin*06 ∩ MIPS has just 305 overlapping edges, *Krogan*06 ∩ MIPS has 359 overlapping edges.

Literature curation strategies analyze thousands of small-scale experiments from the large online biomedical literature (e.g., PubMed) to identify individual PPIs. The analysis is usually done automatically, though manual curation may also contribute. The individual literature-curated PPIs are then collected into a large PPI network. Such literature-curated PPINs generally have fewer errors than PPINs from high-throughput screens with a high error rate. Literature-curated PPINs that represent an accumulation of many small-scale studies may help to correct errors in high-throughput screens, which aim to make discoveries without prior knowledge. While literature-curated PPINs are complementary to high-

Network	Edge overlap with MIPS $\lvert E_{network} \cap E_{MIPS} \rvert$	Edges in network but not in MIPS $\lvert E_{network} - E_{MIPS} \rvert$
Gavin	305	3989
Krogan	359	2225

Table 1: Overlap of high-throughput PPI networks (Gavin and Krogan) with known complexes from the MIPS yeast database. Only those edges were considered where both proteins were present in the PPI network and in MIPS. Symbols denote: E, edges; |·|, set cardinality; ∩, intersection; -, set difference.

throughput screens, they are an unlikely source of new biological knowledge since they are based on what was investigated before (Venkatesan *et al.*, 2009; Cusick *et al.*, 2009).

The potential of PPI networks in medical applications was demonstrated by finding subnetwork markers in a human PPI network, which provide risk prediction of metastasis and classify breast cancer into subtypes (Soon *et al.*, 2010). Changes in the organization of the human PPI network was shown to be a useful indicator for breast cancer prognosis and predicting patient outcome (Wu *et al.*, 2010; Taylor *et al.*, 2009). A PPIN extended with additional datatypes, such as gene-gene coexpression networks, was applied to two glioblastoma (GBM) datasets and candidate oncogenes were projected onto this network. The majorities of GBM candidate oncogenes formed a cluster and were closer than expected by chance. Similar patterns including subnetwork markers for metastasis were found in breast, colorectal and pancreatic cancers (Pujol *et al.*, 2009; Wu *et al.*, 2010; Cerami *et al.,* 2010; van den Akker *et al.*, 2011). All of these medical applications assume reliable PPI networks with a low error rate. Dealing with noise in high-throughput experiments remains a major challenge.

3 What Causes Proteins to Interact?

Often a protein's physical interaction with its partner is mediated by a modular part of the protein's surface, called a structural domain (Jothi *et al.,* 2006). Unraveling molecular details of PPIs requires understanding how proteins interact at the structural level (Pawson & Nash, 2003). There are cases where a domain from one protein interacts (through physical contacts or chemical bonds) with two or more domains from another protein, which almost always co-exist (Hesselberth *et al.*, 2006). Sometimes, the co-occurring domains are fused together to form a single domain in a reference organism. A pair of interacting domains (or proteins) is more likely to share similar functional annotations from the online literature than any random domain pair (Kamburov *et al.*, 2007).

Although relatively simple in isolation, PPIs can be regulated in a dynamic fashion via domain interactions, providing a measuring device to monitor cellular activity. In the PPI pathways that control cellular behavior, for example, catalytic domains frequently induce post-translational protein modifications (such as phosphorylation) that are then recognized by interaction domains (Pawson & Scott, 1997; Pawson, 2003).

Though many PPIs are involved in specific functions and cellular locations, a range of weak interactions are relatively unspecific and often without a meaning. As long as unspecific interactions procure no disadvantage to the organism, they could be tolerated through evolution. Morrison et al. proposed a model where PPIs are explained by an underlying interaction between complementary structural do-

mains; they called this the lock-and-key model (Alberts *et al.*, 2002; Lodish *et al.*, 2000; Morrison *et al.*, 2006).

4 The Importance of Hubs in PPINs

In a PPI network, proteins are represented as nodes. Some nodes interact with many more partners than average; these proteins are called hubs (Albert, 2005). The loss of hubs may cause the breakdown of the PPIN into isolated clusters (Barabasi & Oltvai, 2004). A consequence of highly-connected hub proteins is a PPIN's robustness to random errors, coupled with a high fragility against the removal of the most connected hub proteins. A famous example of a hub protein is the tumor-suppressor p53, which is mutated and inactivated in a high percentage of human tumor types (Jeong *et al,* 2001). The likelihood that a gene is essential (lethal) correlates with the number of PPIs its protein product has.

Since proteins can interact only if they are at the same location at the same time, protein localization plays an important role in PPINs. Vidal *et al*. proposed a well-known model where hubs are distinguished into party hubs and date hubs. While party hubs bind all of their interaction partners simultaneously, date hubs bind them at different times or locations (Han *et al.*, 2004). While party hubs function inside modules, date hubs have an organizational role, connecting processes or modules to one another. The distinction of hubs into party hubs and date hubs has been questioned in the literature. The original proponents confirmed these global properties on literature-curated yeast PPINs (Bertin *et al.,* 2007). However, the opposing side argued that the PPINs used are too small and incomplete (Batada *et al.*, 2006; Batada *et al.*, 2007).

Ekman *et al.* (2006) and Singh *et al.* (2007) tried to find the domain structural basis for the distinction between party and date hubs. They found that date hubs are enriched with long disordered regions, which are important for flexible binding. There is an over-representation of multi-domain proteins among the hubs. Hub proteins, especially party hubs, appear to be more conserved in ancient species.

Over the past 15 years, examples of protein structures have emerged showing that proteins exist in an intrinsically disordered state and function without a stable folded structure; in many cases, lack of structure is actually required for biological function. Haynes *et al*. (2006) argued that intrinsic structural disorder is a common characteristic of hub proteins, since they are more disordered than proteins with one interaction partner. They propose that intrinsic disorder may serve as a determinant of protein interactions for hub proteins in human, yeast, fly and worm PPINs.

Uversky *et al.* proposed three general ways that intrinsically disordered proteins contribute to the high level of hub connectivity. First, intrinsic disorder can serve as the structural basis for hub protein promiscuity. Secondly, intrinsically disordered proteins can bind to structured hub proteins. Thirdly, intrinsic disorder can provide flexible linkers between functional domains with the linkers enabling mechanisms that facilitate binding diversity (Dunker *et al.*, 2005).

5 Evolutionary Conservation of PPIs

Several studies have been conducted to examine the extent of conservation between the PPIs in different organisms. Many signaling circuits embedded in PPI networks are conserved over evolution across species. Sharan *et al*. found 71 PPIN regions that are conserved across C. elegans, D. melanogaster and yeast

S. cerevisiae. These conserved PPIN regions can be useful for predicting new interactions or unknown protein functions (Sharan & Ideker *et al.*, 2005; Sharan *et al*., 2005). The highly connected hub proteins tend to be evolutionarily conserved across species, even in the context of noise in the underlying PPINs (Wuchty *et al*., 2006).

As mentioned above, there are at least two types of interacting surfaces in proteins. Domain-domain interactions are more prevalent in stable protein complexes, whereas domain-disorder interactions are more transient. Domain-disorder interactions evolve much faster than domain-domain interactions. Investigations of intrinsic disorder in proteins showed a considerable proportion of poorly conserved domain-disorder interactions, indicating that the proportion of non-conserved PPIs across species is substantial. The proportion of PPIs that are of the domain-disorder type versus the domain-domain type is not known for any species (Tarassov *et al.*, 2008; Venkatesan *et al.*, 2009; Cusick *et al.*, 2009).

6 Reliability and Coverage of PPINs

Owing to the high error rate of experimental methods, low accuracy and falsely detected interactions remain one of the main problems in dealing with PPINs (Rinner *et al.*, 2007). Huang *et al.* estimated that for yeast, worm and fly screens, the overall false discovery rates (FDRs) are 9.9%, 13.2% and 17.0% and the false negative rates (FNRs) are 51%, 42% and 28% (Huang & Bader, 2009). Hart *et al.* made more pessimistic estimates that owing to a high false positive rate, current PPINs are only 50% complete for yeast and 10% complete for human (Hart *et al.*, 2006).

The reported low coverage and high error rates in PPINs have contributed to significant questioning of the experimental methods used for PPI detection. Vidal et al. concluded that both Y2H and AP-MS data are of high quality, but different reasons cause errors in each datatype. Y2H and AP-MS provide complementary information about a PPIN and both are vital to get a complete picture. They calculated the edge-betweenness for each PPI in a merged network of all available interactions, measuring the number of shortest paths between all protein pairs that traverse a given edge. The higher edge betweenness of PPIs from Y2H shows the tendency of Y2H to detect key PPIs, representing connections between complexes and pathways, more often than AP-MS. They developed an empirical mapping framework that produced a high-quality Y2H network covering ~20% of all yeast PPIs (Yu *et al.,* 2008); they applied this framework to map the estimated ~130,000 human PPIs (Venkatesan *et al*, 2009). Additionally, Vidal et al. concluded that PPIs manually curated from the literature are error prone, as indicated by an extremely low overlap of different curation databases. Occasionally, curator error is responsible for a low reliability of literature curation. However, errors are also due to gene name confusion and the difficulty of extracting accurate information from a long free-text document (Cusick *et al.*, 2009).

PPINs from different high-throughput experiments have low overlap. Gentleman *et al.* argued that the low overlap of datasets is due to low coverage of different methods rather than false positives. The real issues in dealing with low coverage of PPINs involve comparing the methods, detecting noise, interpreting and integrating the data (Gentleman & Huber 2007). Hoffmann and Valencia investigated PPINs resulting from different methods. They argued that while a pairwise comparison of interactions does not reveal similarities between different methods, comparing the connectivities of individual proteins reveals a common tendency between methods manifested as global properties of the PPINs (Hoffmann & Valencia, 2003).

Two approaches for interpreting the results of bait-prey studies are the spoke model vs. the matrix model. The spoke model connects only the bait protein with associated hit proteins, minimizing the false positive PPIs. A matrix model connects all proteins pulled with a bait protein to one another (as a clique) resulting in more false positives, but also more true positive PPIs. Bader and Hogue showed that a spoke model is three times more accurate than a matrix model that connects all proteins (Bader & Hogue, 2002). For a list of PPIN repositories online see the various literature (Cusick *et al.*, 2009; Sanderson, 2009).

7 Statistical Modeling of PPI Networks

A PPI network is typically represented as a graph, where the nodes are proteins and the edges are interactions between proteins in a network. Undirected edges connecting nodes are often used as a model for physical interactions, such that if protein A interacts with B, then B interacts with A. The observed experimental data, however, often display asymmetry: protein A may identify protein B as an interactor when A is used as a prey, but using B as a prey may not detect A. The number of edges connecting a node in an undirected graph is the degree of the node. To represent asymmetric data, one could also use a directed graph model for representing the observed data (Gentleman & Huber, 2007). Several graph statistical models have been used for explaining the connectivities (degree distributions) of PPINs and have made an impact on our understanding of biological networks:

- the Erdos-Renyi model of random graphs,
- scale-free networks following a power law,
- hierarchical modularity (Ravasz *et al.*, 2002; Goldberg *et al.*, 2006).

A statistical model for PPI networks is required for top-down systems biology approaches. Top-down models in systems biology are used for generating hypotheses. An example of top-down systems biology is to analyze a large-scale dataset to find correlations between genes and proteins in an organism's interactome. Top-down approaches can use a general statistical model of the interactome to find correlated molecular behavior in genome-wide studies. However, wet lab experiments cannot be designed and justified based on top-down results alone. Bottom-up systems biology approaches examine the mechanisms of interaction between known components through which functions arise. Bottom-up approaches build detailed models on a particular mechanism that can be simulated computationally. Hypotheses can be integrated into bottom-up models followed by experimental validation in the lab. The experimental data can then be fed back into the top-down approach iteratively to refine the general model used to formulate hypotheses on new molecular mechanisms. This computationally driven experimental biology is relevant to the study of any complex cellular system, such as the development of cancer. However, integrating hypotheses in bottom-up models for verification assumes experimental methods have a low noise rate (Bruggeman & Westerhoff, 2007).

7.1 Random Graphs

A random graph is defined as a fixed number of nodes (proteins), with an edge (interaction) existing between any pair of nodes with independent probability, as Figure 2 shows. In a random graph, the proba-

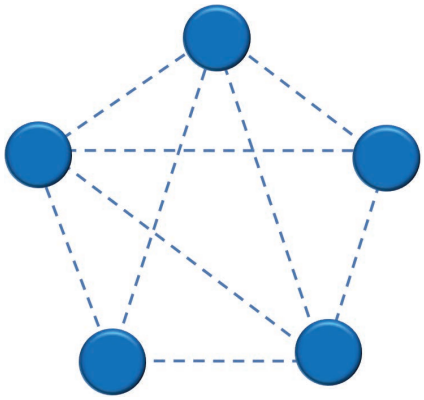


Figure 2: In the Erdos-Renyi model of a random graph, every possible edge occurs with independent probability.

bility that a node has degree k follows the Poisson distribution $f(k \mid \lambda) = e^{-\lambda} \lambda^k / k!$, where λ is the mean degree. The degree distribution is a function, often visualized as a histogram (Goldberg *et al.*, 2007; Chakrabarti & Faloutsos, 2006).

7.2 Scale-free Networks Produced by the Power Law

The degree distribution of a scale-free network follows the power law $P(k) = ck^{-\gamma}$, where $P(k)$ is the fraction of nodes in the network having k connections to other nodes, c is a constant and γ is a constant with a value in the range $2 < \gamma < 3$. In a scale-free network the nodes do not fall into two separate classes corresponding to hubs vs. low-degree nodes, but every degree appears with a frequency given by $P(k)$. This appears as a straight line on a logarithmic plot (Albert, 2005). Low-degree nodes have the highest frequencies and appear most often, while few high-degree nodes (hub proteins) occur (Chakrabarti & Faloutsos, 2006). Figure 3 shows an example of power law distribution.

The origin of the scale-free topology in complex networks can be reduced to two basic mechanisms: growth and preferential attachment. Growth means that the network emerges through the subsequent addition of new nodes. Preferential attachment means that new nodes prefer to link to more connected nodes. Growth and preferential attachment generate hubs through a 'rich-gets-richer' mechanism: the more connected a node is, the more likely it is that new nodes will link to it, which allows the highly connected nodes to acquire new links faster than their less connected nodes. In PPI networks, scale-free topology seems to have its origin in gene duplication. This induces growth in the PPIN because an extra gene that encodes a new protein has the same structure as the original duplicated protein, so they both interact with the same proteins. Ultimately, the proteins that interacted with the original duplicated protein will each gain a new interaction to the new protein. Therefore, proteins with more interactions tend to gain links more often, as it is more likely that they will interact with the duplicated protein (Ravasz *et al.*, 2002; Lima-Mendez & van Helden, 2009).

7.3 Hierarchical Modularity in Networks

Barabasi *et al.* combined the notion of modularity in PPINs with a scale-free network, having a degree distribution following a power law. They proposed a network structure referred to as hierarchical modu-

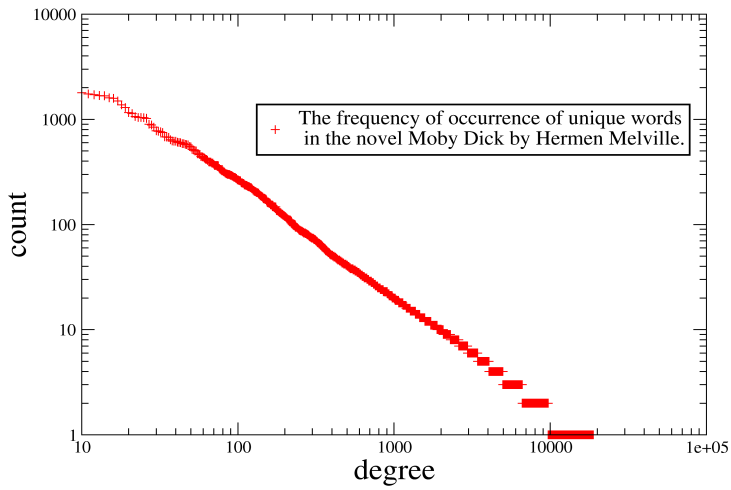


Figure 3: To demonstrate a power law we extracted the unique words in the novel Moby Dick by Herman Melville. We associated with every word a number of edges, representing a word's occurrences in the novel. The plot shows the degree distribution on a log-log scale for the unique words' occurrences in the novel. The continuously decreasing degree distribution shows that low-degree nodes (words occurring a few times) are common (adapted from Clauset *et al*., 2009; Chakrabarti & Faloutsos 2006). In other words, the plot shows that most words occur a few times in the entire novel, but there are a few words that occur repeatedly.

larity. A hierarchical organization, as shown in Figure 4, was proposed as the cause of scale-freeness and a high degree of clustering in networks (Ravasz *et al.*, 2002). They showed that the PPINs of 43 organisms are organized into distinct highly connected modules. The modules combine in a hierarchical manner into larger units, such that their connectivity and degree of clustering follow a power law (a scale-free topology). This hierarchical network architecture has been considered an organizing principle of complex networks (Clauset *et al.*, 2008).

8 Network Motifs and Modules

Network motifs are patterns (sub-graphs) that recur within a network more often than expected by chance. Most networks studied in biology, including PPI networks, seem to be largely composed of a small set of network motifs, which occur repeatedly. Alon *et al*. presented the idea of network motifs after discovering motifs in the gene regulation (transcription) network of the bacteria *E. coli* (Alon, 2007; Shen-Orr *et al.,* 2002). These motifs can be considered as simple building blocks from which the network is composed.

Motifs in PPI networks often result from interactions of proteins in complexes. The abundance of network motifs is partly explained by the enrichment in protein complexes (Albert, 2005; Ma'ayan *et al.*, 2009). The membership of proteins in complexes can be represented with the graph theoretical notion of a bipartite graph or biclique (Andreopoulos *et al.*, 2007; Gentleman & Huber, 2007). Morrison *et al.* pro

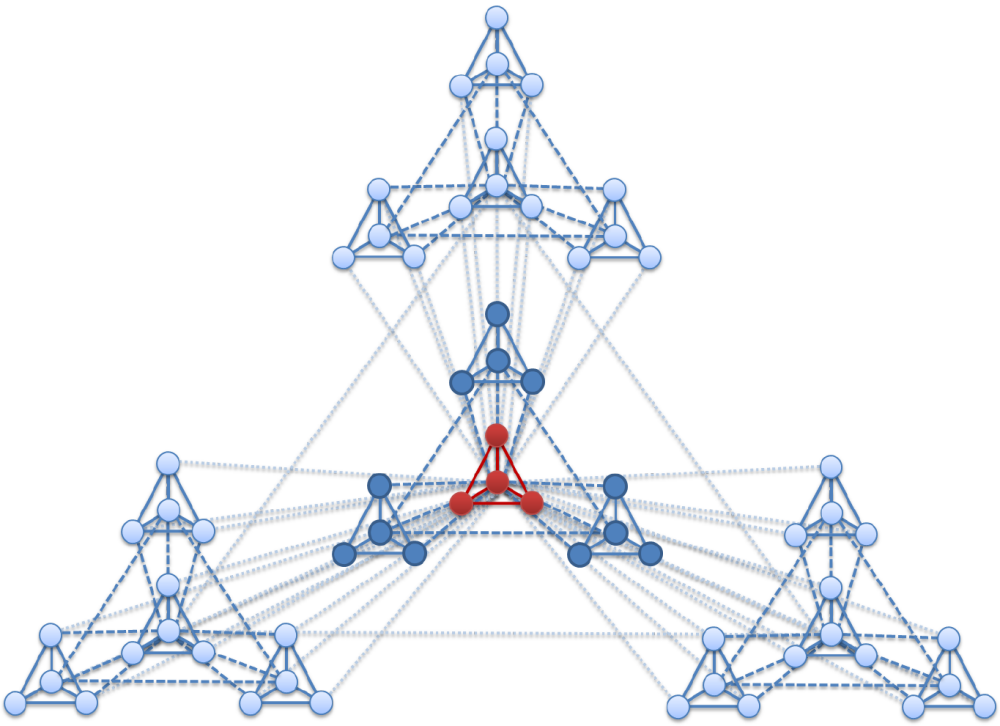
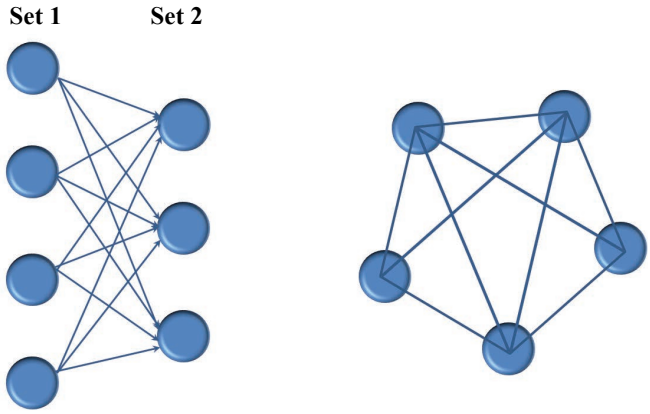


Figure 4: A network with hierarchical modularity. The nodes are organized in groups of small, highly connected modules. For instance, the group of four red nodes shown in the center are connected with solid lines. The modules are then connected in a hierarchical manner into larger units, like the red and blue node groups shown as connected with dashed lines. These are in turn connected to similar node groups via dotted lines. Their connectivity and degree of clustering follow a power law (adapted from Ravasz *et al*. 2002).

posed the lock-and-key model where PPIs are explained by an underlying interaction between complementary structural domains, which leads to modelling complexes as bicliques. They showed their approach could identify bicliques that correspond to known interaction motifs and predict novel biologically relevant motifs (Morrison *et al.*, 2006). Figure 5a shows a biclique, while 5b shows a clique where every node is connected to all other nodes (Chakrabarti & Faloutsos, 2006).

Network motifs are evolutionarily conserved (Wuchty *et al.*, 2006; Giot *et al.*, 2003). Motifs may be produced by convergent evolution of genes, whereby two genes that have similar functions stem from a common-ancestor gene (Alon, 2007). Another cause of motifs is module duplication by evolution. Pereira-Leal et al. observed that at least 20% of complexes in yeast have strong similarity to complexes in other organisms. These complexes may have evolved by duplication retaining the same function as the original complex (Pereira-Leal & Teichmann, 2005).

Triangle motifs are abundant in signal transduction and regulatory networks (Ma'ayan, 2009). Zhang *et al.* integrated multiple interaction types in yeast, including PPIs, genetic interactions, transcriptional regulation, sequence homology, and expression correlation. Then, they examined triangle motifs combining interactions of different types. They proposed that network motifs are signatures of higher-



(a) Bipartite core **(b)** Clique

Figure 5: Indicators of community structure: (a) A 4x3 bipartite core, or biclique, where each node in Set 1 is connected to each node in Set 2. (b) A 5-node clique, where every node is connected to all other nodes.

order network structures that correspond to biological phenomena. For example, a network motif may represent two targets of the same transcription factor bridged by a PPI.

Triangle motifs are abundant in signal transduction and regulatory networks (Ma'ayan, 2009). Zhang *et al.* integrated multiple interaction types in yeast, including PPIs, genetic interactions, transcriptional regulation, sequence homology, and expression correlation. Then, they examined triangle motifs combining interactions of different types. They proposed that network motifs are signatures of higher-order network structures that correspond to biological phenomena. For example, a network motif may represent two targets of the same transcription factor bridged by a PPI.

According to Newman (2006), modularity can be formulated mathematically in PPI networks as the number of edges within groups of proteins minus the number expected in a PPI network of the same size with edges placed at random. Spirin & Mirny (2003) found a high frequency of modules in PPINs, where modules are groups of proteins densely connected internally, but sparsely connected with the rest of the network. This would imply few edges connecting different modules. They found a high frequency of protein complexes manifested as modules, such as splicing machinery, transcription factors, etc. These modules are statistically significant when compared to random graphs and are robust to noise, suggesting that such modules constitute PPIN building blocks. Tamames *et al.* (2007) and Wang & Zhang (2007) argued that PPIN modularity is correlated with the process of reductive evolution, where most of the ancestral genes are lost while other network properties remain unchanged.

9 Finding Protein Complexes

Protein complexes are groups of proteins that interact in the cell at the same time and location. Several approaches used clustering or graph theoretic methods to predict protein complexes in PPI networks by identifying tightly interacting groups of proteins (Andreopoulos *et al.*, 2009; Lubovac *et al.*, 2006; Altaf-Ul-Amin *et al.*, 2006). An early work on identifying protein complexes involved an application of the *k*-

cores algorithm by Bader *et al*. (Bader & Hogue, 2003; Batagelj & Zavernik, 2001). The k-core is computed by pruning all the nodes and their respective edges with degree (number of edges) less than k. That means that if a node u has degree m and it has n neighbors with degree less than k, then u's degree becomes $m - n$ and it will be also pruned if $k > m - n$. As example, consider a cluster of low-degree proteins $\{A, B, C\}$ that is a 2-core or 3-core, but not a 4-core, because A and B have three edges only; k-cores with $k = 4$ cannot find this cluster.

Andreopoulos *et al.* (2007) proposed the MULIC clustering algorithm, which finds bicliques in PPI networks. In the example above, MULIC can find the cluster $\{A, B, C\}$ if all three proteins $\{A, B, C\}$ interact with the same protein partners. MULIC detects such proteins as local interaction partners (mediators) that mediate a module of proteins. Mediator proteins and modules are significantly enriched in gene ontology (GO) annotations, including known functions, cellular processes and locations.

Methods for predicting interactions and complexes in PPI networks may involve finding protein domains believed to interact (Albrecht *et al.*, 2005). Several articles have appeared on predicting PPIs based on their binding sites (Deng *et al.*, 2002; Kim *et al.*, 2002; Sprinzak & Margalit, 2001). These methods generally evaluate a statistical score for the probability of two domains interacting. These scores suggest which protein pairs are most likely to interact; then it is deduced that other protein pairs with these domains are likely to interact. Similarly, Morrison *et al*. (2006) and Li *et al*. (2006) identified bipartite subgraphs in networks, which arise from structural domain–domain interactions.

Methods for finding functional modules in PPINs often use the connectivity of nodes to find dense areas (Chen & Yuan, 2006; Espadaler *et al.*, 2005; Pereira-Leal *et al.*, 2004; Spirin & Mirny, 2003). Some methods predict functional modules based on how many common interaction partners two proteins share (Morrison *et al.*, 2006; Andreopoulos *et al.*, 2007; Andreopoulos *et al.*, 2009; Chua *et al.*, 2006; Okada *et al.*, 2005; Samanta & Liang, 2003). Ding *et al*. (2004) represented PPINs based on an underlying bipartite graph model that allows generating the complex-complex association network. This representation allows viewing the network as consisting of protein complexes that share components. Dunn *et al*. (2005) described separating PPINs into clusters of interconnected proteins, using Girvan and Newman's Edge-Betweenness algorithm (Girvan & Newman, 2002). The detected clusters are enriched in gene ontology (GO) annotations.

Other approaches detect protein complexes on the basis of probabilistic methods. Sharan *et al.* developed a probabilistic model for protein complexes, based on conservation between the yeast S. cerevisiae and the bacteria H. pylori. They used this model for finding conserved complexes by searching for heavy subgraphs in an edge- and node-weighted graph, whose nodes are orthologous protein pairs between two species (Sharan & Ideker *et al.*, 2005; Sharan *et al.*, 2005). Dittrich *et al.* found maximal scoring subnetworks in large PPINs using scalable methods from operations research. They integrated datasets, such as lymphoma microarray data with a large PPIN from the Human Protein Reference Database (HPRD) (Marcus *et al.*, 2008). Liu *et al.* assigned weights to proteins, such that the weight indicates the reliability of the protein-protein interaction. They then proposed a complex prediction algorithm that generates all maximal cliques from the PPIN. This method favors larger clusters, and is robust to random noise as it reduces the impact of unreliable interactions on complex prediction (Wong & Chua, 2009).

10 Network Noise and Finding Errors

Several papers aim to find errors in PPI networks by completing them for missing edges or finding false positives (Yu & Fotouhi, 2006; Valencia & Pazos, 2002; von Mering *et al.*, 2007; Ben-Hur & Noble, 2005; Guo *et al.*, 2008). The approach of Andreopoulos *et al.* (2009) integrates structural information with PPI networks to identify triangle motifs (Andreopoulos *et al.*, 2009). Figure 6 illustrates this approach. PPIs are integrated with complementary datatypes, in particular structural domain-domain interactions (SDDIs), in order to identify triangle motifs representing subnetworks of common functionality and cellular location. The triangles consisting of PPIs and SDDIs at the structural level allow predicting complexes and finding errors in a PPI network. The success of the approach is evaluated by comparing the triangle motifs with known MIPS complexes, resulting in a significant overlap (Mewes *et al.*, 2006).

Several studies collected ensembles of data, such as structural or literature information. Alber *et al*. (2007) collected diverse high-quality data, and analyzed the ensemble to produce a detailed architectural map of a specific protein complex. This work translates the data into spatial restraints, instead of using network motifs. Ramirez *et al.* (2007) assessed the quality and value of publically available human protein network data, by comparing predicted datasets, high-throughput results from yeast two-hybrid screens, and literature-curated protein-protein interactions. This analysis revealed major differences between datasets. Rhodes *et al.* (2005) demonstrated a probabilistic analysis integrating model organism protein interactomes, structural domain data, genome-wide gene expression data and functional annotations that predicted nearly 40,000 human interactions. *Bader et al.* (2004) performed an integrated analysis of proteomics data with data from genetics and gene expression. Huang *et al.* (2004) presented POINT, the “prediction of interactome database”. POINT predicts sets of interacting human proteins by integrating several publicly accessible databases, such as mouse, fruit fly, worm and yeast.

Another large body of work attempts to predict the missing interactions or assign confidences to large noisy interactomes. Some of these use network topology and others use information on structural domain-domain interactions, while others use Bayesian networks or probabilistic measures. Yu *et al.* (2006) described predicting missing PPIs using only the PPI network topology as observed by a high-throughput experiment. The method searches the interactome for defective cliques, nearly complete complexes of pairwise interacting proteins, and predicts the interactions that complete them. Chen *et al.* (2008) proposed using triangles of observed PPIs to predict and validate interactions. Yeast is the only data set large enough to warrant application of this method. Singhal & Resat (2007) presented DomainGA, a computational approach that uses information about structural domain-domain interactions to predict PPIs. This method achieves good prediction for the positive and negative PPIs in yeast. Pitre *et al.* (2006) presented PIPE, a system for predicting PPIs for any target pair of the yeast proteins from their primary structure. Chen *et al.* (2006) introduced a novel measure called IRAP, "interaction reliability by alternative path", for assessing the reliability of PPIs based on the underlying PPI network topology. IRAP measure is effective for discovering reliable PPIs in large noisy PPI networks. Ng *et al.* (2003) proposed an integrative approach that applies structural domain-domain interactions to predict and validate PPIs. Chen & Liu (2005) introduced a random forest of decision trees that is capable of predicting PPIs based on known structural domain-domain interactions. Wu *et al.* (2006) proposed using the similarity between pairs of gene ontology (GO) terms for reconstructing a yeast PPI network based on knowledge of functional associations between the GO annotations.

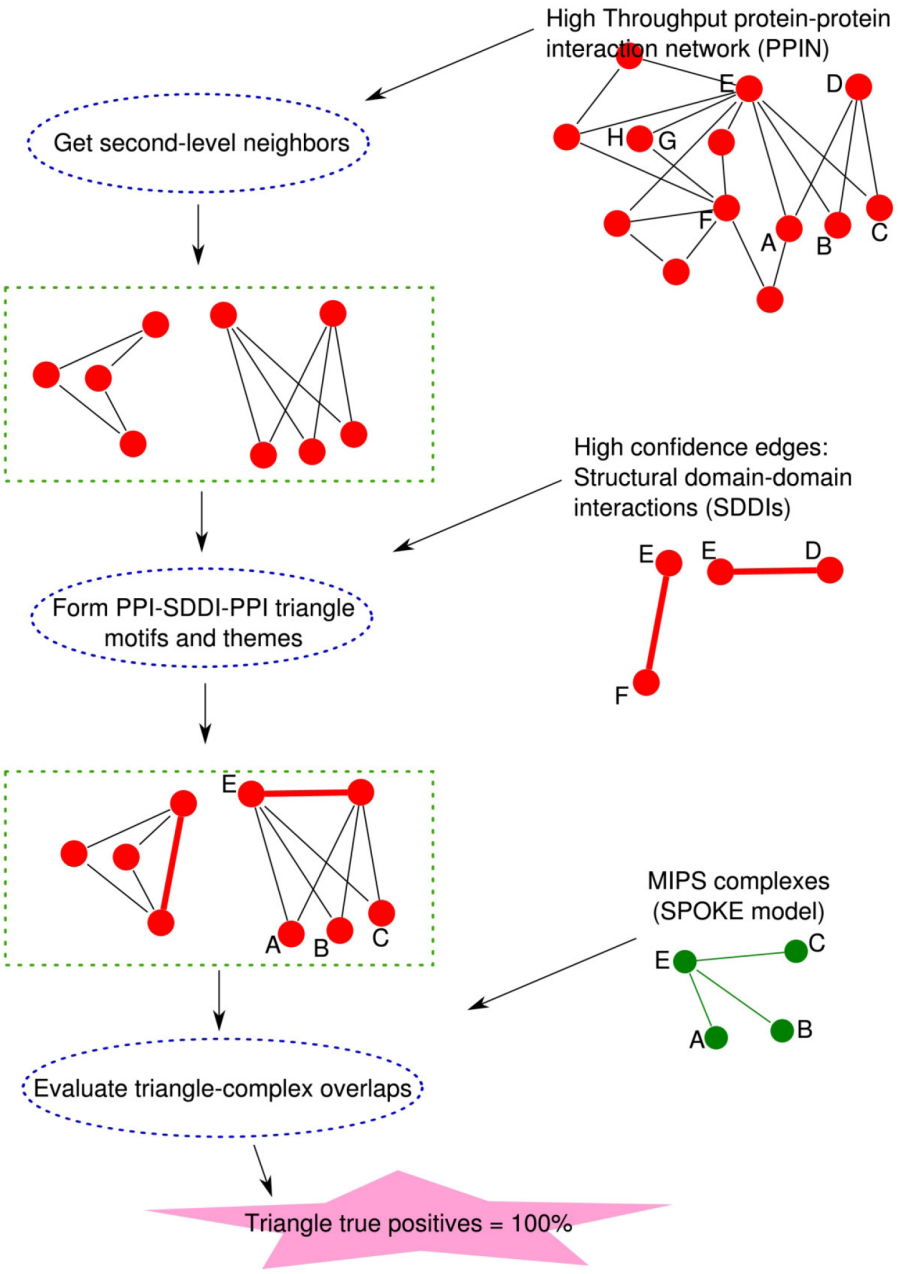


Figure 6: The approach of Andreopoulos *et al*. (2009) denoises networks by building triangles consisting of PPIs and complementary datatypes. It starts by extracting the second-level neighbors from a PPIN. Combining these PPI edges with structural domain-domain interactions (SDDIs, or any other complementary data type) allows building triangle motifs. Then, the triangle motifs are compared with known complexes such as MIPS (Mewes *et al.*, 2006; Andreopoulos *et al.*, 2009).

Jansen *et al.* (2003) developed an approach using Bayesian networks to predict PPIs in yeast. Han *et al.* (2004) proposed PreSPI, a domain combination based PPI prediction approach. PPIs are interpreted as the result of groups of multiple structural domain-domain interactions. This approach also provides an interacting probability for PPIs. Vidal and colleagues used reference sets to calculate the probability that a newly identified PPI is a true biophysical interaction, and assigned confidence scores to all PPIs in interactome networks (Braun, 2009). Yu *et al.* (2009) assigned confidence scores that reflect the reliability of each PPI, by using multiple independent sets of training positives to reduce the bias inherent in using a single training set.

Another body of work has performed large scale analysis of networks, statistical network motif analysis or error estimation. Jin *et al.* (2007) used network motifs to solve the open question about 'party hubs' and 'date hubs' which was raised by previous studies. At the level of network motifs instead of individual proteins, they found two types of hubs, motif party hubs and motif date hubs, whose network motifs display distinct characteristics on biological functions. Zhang *et al.* (2005) observed that different types of networks exhibit different triangle profiles, providing a means for network classification. They extended the network triangle concept to an integrated network of many interaction types. Mathivanan *et al.* (2006) analyzed the major publically available databases that contain literature-curated PPI information for human proteins, finding a large difference in their content. This included public databases such as BIND, DIP, HPRD, IntAct, MINT, MIPS, PDZBase and Reactome (Galperin & Cochrane, 2009). Chiang *et al.* (2007) assessed error statistics in all published large-scale datasets for *S. cerevisiae*.

Collins *et al.* attempted to deal with the noise and false positives in the yeast PPIN derived from high-throughput AP-MS studies (Krogan *et al.*, 2006). They proposed a novel probabilistic metric that takes advantage of the density of high-throughput datasets to provide a measure of the confidence of each PPI. This way, they keep a subset of PPIs that are of higher confidence in BioGRID.

11 Human PPI Networks

A great challenge in the post-genomic era is to construct a complete human PPI network for the more than 20,000 human genes, many of which remain uncharacterized at the moment. Current coverage of the human PPI network is estimated to be around 8-10%, including ~50,000-57,000 binary PPIs (Sanderson, 2009).

Gandhi *et al.* (2006) constructed one of the first integrated human PPI networks by combining HPRD with BIND, DIP, MIPS, MINT and IntAct. Of more than 70,000 human PPIs, only 42 were common to human, worm and fly. Only 16 were common to human, worm, fly and yeast.

Rual *et al.* (2005) presented a high-throughput yeast two hybrid system that screened ~8,100 human open reading frames and detects ~2,800 interactions. This is an initial version of a human binary PPIN. The authors reported that more than 85% of the interactions are not found in PubMed or Google Scholar literature databases, indicating that the interactions are likely to be novel.

Previously noise and low coverage was a great problem in constructing human PPINs. Mathivanan *et al.* (2006) showed that overlap between human PPIs is low despite the presence of the same proteins, and this is true even for databases with similar datatypes, such as literature-derived databases. Moreover, the PPIs that overlap between databases often have dissimilar annotations.

Rhodes *et al.* (2005) proposed a bioinformatics method to predict nearly 40,000 PPIs in human. For this purpose, they integrated model organism PPIs (S. cerevisiae, D. melanogaster, C. elegans) with protein domain data and gene expression data. They validated the predicted PPIs on a test dataset to show a high overlap with known human PPIs.

Especially challenging is the unravelling of PPIs between membrane proteins in humans, as well as extracellular PPIs. Methods such as yeast two-hybrid are not ideal for human PPI networks. A range of two-hybrid methods that can analyze membrane protein and extracellular PPIs have emerged. A robust approach for finding PPIs in membrane proteins is the Membrane Yeast Two Hybrid (MYTH) system, and for extracellular PPIs the AVEXIS system (Sanderson, 2009; Venkatesan *et al.*, 2009; Cusick *et al.*, 2009).

12 Visualization of Biological Networks

We give a brief overview of several visualization tools for biological networks. Cytoscape is a popular open source network visualization tool, which allows the user to see the relationships between nodes and interact with them in a user-friendly manner (Shannon *et al.*, 2003; Merico *et al.* 2009). It supports development of additional plugins for special purposes (Royer *et al.*, 2008). Cytoscape can also be used for combining a PPI network with the results of gene expression profiling for genes of interest (Cline *et al.*, 2007).

Tool	Open Source	Accepts Plugins	Mixed Datatypes	Clustering	Standard PPI Data Standards as Input
Cytoscape	yes	yes	yes	yes	SBML, BioPAX
Osprey	no	no	yes	no	no
Power Graphs	no	no	yes	yes	no
Arena 3D	no	no	yes	yes	no
BioLayout Express3D	no	no	yes	yes	Owl, GraphML, sif, matrix
JClust	no	yes	yes	yes	no
NAViGaTOR	no	no	yes	no	PSI-MI, BioPAX

Table 2: A comparison of different visualization tools for biological networks. Cytoscape supports several standards for representing biological pathways, including BioPAX and SBML. BioPAX is an ontology for representing pathway knowledge (PPIs, metabolic, signaling, gene regulatory pathways), which is used as a data exchange format for biological pathways. SBML is an XML-based format, which is the standard for representing computational models in systems biology today. SBML consists of entities that are acted upon by processes (called reactions). BioCyc contains predicted pathway models for more than 200 organisms in a variety of formats, including SBML and BioPAX. The PSI-MI molecular interaction format is a standard for representing protein interaction data, which is aimed at integrating data from different databases.

Osprey builds graphical representations that are color-coded for gene function and experimental PPI data. Rapid elaboration and organization of network diagrams in a spoke model format can be achieved via a user-friendly interface (Breitkreutz *et al.*, 2003).

Power graphs aim to reduce network complexity by representing a biclique in PPI networks as a single collapsed edge. Power graphs compress up to 90% of the edges in biological networks and are applicable to all types of networks, such as protein interaction, regulatory networks, or homology networks (Royer *et al.*, 2008).

Arena3D introduces a new concept of staggered layers in 3D space. Related data – such as proteins, chemicals, or pathways – can be grouped onto separate layers and arranged via numerous layout algorithms. Data on a layer can be clustered via different algorithms, such as the k-means procedure, affinity propagation, Markov clustering, neighbor joining, tree clustering, or UPGMA ('unweighted pair-group method with arithmetic mean') (Pavlopoulos & O'Donoghue *et al.*, 2008; Pavlopoulos *et al*., 2008).

BioLayout Express3D is a tool for layout, visualization and clustering of large scale networks. In the latest version, the Markov Clustering algorithm (MCL) has become an integral part of BioLayout Express3D for clustering analysis (Freeman *et al.*, 2007).

JClust and Medusa are open source visualization tools that support several cluster analysis algorithms (MCL, MULIC, spectral, RNSC) (Pavlopoulos *et al.*, 2009). NAViGaTOR supports community-developed data formats (PSI-XML, BioPax and GML) (Brown *et al.*, 2009).

13 Applications in Biomedicine and Drug Discovery

PPI networks are used in biomedicine to unravel the molecular basis of disease by studying disease-related subnetworks. PPI networks offer new opportunities for drug research and development. Molecules rarely work alone but rather form part of a series of connected networks, signaling pathways and interactions. Researchers believe that drug development calls for a broader view, which comprehends the properties of the complex systems responsible for the regulation of biological processes (Wu & Stein, 2010; Taylor *et al.*, 2009).

Regarding the metabolic pathways of infections and cancer diseases, the use and increase of knowledge gathered by using PPI networks showed promising results over the last few years (Guda *et al*., 2009, Chuang *et al.*, 2007, Goh *et al*., 2007). The main goal of these networks is the determination of substances that are able to suppress or activate certain interactions that are associated with immune response functionality (Rambaldi *et al.*, 2008; Jonsson & Bates, 2006).

For instance, the correlation between the infection with the pathogenic bacterium *Heliobacter pylori* (*H. pylori*) and the occurrence of various gastro duodenal diseases has been examined in a new way using PPI networks. *H. pylori* infects about 50% of the world population. It is known to cause diseases such as chronic active gastritis in experimental animals and in humans. Many scholars have demonstrated a relationship between *H. pylori* and gastric carcinoma, and the World Health Organization (WHO) and the International Agency for Research on Cancer consensus group have classified *H. pylori* as a definite biological carcinogen (Kim & Kim, 2009).

To examine this correlation by using PPINs, the change of gene-expression during *H. pylori* infection was scanned from online literature databases and translated into proteins. A PPIN was constructed by searching the primary interactions of selected proteins. The constructed PPIN was mathematically

analyzed and its biological function was examined. In addition, the nodes on the network were extended by determining if they had any further functional importance or relation to other proteins. Mathematical analysis of this network showed hub and bottleneck proteins mostly related to immune response. These immune-related proteins interacted on the network with pathways and proteins related to the cell cycle, cell maintenance and proliferation, and transcription regulators. The extension of nodes showed interactions of the immune proteins with cancer related oncogenic proteins. As a result of this study, the detected hub and bottleneck proteins are potential drug targets for gastric inflammation and cancer (Kim & Kim, 2009).

By extending PPINs with other sources of information, such as gene co-expression, protein domain interaction, gene ontology annotations and text-mined protein interactions, multiple pattern similarities were found and gave a new insight into the outcome of several diseases (Srinivasan *et al.*, 2007). For example, such an extended PPIN was applied to two glioblastoma (GBM) datasets and candidate oncogenes were projected onto this network. The majorities of GBM candidate oncogenes formed a cluster and were closer than expected by chance. Network modules with sequence mutations were enriched in known oncogenes, tumor suppressors and signal transduction genes. Similar PPIN patterns were found in breast, colorectal and pancreatic cancers (Pujol *et al.*, 2009; Wu & Stein, 2010; Cerami *et al.*, 2010).

14 Conclusion

This article gave an overview of approaches to analyzing protein-protein interaction networks, as well as uses of PPI networks in biomedical and biological research. PPI networks are prominent in recent efforts to understand the molecular basis of diseases, such as cancer. PPI networks show promise to help decipher cell signaling, which will provide essential information in finding drug targets and curing diseases (Pujol *et al.*, 2009). Important problems still remain open, such as the high level of noise and inconsistency between PPI networks from different experimental studies. One criticism is that PPI networks are often static, as they do not show the dynamic changes in the cellular state over time (Srinivasan *et al.*, 2007). Since PPI networks provide only one viewpoint of the cell, integration with other data types is necessary to get a more complete picture of cellular events (Bapat *et al.*, 2010). Analysis tools, such as Cytoscape, are taking steps towards an integrated interpretation of PPI networks with other data types (Merico *et al.*, 2009).

References

Alon, U. (2007). Network motifs: theory and experimental approaches. Nature Review Genetics 8(6): 450-61.

Alber, F. *et al*. (2007). Determining the architectures of macromolecular assemblies. Nature. 450:683-94.

Alberts, B. *et al*. (2002). Molecular Biology of the Cel. 4th edition. Garland Science.

Albrecht, M. *et al*. (2005). Decomposing protein networks into domain-domain interactions. Bioinformatics 21:220-221.

Albert, R. (2005). Scale-free networks in cell biology. Journal of Cell Science. 118:4947-57.

Aloy, P. (2007). Shaping the future of interactome networks. Genome Biology, 8:316.

Altaf-Ul-Amin, M.d. *et al*. (2006). Development and implementation of an algorithm for detection of protein complexes in large interaction networks. BMC Bioinformatics 7:207.

Andreopoulos, B. *et al*. (2007). Clustering by common friends finds locally significant proteins mediating modules. Bioinformatics, Oxford University Press, 23(9): 1124-1131.

Andreopoulos, B. *et al*. (2009). Triangle network motifs predict complexes by complementing high-error interactomes with structural information. BMC Bioinformatics 10:196.

Andreopoulos, B., An, A., Wang, X. & Schroeder, M. (2009). A roadmap of clustering algorithms: finding a match for a biomedical application. Briefings in Bioinformatics 10 (3): 297-314.

Bader, G.D. & Hogue, C.W.V. (2003). An automated method for finding molecular complexes in large protein interaction networks. BMC Bioinformatics, 4:2.

Bapat, S.A. *et al*. (2010). Gene expression: Protein interaction systems network modeling identifies transformation-associated molecules and pathways in ovarian cancer. Cancer Research 70; 4809.

Barabasi, A.L. & Oltvai, Z.N. (2004). Network biology: understanding the cell's functional organization. Nature Reviews Genetics 5, 101-113.

Batagelj & Zavernik, M. (2001). Cores Decomposition of Networks. Recent Trends in Graph Theory, Algebraic Combinatorics, and Graph Algorithms. Slovenia.

Ben-Hur, A. & Noble, W.S. (2005). Kernel methods for predicting protein-protein interactions. Bioinformatics. 21:i38-46.

Bader, J. *et al*. (2004). Gaining confidence in high-throughput protein interaction networks. Nat Biotechnol. 22:78-85.

Batada, N.N. *et al*. (2006). Stratus not altocumulus: A new view of the yeast protein interaction network. PLoS Biology 4(10): e317.

Batada, N.N. *et al*. (2007). Still stratus not altocumulus: further evidence against the date/party hub distinction. PLoS Biology 5(6):e154.

Bader, G.D. & Hogue, C.W. (2002). Analyzing yeast protein-protein interaction data obtained from different sources. Nature Biotechnology 20(10):991-7.

Breitkreutz, B.J., Stark, C. & Tyers, M. (2003). Osprey: A Network Visualization System. Genome Biology 4(3):R22.

Brown, K.R. *et al*. (2009). NAViGaTOR: Network Analysis, Visualization and Graphing Toronto. Bioinformatics 25(24):3327-3329.

Braun, P. (2009). An experimentally derived confidence score for binary protein-protein interactions. Nat Methods. 6:91-7.

Bruggeman, F.J. & Westerhoff, H.V. (2007). The nature of systems biology. Trends Microbiol. Jan, 15(1): 45-50.

Bertin, N. *et al*. (2007). Confirmation of organized modularity in the yeast interactome. PLoS Biology 5(6):e153.

Cusick, M.E. *et al*. (2009). Literature-curated protein interaction datasets. Nature Methods 6, 39-46.

Chakrabarti, D. & Faloutsos, C. (2006). Graph Mining: Laws, Generators and Algorithms. ACM Computing Surveys (CSUR). Volume 38 Issue 1.

Clauset, A., Moore, C. & Newman, M.E.J. (2008). Hierarchical structure and the prediction of missing links in networks. Nature 453, 98-10.

Collins, S.R. *et al*. (2007). Towards a comprehensive atlas of the physical interactome of Saccharomyces cerevisiae. Molecular and Cellular Proteomics, 6, 439-450.

Cline, M.S. *et al*. (2007). Integration of biological networks and gene expression data using Cytoscape. Nature Protocols 2, 2366 - 2382.

Chuang, H.Y. *et al*. (2007). Network-based classification of breast cancer metastasis. Mol Syst Biol, 3:140.

Chen, J. & Yuan, B. (2006). Detecting functional modules in the yeast protein-protein interaction network. Bioinformatics 22(18): 2283-2290.

Chua, H.N. *et al*. Exploiting indirect neighbours and topological weight to predict protein function from protein-protein interactions. Bioinformatics 2006 22:1623-1630.

Chen, P., Deane, C. & Reinert, G. (2008). Predicting and Validating Protein Interactions Using Network Structure. PLoS Comput Biol. 4.

Chen, J. *et al*. (2006). Increasing confidence of protein interactomes using network topological metrics. Bioinformatics Vol. 22, Num. 16, pp. 1998-2004.

Chen, X & Liu, M. (2005). Prediction of protein-protein interactions using random decision forest framework. Bioinformatics. 21:4394-400.

Chiang, T. *et al*. (2007). Coverage and error models of protein-protein interaction data by directed graph analysis. Genome Biol. 8:R186.

Cerami, E. *et al*. (2010). Automated network analysis identifies core pathways in glioblastoma. PLoS One Feb 12 5(2): e8918.

Clauset, A., Shalizi, C.R. & Newman, M.E.J. (2009). "Power-law distributions in empirical data" . SIAM Review 51(4), 661-703.

Dunker, A.K. *et al*. (2005). Flexible nets: the roles of intrinsic disorder in protein interaction networks. FEBS Journal 272(20):5129-48.

Deng, M. *et al*. (2002). Inferring domain-domain interactions from protein-protein interactions. Genome Res 12:1540-1548.

Ding, C. *et al*. (2004). Multi-protein complex data clustering for detecting protein interactions and functional organizations. Interface 2004: Computational Biology and Bioinformatics. Baltimore, MD, USA.

Dunn, R. *et al*. (2005). The use of edge-betweenness clustering to investigate biological function in protein interaction networks. BMC Bioinformatics 6:39.

Ekman, D. *et al*. (2006). What properties characterize the hub proteins of the protein-protein interaction network of Saccharomyces cerevisiae? Genome Biology 7(6):R45.

Espadaler, J. *et al*. (2005). Detecting remotely related proteins by their interactions and sequence similarity. PNAS;102:7151-7156.

Freeman, T.C., Goldovsky, L. & Brosch, M. (2007). Construction, visualisation, and clustering of transcription networks from microarray expression data. PLoS computational biology, 3(10):2032-2042.

Gentleman, R. & Huber W. (2007). Making the most of high-throughput protein-interaction data. Genome Biology 8(10):112.

Giot, L. *et al*. (2003). A Protein Interaction Map of Drosophila melanogaster. Science 302(5651):1727-36.

Gavin, A. *et al*. (2006). Â Proteome survey reveals modularity of the yeast cell machinery. Nature 440, 631-636.

Goldberg, D.S., Gibson, T.A. & Gabow, A. (2007). Gene and Protein Networks tutorial. ISMB/ECCB, Vienna, Austria.

Gandhi, T.K.B. *et al*. (2006). Analysis of the human protein interactome and comparison with yeast, worm and fly interaction datasets. Nature Genetics 38, 285-293.

Guda, P., Chittur, S.V. & Guda, C. (2009). Comparative analysis of protein-protein interactions in cancer-associated genes. Genomics Proteomics Bioinformatics, 7:25-36.

Girvan, M. & Newman, M. (2002). Community structure in social and biological networks. PNAS 99:7821-7826.

Goh, K.I. *et al*. (2007). The human disease network. Proc Natl Acad Sci USA, 104:8685-8690.

Guo, Y. *et al*. (2008). Using support vector machine combined with auto covariance to predict protein-protein interactions from protein sequences. Nucleic Acids Res. 36:3025-30.

Galperin, M.Y. & Cochrane, G.R. (2009). Nucleic Acids Research annual Database Issue and the NAR online Molecular Biology Database Collection in 2009. Nucleic Acids Res. D1-4.

Hesselberth, J.R. *et al*. (2006). Comparative analysis of Saccharomyces cerevisiae WW domains and their interacting proteins. Genome Biology 7(4):R30 (2006).

Huang, T. (2004). POINT: a database for the prediction of protein-protein interactions based on the orthologous interactome. Bioinformatics. 20:3273-6.

Han, DS. *et al*. (2004). PreSPI: a domain combination based prediction system for protein-protein interaction. Nucleic Acids Res. 32:6312-20.

Han, J.D. *et al*. (2004). Evidence for dynamically organized modularity in the yeast protein-protein interaction network. Nature 430(6995):88-93.

Haynes, C. *et al*. (2006). Intrinsic disorder is a common feature of hub proteins from four eukaryotic interactomes. PLoS Comput Biology 4;2(8):e100.

Huang, H. & Bader, J.S. (2009). Precision and recall estimates for two-hybrid screens. Bioinformatics 25(3): 372-378.

Hart, G.T., Ramani, A.K. & Marcotte, E.M. (2006). How complete are current yeast and human protein-interaction networks? Genome Biology, 7:120.

Hoffmann, R. & Valencia, A. (2003). Protein interaction: same network, different hubs. Trends Genet. 19(12):681-3.

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

Jothi, R. *et al*. (2006). Co-evolutionary analysis of domains in interacting proteins reveals insights into domain-domain interactions mediating protein-protein interactions. J. Mol. Biol., 362, 861-875.

Jansen, R. *et al*. (2003). A Bayesian networks approach for predicting protein-protein interactions from genomic data. Science. 302:449-53.

Jin, G. *et al*. (2007). Hubs with network motifs organize modularity dynamically in the protein-protein interaction network of yeast. PLoS_ONE. 2:e1207.

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

Pujol, A. *et al*. (2009). Unveiling the role of network and systems biology in drug discovery. Cell review. Trends in Pharmacological Sciences. 31(3):115-123.

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

Kamburov, A. *et al*. (2007). Denoising inferred functional association networks obtained by gene fusion analysis. BMC Genomics, 8, 460.

Kim, K.K. & Kim, H.B. (2009). Protein interaction network related to Helicobacter pylori infection response. World J Gastroenterol. 15(36): 4518-4528.

Kim, W. *et al*. (2002). Large scale statistical prediction of protein-protein interaction by potentially interacting domain pair. Genome Inform 13:42-50.

Lodish, H., Berk, A. *et al*. (2000). Molecular Cell Biology. 4th ed., W.H. Freeman, New York.

Lima-Mendez, G .& van Helden, J. (2009). The powerful law of the power law and other myths in network biology. Mol Biosyst 5(12): 1482-93.

Lubovac, Z., Gamalielsson, J. & Olsson, B. (2006). Combining functional and topological properties to identify core modules in protein interaction networks. Proteins 64(4):948-959.

Li, H. *et al*. (2006). Discovering motif pairs at interaction sites from protein sequences on a proteome-wide scale. Bioinformatics 22:989-996.

Ma'ayan, A. (2009). Insights into the organization of biochemical regulatory networks using graph theory analyses. Journal Biological Chemistry 284(9):5451-5.

Ma'ayan, A. *et al*. (2009). SNAVI: Desktop application for analysis and visualization of large-scale signaling networks. BMC Systems Biology, 3:10.

Marcus, T. *et al*. (2008). Identifying functional modules in protein-protein interaction networks: an integrated exact approach. Bioinformatics 24(13): i223-i23.

Mathivanan, S. *et al*. (2006). An evaluation of human protein-protein interaction data in the public domain. BMC Bioinformatics 7(Suppl 5):S19.

Merico, D., Gfeller, D. & Bader, G.D. (2009). How to visually interpret biological data using networks. Nature Biotechnology 27, 921 - 924.

Mathivanan, S. (2006). An evaluation of human protein-protein interaction data in the public domain. BMC Bioinformatics. 7:S19.

Mewes, H. *et al*. (2006**).** MIPS: analysis and annotation of proteins from whole genomes in 2005. *Nucleic Acids Res*, (34 Database)**:**D169-72.

Morrison, J.L. *et al*. (2006). A lock-and-key model for protein-protein interactions. Bioinformatics 22 (16): 2012-2019.

Newman, M.E.J. (2006). Modularity and community structure in networks. PNAS 103(23): 8577-8582.

Ng, S., Zhang, Z. & Tan, S. (2003). Integrative approach for computationally inferring protein domain interactions. Bioinformatics. 19:923-9.

Okada, K. *et al*. (2005). Accurate extraction of functional associations between proteins based on common interaction partners and common domains. Bioinformatics 21:2043-2048.

Pawson, T. & Nash, P. (2003). Assembly of cell regulatory systems through protein interaction domains. Science, 300, 445-452.

Pawson, T. & Scott, J.D. (1997). Signaling through scaffold, anchoring, and adaptor proteins. Science 278, 2075-2080.

Pawson, T. (2003). Organization of cell-regulatory systems through modular-protein-interaction domains. Phil. Trans. R. Soc. Lond. A 2003 361, 1251-1262.

Pereira-Leal, J.B. & Teichmann, S.A. (2005). Novel specificities emerge by stepwise duplication of functional modules. Genome Research 15(4):552-9.

Pavlopoulos, G.A. *et al*. (2009). jClust: a clustering and visualization toolbox. Bioinformatics 25(15):1994-6.

Pavlopoulos, G.A. & O'Donoghue, S.I. *et al*. (2008). Arena3D: visualization of biological networks in 3D. BMC Systems Biology, 2:104.

Pavlopoulos, G.A. *et al*. (2008). A survey of visualization tools for biological network analysis. BioData Mining,Â 1:12.

Pereira-Leal, J.B. *et al*. (2004). Detection of functional modules from protein interaction networks. Proteins 54:49-57.

Pitre, S. *et al*. (2006). PIPE: a protein-protein interaction prediction engine based on the re-occurring short polypeptide sequences between known interacting protein pairs. BMC Bioinformatics. 7:365.

Russell, R.B. & Aloy, P. (2008). Targeting and tinkering with interaction networks. Nature Chemical Biology 4, 666 - 673.

Rinner, O. *et al*. (2007). An integrated mass spectrometric and computational framework for the analysis of protein interaction networks. Nature Biotechnology 25, 345-352.

Ravasz, E. *et al*. (2002). Hierarchical organization of modularity in metabolic networks. Science 297(5586):1551-5.

Rual, J.F. *et al*. (2005). Towards a proteome-scale map of the human protein-protein interaction network. Nature 437, 1173-1178.

Rhodes, D.R. *et al*. (2005). Probabilistic model of the human protein-protein interaction network. Nature Biotechnology 23, 951-959.

Royer, L. *et al*. (2008). Unraveling Protein Networks with Power Graph Analysis. PLoS Comput Biol 4(7): e1000108.

Ramirez, F. *et al*. (2007). Computational analysis of human protein interaction networks. Proteomics. 7:2541-2552.

Rhodes, D. *et al*. (2005). Probabilistic model of the human protein-protein interaction network. Nat Biotechnol. 23:951-9.

Rambaldi, D. *et al*. (2008). Low duplicability and network fragility of cancer genes. Trends Genet, 24:427-430.

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

Schmid, E.M. & McMahon H.T. (2007). Integrating molecular and network biology to decode endocytosis. Nature 448, 883-888.

Simonis, N. *et al*. (2009). Empirically controlled mapping of the C. elegans protein-protein interactome network. Nature Methods 6(1):47-54.

Srinivasan, B.S. *et al*. (2007). Current progress in network research: toward reference networks for key model organisms. Briefings in Bioinformatics 8(5):318-332.

Sharan, R. *et al*. (2005). Identification of protein complexes by comparative analysis of yeast and bacterial protein interaction data. J Comput Biol. 12(6):835-46.

Sharan, R. *et al*. (2005). Cover Article: Conserved patterns of protein interaction in multiple species. Proc Natl Acad Sci USA. 8:102(6) 1974-79.

Sanderson, C.M. (2009). The cartographers toolbox: building bigger and better human protein interaction networks. Briefings in Functional Genomics and Proteomics. 8(1):1-11.

Shen-Orr, S.S. *et al*. (2002). Network motifs in the transcriptional regulation network of *Escherichia coli*. Nature Genetics 31(1): 64-8.

Spirin, V. & Mirny, L.A. (2003). Protein complexes and functional modules in molecular networks. PNAS 100(21): 12123-12128.

Shannon, P. *et al*. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res 2003, 13(11):2498-2504.

Sprinzak, E. & Margalit, H. (2001). Correlated sequence-signatures as markers of protein-protein interaction. J. Mol. Biol 311:681-692.

Spirin, V. & Mirny, L.A. (2003). Protein complexes and functional modules in molecular networks. PNAS 100:12123-12128.

Samanta, M.P. & Liang, S. (2003). Predicting protein functions from redundancies in large-scale protein interaction networks. PNAS;100:12579-12583.

Singhal, M. & Resat, H. (2007). A domain-based approach to predict protein-protein interactions. BMC Bioinformatics. 8:199.

Soon, J., Yoon, B.J. & Dougherty, E.R. (2010). Identification of diagnostic subnetwork markers for cancer in human protein-protein interaction network. BMC Bioinformatics, 11(Suppl 6): S8.

Tarassov, K. *et al*. (2008). An in vivo map of the yeast protein interactome. Science 320(5882): 1465-1470.

Tong, A.H. *et al*. (2002). A combined experimental and computational strategy to define protein interaction networks for peptide recognition modules. Science 295(5553):321-4.

Tamames, J., Moya, A. & Valencia, A. (2007). Modular organization in the reductive evolution of protein-protein interaction networks. Genome Biology 8:R94.

Taylor, I.W. *et al*. (2009). Dynamic modularity in protein interaction networks predicts breast cancer outcome. Nature Biotechnology 27, 199-204.

Venkatesan, K. *et al*. (2009). An empirical framework for binary interactome mapping. Nature Methods 6, 83-90.

Valencia, A & Pazos, F. (2002). Computational methods for the prediction of protein interactions. Curr Opin Struct Biol. 12:368-73.

von Mering, C. *et al*. (2007). STRING 7-recent developments in the integration and prediction of protein interactions. Nucleic Acids Res.:D358-62.

van den Akker, E. *et al*. (2011). Integrating Protein-Protein Interaction Networks with Gene- Gene Co-Expression Networks improves Gene Signatures for Classifying Breast Cancer Metastasis. Journal of Integrative Bioinformatics, 8(2): 188.

Wuchty, S., Barabasi, A.L. & Ferdig, M.T. (2006). Stable evolutionary signal in a Yeast protein interaction network. BMC Evolutionary Biology 30;6:8.

Wang, Z. & Zhang, J. (2007). In Search of the Biological Significance of Modular Structures in Protein Networks. PLoS Comput Biol. 3(6): e107.

Wong, G.L.L. & Chua, H.N. (2009). Complex discovery from weighted PPI networks. Bioinformatics 25(15): 1891-1897.

Wu, G., Feng, X. & Stein, L. (2010). A human functional protein interaction network and its application to cancer data analysis. Genome Biology, 11:R53.

Wu, X. *et al*. (2006). Prediction of yeast protein-protein interaction network: insights from the Gene Ontology and annotations. Nucleic Acids Res. 34:2137-50.

Yu, H *et al*. (2008). High-quality binary protein interaction map of the yeast interactome network. Science. 322(5898):104-10.

Yu, J. & Fotouhi, F (2006). Computational approaches for predicting protein-protein interactions: a survey. J Med Syst.;30:39-44.

Yu, H. *et al*. (2006). Predicting interactions in protein networks by completing defective cliques. Bioinformatics. 22:823-829.

Yu, J. & Finley, R.L.J. (2009). Combining multiple positive training sets to generate confidence scores for protein-protein interactions. Bioinformatics. 25:105-11.

Zhang, L.V. *et al*. (2005). Motifs, themes and thematic maps of an integrated Saccharomyces cerevisiae interaction network. Journal of Biology. 4(2): 6.

Zhang, L. *et al*. (2005). Motifs, themes and thematic maps of an integrated Saccharomyces cerevisiae interaction network. J Biol. 4:6.