Protein-protein interactions: Making sense of networks via graph-theoretic modeling

Nataša Pržulj

The emerging area of network biology is seeking to provide insights into organizational principles of life. However, despite significant collaborative efforts, there is still typically a weak link between biological and computational scientists and a lack of understanding of the research issues across the disciplines. This results in the use of simple computational techniques of limited potential that are incapable of explaining these complex data. Hence, the danger is that the community might begin to view the topological properties of network data as mere statistics, rather than rich sources of biological information. A further danger is that such views might result in the imposition of scientific doctrines, such as scale-free-centric (on the modeling side) and genome-centric (on the biological side) opinions onto this area. Here, we take a graph-theoretic perspective on protein-protein interaction networks and present a high-level overview of the area, commenting on possible challenges ahead.

Keywords:

- graph algorithms; network comparison and alignment; network models; protein function and disease; protein-protein interaction networks

DOI 10.1002/bies.201000044

Department of Computing, Imperial College London, London, UK

Corresponding author:
Nataša Pržulj
E-mail: natasha@imperial.ac.uk

Introduction

Recent advances in experimental biology have enabled the production of large quantities of interaction data between genes and proteins. High-throughput screens, such as yeast two-hybrid (Y2H) assays [1–7], affinity purification coupled to mass spectrometry (AP/MS) [8–11] and synthetic-lethal and suppressor networks [12], have yielded partial networks for humans [5, 6], microbial [13–15] and viral [16] pathogens, as well as many model organisms [1–4, 10–12]. These large networks are offering many interesting and important opportunities for biological and computational scientists. We are currently at a unique time in the history of science when algorithmic and modeling advances applied to these data could contribute to biological understanding, hence potentially impacting therapeutics and public health. The nascent field of network biology faces considerable challenges. First, our current observational data are largely incomplete due to sampling, population averaging, and other biases in data collection, handling and interpretation [17–24]. Also, they are noisy due to bio-technological limitations used for their collection. An example of a protein-protein interaction network (PIN) that illustrates sparsity of the data is presented in Fig. 1. Despite incompleteness and noise, the scientific community has begun analyzing and modeling these data, which yielded interesting and sometimes controversial discoveries [25–35]. For example, it has been questioned whether metabolic networks are ‘‘scale-free’’ or ‘‘scale-rich’’ [25, 26], whether the most connected nodes (‘‘hubs’’) in yeast PIN are lethal [27, 28], whether complex networks are ‘‘self-similar’’ or ‘‘self-dissimilar’’ [29, 30], or whether hubs in yeast PIN can be divided into ‘‘party’’ hubs, those that interact with their partners simultaneously, and ‘‘date’’ hubs, those that interact with different partners at different times or locations [32–35] (see *Network topology and biological function and disease* further below). The controversies often resulted from a lack of understanding of the sampling properties of the data, as well as from the use of computational techniques sensitive to noise [19–21]. Also, a deep significance has initially been attributed to power-law distributions in many biological

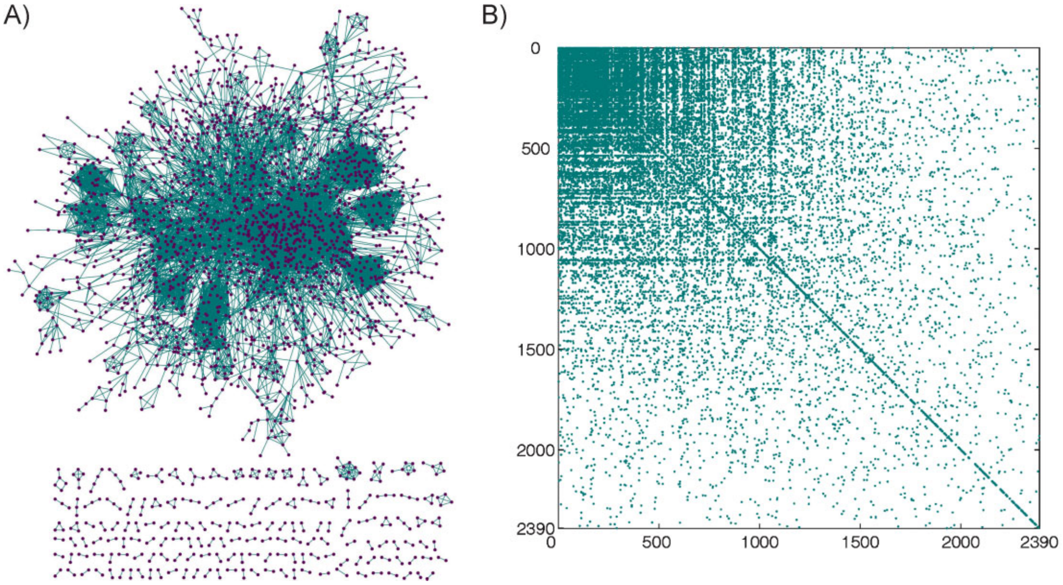

Figure 1. A: An example of a PIN [22]. **B:** The adjacency matrix of the same network illustrating its sparsity.

networks, e.g. suggesting a ''universal architecture'' of complex systems, that has later been challenged [31].

Modeling biological networks is of importance, since only by using models we can obtain a concise summary of the phenomenon at study that might exhibit some unexpected properties that we may want to further explore experimentally. Different modeling paradigms exist. In this manuscript, we focus on graph-theoretic analyses and modeling. It can be argued that such analyses are over-simplified views of networked systems, as they are abstractions that ignore the dynamics (e.g. see [36, 37] for accounts of dynamics). Many protein-protein interactions (PPIs) are naturally undirected, representing so-called ''stable'' interactions in which interacting partners find each other and stay bound (e.g. protein complexes such as ribosome, or hemoglobin), while others are ''transient'' in the sense that the interacting partners bind and unbind at different times and under different conditions (e.g. signaling cascades). Hence, if we had complete knowledge of a PIN, it would need to be represented as a mixed network of directed and undirected edges, with time-dependent components that include many parameters. However, such information is currently unavailable on a systems-level scale, so PINs are static undirected networks that amalgamate stable interactions with all of the currently known transient interactions at all examined time points and experimental conditions. Nevertheless, even simple graph-theoretic analyses of such data have already provided valuable biological insight (see below). Hence, it can be argued that these data contain substantial biological information that is hidden in their complexity, which could be revealed by developing and applying sophisticated graph-theoretic techniques.

The purpose of the manuscript is not to enter the debate about reductionist and deterministic (that philosophers of science tend to consider as naive or simplistic) versus dynamic modeling, issues of the relationship between theory, model, and experiment, or what strictly can be ''predicted'' by a model or a theory (e.g. see refs. [38, 39]). Instead, I argue that one way we can learn about biology is by analyzing and modeling the ''topology'' (also called ''structure'') of biological networks. Hence, I will not refrain from using an analogy with physical sciences and astronomy in the 17th century: even though, thanks to Copernicus, Kepler, Galileo and others, good observational data were available at the time, along with competing theories of whether our solar system was geocentric or heliocentric, it was not until Sir Isaac Newton came up with his laws of universal gravitation and motion that we understood why planets move as they do. Only in the light of these laws did the observations about our solar system become evident and we understood that they are only a part of a much larger phenomenon. Similarly, the main reason to model network data is to understand laws, since only with the help of such laws we can hope to be able to make predictions and reproduce the phenomena. One way of modeling is by using graph theory. Despite imperfect data, properties of a network model have already been used to reduce complexity. For example, they have been used to propose a strategy for optimal interactome detection [40]. Also, we have exploited them for developing efficient algorithms for approximately solving computationally intractable problems (such algorithms are called ''heuristic'' algorithms) [41] and for data de-noising [42].

Network comparisons via network properties

Comparative analyses of network data would enable finding similarities and differences between biological networks as well as knowledge transfer. This would be useful since a lot is often known about a network of a model organism, but very little about networks of other organisms. However, comparing large networks (also called ''graphs'') is computationally infeasible, since any such comparison would involve solving the ''subgraph isomorphism problem'' that asks whether one

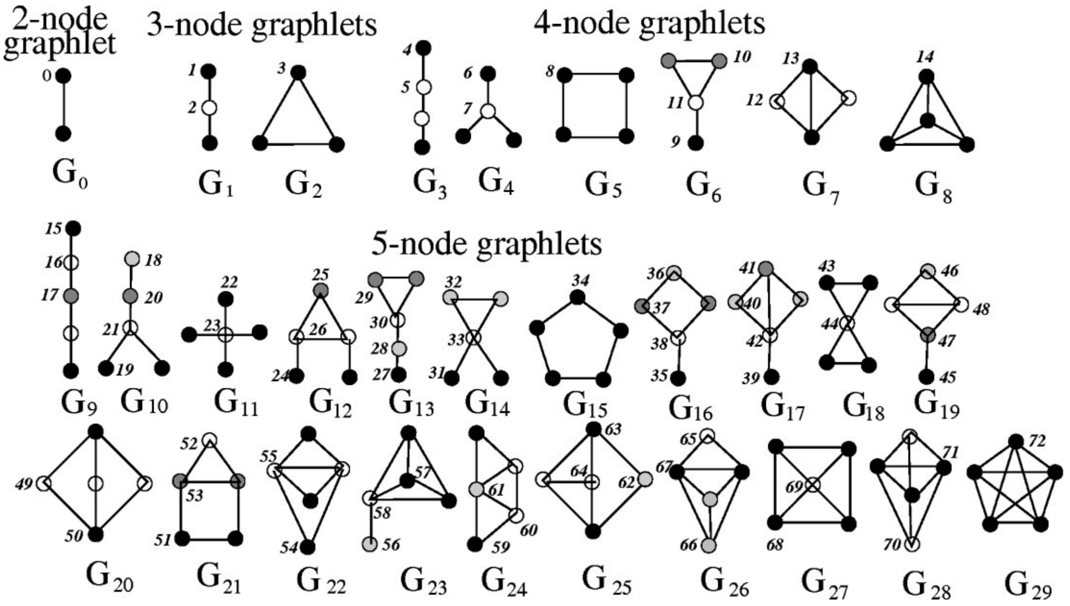

Figure 2. All 2-, 3-, 4-, and 5-node graphlets, G_0, G_1,...,G_{29} and the orbits denoted by 0, 1, 2,...,72 [54]. In each graphlet, nodes belonging to the same orbit are of the same shade.

graph exists as a copy in another graph; unfortunately, this problem is known to be computationally intractable, so no efficient way of solving it exists. Hence, easily computable heuristic approaches are sought.

Some of these easily computable approaches are called ''network properties'' and they can roughly and historically be divided into two groups: top-down macroscopic statistical ''global network properties'' and bottom-up microscopic ''local network properties''. Widely used global network properties include the ''degree distribution'', ''clustering coefficient'', ''clustering spectra'', ''network diameter'', and various forms of network ''centralities'' [43]. Based on these properties, network models (see *Network models* further below) have been proposed for biological networks if the properties of model networks fit those of the data. The ''degree'' of a node is the number of edges that the node touches; hence, the degree is a property of a node. The ''degree distribution'' is a property of a network: it is the distribution of degrees of all nodes in the network; equivalently, it is the probability that a randomly selected node in a network has degree k [commonly denoted by $P(k)$]. Many biological networks, including PINs, have non-Poisson degree distributions with a power-law tail, $P(k) \sim k^{-\gamma}$, $\gamma > 0$; all such networks have been termed ''scale-free'' [44]. Due to its conceptual simplicity, this property has widely been used to ''characterize'' real networks, even though it is a very weak descriptor of network structure [45, 46]. For example, a network consisting of five triangles and a network consisting of one 15-node ring are of the same size, i.e. they have the same number of nodes and edges, and they have the same degree distribution (each node has degree 2), but their topologies are very different. This holds true for other global network properties including the clustering coefficient, which is the measure of ''cliquishness'' in the network, and the network diameter, which measures how ''far spread'' the network is [45]. PINs and other biological and real networks have high clustering coefficients compared to completely random networks, as well as small average diameters, of the order of O(log n), where n is the number of nodes in a network; this is called the ''small-world'' property of a network [47]. However, global network properties of largely incomplete PINs can be misleading, since they describe the network structure produced by the sampling techniques used to obtain these networks rather than the true structure of the currently unknown complete networks [19–21]. Luckily, certain neighborhoods of PINs are well studied, and usually network regions relevant for human disease. Hence, local descriptions applied to the well-studied areas are more appropriate.

The local properties that concern us are based on the notion of a subgraph. A ''subgraph'' (or a ''partial subgraph'') of a network G is a network whose nodes and edges belong to G. An ''induced subgraph'' H of G has to contain all edges of G that connect nodes of H. For example, in graph G_2 in Fig. 2, if we pick all three nodes, we can pick any two edges between them to form a three-node linear path (such as graph G_1 in the same figure); such three-node paths are partial subgraphs of G_2. Since an induced subgraph must contain all edges, the only three-node induced subgraph of G_2 is a triangle. Analogous to sequence motifs, ''network motifs'' have been defined as partial subgraphs that occur in a real network at frequencies much higher than would be expected at random [48–50]. It has been proposed that network motifs are functional building blocks of gene regulation and other biological networks, and that different motifs are characteristic for different types of real networks [48–51]. However, it is unclear what subgraph enrichment should be expected at random, since many different random graph models exist (see *Network models* further below) [52]. Also, when we are characterizing the structure of any network family, we care about induced rather than partial subgraphs [53]. Hence, we define a ''graphlet'' as a small, connected, INDUCED subgraph of a network. We introduce approaches based on the frequencies of occurrence

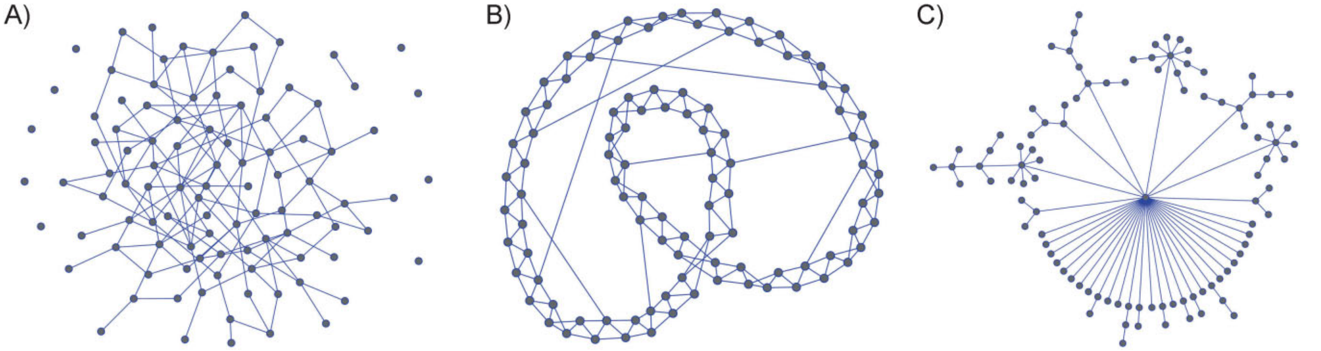

Figure 3. Examples of model networks. **A:** An Erdös-Rényi random graph. **B:** A small-world network. **C:** A scale-free network.

of all graphlets with up to five nodes (presented in Fig. 2), regardless of whether they are over-represented in the data when compared with any model networks [45, 54]. That is, graphlets do not need to be over-represented and this, along with being induced, distinguishes them from network motifs.

By counting the frequency of graphlets across a network, we obtain a statistical characterization of local structure of a network, independent of any network null model, and comparing such frequency distributions gives a measure of structural similarity between networks [45]. We further refined the graphlet idea by noting that in some graphlets, the nodes are distinct from each other. For example, in a ring (cycle) of four nodes, every node looks the same as every other, but in a chain (path) of four nodes, there are two end nodes, and two middle nodes. We formalized this idea by using graph ‘‘automorphism orbits’’ [54] (described below). In this way, we enhance the sensitivity of using graphlets for network analysis and modeling without increasing the computational cost.

Network models

Erdös-Rényi (ER) random graphs are the earliest random graph model. In these graphs, edges are drawn between pairs of nodes uniformly at random with the same probability, p [55]. This model has been extensively studied and many of its properties are mathematically well understood [56]. For this reason, it is a standard model to compare the data against, even though it is not expected to fit the data well. Since ER graphs, unlike PINs, have Poisson degree distributions and low clustering coefficients, other network models have been designed. In ‘‘generalized random graphs’’, the edges are randomly chosen as in ER graphs, but the degree distribution is constrained to match that of the data [57, 58]. ‘‘Small-world’’ networks integrate between regularity and randomness by being regular ring lattices with a small number of randomly rewired edges; hence, they have small diameters and large clustering coefficients [47]. Scale-free (SF) networks include an additional constraint that the degree distribution follows a power-law [44, 46]. Since the degree distributions of many PINs decay as approximate power-laws, many variants of SF network growth models have been proposed, the most notable of which are those based on biologically motivated gene ‘‘duplication and mutation’’ principles [59, 60]. Examples of model networks are presented in Fig. 3.

We have proposed an alternative, biologically motivated models for PINs, based on ‘‘geometric graphs’’ (GEO) [45, 54, 61–63]. Assume we have a collection of points distributed in space. We pick a constant distance ε and say that two points are ‘‘related’’ if they are within ε of each other. This relationship can be represented as a graph, where each point in space is a node and two nodes are connected if they are within distance ε. This is called a ‘‘geometric graph’’; if the points are distributed at random, then it is a ‘‘geometric random graph’’. Illustrations are presented in Fig. 4.

As mentioned above, the degree distribution provides a very weak constraint on the structure of a graph. Graphlet-based methods are much more highly constraining. Using them, we have shown that the data are fit much better by the GEO model than by the SF model [45, 54]. This brings a question of how to decide which property to trust when comparing networks. We have shown that when a series of global and local network properties is used to compare various PINs with various model networks by using several machine learning classifiers, the structure of PINs is the most consistent with the structure of noisy GEO networks [63]. We have further corroborated the fit of GEO by demonstrating that PINs can explicitly be embedded into a low-dimensional geometric space [64]. To build a GEO that corresponds to the PIN, as the distance between proteins we used a function of the shortest path length between them in the PIN. We have devised further refinements of GEO to fit PINs even better, by learning the distribution of proteins in the embedding space [61], or by replicating the principles of gene duplications and mutations mentioned above in a geometric space [62]. Conceptually, the reason for the good fit of GEO to PINs could lie in the observation that all biological entities, including genes and proteins as gene products, exist in some multidimensional biochemical space; currently, it is hard to hypothesize about the nature or dimensionality of that space. Gene duplications and mutations are naturally modeled in the biochemical space: a duplicated gene starts at the same point in the space as its parent and then natural selection acts either to eliminate one, or cause them to slowly separate in the space. This means that the child inherits some of the interactors of its parent, while possibly gaining novel connections as well. The further the child is moved away from its parent in this abstract space, the more different their biochemical properties. Currently, such

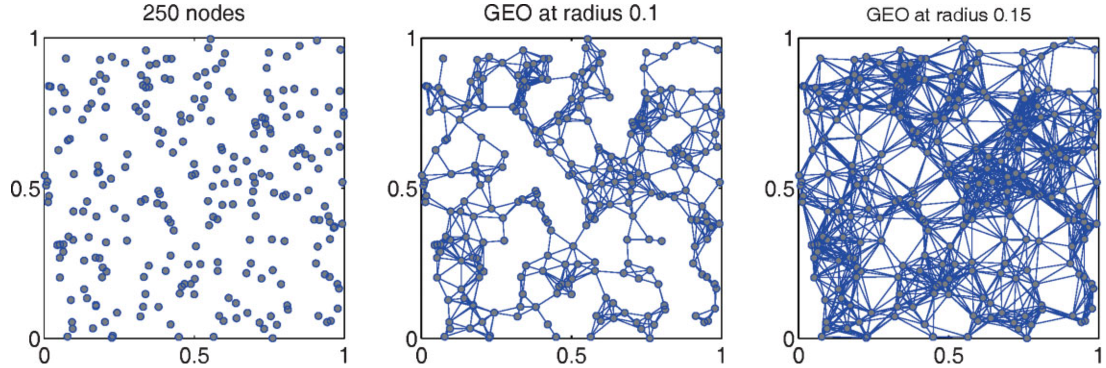

Figure 4. Left panel: 250 points in the unit cube. Middle panel: the resulting geometric graph with a cut-off distance of 0.1. Right panel: the same points, but a different graph resulting from a cut-off distance of 0.15.

GEO models are quite crude mathematical approximations of real biology and their further refinement is needed.

Network alignment

Another way to compare networks is by ‘‘aligning’’ them. Analogous to sequence comparison and alignment that has had a deep impact on our understanding of evolution, biology, and disease, comparison and alignment of biological networks are likely to have a similar impact.

Hence, it has been argued that comparing networks of different organisms in a meaningful way is one of the foremost problems in evolutionary and systems biology [65]. Conceptually, network alignment tries to find the best way to ‘‘fit’’ network *G* into network *H* even if *G* does not exist as an exact subgraph of *H* [65]. A simple example illustrating network alignment is presented in Fig. 5. Unfortunately, in contrast to sequence alignment, any reasonable formulation of this problem makes it computationally infeasible to solve exactly. The reason for this is the underlying subgraph isomorphism problem described above. In addition, PINs and other biological networks contain noise, i.e. missing edges, false edges, or both [66]. Also, biological variation makes it further non-obvious how to measure the ‘‘goodness’’ of the fit between aligned networks.

As for sequence alignments, ‘‘local’’ and ‘‘global’’ network alignments exist. In local alignments, a node can be mapped to several nodes. In contrast, a global network alignment provides a unique alignment from each node in the smaller network to exactly one node in the larger network. However, a disadvantage is that this might lead to non-optimal node pairings in some local network regions. The majority of methods used for alignment have focused on local alignments that aim to find small subnetworks corresponding to pathways or protein complexes conserved in PPI networks of different species [67–70]. The earliest such algorithm is PathBLAST, which searches for high-scoring alignments of pathways between two PINs by taking into account the probabilities that PPIs in a pathway are true PPIs rather than false positives, as well as the homology information derived from sequences of the aligned proteins [67]. PathBLAST identified orthologous pathways between yeast *Saccharomyces cerevisiae* and bacterium *Helicobacter pylori*. A modification of PathBLAST, called NetworkBLAST-M, identified conserved protein complexes in multiple species [71]. Another approach, called MaWISh (Maximum Weight Induced Subgraph), uses the duplication and divergence models for understanding the evolution of protein interactions [72]. Using this model, MaWISh constructs a weighted global alignment graph and tries to find a maximum induced subgraph in it. MaWISh is used to perform pairwise alignments of yeast, worm, and fruitfly PINs. Graemlin is another local network aligner; it gives a score to a possibly conserved module between different networks by computing the log-ratio of the probability that the module is subject to evolutionary constraints and the probability that the module is under no constraints, while taking into account phylogenetic relationships between species whose networks are being aligned [69].

Several global network alignments have been developed [73–79]. The earliest one, IsoRank, follows the intuition that two nodes should be matched only if their neighbors can also be matched [73]. This is formulated as an eigenvector problem and spectral graph algorithms are used to compute scores for topologically aligning node pairs from different networks.

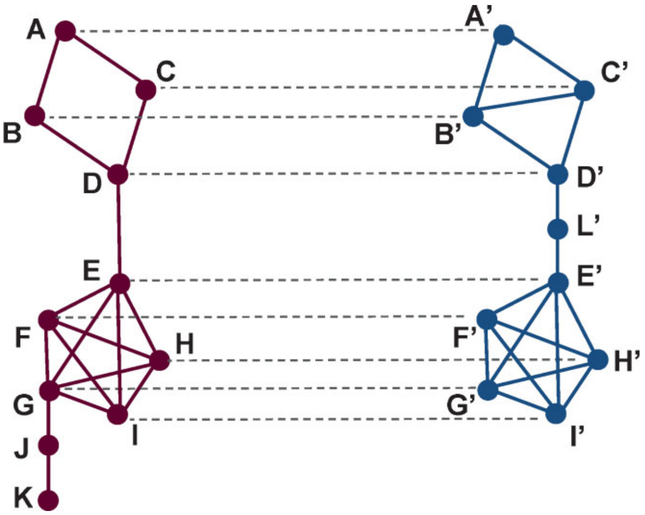

Figure 5. An example of an alignment of two networks.

IsoRank also includes BLAST scores [80] for sequence similarity between proteins in PINs into the node alignment; it utilizes a user-defined weight constant λ that controls for the relative contribution of network similarity, whereas 1-λ controls for the contribution of sequence similarity. These scores are then used in a greedy strategy for constructing an alignment. IsoRank was used to identify functional orthologs between yeast and fly. It was later extended to perform multiple local and global alignments between networks [75, 77]. While this approach combined topology and sequence information, we designed network alignment approaches that use only network topology (see below) [78, 79]. Also, network querying approaches exist that use no topology at all [81]. ''Network querying'' is a type of network comparison that finds, in a PIN, subnetworks similar to the query network, which is usually a protein complex or a pathway. Due to space constraints, we do not survey network querying approaches.

Our global network alignment algorithms that are based solely on local network topology are called GRAAL (GRAph ALigner) [78] and H-GRAAL [79]. Since they rely on topology only, they can align networks of any type, not only biological ones. Both algorithms use ''graphlet degrees'' defined below, which give a highly constraining quantification of the topological similarity between nodes [82]. If we recall that the degree of a node is the number of edges it touches, and that an edge is the only graphlet with two nodes (graphlet G_0 in Fig. 2), we can define a ''graphlet degree'' of node x with respect to each graphlet G_i in Fig. 2, in the sense that the G_i degree of x counts ''how many graphlets of type G_i touch node x'' [54]. Then, the traditional degree is simply the G_0 degree. Since there are 30 graphlets with up to five nodes, this would provide a vector of 30 ''graphlet degrees''. Luckily, specificity can be increased by observing that not all nodes in a graphlet are necessarily topologically equivalent. For example, the middle node in G_1 is topologically distinct from the end nodes of G_1 (Fig. 2). Figure 2 shows all 73 topologically distinct nodes across all graphlets with up to five nodes. Each is called an ''orbit'' (an ''automorphism orbit'', see [54, 82] for details), and we label them 0,. . .,72. The ''graphlet degree vector'' (GDV) or ''GD signature'' of node x hence has 73 elements; element i represents the number of times node x ''touches'' a graphlet at orbit i, across all the graphlets in its neighborhood. Hence, it describes the topology of a node's neighborhood and captures the node's inter-connectivities out to a distance 4. Reaching out to distance 4 from a node might be enough to almost uniquely determine the node's position in a network, since many real-world networks have the ''small-world'' property [47] (described in *Network models* above). Based on finding GD signature-similar nodes across two networks, GRAAL is a seed-and-extend approach, while H-GRAAL is based on the Hungarian algorithm for solving the assignment problem, which is a combinatorial optimization algorithm for finding a maximum weight matching in a weighted bipartite graph. We applied them to yeast and human PINs, and they expose regions of network similarity about an order of magnitude larger than other algorithms. Also, we used them to transfer knowledge from annotated to unannotated parts between aligned networks. Moreover, analogous to sequence alignments, we used the network alignment scores to infer phylogeny [78, 79]. Thus, network alignments could have the potential to provide a completely new, independent source of biological and phylogenetic information.

The reason for developing methods for aligning large biological networks, such as PINs, that use only network topology is twofold. Biological networks describe a part of biological information, just as genetic sequences do. We argue that sequence and topology provide complementary insights into biological knowledge [83]. Analogous to sequence alignment algorithms that do not use biological information external to sequences to perform alignments, using biological information external to network topology might deter from finding biological information that is encoded in network topology. We argue that we can learn the most about biology only after reliable algorithms for analyzing each data type separately are developed and then integrated [78].

Network topology and biological function and disease

The relationship between the topology of PINs and biological function has been the subject of many studies. The aim is to predict function of unannotated proteins [84]. Similarly, the role of PINs in disease has been examined [85]. An early approach found a correlation between protein connectivity, i.e. degree, in a PIN and its essentiality in baker's yeast [27]. However, node degree alone seems to be a poor measure of topology around a node in the network, since the correlation failed on newer PINs [28, 86] and it appears to hold only for literature-curated [87] and smaller in scope Y2H [2] PINs. Note, however, that this might be due to the fact that these data sets are biased toward essential proteins [88]. We examined similar simple correlations between connectivity in a PIN and protein function [89].

Considering high node connectivity as a good measure of topological positioning of a protein in a PIN led to another controversy [32–35]. ''Hubs'' in a PIN have been studied in the context of expression correlation, co-localization, evolutionary rates, and structural perturbation of a PIN upon deletion. Based on these, a distinction between ''party'' hubs and ''date'' hubs described above has been proposed [32, 34]. However, the results could not be reproduced on literature-curated PIN data sets [33, 35]. Hence, in addition to the degree being a weak measure of network topology, the controversy also may have resulted from biases that different techniques for PIN construction impose on PIN topologies [66, 90]. An example of a bias that impacts PIN topology is the one that exists in PINs resulting from AP/MS screens: the interactions in these PINs are typically modeled by using either the ''matrix'' (all proteins associated in the pulled-down protein complex are assumed to directly interact), or the ''spoke'' (the bait protein of the pulled-down complex is assumed to interact with all associated prey proteins, but no other interactions in the complex are assumed) model [24, 91] and the choice of a model has been shown to have an impact on the PIN topology [23].

Because of these controversies linked to using overly simple measures of network topology, such as node degrees, a series of more involved graph-theoretic methods have been developed. Some were based on the assumption that proteins that are closer in a PIN are more likely to have similar function [92–94], others tried to minimize the number of PPIs among different functional categories in a PIN [95, 96], or used cut-based and network flow-based methods [97]. Also, functional homogeneity of groups of proteins that show some type of ''clustering'' or ''coherence'' in a PIN has been examined [11, 89, 98–102]. Graphlet degree vectors (described in *Network alignment* above) have also been used to isolate the topological positions of proteins in PINs and relate them to protein function and involvement in disease [82, 103, 104].

Similarly, human PINs have been examined for topology-disease associations. Again, when only node degrees were used to measure topology, a discrepancy was observed in the sense that some groups reported that genes (and proteins) involved in disease tend to have high degrees in PINs [105, 106], while others contradicted that conclusion [107]. Apart from this, general observations were that disease-causing proteins are closer together and are centrally positioned within the PIN [105–107]. However, these results might be biased, since disease-causing proteins may exhibit these properties simply because they have been better studied than nondisease proteins [85]. More constraining measures of network topology, such as graphlet degree vectors, were also utilized and revealed that topological similarities of known regulators of melanogenesis in the human PIN can be used to predict novel melanogenesis regulators involved in cancer [104]. Also, by simulating propagation of an abstract functional ''flow'' (where ''function'' means known involvement of a protein in a disease) from disease-causing proteins to their neighbors in the PIN was used to score the strength of disease association of proteins and protein complexes [108]; note that this has nothing to do with network dynamics, it is a method applied on a (static) PIN to rank and predict involvement of a protein in a disease. An approach that combines different data sets with PINs to predict cancer genes has also been proposed [109].

A general conclusion is that the relationship between network topology and biological function and involvement in disease is far from being random, even though we are currently not capable of providing complete mathematical characterizations of those relationships.

Outlook

The nascent research area of network biology is already in turmoil. It gathers researchers from various disciplines who have different levels of biological and quantitative understanding. Hence, the use of simple measures of network topologies, such as node degrees, and modeling of simple biological phenomena are preferred. However, as demonstrated above, such practices result in inconclusive or controversial results clearly demonstrating the need for developing better algorithmic and mathematical tools. Discouraged by such controversies, some sub-communities have started developing aversions toward ''overly complicated'' methods and new models. Such positions could potentially lead to the emergence of overly simplistic doctrines that could further hinder this emerging research area. For example, widely publicized use of simple computational techniques, such as degree distributions, on early and noisy PINs has contributed to the propensity toward SF-centric modeling of complex biological networks [31], which has since been shown to be far too simplistic a model for such networks.

Another example involves the genome-centric view of biological systems. Analyses of genetic sequence data have certainly revolutionized our view of biology. However, they have not provided us with complete understanding of biological systems and we should keep being open toward new scientific horizons. In particular, even though PIN data are new and currently noisy and incomplete, and even though most mathematical methods currently applied to them are rather primitive, there is already evidence documenting that they can reveal biological information that could not have been inferred from genetic sequence, at least not by using the currently available sequence analysis tools. Despite these results, some members of the community have kept questioning the value of the network data with unsubstantiated claims challenging relatedness of network topology and biological function. At a recent meeting, this led to a publicly asked question of whether the community ''should keep analyzing PIN data''. Needless to say, the danger is that such questions would lead to scientific censorship that the area might already be starting to experience. For instance, studies showing that PIN data reveal biological information that cannot be obtained from genetic sequence can be viewed as being wrong, since they are not in agreement with the well-established sequence-based beliefs. Conversely, if PIN-based studies are in agreement with what sequences tell us, then they are often regarded as useless, since they only confirm what we can get from sequences. Hence, PIN analyses are bound to lose in such an unfairly played game. Moreover, such views might have negative impacts on the availability of funding resources needed for continued data collection, as well as for the accompanying development of reliable computational and modeling methods. Both genome analysis and network topology analysis are major domains of modern bioinformatics, and playing either one against each other, which happens sometimes at bioinformatics meetings, is akin to getting lost in local politics when there is an external threat. The debate between genome and PIN analysis is, when put in larger perspective of new challenges in systems biology, after all quite marginal.

Finally, there is somewhat of a clash within the computational research community. Even though finding statistical patterns in the data has surely proved its value, such approaches do not provide mechanistic descriptions of how things actually work. Hence, despite the prevalence and ease of use of statistical tools, we should keep searching for descriptive models that would provide understanding and enable reproduction of the phenomena.

Acknowledgments

Funding was received from NSF CAREER IIS-0644424 and NSF CDI OIA-1028394 grants.

References

1. **Ito T**, **Tashiro K**, **Muta S**, **Ozawa R**, et al. 2000. Toward a protein-protein interaction map of the budding yeast: A comprehensive system to examine two-hybrid interactions in all possible combinations between the yeast proteins. *Proc Natl Acad Sci USA* **97**: 1143–7.

2. **Uetz P**, **Giot L**, **Cagney G**, **Mansfield TA**, et al. 2000. A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*. *Nature* **403**: 623–7.
3. **Li S**, **Armstrong C**, **Bertin N**, **Ge H**, et al. 2004. A map of the interactome network of the metazoan *C. elegans*. *Science* **303**: 540–3.
4. **Giot L**, **Bader J**, **Brouwer C**, **Chaudhuri A**, et al. 2003. A protein interaction map of *Drosophila melanogaster*. *Science* **302**: 1727–36.
5. **Stelzl U**, **Worm U**, **Lalowski M**, **Haenig C**, et al. 2005. A human protein-protein interaction network: A resource for annotating the proteome. *Cell* **122**: 957–68.
6. **Rual J-F**, **Venkatesan K**, **Hao T**, **Hirozane-Kishikawa T**, et al. 2005. Towards a proteome-scale map of the human protein-protein interaction network. *Nature* **437**: 1173–8.
7. **Simonis N**, **Rual J-F**, **Carvunis A-R**, **Tasan M**, et al. 2009. Empirically controlled mapping of the *Caenorhabditis elegans* protein-protein interactome network. *Nat Methods* **6**: 47–54.
8. **Gavin AC**, **Bosche M**, **Krause R**, **Grandi P**, et al. 2002. Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* **415**: 141–7.
9. **Ho Y**, **Gruhler A**, **Heilbut A**, **Bader GD**, et al. 2002. Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. *Nature* **415**: 180–3.
10. **Gavin A**, **Aloy P**, **Grandi P**, **Krause R**, et al. 2006. Proteome survey reveals modularity of the yeast cell machinery. *Nature* **440**: 631–6.
11. **Krogan N**, **Cagney G**, **Yu H**, **Zhong G**, et al. 2006. Global landscape of protein complexes in the yeast *Saccharomyces cerevisiae*. *Nature* **440**: 637–43.
12. **Tong AHY**, **Lesage G**, **Bader GD**, **Ding H**, et al. 2004. Global mapping of the yeast genetic interaction network. *Science* **303**: 808–13.
13. **Rain J-D**, **Selig L**, **De Reuse H**, **Battaglia V**, et al. 2001. The protein-protein interaction map of *Helicobacter pylori*. *Nature* **409**: 211–5.
14. **Parrish JR**, **Yu J**, **Liu G**, **Hines JA**, et al. 2007. A proteome-wide protein interaction map for *Campylobacter jejuni*. *Genome Biol* **8**: R130.
15. **LaCount DJ**, **Vignali M**, **Chettier R**, **Phansalkar A**, et al. 2005. A protein interaction network of the malaria parasite *Plasmodium falciparum*. *Nature* **438**: 103–7.
16. **Chatr-aryamontri A**, **Ceol A**, **Peluso D**, **Nardozza A**, et al. 2009. Virusmint: A viral protein interaction database. *Nucleic Acids Res* **37**: D669–D673.
17. **von Mering C**, **Krause R**, **Snel B**, **Cornell M**, et al. 2002. Comparative assessment of large-scale data sets of protein-protein interactions. *Nature* **417**: 399–403.
18. **de Silva E**, **Stumpf M.** 2005. Complex networks and simple models in biology. *J R Soc Interface* **2**: 419–30.
19. **Stumpf MPH**, **Wiuf C**, **May RM.** 2005. Subnets of scale-free networks are not scale free: Sampling properties of networks. *Proc Natl Acad Sci USA* **102**: 4221–4.
20. **Han JDH**, **Dupuy D**, **Bertin N**, **Cusick ME**, et al. 2005. Effect of sampling on topology predictions of protein-protein interaction networks. *Nat Biotechnol* **23**: 839–44.
21. **de Silva E**, **Thorne T**, **Ingram P**, **Agrafioti I**, et al. 2006. The effects of incomplete protein interaction data on structural and evolutionary inferences. *BMC Biol* **4**: 1–13.
22. **Collins S**, **Kemmeren P**, **Zhao X-C**, **Greenblatt J**, et al. 2007. Toward a comprehensive atlas of the physical interactome of *Saccharomyces cerevisiae*. *Mol Cell Proteomics* **6**: 439–50.
23. **Hakes L**, **Pinney J**, **Robertson DL**, **Lovell SC.** 2008. Protein-protein interaction networks and biology–what's the connection? *Nat Biotechnol* **26**: 69–72.
24. **Wodak S**, **Pu S**, **Vlasblom J**, **Seraphin B.** 2009. Challenges and rewards of interaction proteomics. *Mol Cell Proteomics* **8**: 3–18.
25. **Jeong H**, **Tombor B**, **Albert R**, **Oltvai ZN**, et al. 2000. The large-scale organization of metabolic networks. *Nature* **407**: 651–4.
26. **Tanaka R.** 2005. Scale-rich metabolic networks. *Phys Rev Lett* **94**: 168101.
27. **Jeong H**, **Mason SP**, **Barabási A-L**, **Oltvai ZN.** 2001. Lethality and centrality in protein networks. *Nature* **411**: 41–2.
28. **Coulomb S**, **Bauer M**, **Bernard D**, **Marsolier-Kergoat M-C.** 2005. Gene essentiality and the topology of protein interaction networks. *Proc Biol Sci B* **272**: 1721–5.
29. **Song C**, **Havlin S**, **Makse HA.** 2005. Self-similarity of complex networks. *Nature* **433**: 392–5.
30. **Itzkovitz S**, **Levitt R**, **Kashtan N**, **Milo R**, et al. 2005. Coarse graining and self dissimilarity of complex networks. *Phys Rev E* **71**: 016127.
31. **Keller EF.** 2005. Revisiting ''scale-free'' networks. *BioEssays* **27**: 11060–8.
32. **Han J-D**, **Bertin N**, **Hao T**, **Goldberg DS**, et al. 2004. Evidence for dynamically organized modularity in the yeast protein-protein interaction network. *Nature* **430**: 88–93.
33. **Batada NN**, **Reguly T**, **Breitkreutz A**, **Boucher L**, et al. 2006. Stratus not altocumulus: A new view of the yeast protein interaction network. *PLoS Biol* **4**: e317.
34. **Bertin N**, **Simonis N**, **Dupuy D**, **Cusick ME**, et al. 2007. Confirmation of organized modularity in the yeast interactome. *PLoS Biol* **5**: e153.
35. **Batada NN**, **Reguly T**, **Breitkreutz A**, **Boucher L**, et al. 2007. Still stratus not altocumulus: Further evidence against the date/party hub distinction. *PLoS Biol* **5**: e154.
36. **Huang S.** 2004. Back to the biology in systems biology: What can we learn from biomolecular networks? *Brief Funct Genomic Proteomic* **2**: 279–97.
37. **Gunawardena J.** 2010. Models in computational systems biology. In Lodhi HM, Muggleton SH ed; *Elements of Computational Systems Biology*, Wiley. p 65–90.
38. **Keller EF.** 2000. Models of and models for: Theory and practice in contemporary biology. *Philos Sci* **67**: S72–S86.
39. **Dougherty ER**, **Braga-Neto U.** 2006. Epistemology of computational biology: Mathematical models and experimental prediction as the basis of their validity. *J Biol Sys* **14**: 65–90.
40. **Lappe M**, **Holm L.** 2004. Unraveling protein interaction networks with near-optimal efficiency. *Nat Biotechnol* **22**: 98–103.
41. **Pržulj N**, **Corneil DG**, **Jurisica I**. 2006. Efficient estimation of graphlet frequency distributions in protein-protein interaction networks. *Bioinformatics* **22**: 974–80.
42. **Kuchaiev O**, **Rasajski M**, **Higham D**, **Pržulj N.** 2009. Geometric denoising of protein-protein interaction networks. *PLoS Comput Biol* **5**: e1000454.
43. **Newman MEJ.** 2003. The structure and function of complex networks. *SIAM Rev* **45**: 167–256.
44. **Barabási A-L**, **Albert R.** 1999. Emergence of scaling in random networks. *Science* **286**: 509–12.
45. **Pržulj N**, **Corneil DG**, **Jurisica I.** 2004. Modeling interactome: Scale-free or geometric? *Bioinformatics* **20**: 3508–15.
46. **Li L**, **Alderson D**, **Tanaka R**, **Doyle JC**, et al. 2005. Towards a theory of scale-free graphs: Definition, properties, and implications. *Internet Math* **4**: 431–523.
47. **Watts DJ**, **Strogatz SH.** 1998. Collective dynamics of 'small-world' networks. *Nature* **393**: 440–2.
48. **Milo R**, **Shen-Orr SS**, **Itzkovitz S**, **Kashtan N**, et al. 2002. Network motifs: Simple building blocks of complex networks. *Science* **298**: 824–7.
49. **Shen-Orr SS**, **Milo R**, **Mangan S**, **Alon U.** 2002. Network motifs in the transcriptional regulation network of *Escherichia coli*. *Nat Genet* **31**: 64–8.
50. **Milo R**, **Itzkovitz S**, **Kashtan N**, **Levitt R**, et al. 2004. Superfamilies of evolved and designed networks. *Science* **303**: 1538–42.
51. **Alon U.** 2007. Network motifs: Theory and experimental approaches. *Nat Rev Genet* **8**: 450–61.
52. **Artzy-Randrup Y**, **Fleishman SJ**, **Ben-Tal N**, **Stone L.** 2004. Comment on ''Network motifs: Simple building blocks of complex networks'' and ''Superfamilies of evolved and designed networks''. *Science* **305**: 1107.
53. **Brandstadt A**, **Van Bang L**, **Spinrad JP.** 1999. *Graph classes: A survey*. Philadelphia, PA 19104-688: SIAM Monographs on Discrete Mathematics and Applications.
54. **Pržulj N.** 2007. Biological network comparison using graphlet degree distribution. *Bioinformatics* **23**: e177–183.
55. **Erdös P**, **Rényi A.** 1959. On random graphs. *Publicationes Mathematicae* **6**: 290–7.
56. **Bollobas B.** 1985. *Random Graphs*. Academic, London.
57. **Newman MEJ**, **Strogatz SH**, **Watts DJ.** 2001. Random graphs with arbitrary degree distributions and their applications. *Phys Rev E* **64**: 026118–1.
58. **Aiello W**, **Chung F**, **Lu L.** 2001. A random graph model for power law graphs. *Exp Math* **10**: 53–66.
59. **Wagner A.** 2003. How the global structure of protein interaction networks evolves. *Proc Biol Sci* **270**: 457–66.
60. **Pastor-Satorras R**, **Smith E**, **Sole RV.** 2003. Evolving protein interaction networks through gene duplication. *J Theor Biol* **222**: 199–210.
61. **Kuchaiev O**, **Pržulj N.** 2009. Learning the structure of protein-protein interaction networks. *2009 Pacific Symposium on Biocomputing (PSB)*.
62. **Pržulj N**, **Kuchaiev O**, **Stevanovic A**, **Hayes W.** 2010. Geometric evolutionary dynamics of protein interaction networks. *2010 Pacific Symposium on Biocomputing (PSB).*

63. **Memisević V**, **Milenković T**, **Pržulj N.** 2010. An integrative approach to modeling biological networks. *J Integr Bioinform* **7**: 120.
64. **Higham D**, **Rašajski M**, **Pržulj N**. 2008. Fitting a geometric graph to a protein-protein interaction network. *Bioinformatics* **24**: 1093–9.
65. **Sharan R**, **Ideker T**. 2006. Modeling cellular machinery through biological network comparison. *Nat Biotechnol* **24**: 427–33.
66. **Venkatesan K**, **Rual J-F**, **Vazquez A**, **Stelzl U**, et al. 2009. An empirical framework for binary interactome mapping. *Nat Methods* **6**: 83–90.
67. **Kelley BP**, **Bingbing Y**, **Lewitter F**, **Sharan R**, et al. 2004. Path-BLAST: A tool for alignment of protein interaction networks. *Nucleic Acids Res* **32**: 83–8.
68. **Berg J**, **Lassig M.** 2004. Local graph alignment and motif search in biological networks. *Proc Natl Acad Sci USA* **101**: 14689–94.
69. **Flannick J**, **Novak A**, **Balaji S**, **Harley H**, et al. 2006. Graemlin general and robust alignment of multiple large interaction networks. *Genome Res* **16**: 1169–81.
70. **Berg J**, **Lassig M.** 2006. Cross-species analysis of biological networks by Bayesian alignment. *Proc Natl Acad Sci USA* **103**: 10967–72.
71. **Sharan R**, **Suthram S**, **Kelley RM**, **Kuhn T**, et al. 2005. Conserved patterns of protein interaction in multiple species. *Proc Natl Acad Sci USA* **102**: 1974–9.
72. **Koyutürk M**, **Kim Y**, **Topkara U**, **Subramaniam S**, et al. 2006. Pairwise alignment of protein interaction networks. *J Comput Biol* **13**: 182–99.
73. **Singh R**, **Xu J**, **Berger B.** 2007. Pairwise global alignment of protein interaction networks by matching neighborhood topology. In *Research in Computational Molecular Biology*. Springer. p 16–31.
74. **Flannick J**, **Novak AF**, **Do CB**, **Srinivasan BS**, et al. 2008. Automatic parameter learning for multiple network alignment. *Proceedings of RECOMB*, 214–31.
75. **Singh R**, **Xu J**, **Berger B**. 2008. Global alignment of multiple protein interaction networks. *2008 Pacific Symposium on Biocomputing (PSB)*, p. 303–314.
76. **Zaslavskiy M**, **Bach F**, **Vert JP.** 2009. Global alignment of protein-protein interaction networks by graph matching methods. *Bioinformatics* **25**: i259–267.
77. **Liao C-S**, **Lu K**, **Baym M**, **Singh R**, et al. 2009. IsoRankN: Spectral methods for global alignment of multiple protein networks. *Bioinformatics* **25**: i253–258.
78. **Kuchaiev O**, **Milenkovic T**, **Memisevic V**, **Hayes W**, et al. 2010. Topological network alignment uncovers biological function and phylogeny. *J R Soc Interface* **7**: 1341–54.
79. **Milenkovic T**, **Ng WL**, **Hayes W**, **Pržulj N.** 2010. Optimal network alignment with graphlet degree vectors. *Cancer Inform* **9**: 121–37.
80. **Altschul SF**, **Gish W**, **Miller W**, **Lipman DJ.** 1990. Basic local alignment search tool. *J Mol Biol* **215**: 403–10.
81. **Bruckner S**, **Huffner F**, **Karp RM**, **Shamir R**, et al. 2009. Topology-free querying of protein interaction networks. *Proceedings of RECOMB 2009* 5541 of LNCS: 74–89.
82. **Milenković T**, **Pržulj N.** 2008. Uncovering biological network function via graphlet degree signatures. *Cancer Inform* **6**: 257–73.
83. **Memisević V**, **Milenković T**, **Pržulj N.** 2010. Complementarity of network and sequence information in homologous proteins. *J Integr Bioinform* **7**: 135.
84. **Sharan R**, **Ulitsky I**, **Shamir R.** 2007. Network-based prediction of protein function. *Mol Syst Biol* **3**: 1–13.
85. **Sharan R**, **Ideker T.** 2008. Protein networks in disease. *Genome Res* **18**: 644–52.
86. **Yu H**, **Braun P**, **Yildirim MA**, **Lemmens I**, et al. 2008. High-quality binary protein interaction map of the yeast interactome networks. *Science* **322**: 104–10.
87. **Reguly T**, **Breitkreutz A**, **Boucher L**, **Breitkreutz BJ**, et al. 2006. Comprehensive curation and analysis of global interaction networks in *Saccharomyces cerevisiae*. *J Biol* **5**: 11.
88. **Ratmann O**, **Wiuf C**, **Pinney JW.** 2009. From evidence to inference: Probing the evolution of protein interaction networks. *HFSP J* **3**: 290–306.
89. **Pržulj N**, **Wigle D**, **Jurisica I.** 2004. Functional topology in a network of protein interactions. *Bioinformatics* **20**: 340–8.
90. **Cusick ME**, **Yu H**, **Smolyar A**, **Venkatesan K**, et al. 2009. Literature curated protein interaction datasets. *Nat Methods* **6**: 39–46.
91. **Bader GD**, **Hogue CWV.** 2002. Analyzing yeast protein-protein interaction data obtained from different sources. *Nat Biotechnol* **20**: 991–7.
92. **Hishigaki H**, **Nakai K**, **Ono T**, **Tanigami A**, et al. 2001. Assessment of prediction accuracy of protein function from protein-protein interaction data. *Yeast* **18**: 523–31.
93. **Schwikowski B**, **Uetz P**, **Fields A**. 2000. A network of protein-protein interactions in yeast. *Nat Biotechnol* **18**: 1257–61.
94. **Chua H**, **Sung W**, **Wong L.** 2006. Exploiting indirect neighbours and topological weight to predict protein function from protein-protein interactions. *Bioinformatics* **22**: 1623–30.
95. **Vazquez A**, **Flammini A**, **Maritan A**, **Vespignani A.** 2003. Global protein function prediction from protein-protein interaction networks. *Nat Biotechnol* **21**: 697–700.
96. **Karaoz U**, **Murali T**, **Letovsky S**, **Zheng Y**, et al. 2004. Whole-genome annotation by using evidence integration in functional-linkage networks. *Proc Natl Acad Sci USA* **101**: 2888–93.
97. **Nabieva E**, **Jim K**, **Agarwal A**, **Chazelle B**, et al. 2005. Whole-proteome prediction of protein function via graph-theoretic analysis of interaction maps. *Bioinformatics* **21**: i302–310.
98. **Bader GD**, **Hogue CWV.** 2003. An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinformatics* **4**: 2.
99. **Sharan R**, **Ideker T**, **Kelley BP**, **Shamir R**, et al. 2004. Identification of protein complexes by comparative analysis of yeast and bacterial protein interaction data. *Proceedings of RECOMB*, **2004**: 282–9.
100. **King AD**, **Pržulj N**, **Jurisica I.** 2004. Protein complex prediction via cost-based clustering. *Bioinformatics* **20**: 3013–20.
101. **Jung SH**, **Hyun B**, **Jang W.-H**, **Hur H.-Y**, et al. 2010. Protein complex prediction based on simultaneous protein interaction network. *Bioinformatics* **26**: 385–91.
102. **Kaake RM**, **Milenković T**, **Pržulj N**, **Kaiser P**, et al. 2010. Characterization of cell cycle specific protein interaction networks of the yeast 26s proteasome complex by the QTAX strategy. *J Proteome Res* **9**: 2016–29.
103. **Guerrero C**, **Milenković T**, **Pržulj N**, **Kaiser P**, et al. 2008. Characterization of the proteasome interaction network using a QTAX-based tag-team strategy and protein interaction network analysis. *Proc Natl Acad Sci USA* **105**: 13333–8.
104. **Milenković T**, **Memisević V**, **Ganesan AK**, **Pržulj N.** 2010. Systems-level cancer gene identification from protein interaction network topology applied to melanogenesis-related interaction networks. *J R Soc Interface* **7**: 423–37.
105. **Wachi S**, **Yoneda K**, **Wu R.** 2005. Interactome-transcriptome analysis reveals the high centrality of genes differentially expressed in lung cancer tissues. *Bioinformatics* **21**: 4205–8.
106. **Jonsson PF**, **Bates PA.** 2006. Lobal topological features of cancer proteins in the human interactome. *Bioinformatics* **22**: 2291–7.
107. **Goh KI**, **Cusick ME**, **Valle D**, **Childs B**, et al. 2007. The human disease network. *Proc Natl Acad Sci USA* **104**: 8685–90.
108. **Vanunu O**, **Magger O**, **Ruppin E**, **Shlomi T**, et al. 2010. Associating genes and protein complexes with disease via network propagation. *PLoS Comput Biol* **6**: e1000641.
109. **Aragues R**, **Sander C**, **Oliva B.** 2008. Predicting cancer involvement of genes from heterogeneous data. *BMC Bioinformatics* **9**: 172.