Contents

Chapter 1

Computational methods for analyzing and modeling biological networks

Nataša Pržulj and Tijana Milenković
Department of Computer Science
University of California, Irvine
Irvine, CA 92697-3435, USA
Phone: +1-949-824-7312
Fax: +1-949-824-4056
E-mail: natasha@ics.uci.edu; tmilenko@ics.uci.edu

Large networks have been used to model and analyze many real world phenomena including biomolecular systems. Although graph theoretic modeling in systems biology is still in its infancy, network-based analyses of cellular systems have already been used to address many important biological questions. We survey methods for analyzing, modeling, and comparing large biological networks that have given insights into biological function and disease. We focus on protein-protein interaction (PPI) networks, since proteins are important macromolecules of life and understanding the collective behavior of their interactions is

of importance. After discussing the major challenges in the field, we survey network properties, measures used to characterize and compare complex networks. We also give an overview of network models for PPI networks. We discuss to what extent each of the models fits PPI networks and demonstrate that geometric random graphs provide the best fit. We also provide an overview of available network alignment methods and discuss their potential in predicting function of individual proteins, protein complexes, and larger cellular machines. Finally, we present a method that establishes a link between the topological surrounding of a node in a PPI network and its involvement in performing biological functions and in disease.

1.1 Motivation

Recent technological advances in experimental biology have yielded large amounts of biological network data. Many other real-world phenomena have also been described in terms of large *networks* (also called *graphs*), such as various types of social and technological networks. Thus, understanding these complex phenomena has become an important scientific problem that has lead to intensive research in network modeling and analyses. The hope is that utilizing such systems-level approaches to analyzing and modeling complex biological systems will provide insights into biological function, evolution, and disease.

1.1.1 Types of biological networks and availability of data sets

Biological networks come in a variety of forms. They include protein-protein interaction networks, transcriptional regulation networks, metabolic networks, signal transduction networks, protein structure networks, and networks summarizing neuronal connectivities. These networks differ in whether their nodes represent biomolecules such as genes, proteins or metabolites, and whether their edges indicate functional, physical or chemical interactions between the corresponding biomolecules. Studying biological networks at these various granularities could provide valuable insight into the inner working of cells and might lead to

important discoveries about complex diseases. However, it is the proteins that execute the genetic program and that carry almost all biological processes. Thus, we primarily focus on protein-protein interaction (PPI) networks, in which nodes correspond to proteins, and undirected edges represent physical interactions amongst them. Nevertheless, methods and tools presented in this chapter can also be easily applied to other types of biological networks.

There exist a variety of methods for obtaining these rich biological network data, such as yeast 2-hybrid (Y2H) screening, protein complex purification methods using mass-spectrometry (e.g., tandem affinity purification (TAP) and high-throughput mass-spectrometric protein complex identification (HMS-PCI)), correlated messenger RNA (mRNA) expression profiles, genetic interactions, or in silico interaction prediction methods. Y2H and mass spectrometry techniques aim to detect physical binding between proteins, whereas genetic interactions, mRNA coexpression, and in silico methods seek to predict functional associations, for example, between a transcriptional regulator and the pathway it controls. For a more detailed survey of these biochemical methods, see (1).

Numerous datasets resulting from small- and large-scale screens are now publicly available in several databases including: Saccharomyces Genome Database (SGD)[1], Munich Information Center for Protein Sequences (MIPS)[2], the Database of Interacting Proteins (DIP)[3], the Molecular Interactions Database (MINT)[4], the Online Predicted Human Interaction Database (OPHID)[5], Human Protein Reference Database (HPRD)[6], and the Biological General Repository for Interaction Datasets (BioGRID)[7]. For a more detailed survey of these databases, see (1).

[1]http://www.yeastgenome.org/
[2]http://mips.gsf.de/
[3]http://dip.doe-mbi.ucla.edu/
[4]http://mint.bio.uniroma2.it/mint/
[5]http://ophid.utoronto.ca/ophid/
[6]http://www.hprd.org/
[7]http://www.thebiogrid.org/

1.1.2 Major challenges

Major challenges when studying biological networks include network analyses, comparisons, and modeling, all aimed at discovering a relationship between network topology on one side and biological function, disease, and evolution on the other.

Since proteins are essential macromolecules of life, we need to understand their function and their role in disease. However, the number of functionally unclassified proteins is large even for simple and well studied organisms such as baker's yeast. Moreover, it is still unclear in what cellular states serious diseases, such as cancer, occur. Methods for determining protein function and identifying disease genes have shifted their focus from targeting specific proteins based solely on sequence homology to analyses of the entire proteome based on PPI networks (2). Since proteins interact to perform a certain function, PPI networks by definition reflect the interconnected nature of biological processes. Therefore, analyzing structural properties of PPI networks may provide useful clues about the biological function of individual proteins, protein complexes, pathways they participate in, and larger subcellular machines (2; 3). Additionally, recent studies have been investigating associations between diseases and network topology in PPI networks and have shown that disease genes share common topological properties (4; 5; 3). Finding this relationship between PPI network topology and biological function and disease remains one of the most challenging problems in the post-genomic era.

Analogous to genetic sequence comparison, comparing large cellular networks will revolutionize biological understanding. However, comparing large networks is computationally infeasible due to NP-completeness of the underlying subgraph isomorphism problem. Note that even if the subgraph isomorphism was feasible, it would not find a practical application in biological network comparisons, since biological networks are extremely unlikely to be isomorphic. Thus, large network analyses and comparisons rely on heuristics, commonly called *network properties*. These properties are roughly categorized into *global* and *local*. The most widely used global properties are the *degree distribution*, the *average clustering coefficient*, the *clustering spectrum*, the *average diameter* and the *spectrum of shortest*

path lengths (6). Local properties include *network motifs*, small over-represented subgraphs (7), and two measures based on *graphlets*, small induced subgraphs of large networks: the *relative graphlet frequency distance (RGF-distance)*, which compares the frequencies of the appearance of graphlets in two networks (8), and the *graphlet degree distribution agreement (GDD-agreement)*, which is a graphlet-based generalization of the degree distribution (9). These properties have been used to compare biological networks against model networks and to find well-fitting network models for biological networks (8; 9; 10), as well as to suggest biological function of proteins in PPI networks (2). The properties of models have further been exploited to guide biological experiments and discover new biological features, as well as to propose efficient heuristic strategies in many domains, including time- and cost-optimal detection of the human interactome.

Modeling biological networks is of crucial importance for any computational study of these networks. Only a well-fitting network model that precisely reproduces the network structure and laws through which the network has emerged can enable us to understand and replicate the biological processes and the underlying complex evolutionary mechanisms in the cell. Various network models have been proposed for real-world biological networks. Starting with Erdös-Rényi random graphs (11), network models have progressed through a series of versions designed to match certain properties of real-world networks. Examples include random graphs that match the degree distribution of the data (12), network growth models that produce networks with scale-free degree distributions (13) or small network diameters (14), geometric random graphs (15), or networks that reproduce some biological and topological properties of real biological networks (e.g., stickiness model (10)). An open-source software tool called *GraphCrunch* (16) implements the latest research on biological network models and properties and compares real-world networks against a variety of network models with respect to a wide range of network properties.

1.2 Network properties

Global network properties give an overall view of a network. The most commonly used global network properties are the degree distribution, the average network diameter, the spectrum of shortest path lengths, the average clustering coefficient, and the clustering spectrum. The *degree* of a node is the number of edges incident to the node. The *degree distribution, P(k)*, describes the probability that a node has degree k. The smallest number of links that have to be traversed to get from a node x to a node y in a network is called the *distance* between nodes x and y and a path through the network that achieves this distance is called the *shortest path* between nodes x and y. The average of shortest path lengths over all pairs of nodes in a network is called the *average network diameter*. The *spectrum of shortest path lengths* is the distribution of shortest path lengths between all pairs of nodes in a network. The *clustering coefficient* of a node z in a network, C_z, is defined as the probability that two nodes x and y which are connected to the node z are themselves connected. The average of C_z over all nodes z of a network is the *average clustering coefficient, C,* of the network; it measures the tendency of the network to form highly interconnected regions called clusters. The distribution of the average clustering coefficients of all nodes of degree k in a network is the *clustering spectrum, C(k)*.

However, global network properties are not detailed enough to capture complex topological characteristics of real-world networks. Network properties should encompass large number of constraints, in order to reduce degrees of freedom in which networks being compared can vary. Thus, more constraining measures of local network structure have been introduced. Local properties include *network motifs*, small over-represented subgraphs (7), and highly constraining measures of local structural similarities between two networks: RGF-distance (8) and GDD-agreement (9). RGF-distance and GDD-agreement are based on *graphlets*, small connected non-isomorphic induced subgraphs of large networks (8). Note that graphlets are different from network motifs since they must be induced subgraphs (motifs are partial subgraphs) and since they do not need to be over-represented in the data when compared to "randomized" networks. An *induced subgraph* of a graph G on a subset S

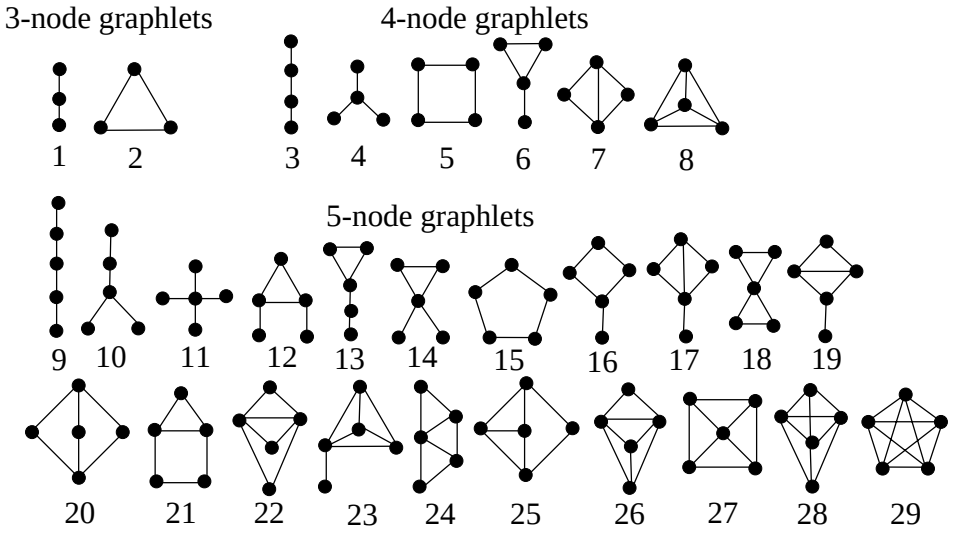

Figure 1.1: All 3-node, 4-node and 5-node graphlets (8).

of nodes of G is obtained by taking S and all edges of G having both end-points in S; *partial subgraphs* are obtained by taking S and some of the edges of G having both end-points in S. Thus, graphlets avoid both main criticisms associated with network motifs: (1) they do not depend on the randomization scheme, since they do not need to be over-represented in the data when compared with randomized networks, and (2) they are not susceptible to over-counting as motifs are, since graphlets are induced while motifs are partial subgraphs. Since the number of graphlets on n nodes increases super-exponentially with n, RGF-distance and GDD-agreement computations are currently based on 3-5-node graphlets (presented in Figure 1.1).

RGF-distance compares the frequencies of the appearance of all 3-5-node graphlets in two networks (see (8) for details). If networks being compared have the same number of nodes and edges, the frequencies of occurrence of the only 1-node graphlet (a node) and the only 2-node graphlet (an edge) are also taken into account by this measure. Thus, RGF-distance encompasses 31 similarity constraints by examining the fit of 31 graphlet frequencies.

GDD-agreement generalizes the notion of the degree distribution to the spectrum of *graphlet degree distributions (GDDs)* in the following way (9). The degree distribution measures the number of nodes of degree k, i.e., the number of nodes "touching" k edges, for each value of k. Note that an edge is the only graphlet with two nodes (graphlet denoted by G_0

in Figure 1.2). GDDs generalize the degree distribution to other graphlets: they measure for each graphlet G_i, $i \in 0, 1,..., 29$, (illustrated in Figure 1.2) the number of nodes "touching" k graphlets G_i at a particular node. A node at which a graphlet is "touched" is topologically relevant, since it allows us to distinguish between nodes "touching", for example, a copy of graphlet G_1 in Figure 1.2 at an end-node, or at the middle node. This is summarized by *automorphism orbits* (or just *orbits*, for brevity), as illustrated in Figure 1.2: for graphlets G_0, G_1, ..., G_{29}, there are 73 different orbits, numerated from 0 to 72 (see (9) for details). For each orbit j, the j^{th} *GDD*, i.e., the distribution of the number of nodes "touching" the corresponding graphlet at orbit j, is measured. Thus, the degree distribution is the 0^{th} GDD. The j^{th} *GDD-agreement* compares the j^{th} GDDs of two networks (see (9) for details). The total GDD-agreement between two networks is the arithmetic or the geometric average of the j^{th} GDD-agreements over all j (henceforth arithmetic and geometric averages are denoted by "amean" and "gmean", respectively). GDD-agreement is scaled to always be between 0 and 1, where 1 means that two networks are identical with respect to this property. By calculating the fit of each of the 73 GDDs of the networks being compared, GDD-agreement encompasses 73 similarity constraints. Furthermore, each of these 73 constraints enforces a similarity of two distributions, additionally restricting the ways in which the networks being compared can differ. (Note that the degree distribution is only one of these 73 constraints.) Therefore, GDD-agreement is a very strong measure of structural similarity between two networks. RGF-distance and GDD-agreement were used to discover a new, well-fitting, geometric random graph model of PPI networks (8; 9) (see Section 1.3.2).

1.3 Network models

1.3.1 Survey of network models

The most commonly used network models for PPI networks include: Erdös-Rényi random graphs (11), random graphs with the same degree distribution as the data (12), scale-free networks (13; 17), geometric random graphs (15), and models incorporating complexities of

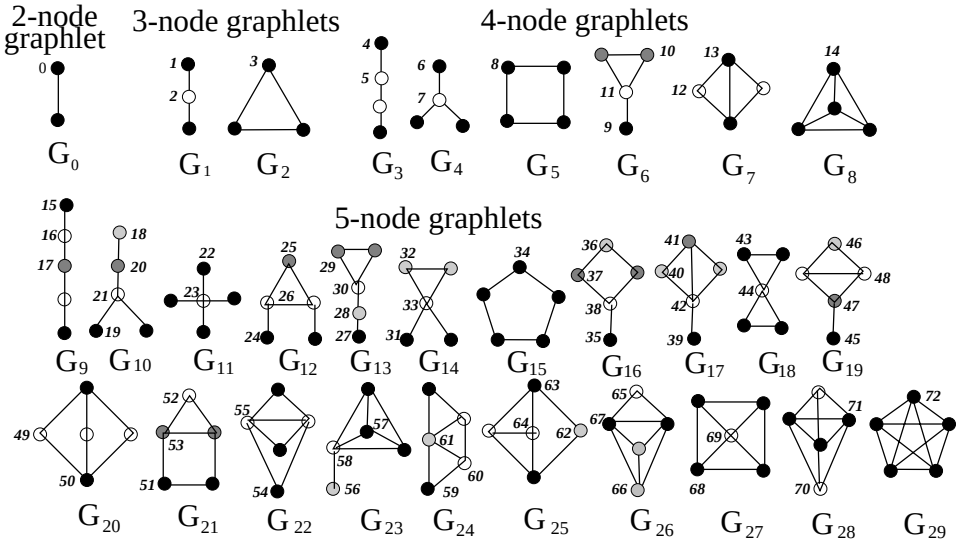

Figure 1.2: The thirty 2-, 3-, 4-, and 5-node graphlets $G_0, G_1, \ldots, G_{29}$ and their automorphism orbits $0, 1, 2, \ldots, 72$. In a graphlet G_i, $i \in \{0, 1, \ldots 29\}$, nodes belonging to the same orbit are of the same shade (9).

binding domains of proteins (e.g., stickiness-index based model networks (10)).

Erdös-Rényi random graphs ("ER") are based on the principle that the probability that there is an edge between any pair of nodes is distributed uniformly at random. Erdös and Rényi have defined several variants of the model. The most commonly studied one is denoted by $G_{n,p}$, where each possible edge in the graph on n nodes is present with probability p and absent with probability $1 - p$. These networks have small diameters, Poisson degree distributions, and low clustering coefficients, and thus do not provide a good fit to real-world PPI networks which typically have small diameters, but power-law degree distributions and high clustering coefficients. Random graphs with the same degree distribution as the data ("ER-DD") capture the degree distribution of a real-world network while leaving all other aspects as in Erdös-Rényi random model. They can be generated by using the "stubs method": the number of "stubs" (to be filled by edges) is assigned to each node in the model network according to the degree distribution of the real-world network; edges are created between pairs of nodes picked at random; after an edge is created, the number of "stubs" left available at the corresponding "end-nodes" of the edge is decreased by one. Thus, these networks preserve the degree distribution and small diameters of the real-world networks.

However, this model also produces networks with low clustering coefficients and thus other network models have been sought. One such example are small-world networks (14). These networks are created from regular ring lattices by random rewiring of a small percentage of their edges. However, although these networks have high clustering coefficients and small diameters, they fail to reproduce power-law degree distributions of real-world networks.

Scale-free networks are characterized by power-law degree distributions. One such model is the Barabási-Albert preferential attachment model ("SF-BA") (13), in which newly added nodes preferentially attach to existing nodes with probability proportional to the degree of the target node. Other variants focused on modeling PPI networks include scale-free network models constructed by mimicking "gene duplications and mutations" (17). In these model networks, connectivity of some nodes is significantly higher than for the other nodes, resulting in power-law degree distribution. Although the average diameter is small, they typically still have low clustering coefficients.

High clustering coefficients of real-world networks are well reproduced by geometric random graphs ("GEO") that are defined as follows: nodes correspond to uniformly randomly distributed points in a metric space and edges are created between pairs of nodes if the corresponding points are close enough in the metric space according to some distance norm (15). For example, 3-dimensional Euclidean boxes ("GEO-3D") and the Euclidean distance norm have been used to model PPI networks (8; 9). Although this model creates networks with high clustering coefficients and small diameters, it still fails to reproduce power-law degree distributions of real-world PPI networks. Instead, geometric random graphs have Poisson degree distribution. However, it has been argued that power-law degree distributions in PPI networks are an artifact of noise present in them.

Finally, "stickiness network model" ("STICKY") is based on stickiness indices, numbers that summarize node connectivities and thus also the complexities of binding domains of proteins in protein-protein interaction (PPI) networks. The probability that there is an edge between two nodes in this network model is directly proportional to the stickiness indices of nodes, i.e., to the degrees of their corresponding proteins in real-world PPI networks (see (10) for details). Networks produced by this model have the expected degree distribution

of a real-world network. Additionally, they mimic well the clustering coefficients and the diameters of real-world networks.

1.3.2 An optimized null model for PPI networks

Modeling bio-chemical networks is a vibrant research area. The choice of an appropriate null model can have important implications for many graph-based analyses of these networks. For example, the use of an adequate null model is vital for structural motif discovery, which requires comparing real-world networks with randomized ones. Using an inappropriate network null model may identify as overrepresented (underrepresented) subgraphs that otherwise would not have been identified. Another example is that a good null model can be used to guide biological experiments in a time- and cost-optimal way and thus minimize the costs of interactome detection by predicting the behavior of a system. Since incorrect models lead to incorrect predictions, it is vital to have as accurate a model as possible.

Geometricity of PPI networks

Also, as new biological network data becomes available, we must ensure that the theoretical models continue to accurately represent the data. The scale-free model has been assumed to provide such a model for PPI networks. However, in the light of new PPI network data, several studies have started questioning the wellness of fit of scale-free network model. For example, Pržulj et al. (8) and Pržulj (9) have used two highly constraining measures of local network structures to compare real-world PPI networks to various network models and have shown compelling evidence that the structure of yeast PPI networks is closer to the geometric random graph model than to the widely accepted scale-free model. Furthermore, Higham et al. (18) have designed a method for embedding networks into a low-dimensional Euclidean space and demonstrated that PPI networks can be embedded and thus have a geometric graph structure (see below).

In search of a well fitting null model for biological networks, one has to consider biological properties of a system being modeled. Geometric random graph model of PPI networks is

biologically motivated. Genes and proteins as their products exist in some highly-dimensional biochemical space. Currently accepted paradigm is that evolution happens through a series of gene duplication and mutation events. Thus, after a parent gene is duplicated, the child gene is at the same position in the biochemical space as the parent and therefore inherits interactions with all interacting partners of the parent. Evolutionary optimization then acts on the child gene to either become obsolete and disappear from the genome, or to mutate distancing itself somewhat from the parent, but preserving some of the parent's interacting partners (due to proximity to the parent in the biochemical space) while also establishing new interactions with other genes (due to the short distance of the mutated child from the parent in the space). Similarly, in geometric random graphs, the closer the nodes are in a metric space, the more interactors they will have in common, and vice-versa. Thus, the superior fit of geometric random graphs to PPI networks over other random models is not surprising.

A well-fitting null model should generate graphs that closely resemble the structure of real networks. This closeness in structure is reflected across a wide range of statistical measures, i.e., network properties. Thus, testing the fit of a model entails comparing model-derived random graphs to real networks according to these measures. Global network properties, such as the degree distribution, may not be detailed enough to capture the complex topological characteristics of PPI networks. The more constraining the measures are, the fewer degrees of freedom exist in which the compared networks can vary. Thus, by using highly constraining measures, such as RGF-distance and GDD-agreement, a better-fitting null model can be found.

RGF-distance was used to compare PPI networks of yeast *S. cerevisiae* and fruitfly *D. melanogaster* to a variety of network models and to show the supremacy of the fit of geometric random graph model to these networks over three other random graph models. Pržulj et al. (8) compared the frequencies of the appearance of all 3-5-node graphlets in these PPI networks with the frequencies of their appearance in four different types of random networks of the same size as the data: ER, ER-DD, SF-BA, and GEO. Furthermore, several variants of the geometric random graphs were used, depending on the dimensionality of the

Euclidean space chosen to generate them: two-dimensional (GEO-2D), three-dimensional (GEO-3D), and four-dimensional (GEO-4D) geometric random graphs. Four real-world PPI networks were analyzed: high-confidence and lower-confidence yeast PPI networks, and high-confidence and low-confidence fruitfly PPI networks. Pržulj et al. (8) computed RGF-distances between these real-world PPI networks and the corresponding ER, ER-DD, SF-BA and GEO random networks. They found that the GEO random networks fitted the data an order of magnitude better than other network models in the higher-confidence PPI networks, and less so (but still better) in the more noisy PPI networks. The only exception was the noisy fruitfly PPI network which exhibited scale-free behavior. It was hypothesized that this behavior of the graphlet frequency parameter was the consequence of a large amount of noise present in this network.

Since currently available PPI data sets are incomplete, i.e., have a large percentage of false negatives or missing interactions, and thus are expected to have higher edge densities, Pržulj et al. also compared the high-confidence yeast PPI network against 3-dimensional geometric random graphs with the same number of nodes, but about three and six times as many edges as the PPI network, respectively. By making the GEO-3D networks corresponding to this PPI network about six times as dense as the PPI network, the closest fit to the PPI network with respect to RGF-distance was observed. Additionally, to address the existence of noise, i.e., false positives, in PPI networks, the high-confidence yeast PPI network was perturbed by randomly adding, deleting and rewiring 10%, 20% and 30% of edges and RGF-distances between the perturbed networks and the PPI network were computed. The study demonstrated that graphlet frequencies were robust to these random perturbations, thus further increasing the confidence in PPI networks having geometric network structure.

Geometric structure of PPI networks has also been confirmed by GDD-agreements between PPI and model networks drawn from several different random graph models (8): ER, ER-DD, SF-BA, and GEO-3D. Several PPI networks of each of the following four eukaryotic organisms were examined: yeast *S. cerevisiae*, frutifly *D. melanogaster*, nematode worm *C. elegans*, and human. The total of fourteen PPI networks originating from different sources, obtained with different interaction detection techniques (such as Y2H, TAP, or HMS-PCI, as

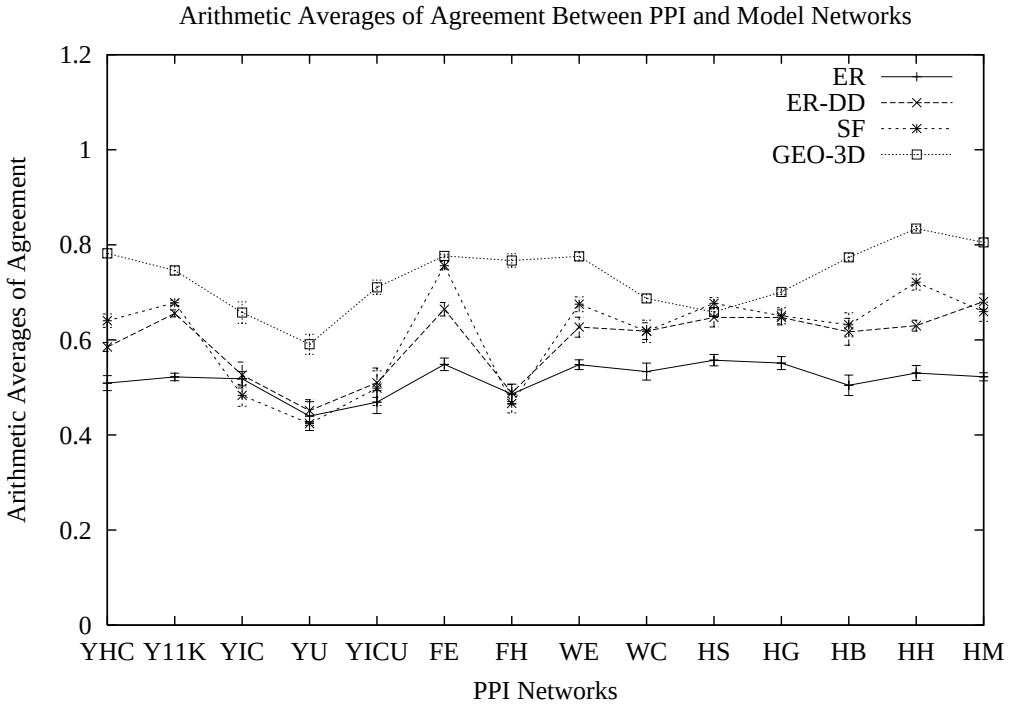

Figure 1.3: GDD-agreements between the fourteen PPI networks of four organisms and their corresponding model networks. Labels on the horisontal axis denote the 14 PPI networks (see (9) for details). Averages of GDD-agreements between 25 model networks and the corresponding PPI network are presented for each random graph model and each PPI network, i.e., at each point in the Figure; the error bar around a point is one standard deviation below and above the point (in some cases, error bars are barely visible, since they are of the size of the point). The figure is taken from (9).

well as human curation), and of different interaction confidence levels were analyzed. GEO-3D network model showed the highest GDD-agreement for all but one of the fourteen PPI networks; for the remaining network, GDD-agreements between GEO-3D, SF, and ER-DD models and the data were about the same (see Figure 1.3).

Additionally, an algorithm that directly tests whether PPI networks are geometric has been proposed (18). It does so by embedding PPI networks into a low dimensional Euclidean space. If a geometric network model fits the PPI network data, then it is expected that PPI networks can be embedded into some space. Geometric random graphs constructed from Euclidean space are chosen as a proof of concept. These graphs in 2-dimensional space are

generated by placing N nodes uniformly at random in the unit square, and by connecting two nodes by an edge if they are within a given Euclidean distance. The 3- and 4-dimensional cases are defined analogously. The task is then to embed the proteins into n-dimensional Euclidean space for $n = 2, 3, 4$, given only their PPI network connectivity information. The algorithm is based on Multi-Dimensional Scaling, with shortest path lengths between protein pairs in a PPI network playing the role of the Euclidean distances between the corresponding nodes embedded in the n-dimensional Euclidian space. After proteins are embedded, a radius r is chosen and each node is connected to all other nodes that are at most at distance r from that node. Each choice of a radius thus corresponds to a different geometric graph. By varying the radius, specificity and sensitivity are measured and the overall goodness of fit is judged by computing the areas under the Receiver Operator Characteristic (ROC) curve, with higher values indicating a better fit.

The algorithm was applied to nineteen real-world PPI networks of yeast, fruitfly, worm, and human obtained from different sources, as well as to artificial networks generated using seven types of random graph models: ER, ER-DD, GEO (GEO-2D, GEO-3D, and GEO-4D), SF-BA, and STICKY. These networks were embedded into 2-dimensional (2D), 3-dimensional (3D), and 4-dimensional (4D) Euclidean space. The resulting areas under the ROC curve (AUCs) were high for all PPI networks. The highest AUC value was obtained for embedding the high-confidence yeast (YHC) PPI network. The authors focused their further analyses on YHC network. Random graphs of the same size as YHC network drawn from the seven network models were embedded into 2D, 3D and 4D space. For geometric random networks, AUCs were very high, with values above 0.9. For non-geometric networks AUCs were below 0.78. Since PPI networks are noisy, to test whether PPI networks had a geometric structure, the authors added noise to GEO-3D networks by randomly rewiring 10%, 20% and 30% of their edges. These rewired networks were then embedded into 2D, 3D and 4D space, and their AUCs were computed. The values of AUCs for the 10% rewired GEO-3D networks were very similar to those for real-world networks. Thus, by embedding networks into low dimensional Euclidean space, the authors performed a direct test of whether PPI networks had a geometric graph structure. The results yielded support to the results of previous

studies and to the hypothesis that the structure of currently available PPI networks was consistent with the structure of noisy geometric graphs (18).

"Stickiness" of PPI networks

Another biologically motivated "stickiness index"-based network model has been proposed for PPI networks (10). It is commonly considered that proteins interact because they share complimentary physical aspects, a concept that is consistent with the underlying biochemistry. These physical aspects are referred to as binding domains (10). Stickiness-index-based network model is based on "stickiness indices" of proteins in PPI networks, where a stickiness index of a protein is a single number that is based on its normalized degree and it summarizes the abundance and popularity of binding domains on the protein. The model assumes that a high degree of a protein implies that the protein has many binding domains and/or its binding domains are commonly involved in interactions. Additionally, the model considers that a pair of proteins is more likely to interact (i.e., share complementary binding domains) if both proteins have high stickiness indices, and less likely to interact if one or both have a low stickiness index. Thus, according to this model, the probability that there exist an edge between two nodes in a random graph is the product of the two stickiness indices of the corresponding proteins in the PPI network (see (10) for details). The resulting model networks are guaranteed to have the expected degree distributions of real-world networks.

To examine the fit of this network model to real-world PPI networks, as well as to compare its fit against the fit of other network models, a variety of global and local network properties were used (10): the degree distribution, clustering coefficient, network diameter, and RGF-distance. In addition to the stickiness-index-based network model (STICKY), model networks were also drawn from the following network models: ER, ER-DD, SF-BA, and GEO-3D. The fit of fourteen real-world PPI networks of four organisms (yeast, fruitfly, worm, and human) to each of these five network models was evaluated with respect of all above mentioned network properties. With respect to RGF-distance, the stickiness model showed an improved fit over all other network models in ten out of fourteen tested PPI networks. It showed as good results as the GEO-3D model in one and was outperformed by

the GEO-3D model in three PPI networks. In addition, this model reproduced well global network properties such as the degree distribution, the clustering coefficients, and the average diameters of PPI networks. Thus, this model using biologically motivated assumptions mentioned above clearly outperforms scale-free network models such as SF-BA and ER-DD that also match the degree distribution of a real-world PPI network.

1.4 Network comparison and alignment

Just as comparative genomics has led to an explosion of knowledge about evolution, biology, and disease, so will comparative proteomics. As more biological network data is becoming available, comparative analyses of these networks across species are proving to be valuable, since such systems biology types of comparisons may lead to exciting discoveries in evolutionary biology. For example, comparing networks of different organisms might provide deeper insights into conservation of proteins, their function, and protein-protein interactions through evolution. Conceptually, network comparison is the process of contrasting two or more interaction networks, representing different species, conditions, interaction types, or time points, aimed at answering some fundamental biological questions.

1.4.1 Types of network comparison methods

Three different types of comparative methods exist (19). The most common methods are *network alignments*. An alignment is achieved by constructing a mapping between the nodes of networks being compared, as well as between the corresponding interactions. In this process, topologically and functionally similar regions of biological networks are discovered. Depending on the properties of mappings, network alignment can be local or global. Most of the research in previous years has been focused on local alignments. With local network alignment algorithms, optimal mappings are chosen independently for each local region of similarity. With global network alignments, one optimal mapping for the entire network is constructed, even though this may imply less perfect alignments in some local regions.

The second type of a method is *network integration*, the process of combining different

networks and encompassing interactions of different types over the same set of elements to study their interrelations. Each type of a network provides an insight into a different slice of biological information, and thus, integrating different network types could provide a more comprehensive picture of the overall biological system under study (19). The major difference from network alignment is the following. Networks to be integrated are defined over the same set of elements and integration is achieved by merging them into a single network with multiple types of interactions, each drawn from one of the original networks. A fundamental problem is then to identify in the merged network functional modules that are supported by interactions of multiple types.

Finally, the third type of comparison is *network querying*, in which a network is searched for subnetworks that are similar to a subnetwork query of interest. Network alignment and integration are focused on *de novo* discovery of biologically significant regions embedded in a network, based on the assumption that regions supported by multiple networks are functionally important. In contrast, network querying searches for a subnetwork that is previously known to be functional. The goal is to identify subnetworks in a given network that are similar to the query. However, network querying tools are still in their infancy since they are currently limited to sparse topologies, such as paths and trees.

1.4.2 Algorithms for network alignment

Here, we focus on network alignment methods that have been applied to PPI networks. Due to the subgraph isomorphism problem, the problem of network alignment is computationally hard and thus heuristic approaches have been sought.

Conceptually, to perform network alignment, a merged representation of the networks being compared, called a network alignment graph, is created (19). In a network alignment graph, the nodes represent sets of "similar" molecules, one from each network, and the links represent conserved molecular interactions across the different networks. There are two core challenges involved in performing network alignment and constructing the network alignment graph. First, a scoring framework that captures the "similarities" between nodes

originating in different networks must be defined. Then, a way to rapidly identify high-scoring alignments (i.e., conserved functional modules) from among the exponentially large set of possible alignments needs to be specified. Due to the computational complexity, a greedy algorithm needs to be used for this purpose. Methods for network alignment differ in these two challenges, depending how they define the similarity scores between protein pairs and what greedy algorithm they use to identify conserved subnetworks.

The problem of network alignment has been approached in different ways and a variety of algorithms have been developed. Unlike the majority of algorithms focusing on pairwise network alignments, newer approaches have tried to address the problem of aligning networks belonging to multiple organisms; note that multiple network alignment represents a challenge since computational complexity increases dramatically with the number of networks. Additionally, instead of performing local alignments, algorithms for global network alignment have emerged. Finally, whereas previous studies have focused on network alignment based solely on biological functional information such as protein sequence similarity, recent studies have been combining the functional information with network topological information (19).

In the most simple case, the similarity of a protein pair, where one protein originates from each of the networks being aligned, is determined solely by their sequence similarity. Then, the top scoring protein pairs, typically found by applying BLAST to perform all-to-all alignment between sequences of proteins originating in different networks, are aligned between the two networks. The most simple network alignment then identifies pairs of interactions in PPI networks, called interologs, involving two proteins in one species and their best sequence matches in another species. However, beyond this simple identification of conserved protein interactions, it is more interesting to identify network *subgraphs* that might have been conserved across species. Algorithms for network alignment are still in their early development, since they are currently limited to identifying conserved pathways or complexes. Methods for detecting larger and denser structures are unquestionably needed.

An algorithm called *PathBLAST*[8] searches for high-scoring pathway alignments between

[8]http://www.pathblast.org/

two networks under study. Pathway alignments are scored as follows. The likelihood of a pathway match is computed by taking into account both the probabilities of true homology between proteins that are aligned on the path and the probabilities that the protein-protein interactions that are present in the path are real, i.e., not false-positive errors. The score of a path is thus a product of independent probabilities for each aligned protein pair and for each protein interaction. The probability of a protein pair is based on the BLAST E-value of aligning sequences of the corresponding proteins, whereas the probability of a protein interaction is based on the false-positive rates associated with interactions. The PathBLAST method has been extended to detect conserved protein clusters rather than paths, by deploying a likelihood-based scoring scheme that weighs the denseness of a given subnetwork versus the chance of observing such network substructure at random. Moreover, it has also been extended to allow for the alignment of more than two networks.

MaWISh[9] is a method for pairwise local alignment of PPI networks implementing an evolution-based scoring scheme to detect conserved protein clusters. This mathematical model extends the concepts of evolutionary events in sequence alignment to that of duplication, match, and mismatch in network alignment. The method evaluates the similarity between graph structures through a scoring function that accounts for these evolutionary events. Each duplication is associated with a score that reflects the divergence of function between the two proteins. The score is based on the protein sequence similarity and is computed by BLAST. A match corresponds to a conserved interaction between two orthologous protein pairs. A mismatch, on the other hand, is the lack of an interaction in the PPI network of one organism between a pair of proteins whose orthologs interact in the other organism. A mismatch may correspond to the emergence of a new interaction or the elimination of a previously existing interaction in one of the species after the split, or to an experimental error. After each match, mismatch, and duplication is given a score, the optimal alignment is defined a set of nodes with the maximum score, computed by summing all possible matches, mismatches, and duplications in the given set of nodes.

[9]http://www.cs.purdue.edu/homes/koyuturk/mawish/

Graemlin[10] has been introduced as the first method capable of multiple alignment of an arbitrary number of networks, supporting both global and local search, and being capable of searching for dense conserved subnetworks of an arbitrary structure. Graemlin's purpose is to search for evolutionarily conserved functional modules across species. The method supports five types of evolutionary events: protein sequence mutations, protein insertions and deletions, protein duplications, and protein divergences (a divergence being inverse of a duplication). The module alignment score is computed by deploying two models that assign probabilities to the evolutionary events: the alignment model that assumes that a module is subject to evolutionary constraint, and the random model that assumes that the proteins are under no constraints. Then, the score of the alignment is the log-ratio of the two probabilities, which is a common method for scoring sequence alignments.

Finally, we conclude with *ISORANK* (20), a method initially designed for pairwise global alignment of PPI networks that has later been extended to allow for multiple local alignment of these networks. The initial version of ISORANK maximizes the overall match between the two networks by using both biological (i.e., BLAST-computed protein sequence similarity) and topological (i.e., protein interaction) information. Given two networks, the output of the algorithm is the maximum common subgraph between the two graphs, i.e., the largest graph that is isomorphic to subgraphs of both networks. The algorithm works in two stages. It first associates a score with each possible match between nodes of the two networks. The scores are computed using the intuition that two nodes, one from each network, are a good match if their respective neighbors also match well with each other. The method captures not only local topology of nodes, but also non-local influences on the score of a protein pair: the score of the protein pair depends on the score of the neighbors of the two nodes, and the latter, in turn, depend on the neighbors of their neighbors, and so on. The incorporation of other information, e.g. BLAST scores, into this model is straightforward. The second stage constructs the mapping by extracting from all protein pairs the high-scoring matches by applying the repetitive greedy strategy of identifying and outputting the highest scoring

[10]http://graemlin.stanford.edu/

pair and removing all scores involving any of the two identified nodes was more efficient. The later version of ISORANK is the direct generalization of the algorith to support multiple networks.

These methods have been used to identify network regions, protein complexes, and functional modules that have been conserved across species. Additionally, since a conserved subnetwork that contains many proteins of the same known function suggests that the remaining proteins also have that function, the network alignment methods have been used to predict new protein functions. Similarly, functions for unannotated proteins in one species have been predicted based on the functions of their aligned annotated partners in the other species. Finally, since proteins that are aligned together are likely to be functional orthologs, orthology relationships across multiple species have been predicted.

1.5 From structure to function in biological networks

1.5.1 Protein function prediction

We have illustrated how network comparisons across species can help in functional annotation of individual proteins and in identiciation of network regions representing functional modules or protein complexes. Similarly, biological function of uncharacterized proteins in a PPI network of a single species can be determined from the function of other, well described proteins from the same network. It has been shown that proteins that are closer in a network are more likely to perform the same function (2). Similarly, proteins with similar topological neighborhoods show tendency to have similar biological characteristics (3). Moreover, cancer genes have beene shown to have greater connectivities and centralities compared to non-cancer genes (5). Thus, various network structural properties such as the shortest path distances between proteins, network centralities, or graphlet-based descriptions of protein's topological neighborhoods can suggest their involvement in certain biological functions and disease. Since defining a relationship between PPI network topology and biological function and disease and inferring protein function from it is considered to be one of the major

challenges in the post-genomic era (2), various approaches for determining protein function from PPI networks have been proposed. They can be categorized into two major classes: direct and cluster-based methods (2).

The methods in the first class infer the function of an individual protein by exploring the premise that proteins that are closer in the network are more likely to share a function. The simplest method of this type is the "majority rule" that investigates only the direct neighborhood of a protein, and annotates it with the most common functions among its annotated neighbors. However, this approach does not assign any significance values to predicted functions. Additionally, it considers only nodes directly connected to the protein of interest and thus, only very limited topology of a network is used in the annotation process. Finally, it fails to differentiate between proteins at different distances from the target protein. Other approaches have tried to overcome these limitations by observing n-neighborhood of an annotated protein, where n-neighborhood of a protein is defined as a set of proteins that are at most at distance n from the target protein. Then, the protein of interest is assigned the most significant function (with the highest χ-square value) among functions of all n-neighboring proteins. This approach thus covers larger portion of the network compared to the simple majority rule. Additionally, it assigns confidence scores to the predicted functions. However, it still fails to distinguish between proteins at different distances from the protein of interest. A forthcoming study has addressed this approach by assigning different weights to proteins at different distances from the target protein. However, this method observes only 1- and 2-neighborhoods of proteins, thus again covering only their local topologies.

For this reason, several global optimization-based function prediction strategies have been proposed. For example, any given assignment of functions to the whole set of unclassified proteins in a network is given a score, counting the number of interacting pairs of nodes with no common function; the functional assignment with the lowest score maximizes the presence of the same function among interacting proteins. Since this method again fails to distinguish between proteins at different distances from the protein of interest, thus not rewarding local proximity, a network-flow-based method that considers both local and global effects has been proposed. According to this method, each functionally annotated protein

in the network is considered as the source of a "functional flow". Then, the spread of the functional flow through the network is simulated over time, and each unannotated protein is assigned a score for having the function based on the amount of flow it received during the simulation.

Approaches of the second type are exploiting the existence of regions in PPI networks that contain a large number of connections between the constituent proteins. These dense regions are a sign of a common involvement of those proteins in certain biological processes and therefore are feasible candidates for biological complexes or functional modules. Thus, these approaches try to identify clusters in a network and then, instead of predicting functions of individual proteins, they assign an entire cluster with a function based on the functions of its annotated members. Various approaches for identifying these functionally enriched modules solely from PPI network topology have been defined. For example, the highly connected subgraphs and the restricted neighborhood search clustering (RNSC) algorithms have been used to detect complexes in PPI networks (21; 22). However, with some of the cluster-based methods (e.g., MCODE[11]), the number of clusters or the size of the sought clusters need to be provided as input (2).

Additionally, several iterative hierarchical-clustering-based methods that form clusters by computing the similarities between protein pairs have been proposed. Thus, the key decision with these methods is the choice of the appropriate similarity measures between protein pairs. The most intuitive network topology-based measure is to use pairwise distances between proteins in the network (2): the smaller the distance between the two proteins in the PPI network is, the more "similar" they are, and thus, the more likely they are to belong to the same cluster. In other words, module members are likely to have similar shortest path distance profiles. However, global network properties such as pairwise shortest path distances might not be detailed enough and more constraining measures of node similarities are necessary.

Here, we present a sensitive graph theoretic method for comparing local network struc-

[11]http://baderlab.org/Software/MCODE

tures of protein neighborhoods in PPI networks that demonstrates that in PPI networks, biological function of a protein and its local network structure are closely related (3). The method summarizes the local network topology around a protein in a PPI network into a vector of "graphlet degrees" called the "signature of a node" (i.e., "the signature of a protein") and computes the "signature similarities" between all protein pairs (see Section 1.5.1 below for details). Proteins with topologically similar network neighborhoods are then grouped together under this measure and the resulting protein groups have been shown to belong to the same protein complexes, perform the same biological functions, are localized in the same subcellular compartments, and have the same tissue expressions. This has been verified for PPI networks of a unicellular and a multicellular eukaryotic organisms of yeast and human, respectively. Thus, it is hypothesized that PPI network structure and biological function are closely related in other eukaryotic organisms as well. Next, since the number of functionally unclassified proteins is large even for simple and well studied organisms such as baker's yeast *Saccharomyces cerevisiae*, Milenković and Pržulj (3) have described how to apply their technique to predict a protein's membership in protein complexes, biological functional groups, and subcellular compartments for yet unclassified yeast proteins. Additionally, they have shown how this method can be used for identification of new disease genes, demonstrating that it can provide valuable guidelines for future experimental research.

Graphlet degree signatures

Similar to graphlet degree distributions (GDDs) described in Section 1.2, this node similarity measure generalizes the degree of a node, which counts the number of edges that a node touches, into a vector of *graphlet degrees (GDs)*, counting the number of graphlets that a node touches. The method counts the number of graphlets touching a node for all 2-5-node graphlets (these graphlets are denoted by G_0, G_1, ..., G_{29} in Figure 1.2); counts involving larger graphlets become computationally infeasible for large networks. For example, an outer (black) node in graphlet G_9 touches graphlets G_0, G_1, G_3, and G_9 once, and it touches no other graphlets. Clearly, the degree of a node is the first coordinate in this vector, since an edge (graphlet G_0) is the only 2-node graphlet. This vector is called the *signature* of a node.

Due to the existence of 73 automorphism orbits for 2-5-node graphlets (described in Section 1.2), the signature vector of a node has 73 coordinates. For example, a node at orbit 15 in graphlet G_9 touches orbits 0, 1, 4, and 15 once, and all other orbits zero times. Thus, its signature will have 1s in the 0^{th}, 1^{st}, 4^{th}, and 15^{th} coordinate, and 0s in the remaining 69 coordinates.

Node *signature similarities* are computed as follows. A 73-dimensional vector W is defined to contain the weights w_i corresponding to orbits $i \in \{0, \ldots, 72\}$. Different weights are assigned to different orbits for the reasons illustrated below. For example, the differences in orbit 0 (i.e., in the degree) of two nodes will automatically imply the differences in all other orbits for these nodes, since all orbits contain, i.e., "depend on", orbit 0. Similarly, the differences in orbit 3 (the triangle) of two nodes will automatically imply the differences in all other orbits of the two nodes that contain orbit 3, such as orbits 14 and 72. This is generalized to all orbits. Thus, higher weights need to be assigned to "important" orbits, those that are not affected by many other orbits, and lower weights need to be assigned to "less important" orbits, those that depend on many other orbits. By doing so, the redundancy of an orbit contained in other orbits is removed. For details on computing orbit weights, see (3).

For a node u, u_i denotes the i^{th} coordinate of its signature vector, i.e., u_i is the number of times node u touches orbit i. The distance $D_i(u, v)$ between the i^{th} orbits of nodes u and v is then defined as: $D_i(u, v) = w_i \times \frac{|log(u_i+1)-log(v_i+1)|}{log(max\{u_i,v_i\}+2)}$. The authors use log in the numerator because the i^{th} coordinates of signature vectors of two nodes can differ by several orders of magnitude and the distance measure should not be entirely dominated by these large values (3). Also, by using these logarithms, they take into account the relative difference between u_i and v_i instead of the absolute difference. They scale D_i to be in [0, 1) by dividing with the value of the denominator in the formula for $D_i(u, v)$. The total distance $D(u, v)$ between nodes u and v is then defined as: $D(u, v) = \frac{\sum_{i=0}^{72} D_i}{\sum_{i=0}^{72} w_i}$. Clearly, the distance $D(u, v)$ is in [0, 1), where distance 0 means the identity of signatures of nodes u and v. Finally, the *signature similarity*, $S(u, v)$, between nodes u and v is: $S(u, v) = 1 - D(u, v)$. For example, the two outer (black) nodes at orbit 15 in graphlet G_9 have the same signatures, and thus, their total distance is 0 and their signature similarity is 1.

Clusters in a PPI network are formed as follows: for a node of interest, construct a cluster containing that node and all nodes in a network that have similar signatures to it. This is repeated for each node in the PPI network. Thus, nodes u and v will be in the same cluster if their signature similarity $S(u, v)$ is above a chosen threshold. Thresholds are determined experimentally to be in 0.9-0.95. For thresholds above these values, only a few small clusters are obtained, especially for smaller PPI networks, indicating too high stringency in signature similarities. For thresholds bellow 0.9, the clusters are very large, especially for larger PPI networks, indicating a loss of signature similarity. To illustrate signature similarities and the choices of signature similarity thresholds, Figure 1.4 presents the signature vectors of yeast proteins in a PPI network with signature similarities above 0.90 (Figure 1.4 A) and below 0.40 (Figure 1.4 B); signature vectors of proteins with high signature similarities follow the same pattern, while those of proteins with low signature similarities have very different patterns.

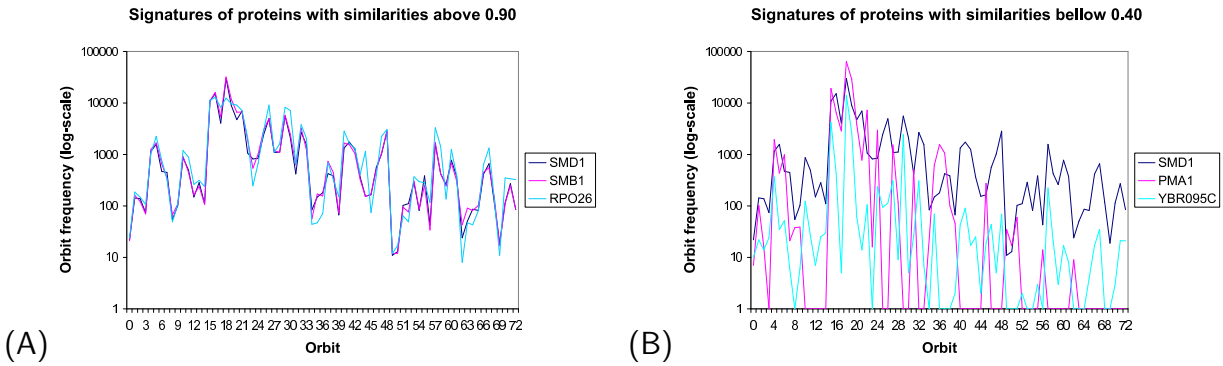

Figure 1.4: Signature vectors of proteins with signature similarities: (A) above 0.90; and (B) below 0.40. The 73 orbits are presented on the abscissa and the numbers of times that nodes touch a particular orbit are presented on the ordinate in log scale. In the interest of the aesthetics of the plot, 1 is added to all orbit frequencies to avoid the log-function to go to infinity in the case of orbit frequencies of 0.

The method described above was applied to six yeast and three human PPI networks of different sizes and different confidence levels (3). After the creation of clusters in each network, the clusters were searched for common *protein properties*. In yeast PPI networks, protein properties included protein complexes, functional groups, and subcellular localiza-

tions described in MIPS[12]; in human PPI networks, they involved biological processes, cellular components, and tissue expressions described in HPRD[13]. Classification schemes and the data for the protein properties were downloaded from MIPS and HPRD in November 2007. For each of the protein property classification schemes, two levels of strictness were defined: the *strict* scheme that uses the most specific MIPS annotations, and the *flexible* one that uses the least specific ones. For example, for a protein complex category annotated by *510.190.900* in MIPS, the strict scheme returns *510.190.900*, and the flexible one returns *510*.

In the clusters, the sizes of the largest common categories for a given protein property were measured as a percentage of the cluster size and this was refered to as the *hit-rate*. Clearly, a protein can belong to more than one protein complex, be involved in more than one biological function, or belong to more than one subcellular compartment. Thus, it is possible to have an overlap between categories, as well as more than one largest category in a cluster for a given protein property. A *miss-rate* was also defined as a percentage of the nodes in a cluster that are not in any common category with other nodes in the cluster. For each of the PPI networks, the corresponding protein properties, and the two schemes (the strict and the flexible), the percentage of clusters having given hit- and miss-rates was measured. The hit- and miss-rates were binned in increments of 10%. The method identified high hit-rates and low-miss rates in a large percentage of clusters in all six yeast and three human PPI networks, for all protein properties (see (3) for details). For example, for protein complexes in the high-confidence yeast PPI network, 44% of clusters had 100% hit-rate according to the flexible scheme. This additionally validated the method, since PPIs in this network have been obtained mainly by TAP and HMS-PCI, which are known to favor protein complexes. Thus, the method established a strong relationship between a biological function of a protein and the network structure around it in a PPI network.

To evaluate the effect of noise in PPI networks on the accuracy of the method, the authors compared the results for the high-confidence and the lower-confidence yeast networks.

[12]http://mips.gsf.de/
[13]http://www.hprd.org/

As expected, clusters in a noisier network had lower hit-rates then the clusters in a high-confidence network. However, low miss-rates were still preserved in clusters of both networks for all three protein properties, indicating the robustness of the method to noise present in PPI networks.

Furthermore, the statistical significance of the results was examined by the following two analyses: (a) random re-shuffling of node labels in PPI networks and comparisons of the results with those obtained from such randomized networks; and (b) computing p-values for obtaining the observed homogeneous protein properties within clusters. Hit-rates for all protein properties for the PPI networks were higher than for randomized networks, thus indicating that the signature-based algorithm captured the true biological signal from the topology of PPI networks. Similarly, miss-rates for the data were lower than for randomized networks with respect to all protein properties in PPI networks. This was especially true for protein complexes and biological function in yeast. Only for subcellular localization, miss-rates in the data and in randomized networks were comparable, but hit-rates were still higher in the data than in the randomized networks. The main reason for observing similarity in miss rates for subcellular localization in the data and randomized networks was the large sizes of most subcellular localization categories. Therefore, it was expected that most proteins in a cluster will be in a common localization category with at least one other protein in a cluster, which led to low miss-rates.

Additionally, the probability that a given cluster was enriched by a given category merely by chance was computed, following the hypergeometric distribution (as in (22)). The total number of proteins in all clusters was denoted by $|N|$, the size of cluster C by $|C|$, and the number of proteins in cluster C that were enriched by category P by k, where category P contained $|P|$ out of $|N|$ proteins. Thus, the hit-rate of cluster C was $k/|C|$, and the p-value for cluster C and category P, i.e., the probability of observing the same or higher hit-rate, was: $p\text{-value} = 1 - \sum_{i=0}^{k-1} \frac{\binom{|P|}{i}\binom{|N|-|P|}{|C|-i}}{\binom{|N|}{|C|}}$. The p-value of a cluster was considered to be its smallest p-value over all categories for a given protein property. p-values for all signature-based clusters with hit rates of at least 50% were computed for all yeast and human PPI networks and for all protein properties. Then, the percentage of clusters (out of the

total number of clusters with hit-rates of at least 50%) having a given p-value was found. Depending on a method and its application, sensible cut-offs for p-values were reported to range from 10^{-2} to 10^{-8} (22).

For yeast, with respect to both the strict and the flexible scheme, low p-values of $O(10^{-2})$ or lower were observed for protein complexes and biological functions, whereas for subcellular localizations, a percentage of clusters had higher p-values of $O(10^{-1})$. As explained above, since subcellular localization categories typically contained a large number of proteins, this high p-values were expected. Since the flexible scheme by definition meant that a larger number of proteins was contained within each category than in the strict scheme, somewhat higher, but still low enough, p-values were observed for flexible than for the strict scheme. For human, a significant percentage of clusters had high p-values, for all three human protein properties. The reason for this was the same as for subcellular localization in yeast: many proteins existed within each category in all three human protein properties. However, a certain percentage of clusters had low p-values. For example, for tissue expressions, about 50% of clusters in all three human PPI networks had p-values of $O(10^{-2})$ or lower. Therefore, although p-values varied depending on a given protein property and the size of its categories, the algorithm identified clusters in which true biological signal was captured. This was especially true for protein complexes and biological functions in all six yeast PPI networks, with respect to both the strict and the flexible scheme.

This technique can also be applied to predict protein properties of yet unclassified proteins by forming a cluster of proteins that are similar to the unclassified protein of interest and assigning it the most common properties of the classified proteins in the cluster. The authors did this for all 115 functionally unclassified yeast proteins from MIPS that had degrees higher than four in any of the six yeast PPI networks that they analyzed (the list of predictions is in (3)). Note that a yeast protein can belong to more than one yeast PPI network that was analyzed. Thus, biological functions that such proteins perform can be predicted from clusters derived from different yeast PPI networks. An overlap of the predicted protein functions obtained from multiple PPI networks for the same organism was observed, additionally confirming the validity of predictions. Furthermore, there existed an

overlap between protein function predictions produced by this method and those of others. Finally, the predictions were verified in the literature.

The graphlet degree signatures-based method has several advantages over direct approaches for protein function prediction (described above). Not only that it assigns a confidence score to each predicted annotation (in terms of hit- and miss-rates), but also for doing that it takes into account up to 5-neighborhoods of a node along with their interconnectivities, since it is based on 2-5-node graphlets. Additionally, although the signature of a node describes its "5-deep" local neighborhood, due to typically small diameters of PPI networks, it is possible that 2-5-node-graphlet-based signatures capture the full, or almost full topology of these networks.

Also, the graphlet degree signatures method belongs to the group of clustering-based approaches. However, unlike other methods of this type that define a cluster as a dense interconnected region of a network (typically enriching for a biological function, as described above), this method defines a cluster as a set of nodes with similar topological signatures. Thus, nodes belonging to the same cluster do not need to be connected or belong to the same part of the network. Additionally, whereas other approaches typically assign the function to the entire cluster, since this method forms a cluster for each protein in the PPI network individually, it assigns function(s) to individual proteins. Moreover, the method does not require the number of clusters or their size to be predefined, unlike some of the other above mentioned approaches. Furthermore, to create pairwise similarities between protein pairs, this method uses highly constraining local-topology-based measure of similarity of proteins' signatures, unlike other studies that use only global network properties; additionally, this clustering method is not hierarchical, and it allows for overlap between clusters.

It is difficult to perform direct comparisons of the performance of all methods described above. Attempts to perform a comparison of several cluster-based methods have been made. However, due to the different performance measures across different studies, fundamental differences between different annotation types, and the lack of the golden standards for functional annotation, any comprehensive comparison is very difficult (2). For example, only this study has used hit- and miss-rates to measure the success of the results of the method.

Additionally, some studies have used the MIPS[14] annotation catalogs, whereas other studies have used Gene Ontology[15] as the annotation source, and some annotations that exist in one data source might not exist in the other. Moreover, to our knowledge, the above described study by Milenković and Pržulj (3) is the only one that related the PPI network structure to all of the following: protein complexes, biological functions, and subcellular localizations for yeast, and cellular components, tissue expressions, and biological processes for human, thus making it impossible to do comparisons with other methods from this aspect. However, even if they were unable to quantify their results with respect to other studies, Milenković and Pržulj provided other indications of the correctness of their approach (see above).

The graphlet signatures-based method is easily extendible to include larger graphlets, but this would increase the computational complexity. The complexity is currently $O(|V|^5)$ for a graph $G(V, E)$, since the searches for graphlets with up to 5 nodes are performed. Nonetheless, since the algorithm is "embarrassingly parallel", i.e., can easily be distributed over a cluster of machines, extending it to larger graphlets is feasible. In addition to the design of the signature similarity measure as a number in (0, 1], this makes the technique usable for much larger networks.

1.5.2 Disease gene identification

In addition to protein function prediction, several studies have investigated associations between diseases and PPI network topology. Radivojac et al. (4) have tried to identify candidate disease genes from a human PPI network by encoding each gene in the network based on the distribution of shortest path lengths to all genes associated with disease or having known functional annotation. Additionally, Jonsson and Bates (5) analyzed network properties of cancer genes and demonstrated greater connectivity and centrality of cancer genes compared to non-cancer genes indicating an increased central role of cancer genes within the interactome. However, these studies have been mainly based on global network properties, which might not be detailed enough to encompass complex topological characteristics of

[14] http://mips.gsf.de/
[15] http://www.geneontology.org/

disease genes in the context of PPI networks.

Similarly, graphlet degree signature-based method has also been applied to disease genes. A set of genes implicated in genetic diseases available from HPRD[16] was examined. To increase coverage of PPIs, the human PPI network that was analyzed was the union of the human PPI networks from HPRD, BIOGRID, and Rual et al. (23), which consisted of 41,755 unique interactions amongst 10,488 different proteins. There were 1,491 disease genes in this PPI network out of which 71 were cancer genes. If network topology is related to function, then it is expected that genes implicated in cancer might have similar graphlet degree signatures. To test this hypothesis, Milenković and Pržulj looked for all proteins with a signature similarity of 0.95 or higher with protein TP53. The resulting cluster contained ten proteins, eight of which were disease genes; six of these eight disease genes were cancer genes (TP53, EP300, SRC, BRCA1, EGFR, and AR). The remaining two proteins in the cluster were SMAD2 and SMAD3 which are members of TGF-beta signaling pathway whose deregulation contributes to the pathogenesis of many diseases including cancer. The striking signature similarity of this 10-node cluster is depicted in Figure 1.5.

1.6 Software tools for network analyses and modeling

The recent explosion in biological and other real-world network data has created the need for improved tools for large network analyses. In addition to well established *global* network properties, several new mathematical techniques for analyzing *local* structural properties of large networks have been developed (see Section 1.2). Adequate null-models for biological networks have been sought in many research domains and various network models have been proposed (see Section 1.3.1). Network properties are used to assess the fit of network models to the data.

Computing global network properties is computationally and conceptually easy and various software tools are available for this purpose, such as tYNA[17] and pajek[18]. However,

[16]http://www.hprd.org/
[17]http://tyna.gersteinlab.org/tyna/
[18]http://vlado.fmf.uni-lj.si/pub/networks/pajek/

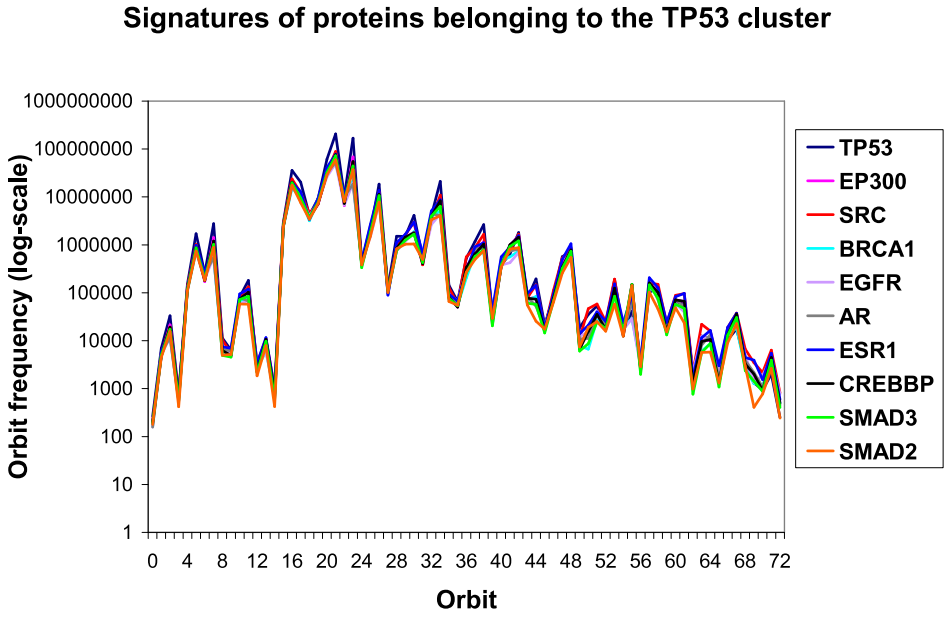

Figure 1.5: Signature vectors of proteins belonging to the TP53 cluster. The cluster is formed using the threshold of 0.95. The axes have the same meaning as in Figure 1.4.

none of the tools have built-in capabilities to compare real-world networks against a series of network models based on these properties. Furthermore, the computational challenge is in finding local properties. Currently available software packages that find local network properties focus on searching for network motifs. These tools include mfinder[19], MAVisto[20], and FANMOD[21]. Until recently, there did not exist a publicly available, open-source software tool that computed local properties other than network motifs. Thus, Milenković et al. have introduced GraphCrunch[22] (16), a software tool that finds well-fitting network models by comparing large real-world networks against random graph models according to various global and local network structural similarity measures.

GraphCrunch has unique capabilities of finding computationally expensive RGF-distance and GDD-agreement measures. In addition, it computes several standard global network measures and therefore supports the largest variety of network measures thus far. More

[19]http://www.weizmann.ac.il/mcb/UriAlon/groupNetworkMotifSW.html
[20]http://mavisto.ipk-gatersleben.de/
[21]http://www.minet.uni-jena.de/~wernicke/motifs/index.html
[22]http://www.ics.uci.edu/~bio-nets/graphcrunch/. Labeled as "Highly Accessed" by *BMC Bioinformatics*.

specifically, while some of the other above mentioned software tools compute only local network properties, others that analyze both local and global properties offer fewer functions than GraphCrunch does. The main purpose of mfinder, MAVisto, and FANMOD is motif search; they do not compute global network properties. On the other hand, pajek focuses on global network properties and has very limited local network analysis capabilities; its search for subgraphs is limited to 3-4-node rings. tYNA's global and local network analyses are limited: it calculates the statistics of global network properties and focuses on three network motif types only. Unlike any of these software packages, GraphCrunch uses all of the 2-5-node graphlets for computing its two highly constraining graphlet-based local network properties, GDD-agreement (9) and RGF-distance (8), along with five standard global properties. The properties currently supported by GraphCrunch are presented in Table 1.1 and described in Section 1.2.

Furthermore, GraphCrunch uses all of these properties for comparing real-world networks against a series of network models. Five network models are currently supported by GraphCrunch. They are presented in Table 1.2 and explained in Section 1.3.1. Although mfinder, FANMOD and pajek offer more than one network model (MAVisto does not), none of these tools supports a variety of network models as GraphCrunch does. Note that tYNA does not generate random models at all and it searches for subgraphs in real-world networks only. Furthermore, GraphCrunch determines the fit of various network models to real-world networks with respect to an array of global and local network properties; none of the other currently available network analysis software tools have this functionality.

Finally, although mfinder and FANMOD both include an option of using random subgraph sampling heuristics for speeding up the computing time, a feature that GraphCrunch currently does not support, GraphCrunch's exhaustive graphlet counting is very competitive. Moreover, GraphCrunch is easily extendible to include additional network measures and models and it has built-in parallel computing capabilities allowing for a user specified list of machines on which to perform compute intensive searches for local network properties. This feature that is not supported by Mfinder, FANMOD, MAVisto, tYNA, or pajek. This functionality will become crucial as biological network data sets grow.

Global Properties:
Degree distribution
Clustering coefficient
Clustering spectrum
Average diameter
Spectrum of shortest path lengths
Local Properties:
Relative graphlet frequency distance (RGF-distance)
Graphlet degree distribution agreement (GDD-agreement)

Table 1.1: Network properties currently supported by GraphCrunch.

Models:
Erdös-Rényi random graphs ("ER")
Random graphs with the same degree distribution as the data ("ER-DD")
Scale-free Barabási-Albert model graphs ("SF-BA")
N-dimensional geometric random graphs ("GEO-nd"; default: "GEO-3d")
Stickiness model graphs ("STICKY")

Table 1.2: Network models currently supported by GraphCrunch.

Many network analysis and modeling software tools are already available. However, as biological data becomes larger and more complete, the need for improving these tools and the algorithms that they implement will continue to rise.

1.7 Concluding remarks

Biological networks research is still in its infancy, but has already become a vibrant research area that is likely to have impacts onto biological understanding and therapeutics. As such, it is rich in open research problems that we are currently only scratching a surface of. Many

new, unforseen problems will keep emerging. Thus, the field is likely to stay at the top of scientific endevor in the years to come.

List of Figures

Bibliography

[1] Pržulj N: **Graph Theory Analysis of Protein-Protein Interactions**. In *Knowledge Discovery in Proteomics*. Edited by Jurisica I, Wigle D, CRC Press 2005:73–128.

[2] Sharan R, Ulitsky I, Ideker T: **Network-based prediction of protein function**. *Molecular Systems Biology* 2007, **3**(88).

[3] Milenković T, Pržulj N: **Uncovering Biological Network Function via Graphlet Degree Signatures**. *Cancer Informatics* 2008, **4**:257–273.

[4] Radivojac P, Peng K, Clark WT, Peters BJ, Mohan A, Boyle SM, Mooney SD: **An integrated approach to inferring gene-disease associations in humans**. *Proteins* 2008, :in press.

[5] Jonsson P, Bates P: **Global topological features of cancer proteins in the human interactome**. *Bioinformatics* 2006, **22**(18):2291–2297.

[6] Newman MEJ: **The structure and function of complex networks**. *SIAM Review* 2003, **45**(2):167–256.

[7] Milo R, Shen-Orr SS, Itzkovitz S, Kashtan N, Chklovskii D, Alon U: **Network motifs: simple building blocks of complex networks**. *Science* 2002, **298**:824–827.

[8] Pržulj N, Corneil DG, Jurisica I: **Modeling Interactome: Scale-Free or Geometric?** *Bioinformatics* 2004, **20**(18):3508–3515.

[9] Pržulj N: **Biological Network Comparison Using Graphlet Degree Distribution**. *Bioinformatics* 2006, **23**:e177–e183.

[10] Pržulj N, Higham D: **Modelling protein-protein interaction networks via a stick-**

iness index. *Journal of the Royal Society Interface* 2006, **3**(10):711–716.

[11] Erdös P, Rényi A: **On random graphs**. *Publicationes Mathematicae* 1959, **6**:290–297.

[12] Molloy M, Reed B: **A critical point of random graphs with a given degree sequence**. *Random Structures and Algorithms* 1995, **6**:161–180.

[13] Barabási AL, Albert R: **Emergence of scaling in random networks**. *Science* 1999, **286**(5439):509–512.

[14] Watts DJ, Strogatz SH: **Collective dynamics of 'small-world' networks**. *Nature* 1998, **393**:440–442.

[15] Penrose M: *Geometric Random Graphs*. Oxford University Press 2003.

[16] Milenković T, Lai J, Pržulj N: **GraphCrunch: a tool for large network analyses**. *BMC Bioinformatics* 2008, **9**(70).

[17] Vazquez A, Flammini A, Maritan A, Vespignani A: **Modeling of Protein Interaction Networks**. *ComPlexUs* 2001, **1**:38–44.

[18] Higham D, Rašajski M, Pržulj N: **Fitting a geometric graph to a proteinprotein interaction network**. *Bioinformatics* 2008, **24**:1093–1099.

[19] Sharan R, Ideker T: **Modeling cellular machinery through biological network comparison**. *Nature Biotechnology* 2006, **24**(4):427–433.

[20] Singh R, Xu J, Berger B: **Pairwise Global Alignment of Protein Interaction Networks by Matching Neighborhood Topology**. In *Research in Computational Molecular Biology*, Springer 2007:16–31.

[21] Pržulj N, Wigle D, Jurisica I: **Functional topology in a network of protein interactions**. *Bioinformatics* 2004, **20**(3):340–348.

[22] King AD, Pržulj N, Jurisica I: **Protein complex prediction via cost-based clustering**. *Bioinformatics* 2004, **20**(17):3013–3020.

[23] Rual JF, Venkatesan K, Hao T, Hirozane-Kishikawa T, Dricot A, Li N, Berriz GF, Gibbons FD, Dreze M, Ayivi-Guedehoussou N, Klitgord N, Simon C, Boxem M,

Milstein S, Rosenberg J, Goldberg DS, Zhang LV, Wong SL, Franklin G, Li S, Albala JS, Lim J, Fraughton C, Llamosas E, Cevik S, Bex C, Lamesch P, Sikorski RS, Vandenhaute J, Zoghbi HY, Smolyar A, Bosak S, Sequerra R, Doucette-Stamm L, Cusick ME, Hill DE, Roth FP, Vidal M: **Towards a proteome-scale map of the human protein-protein interaction network**. *Nature* 2005, **437**:1173–78.