Knowledge Discovery in Proteomics: Graph Theory Analysis of Protein-Protein Interactions

Nataša Pržulj[1] (Editors: Igor Jurisica and Dennis Wigle)

[1]Department of Computer Science, University of Toronto,
10 King's College Road, Toronto, ON, M5S 3G4, Canada

January 3, 2005

Understanding protein-protein interactions is an important problem in proteomics. It is widely believed that studying networks of these interactions will provide valuable insight about the inner working of cells, and will lead to important insights into complex diseases. The recent deluge of experimental protein-protein interaction data available has made graph theory approaches an important part of computational biology and the knowledge discovery process.

Protein-protein interaction (PPI) networks are commonly represented as graphs, with nodes corresponding to proteins and edges representing PPIs. An example of a PPI network constructed in this way is presented in Figure 1. In general, these networks have directed edges and varying length; however, most of current PPI networks are undirected and represent only binary interactions.

Using high-throughput (HTP) techniques such as mass spectrometry (described in Chapter **??**), and yeast 2-hybrid screening, a large volume of experimental PPI data has been generated. For example, the yeast *S. cerevisiae* contains over $6,000$ proteins, and over 78,000 PPIs have now been identified. The analogous networks for mammals are expected to be much larger. For example, humans are expected to have around $120,000$ proteins, and thus approximately 10^6 PPIs. However, some of the largest public data sets of human PPIs currently available, such as the Database of Interacting Proteins [197], and A Molecular INTeraction database [201], contain less than $2,500$ interactions. Recently, a new manually created human PPI data set has become available [155], comprising about $13,000$ interactions; however, this is in a format not amenable to further automated data analysis

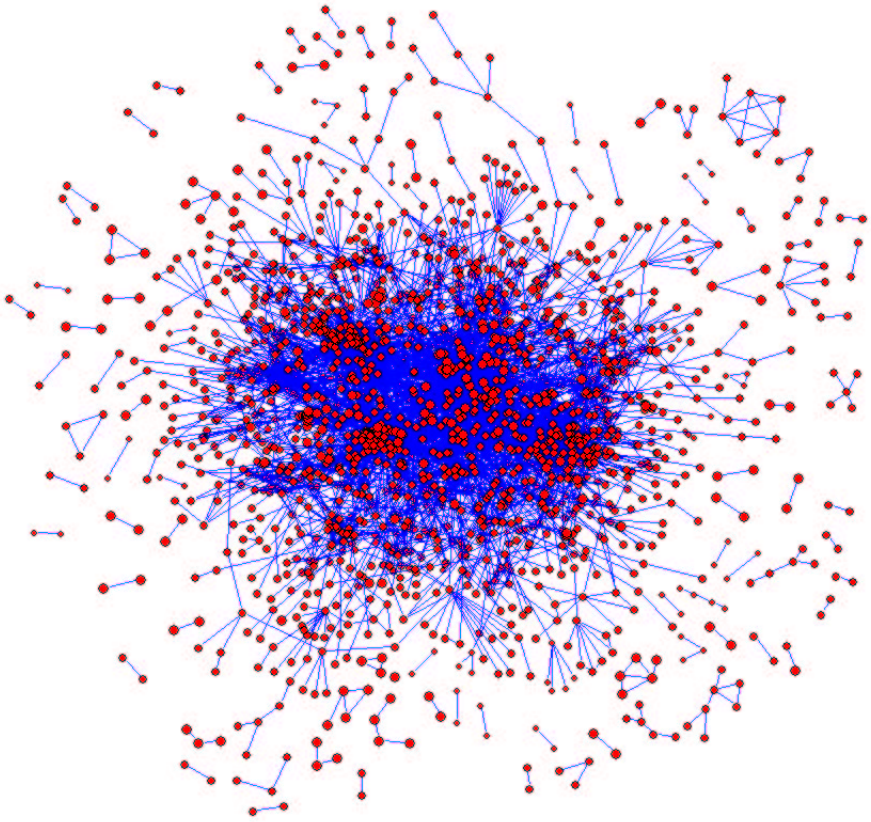

Figure 1: The PPI network constructed on $11,000$ yeast interactions [187] involving $2,401$ proteins.

(at the time of writing). Another new data set comprising predicted protein interactions called HPID [81] is also available only for manual and visual browsing, and thus not directly suitable for knowledge discovery approaches. A different approach to generate large numbers of putative human protein interactions is to use model organism data, and map to human orthologues using sequence homology. This enables us to generate a much larger human PPI data set, currently almost $50,000$ interactions [42]. An important aspect of OPHID [42] is effective access to data to facilitate retrieval, visualization and analysis. Thus, multiple output formats (including PSI [85]) are supported. Without such flexibility, bioinformatic use of PPI data for knowledge discovery would be much harder.

PPI data sets provide both an opportunity and a challenge. Analyzing these networks may provide useful clues about the function of individual proteins, protein complexes, signaling pathways, and larger subnetworks. However, the data volume and noise within it renders many algorithms for its analysis intractable.

One of the goals of systems biology is to explain relationships between structure, function, and regulation of molecular networks by combining theoretical and experimental approaches, as described in Chapter **??**. Graph theory is an integral part of this process, as it enables us to analyze structural properties of PPI networks, and link them to other information, such as function. Using this analysis leads to building predictive models for hypothesis generation, and thus more efficient and effective experiment planning.

After describing graph theoretic and biological terminology used in the PPI literature, we will introduce three large research areas necessary for understanding the issues arising in studying PPI networks:

1. mathematical models of large networks and the most important properties of these models;

2. PPI identification methods, publicly available PPI data sets, some of the biological structures embedded in the PPI networks and methods used for their detection, and the mathematical properties of the currently available PPI networks;

3. recent graph theoretic algorithms that have successfully been used in biological applications, and which may be used to identify biological structures in PPI networks.

1 Graph Theoretic Terminology

A *graph* is a collection of points with lines connecting pairs of points [193]. The points are called *nodes* or *vertices*, and the lines are called *edges*. A graph is usually denoted by G, or by $G(V, E)$, where V is the set of nodes and $E \subseteq V \times V$ is the set of edges of G. We often use n to represent number of nodes, $|V|$, and m to represent number of edges, $|E|$. We also use $V(G)$ to represent the set of nodes of a graph G, and $E(G)$ to represent the set of edges of a graph G.

A graph is *undirected* if its edges (node pairs) are undirected, and otherwise it is *directed*. A graph is *weighted* if there is a weight function associated with its edges, or nodes. A graph is *complete* if it has an edge between every pair of nodes. Such a graph is also called a *clique*. A complete graph on n nodes is commonly denoted by K_n. A graph G is *bipartite* if its node set can be partitioned into two sets, A and B, such that every edge of G has one node in A and the other in B.

Nodes joined by an edge are called *adjacent*. A *neighbor* of a node v is a node adjacent to v. We denote by $N(v)$ the set of neighbors of node v (called the *neighborhood* of v), and by $N[v]$ the *closed neighborhood* of v, which is defined as $N[v] = N(v) \cup \{v\}$. The *degree* of a node is the number of edges incident with the node. In directed graphs, an *in-degree* of a node is the number of edges ending at the node, and the *out-degree* is the number of edges originating at the node.

A *path* in a graph is a sequence of nodes and edges, such that a node belongs to the edges before and after it and no nodes are repeated. A path with k nodes is commonly denoted by P_k. The path *length* is the number of edges in the path. The shortest path length between nodes u and v is commonly denoted by $d(u, v)$. The *diameter* of a graph is the maximum of $d(u, v)$ over all nodes u and v. If a graph is disconnected, we assume that its diameter is equal to the maximum of the diameters of its connected components.

A *subgraph* of G is a graph whose all nodes and edges belong to G. An *induced subgraph* H of G, denoted by $H \lhd G$, is a subgraph of G on $V(H)$ nodes, such that $E(H)$ consists of all edges of G that connect nodes of $V(H)$. The *minimum edge cut* of a graph G is the set of edges S, such that $|S|$ is of minimum size over all sets of edges that disconnect the graph upon removal. The minimum number of edges whose deletion disconnects G is called *edge connectivity*. A graph is k-edge-connected if its edge connectivity is $\geq k$. In a weighted graph, the minimum weight edge cut of a graph can be defined.

2 Biological Terminology

Proteins are key components of cellular machinery. They play multiple roles – transferring signals, controlling the function of enzymes, and regulating production and activities in the cell. To do this, they interact with other proteins, DNA, and other molecules. Some of the protein interactions are permanent, while others are transient and happen only during certain cellular processes. Groups of proteins that together perform a certain cellular task are called *protein complexes*. There is some evidence to suggest that protein complexes correspond to complete or "nearly complete" subgraphs of PPI networks (see Section 4.3.1 and [16, 160, 108]).

A *domain* is part of a protein that has its own unique binding properties or function. The combination of domains in a protein determines its overall function. Examples of protein function include cell growth and maintenance, signal transduction, transcription, translation, metabolism, and others. These are systematically described in Gene Ontology [51]. Many domains mediate protein interactions with other biomolecules. Returning to knowledge management terminology from Chapter **??**, there is a many-to-many relationship between proteins and domains: A protein may have several different domains and the same domain may be found in different proteins.

A *molecular pathway* is a directed sequence of molecular reactions involved in cellular processes. Modeling them in PPI networks will lead to adding directionality and causality, and thus will be necessary for simulations. Shortest paths in PPI networks have been used to model pathways (see section 4.3.2 and [160]).

Homology is a relationship between two biological features (here we consider genes, or proteins) that have a common ancestor. The two subclasses of homology are *orthology* and *paralogy*. Two genes are *orthologous* if they have evolved from a common ancestor by speciation; they often have the same function, taken over from the precursor gene in the species of origin. Orthologous gene products are believed to be responsible for essential cellular activities. In contrast, *paralogous* proteins have evolved by gene duplication; they either diverge functionally, or all but one of the versions is lost.

3 Large Network Models

Describing real-world phenomena by a network may improve our understanding of the phenomena, and allow for simulations and predictions. Diverse complex systems can effectively be described by large networks. For example:

- the cell, where we model genes/proteins/metabolites by nodes and their interactions by edges;
- the Internet, which is a complex network of routers and computers connected by various physical or wireless links;
- the World Wide Web, which is a virtual network of Web pages connected by hyper-links;
- networks of infection spread, which represent spread of biological or computer viruses;
- electronic circuits, which represent connection among microelectronic components;
- food chain webs, comprising networks of food-dependent linkages among animals;
- human collaboration networks, which link scientists based on collaboration or actors based on appearance in the same movies, etc.

Several possible network models can be created for given real-world phenomena. However, the choice of a model is not arbitrary; the model must resemble properties of a true network. This leads to two challenges: 1) objectively describing characteristic properties of complex, real-world networks, and 2) defining network models that maximize overlap of characteristic network properties. As a result, it is likely that these models will change, as we generate more data and improve our understanding of given phenomena.

Despite progress, the field is still in its early phase. The emergence of the Internet, the World Wide Web, and cellular function data has made a significant impact on the modeling of large networks. Network theory has consequently become an important area of research on its own. Several articles give good surveys of large network models [5, 146, 147, 180].

3.1 Properties of Large Networks

One can study *global* and *local* properties of networks. While global properties provide an overall view of a given network, they fail to describe intricate differences among networks. Local properties measure small, local sub-structures or patterns, called *motifs* [135, 171, 199] or *graphlets* [156]. The main advantage of the local properties is evident when we study networks with incomplete node and edge sets. The reason is that while the local structures of these networks are more likely to be complete, the global properties are highly biased. For example, PPI networks are still under-studied and thus global properties cannot truly describe these networks, when, for example, instead of millions of interactions for human, we only have tens of thousands available now.

So far, the greatest research focus and progress has been made on studying global properties, such as: diameter [6], clustering [84], and degree distribution [149].

As mentioned in Section 1, the *diameter* is a maximum of shortest path lengths between any two nodes in the network. Despite their large sizes, most real-world networks have small diameters. This property is often referred to as the *small-world* property [192].

A network shows *clustering* (or *network transitivity*) if the probability of a pair of nodes being adjacent is higher when the two nodes have a common neighbor. *Clustering coefficient* C is defined as the average probability that two neighbors of a given node are adjacent [192]. More formally, if a node v in the network has d_v neighbors, the ratio between the number of edges E_v between the neighbors of v, and the largest possible number of edges between them, $\frac{d_v(d_v-1)}{2}$, is called the clustering coefficient of node v, and is denoted by C_v:

$$C_v = \frac{2E_v}{d_v(d_v-1)}.$$

The clustering coefficient C of the whole network is the average of C_vs for all nodes v in the network. Complex, real-world networks exhibit a large degree of clustering, i.e., their clustering coefficient is large.

The *degree distribution* characterizes the distribution of degrees in a network. Lets denote by $P(k)$ the probability that a randomly selected node of a network has degree k. Most large real-world networks have non-Poisson degree distributions. For example, a large number of these networks has the degree distribution with a power-law tail, $P(k) \approx k^{-\gamma}$. Such networks are called *scale-free* [20].

Measuring global properties of real-world, complex networks led to proposing multiple models of these networks. Proposed in late 1950s [64], *random graphs* represent the simplest model of a complex network, yet are still an active research area (see below). The *small-world model* [192] was motivated by clustering, and it interpolates between the highly clustered regular ring lattices (defined below) and random graphs. The *scale-free* model [20] was motivated by the discovery of the power-law degree distribution. The *geometric random graph model* [75, 80, 154] has recently been used to model real-world networks, such as wireless communication [49, 79, 32], electrical power-grid [134], protein structure [134], and PPI networks [156].

Measuring local properties follows a bottom-up approach by focusing on finding small, over-represented patterns in a network [94, 134, 135, 171]. In this approach, *motifs* of a network are identified as small subgraphs of a large network that appear significantly more frequently than in the randomized network. Not surprisingly, different types of real-world networks have different motifs [135]. Furthermore, different real-world evolved and designed networks have been grouped into superfamilies according to their local structural properties [134].

A slightly different approach to measuring local network structure has recently been proposed. *Graphlets* have been defined as small, induced subgraphs of a large network [156]. They do not need to be over-represented in a real-world network and this, along with being induced, distinguishes them from motifs. The distribution of graphlet frequencies in real-world networks can be close to those of model networks pointing to limitations of previous models and suggesting new ways to model real-world networks [134, 157].

3.2 Random Graphs

Random graphs are based on the principle that the probability that there is an edge between any pair of nodes (denoted by p) is distributed uniformly at random. Thus, a random graph on n nodes has approximately $\frac{n(n-1)}{2}p$ edges, distributed uniformly at random.

Random graphs represent the earliest model of a complex network. Since the 1950s, large networks with no apparent design have been modeled by random graphs. The pioneering work of Erdös and Rényi [64, 65, 66] led to this field becoming a significant research area (for details see survey in [34]).

Erdös and Rényi defined several versions of the model, out of which 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 $1-p$. The properties of $G_{n,p}$ are often expressed in terms of the

average degree z of a node. The average number of edges in the graph $G_{n,p}$ is $\frac{n(n-1)}{2}p$, each edge contains two nodes, and thus the average degree of a node is:

$$z = \frac{n(n-1)p}{n} = (n-1)p,$$

which is approximately equal to np for large n.

These graphs have many properties that can be calculated exactly in the limit of large n, which makes them appealing as models of real networks. Thus, the following terminology is commonly used in the literature on random graphs. It is said that *almost all* random graphs (or *almost every* random graph) on n nodes have a property X, if the probability $Pr(X)$ that a graph has the property X satisfies $\lim_{n\to\infty} Pr(X) = 1$. A graph on n nodes *almost always*, or *almost surely*, satisfies a property X, if $\lim_{n\to\infty} Pr(X) = 1$.

Examples of properties that can be calculated exactly in the limit of large n include the following. When the number of edges (m) is small, the graph is likely to be fragmented into many small connected components having node sets of size at most $O(\log n)$. As m increases the components grow at first by linking to isolated nodes, and later by fusing with other components. A transition happens at $m = \frac{n}{2}$, when many clusters cross-link spontaneously to form a unique largest component called the *giant component*, whose node set size is much larger than the node set sizes of any other component. The giant component contains $O(n)$ nodes, while the second largest component contains $O(\log n)$ nodes. Furthermore, the shortest path length between pairs of nodes in the giant component grows with $\log n$ (more details are given later), and thus, these graphs are small worlds. This result is typical for random-graph theory whose main goal usually is to determine at what probability p a certain graph property is most likely to appear. Erdös and Rényi's greatest discovery was that many important properties, such as the emergence of the giant component, appear quite suddenly, i.e., at a given probability either almost all graphs have some property, or almost no graphs have it.

The probability $P(k)$ of a given node in a random graph on n nodes having degree k is given by the binomial distribution:

$$P(k) = \left(\begin{array}{c} n-1 \\ k \end{array} \right) p^k (1-p)^{n-1-k},$$

which in the limit where $n \gg kz$ becomes the Poisson distribution:

$$P(k) = \frac{z^k e^{-z}}{k!}.$$

Both of these distributions are strongly peaked around the mean z and have a tail that decays rapidly as $1/k!$. Minimum and maximum degrees of random graphs are finite for a large range of p. For instance, if $p \approx n^{-1-1/k}$, almost no random graph has nodes with degree higher than k. However, if:

$$p = \frac{\ln n + k \ln(\ln n) + c}{n},$$

almost every random graph has a minimum degree of at least k. If $pn/\ln n \to \infty$, the maximum degree of almost all random graphs has the same order of magnitude as the average degree. Thus, a typical random graph has rather homogeneous degrees.

Random graphs tend to have small diameters. Random graphs on n nodes and probability p have a narrow range of diameters, usually concentrated around $\frac{\ln n}{\ln np} = \frac{\ln n}{\ln z}$ [48]. However, for $z = np < 1$, a typical random graph is composed of isolated trees and its diameter equals to the diameter of a tree. If $z > 1$, the giant component emerges, and the diameter of the graph is equal to the diameter of the giant component if $z > 3.5$, and is proportional to $\frac{\ln n}{\ln k}$. If $z \geq \ln n$, almost every random graph is totally connected and the diameters of these graphs on n nodes and with the same z are concentrated on a few values around $\frac{\ln n}{\ln z}$. The average path length also scales with the number of nodes as $\frac{\ln n}{\ln z}$, which is a reasonable estimate for average path lengths of many real-world networks [146].

Random graphs have served as idealized models of gene networks [105], ecosystems [127], and the spread of infectious diseases [104] and computer viruses [107]. Although, random graph models reasonably approximate the corresponding properties of these real-world networks, they still differ from them in two fundamental ways:

1. **Degree distribution:** Real-world networks appear to have power-law degree distributions [6, 68, 78, 4, 20], i.e., a small but not negligible faction of their nodes has a very large degree. These degree distributions differ from the rapidly decaying Poisson degree distribution, and they have profound effects on the behavior of the network. Examples of degree distributions of real-world networks are presented in Figure 2.

2. **Clustering properties:** While real-world networks have strong clustering, the Erdös and Rényi model does not [192, 191]. The prob-

abilities of pairs of nodes being adjacent in Erdös-Rényi random graphs are by definition independent, i.e., the probability of two nodes being adjacent is the same regardless of whether they have a common neighbor. Thus, the clustering coefficient for a random graph is $C = p$. Table 1 illustrates this by comparing clustering coefficients of real-world and random networks [146].

Since random graphs do not provide an adequate model for real-world networks with respect to degree distributions and network clustering properties, we will review other network models, which fit the real-world networks better.

Table 1: Clustering coefficients, C, for a number of different networks; n is the number of node, z is the mean degree. Taken from [146].

Network	n	z	C measured	C for random graph
Internet [153]	6,374	3.8	0.24	0.00060
World Wide Web (sites) [2]	153,127	35.2	0.11	0.00023
power grid [192]	4,941	2.7	0.080	0.00054
biology collaborations [140]	1,520,251	15.5	0.081	0.000010
mathematics collaborations [141]	253,339	3.9	0.15	0.000015
film actor collaborations [149]	449,913	113.4	0.20	0.00025
company directors [149]	7,673	14.4	0.59	0.0019
word co-occurrence [90]	460,902	70.1	0.44	0.00015
neural network [192]	282	14.0	0.28	0.049
metabolic network [69]	315	28.3	0.59	0.090
food web [138]	134	8.7	0.22	0.065

3.3 Generalized Random Graphs

This model captures power-law degree distribution in a graph, while leaving all other aspects as in the random graph model. That is, the edges are randomly chosen with the constraint that the degree distribution is restricted to a power law. A systematic analysis of these scale-free random networks showed that there is a threshold value of γ in the degree distribution $P(k) \approx k^{-\gamma}$, at which the properties of these networks suddenly change.

Generating a random graph with a non-Poisson degree distribution is relatively simple and has been discussed in a number of papers starting with [30]. Given a degree distribution (as a degree sequence), one can generate

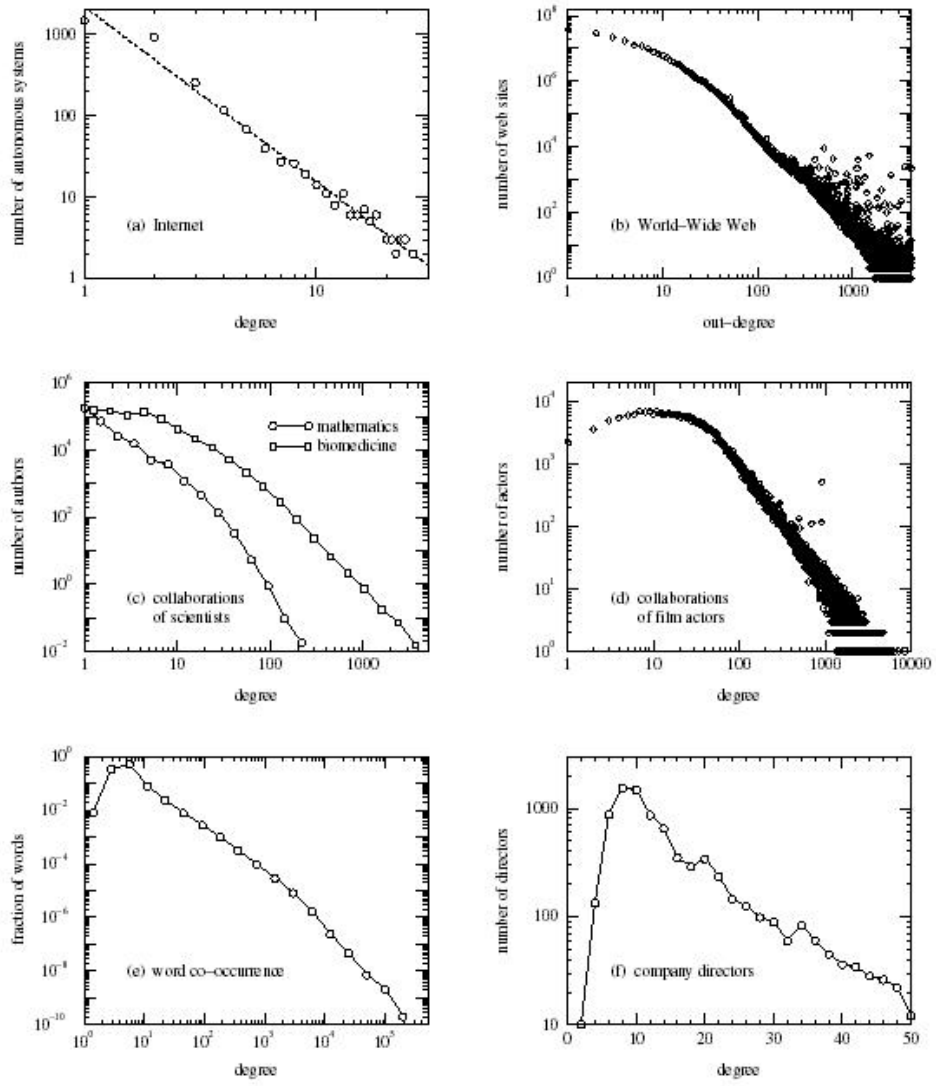

Figure 2: Degree distributions for different networks. (a) Physical connections between autonomous systems on the Internet in 1997 [68], (b) a 200 million page subset of the World Wide Web in 1999 [41], (c) collaborations between biomedical scientists and between mathematicians [140] [141], (d) collaborations of film actors [11], (e) co-occurrence of words in English [90], (f) board membership of directors of Fortune 1000 companies for year 1999 [149]. Taken from [146].

a random graph by assigning to a node i a degree k_i from the given degree sequence, and then choosing pairs of nodes uniformly at random to make edges so that the assigned degrees remain preserved. When all degrees have been used up to make edges, the resulting graph is a random member of the set of graphs with the desired degree distribution. Naturally, the sum of degrees has to be even to successfully complete the above algorithm. Note that this method does not allow the clustering coefficient to be specified, which is one of the crucial properties of these graphs that makes it possible to solve exactly for many of their properties in the limit of large n. For example, if want to find the mean number of second neighbors of a randomly chosen node in a graph with clustering, we have to account for the fact that many of the second neighbors of a node are also its first neighbors as well. However, in a random graph without clustering, the probability that a second neighbor of a node is also its first neighbor behaves as $\frac{1}{n}$, regardless of the degree distribution, and thus can be ignored in the limit of large n [146].

It has been proven that almost all random graphs with a fixed degree distribution, and no nodes of degree smaller than 2, have a unique giant component [116]. There is a simple condition for the birth of the giant component, as well as an implicit formula for its size [136, 137]. More specifically, for $n \gg 1$ and $P(k) = \frac{d_k}{n}$, Q is defined as:

$$Q = \sum_{k=1}^{\infty} P(k)k(k-2),$$

and it was shown that if $Q < 0$ the graph almost always consists of many small components, the average component size almost always diverges as $Q \to 0^-$, and a giant component almost surely emerges for $Q > 0$, under the condition that the maximum degree is less than $n^{\frac{1}{4}}$.

One can apply these results to a random graph model for scale-free networks. Aiello, Chung and Lu showed that for a power-law $P(k)$, the condition on Q implies that a giant component exists if and only if $\gamma < 3.47875... = \gamma_0$ [3]. They also observed several interesting properties for different values of γ:

- When $\gamma > \gamma_0$, the random graph is disconnected and made of independent finite clusters.
- When $\gamma < \gamma_0$, there is almost surely a unique infinite cluster.
- When $2 \leq \gamma < \gamma_0$, the second largest component almost surely has a size of the order of $\ln n$.

- When $1 < \gamma < 2$, every node with degree greater than $\ln n$ almost surely belongs to the infinite cluster, and the size of second largest component does not increase as the size of the graph goes to infinity. Thus, the fraction of nodes in the infinite cluster approaches 1 as the system size increases meaning that the graph becomes totally connected in the limit of infinite system size.

- When $0 < \gamma < 1$, the graph is almost surely connected.

Using the mathematics of generating functions [194], one can calculate exactly many statistical properties of these graphs in the limit of large n [149], such as the emergence of the giant component, the size of the giant component, the average distribution of the sizes of the other components, the average numbers of nodes at a certain distance from a given node, the clustering coefficient, the typical distance between a pair of nodes in a graph, etc. As described in [149], one can define the generating function

$$G_0(x) = \sum_{k=0}^{\infty} P(k)x^k$$

for the probability distribution of node degrees k, where the distribution $P(k)$ is assumed to be normalized so that $G_0(1) = 1$. Then, the condition for the emergence of the giant component is $\sum_k k(k-2)P(k) = 0$ [136, 149] (a positive sum leads to the appearance of a giant cluster). The size of the giant component is $S = 1 - G_0(u)$, where u is the smallest non-negative real solution of the equation $u = G_1(u)$ [137, 149].

This theory has been applied to the modeling of collaboration graphs, which are bipartite, and the World Wide Web, which is directed [149]. It has been shown that this theory gives good order of magnitude estimates of the properties of known collaboration graphs of business people, scientists, and movie actors, although there are measurable differences between theory and data that point to the presence of interesting effects, such as sociological effects in collaboration networks [149].

3.4 Small-world Networks

Networks of many biological, social, and artificial systems often exhibit *small-world* topology, i.e., a small-world character along with unusually large clustering coefficients independent of network size. Watts and Strogatz proposed this one-parameter model of networks in order to interpolate between an ordered finite-dimensional lattice and a random graph [192]. They start

from a ring lattice with n nodes and m edges in which every node is adjacent to "its first k neighbors" on the ring (an illustration is presented in Figure 3), and "re-wire" each edge at random with probability p, not allowing for self-loops and multiple edges. This process introduces $\frac{pnk}{2}$ "long-range" edges. Thus, the graph can be "tuned" between regularity ($p = 0$) and disorder ($p = 1$).

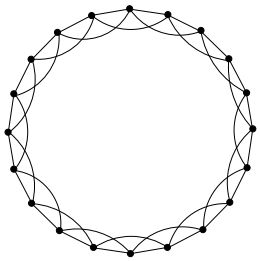

Figure 3: A regular ring lattice for $k = 2$.

Structural properties of these graphs have been quantified by their *characteristic path length* $L(p)$ (the shortest path length between two nodes averaged over all pairs of nodes), and the clustering coefficient $C(p)$ as functions of re-wiring probability p [192]. Remember that if:

$$C_v = \frac{|E(N(v))|}{\frac{1}{2}|N(v)|(|N(v)| - 1)},$$

then C is the average of C_v over all nodes v.

Watts and Strogatz established that the regular lattice at $p = 0$ is a highly clustered "large world" in which L grows linearly with n, since:

$$L(0) \approx \frac{n}{2k} \gg 1 \text{ and } C(0) \approx \frac{3}{2}.$$

On the other hand, as $p \to 1$ the model converges to a random graph, which is a poorly clustered "small world," where L grows logarithmically with n, since:

$$L(1) \approx \frac{\ln n}{\ln k}, \quad \textit{and} \quad C(1) \approx \frac{k}{n}.$$

Note that these limiting cases do not imply that large C is always associated with large L, and small C with small L. On the contrary, even the small amount of re-wiring transforms the network into a small-world with short paths between any two nodes, like in the giant component of a random graph, but at the same time such a network is much more clustered than a random graph.

This is in agreement with the characteristics of real-world networks. For example, the collaboration graph of actors in feature films, the neural network of the nematode worm *C. elegans*, and the electrical power grid of the western United States all have a small-world topology [192]. Thus, this model is generic for many large, sparse networks found in nature [192].

This pioneering work led to developing a large area of research, and many additional empirical examples of small-world networks have been documented [20, 99, 188, 11, 177, 174]. Graphs associated with many different search problems have a small-world topology, and the cost of solving them can have a heavy-tailed distribution [190]. This is due to the fact that local decisions in a small-world topology quickly propagate globally. One can use randomization and restarts to eliminate these heavy tails [190]. Finding a short chain of acquaintances linking oneself to a random person using only local information is solvable only for certain kinds of small worlds [110] (this problem was originally posed by Milgram's sociological experiment [133]).

Significant research in small-world networks exists outside computer science as well. Epidemiologists have studied how local clustering and global contacts together influence the spread of infectious disease, trying to make vaccination strategies and understand evolution of virulence [19, 36, 106, 189]. Neurobiologists have asked about possible evolutionary significance of small-world neural topology. They have argued that small-world topology combines fast signal processing with coherent oscillations [115], and thus was selected by adaptation to rich sensory environments and motor demands [26].

One of the areas most active in research of small-world networks is within statistical physics (a good overview can be found in [144]). A variant of the Watts-Strogatz model was proposed by Newman and Watts [142, 143] in which no edges are removed from the regular lattice and new edges are added between randomly chosen pairs of nodes. This model is easier to analyze, since it does not lead to the formation of isolated components, which could happen in the original model. The formula for a characteristic path length in these networks has been derived to be [148]:

$$L(p) = \frac{n}{k} f(nkp),$$

where

$$f(x) \approx \frac{1}{2\sqrt{x^2+2x}} \tanh^{-1} \frac{x}{\sqrt{x^2+2x}}.$$

This solution is asymptotically exact in the limits of large n and when either $nkp \to \infty$, or $nkp \to 0$ (large or small number of shortcuts). This result

can be improved by finding a rigorous distributional approximation for $L(p)$, together with a bound on the error [24].

Small-world networks have a relatively high clustering coefficient. In a regular lattice ($p = 0$), the clustering coefficient does not depend on the lattice size, but only on its topology. It remains close to $C(0)$ up to relatively large values of p as the network gets randomized. A slightly different, but equivalent definition of C, $C'(p)$, is defined as the fraction between the mean number of edges between the neighbors of a node and the mean number of possible edges between those neighbors [25]. It was used to derive a formula for $C(p)$ [25]. Starting with a regular lattice with a clustering coefficient $C(0)$, and observing that for $p > 0$ two neighbors of a node v that were connected at $p = 0$ are still neighbors of v and connected by an edge with probability $(1-p)^3$, since there are three edges that need to remain intact, it follows that $C'(p) \approx C(0)(1-p)^3$. The deviation of $C(p)$ from this expression is small and goes to zero as $n \to \infty$ [25]. The corresponding expression for the Newman-Watts model [145] is:

$$C'(p) = \frac{3k(k-1)}{2k(2k-1) + 8pk^2 + 4p^2k^2}.$$

The degree distribution of small-world networks is similar to that of a random graph. In the Watts-Strogatz model for $p = 0$, each node has the same degree k. A non-zero p introduces disorder in the network and widens the degree distribution while still maintaining the average degree equal to k. Since only one end of an edge gets re-wired, $\frac{pnk}{2}$ edges in total, each node has degree at least $\frac{k}{2}$ after re-wiring. Thus, for $k > 2$ there are no isolated nodes. For $p > 0$, the degree k_v of a node v can be expressed as $k_v = \frac{k}{2} + c_v$ [25], where c_v is divided into two parts, $c_v = c_v^1 + c_v^2$, so that $c_v^1 \leq \frac{k}{2}$ edges have been left in place with probability $1 - p$, and c_v^2 edges have been re-wired towards v, each with probability $\frac{1}{n}$. For large n the probability distributions for c_v^1 and c_v^2 are:

$$P_1(c_v^1) = \binom{\frac{k}{2}}{c_v^1}(1 - p^{c_v^1})p^{\frac{k}{2} - c_v^1}$$

and

$$P_2(c_v^2) = \binom{\frac{pnk}{2}}{c_v^2}(\frac{1}{n})^{c_v^2}(1 - \frac{1}{n})^{\frac{pnk}{2} - c_v^2} \approx \frac{(pk/2)^{c_v^2}}{c_v^2!}e^{-pk/2}.$$

Combining these two factors, the degree distribution is:

$$P_p(c) = \sum_{n=0}^{min(c-k/2, k/2)} \binom{k/2}{n}(1-p)^n p^{k/2-n} \times \frac{(pk/2)^{c-k/2-n}}{(c-k/2-n)!}e^{-pk/2},$$

for $c \geq \frac{k}{2}$. As p grows, the distribution becomes broader, but it stays strongly peaked at the average degree with an exponentially decaying tail.

3.5 Scale-free Networks

In many real networks connectivity of some nodes is significantly higher than for the other nodes. For example, the degree distributions of the Internet backbone [68], metabolic reaction networks [99], the telephone call graph [1], and the World Wide Web [41] decay as a power law $P(k) \approx k^{-\gamma}$, with the exponent $\gamma \approx 2.1 - 2.4$. This form of heavy-tailed distribution would imply an infinite variance, but in reality there are only a few nodes with many links, such as search engines for the World Wide Web.

The earliest work on the theory of scale-free networks dates back to 1955 [172], but it has recently been rediscovered [20, 21, 38]. A heavy-tailed degree distribution in these networks emerges automatically from a stochastic growth model, in which new nodes are added continuously and they preferentially attach to existing nodes with probability proportional to the degree of the target node [21]. That is, high-degree nodes become of even higher degree with time and the resulting degree distribution is $P(k) \approx k^{-3}$. Further, if either the growth, or the preferential attachment is eliminated, the resulting network does not exhibit scale-free properties [21]. Thus, both the growth and preferential attachment are needed simultaneously to produce the power-law distribution observed in real networks.

The average path length in the Barabasi-Albert network [21] is smaller than in a random graph, indicating that a heterogeneous scale-free topology is more efficient in bringing nodes close together than the homogeneous random graph topology. Analytical results show that the average path length, ℓ, satisfies $\ell \approx \frac{\ln n}{\ln \ln n}$ [35]. Interestingly, while in random graph models with arbitrary degree distribution the node degrees are uncorrelated [3, 149], nontrivial correlations develop spontaneously between the degrees of connected nodes in the Barabasi-Albert model [111]. There has been no analytical prediction for the clustering coefficient of the Barabasi-Albert model.

It has been observed that the clustering coefficient of a scale-free network is about five times higher than that of a random graph, and that this factor slowly increases with the number of nodes [5]. However, the clustering coefficient of the Barabasi-Albert model decreases with the network size approximately as $C \approx n^{-0.75}$, which is a slower decay than the $C = \frac{<k>}{N}$ for random graphs, where $< k >$ denotes the average degree, but is still different from the small-world models in which C is independent of n.

The Barabasi-Albert model is a minimal model that captures the mech-

anisms responsible for the power-law degree distributions observed in real networks. There are discrepancies between this model and real networks. While the exponent of the predicted power-law distribution for the model is fixed, real networks have measured exponents varying between 1 and 3.

This discrepancy has led to an increased interest in addressing network evolution questions. The theory of evolving networks offers insights into network topology and its evolution. More sophisticated models including the effects of adding or re-wiring edges, allowing nodes to age so that they can no longer accept new edges, or varying the form of preferential attachment have been developed [4, 56, 112].

In addition to scale-free degree distributions, these generalized models also predict exponential and truncated power-law degree distribution in some parameter regimes. Scale-free networks are resistant to random failures due to a few high-degree "hubs" dominating their topology: any node that fails probably has a small degree, and thus does not severely affect the rest of the network [7]. However, such networks are vulnerable to deliberate attacks on the hubs. These intuitive ideas have been confirmed numerically [7, 41] and analytically [44, 50] by examining how the average path length and size of the giant component depend on the number and degree of the removed nodes. Implications have been made for the resilience of the Internet [38], the design of therapeutic drugs [99], and the evolution of metabolic networks [99, 188].

To generate networks with scale-free topologies in a deterministic, rather than stochastic way, Barabasi, Ravasz, and Vicsek have introduced a simple model, which they solved exactly showing that the tail of the degree distribution of the model follows a power law [23]. The first steps of the construction are presented in Figure 4. The construction can be viewed as follows. The starting point is a P_3. In the next iteration, add two more copies of a P_3 and connect the mid-point of the initial P_3 with the outer nodes of the two new P_3s. In the next step, make two copies of the 9-node module constructed in the previous step, and connect "end" nodes of the two new copies to the "middle" node of the old module (as presented in Figure 4). This process can continue indefinitely. The degree distribution of such a graph behaves as $P(k) \approx k^{\frac{\ln 3}{\ln 2}}$. An additional property that these networks have is the hierarchical combination of smaller modules into larger ones. Thus, these networks are called "hierarchical".

Another deterministic graph construction, called "pseudo-fractal", has been proposed to model evolving scale-free networks [55]. The scheme of the growth of the scale-free pseudo-fractal graph is presented in Figure 5.

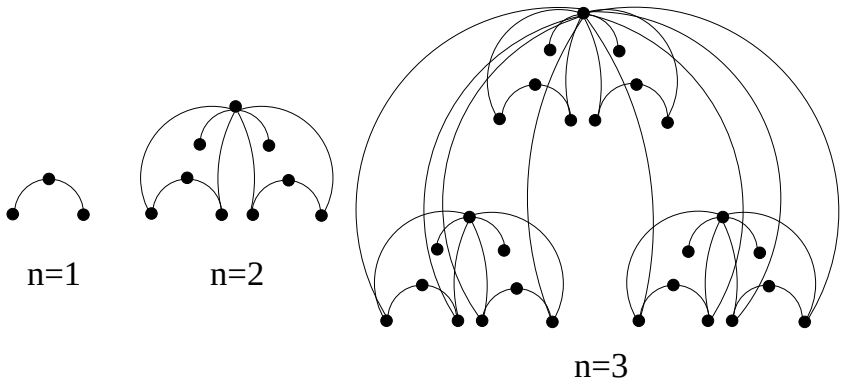

Figure 4: Scheme of the growth of a scale-free deterministic hierarchical graph. Adapted from [23].

The degree distribution of this graph can be characterized by a power law with exponent $\gamma = 1 + \ln 3 / \ln 2 \approx 2.585$, which is close to the distribution of real growing scale-free networks. All main characteristics of the graph were determined both exactly and numerically [55]. For example, the shortest path length distribution follows a Gaussian of width $\approx \sqrt{\ln n}$ centered at $\bar{l} \approx \ln n$ (for $\ln n \gg 1$), clustering coefficient of a degree k node follows $C(k) = 2/k$, and the eigenvalue spectrum of the adjacency matrix of the graph follows a power-law with the exponent $2 + \gamma$.

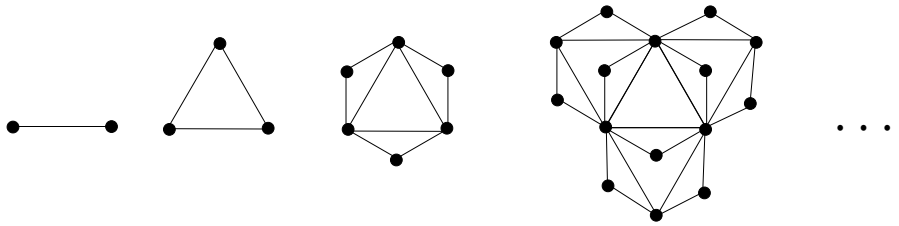

Figure 5: Scheme of the growth of the scale-free pseudofractal graph. The graph at time step $t + 1$ can be constructed by "connecting together" three graphs from step t. Adapted from [55].

4 Protein Interaction Networks

With DNA sequence becoming available for an increasing number of organisms, there is a growing interest in correlating the genome with the proteome

to explain biological function and to develop new, effective protein-targeting drugs. One of the key questions in proteomics today is identifying interactions of proteins and other molecules. The hope is to exploit this information to understand both the disease and healthy states, and to use this information for developing new therapeutic approaches.

Different methods have been used to identify protein interactions, including diverse biochemical and computational approaches. A survey of biochemical methods used to identify proteins and PPIs can be found in [152]. Unfortunately, most of these methods are lab-intensive, have strong biases, and low accuracy. Despite the resulting low confidence in the identified interactions, maps of PPIs are being constructed and their analysis is increasingly attracting attention. Many laboratories throughout the world are contributing to one of the ultimate goals of the biological sciences – the creation and understanding of a full molecular map of a cell. The computational focus is to analyze currently available networks of PPIs.

Despite many shortcomings of representing a PPI network using the standard mathematical representation of a network, with proteins being represented by nodes and interactions being represented by edges, this has been the only mathematical model used so far to represent and analyze these networks. This model does not address the following important properties of PPI data sets:

1. There is a large percentage of false-positive interactions in these data sets. For example, a common class of false-positive PPIs happens when in reality proteins indirectly interact, i.e., through mediation of one or more other molecules, but an experimental method detects this as a direct physical interaction. This may be a reason why very dense subnetworks are being detected inside PPI networks.

2. False-negative interactions are also present in these networks resulting from imperfect experimental techniques used to identify interactions. Since different biochemical methods have different biases, they detect different subsets of false-negatives. Thus, finding "high confidence" PPIs as overlap of multiple data sets may discard real interactions.

3. The following PPI properties are not captured by this model: spatial and temporal information, information about experimental conditions, strength of the interactions, number of experiments confirming the interactions, etc. Without this information, no true modeling and simulation can be performed. This has partially been addressed by von Mering *et al.* who classified PPIs into groups depending on the

number of experiments that detected a specific PPI; they call this a *level of confidence* that a given PPI is a true interaction [187].

Next, we give an overview of recent PPI identification methods, currently available PPI data sets and their repositories, biological structures contained in the PPI networks (such as protein complexes and pathways), and computational methods used to identify them in PPI networks. Then we focus on surveying the literature on PPI network properties and structure. We point to open problems and future research directions in the area of PPI networks in Section 6.

4.1 PPI Identification Methods

The lists of genes and encoded proteins are becoming available for an increasing number of organisms. Databases, such as Ensembl [87] and GenBank [31] (described in the next section) contain publicly available DNA sequences for more than $105,000$ organisms, including whole genomes of many organisms in all three domains of life, bacteria, archea, and eukaryota, as well as their protein data. In parallel to the increasing number of genomes becoming available, high-throughput PPI detection methods have been producing a huge amount of interaction data. Such methods include yeast 2-hybrid systems [91, 92, 185], protein complex purification methods using mass spectrometry [72, 86], correlated messenger RNA (m-RNA) expression profiles [88], genetic interactions [132], and *in silico* interaction predictions derived from gene fusion [62], gene neighborhood [54], and gene co-occurrences or phylogenetic profiles [89]. None of these PPI detection methods is perfect and the rates of false positives and false negatives vary from method to method. A brief summary describing these methods can be found in [187]. Next we outline the main characteristics of each of these methods.

Yeast 2-hybrid assay is an *in vivo* technique involving fusing one protein to a DNA-binding domain and the other to a transcriptional activator domain. An interaction between the two proteins is detected by the formation of a functional transcription factor. This technique detects even transient and unstable interactions. However, it is not related to the physiological setting. Also, the method will not detect interactions where three or more proteins need to be involved.

Mass spectrometry of purified complexes involves tagging individual proteins which are used as hooks to biochemically purify whole protein complexes. The complexes are separated and their components identified by

mass spectrometry. There are two protocols, tandem affinity purification (TAP) [72, 163], and high-throughput mass-spectrometric protein complex identification (HMS-PCI) [86, 120]. This technique detects real complexes in their physiological settings and enables a consistency check by tagging several members of a complex at the same time. However, its drawbacks are that it might miss some complexes that are not present under the given conditions, tagging can disturb complex formation, and loosely associated components can be washed off during purification. More details are presented in Chapter **??**.

Correlated m-RNA expression (synexpression) involves measuring m-RNA levels under many different cellular conditions and grouping genes that show a similar transcriptional response to these conditions. The groups that encode physically interacting proteins were shown to frequently exhibit this behavior [73]. This is an indirect *in vivo* technique, which has a much broader coverage of cellular conditions than other techniques. However, it is very sensitive to parameter choices and clustering methods used during the analysis, and thus is not very accurate for predicting direct physical interactions.

Genetic interactions is an indirect *in vivo* technique that involves detecting interactions by observing the phenotypic results of gene mutations. An example of a genetic interaction is *synthetic lethality*, which involves detecting pairs of genes that cause lethality when mutated at the same time. These genes are frequently functionally associated, and thus their encoded proteins may physically interact. However, synthetically lethal genes may also represent alternative cellular wiring, and not a protein interaction.

In silico predictions through genome analysis involve screening whole genomes for the following types of interaction evidence: 1) finding conserved operons in prokaryotic genomes which often encode interacting proteins [54]; 2) finding similar phylogenetic profiles, since interacting proteins often have similar phylogenetic profile, i.e., they are either both present, or both absent from a fully sequenced genome [89]; 3) finding proteins that are found fused into one polypeptide chain [62]; 4) finding structural and sequence motifs within the protein-protein interfaces of known interactions that allow the construction of general rules for protein interaction interfaces [100, 101].

In silico methods are fast and relatively inexpensive techniques whose coverage expands with more and more organisms becoming fully sequenced. However, they require orthology between proteins and fail when orthology relationships are not clear. Also, these methods generally favor high sensitivity at the cost of low specificity, and thus they produce large number of false positives.

4.2 Public Data Sets

Vast amounts of biological data that are being generated by many HTP methods are deposited in numerous databases. Different PPI databases contain PPIs from different single experiments, HTP experiments, and literature sources. PPIs resulting from the most recent studies are usually only available on the journal web sites where the corresponding papers appeared. Standardization efforts to represent and organize PPI databases [85] will eventually improve public data sets. However, due to relatively high flexibility in the data format proposed, data curation and integration will remain challenging.

Here we briefly introduce the main databases, including nucleotide sequence, protein sequence, and PPI databases. Nucleotide and protein sequence databases do not suffer from the lack of standardization that is present in PPI databases. A recent comprehensive list of major molecular biology databases can be found in [28].

The largest nucleotide sequence databases are EMBL[1] [179], GenBank[2] [31], and DDBJ[3] [182]. They contain sequences from the literature as well as those submitted directly by individual laboratories. These databases store information in a general manner for all organisms. Organism specific databases exist for many organisms. For example, the complete genome of bakers yeast and related yeast strains can be found in Saccharomyces Genome Database (SGD)[4] [58]. FlyBase[5] [12] contains the complete genome of the fruit fly *Drosophila melanogaster*. It is one of the earliest model organism databases. Ensembl[6] [87] contains the draft human genome sequence along with its gene prediction and large scale annotation. It currently contains over 4,300 megabases and 29,000 predicted genes, as well as information about predicted genes and proteins, protein families, domains etc. Ensembl is not only free, but is also open source.

SwissProt[7] [18] and Protein Information Resource (PIR)[8] [128] are two major protein sequence databases. They are both manually curated and contain literature links. They exhibit a large degree of overlap, but still contain many sequences that can be found in only one of them. SwissProt

[1] http://www.ebi.ac.uk/embl/
[2] http://www.ncbi.nlm.nih.gov/Genbank/
[3] http://www.ddbj.nig.ac.jp/
[4] http://genome-www.stanford.edu/Saccharomyces/
[5] http://flybase.bio.indiana.edu/
[6] http://www.ensembl.org/
[7] http://www.ebi.ac.uk/swissprot/
[8] http://pir.georgetown.edu/

maintains a high level of annotation for each protein including its function, domain structure, and post-translational modification information. It contains over 101,000 curated protein sequences. Computationally derived translations of EMBL nucleotide coding sequences that have not yet been integrated into the SwissProt resource can be found in Trembl[9]. The Non-Redundant Database (NRDB)[10] merges two sequences into a representative sequence if they exhibit a large degree of similarity. This is useful when a large scale computational analysis needs to be performed.

There is a growing number of public databases comprising PPI data for one or multiple organisms, including the following major resources:

- The Munich Information Center for Protein Sequences (MIPS)[11] provides high quality curated genome related information, such as PPIs, protein complexes, pathways etc., for several organisms [132].

- Yeast Proteomics Database (YPD)[12] is a curated database, comprising bakers yeast, *S. cerevisiae*, protein information, including their sequence and genetic information, related proteins, PPIs, complexes, literature links, etc. [52].

- The Database of Interacting Proteins (DIP)[13] is a curated database containing information about experimentally determined PPIs. It catalogues around $11,000$ unique interactions between $5,900$ proteins from over 80 organisms including yeast and human [197]. However, it still maintains only 988 human PPIs, compared to the $15,358$ interactions in *S. cerevisiae*, $13,178$ for *D. melanogaster*, and $5,500$ for *C. elegans*.

- Human Reference Protein Database (HRPD)[14] is a large-scale effort to manually catalogue many known human PPIs related to diseases in a database. Currently, HPRD comprises $12,641$ human protein interactions derived from literature sources [155]. However, it is currently in a form not suitable for large-scale analysis, although the trend is to provide export in PSI format [85].

[9] http://www.ebi.ac.uk/trembl/
[10] http://www.ebi.ac.uk/ holm/nrdb90
[11] http://mips.gsf.de
[12] http://www.incyte.com/sequence/proteome/databases/YPD.shtml
[13] http://dip.doe-mbi.ucla.edu/
[14] http://www.hprd.org/

- The Biomolecular Interaction Network Database (BIND)[15] archives biomolecular interaction, complex, and pathway information [14]. BIND stores interactions of diverse biological objects: a protein, RNA, DNA, molecular complex, small molecule, photon, or gene. This database includes Pajek [27] as a network visualization tool. It includes a network clustering tool as well, called the Molecular Complex Detection (MCODE) algorithm [16] (described in section 4.3), and a functional alignment search tool (FAST) [14] (details of FAST are not yet available in the literature), which displays the domain annotation for a group of functionally related proteins.

- The General Repository for Interaction Datasets (GRID)[16] stores genetic and physical interactions. It contains interactions from several genome and proteome wide studies, as well as the interactions from MIPS and BIND databases [39]. It also provides a powerful network visualization tool called Osprey [40].

- A Molecular INTeraction database (MINT)[17] contains about $2,500$ interactions curated manually from the literature [201].

- Online Predicted Human Interaction Database (OPHID)[18], is a web-based database of predicted interactions between human proteins [42]. It incorporates the known interactions from DIP, MINT and HPRD, and combines them with predictions made from *S. cerevisiae*, *C. elegans*, *D. melanogaster*, and *M. musculus*. We have done extensive investigation into biological evidence to support these predictions. OPHID currently contains $22,487$ predicted interactions, including $1,323$ that are supported by multiple pieces of evidence (domains, co-expression and GO terms), while another $7,955$ are supported by a single piece of evidence. The OPHID database can be queried using single IDs or in a batch mode, and results can be displayed as text, HTML, or visualized using our custom visualization tool. In addition, the entire database is available in tab-delimited text or PSI-compliant XML format [85].

As experimentally derived PPIs include both false positives and negatives, and individual detection methods have different sensitivity and specificity, it is important to assess the quality of these datasets.

[15] http://www.binddb.org/
[16] http://biodata.mshri.on.ca/grid/
[17] mint.bio.uniroma2.it/mint/
[18] http://ophid.utoronto.ca

Von Mering *et al.* have performed a systematic synthesis and evaluation of PPIs detected by major high-throughput PPI identification methods for yeast *S. Cerevisiae* [187], a model organism relevant to human biology [165]. They integrated 78,390 interactions between 5,321 yeast proteins, out of which only 2,455 are supported by more than one PPI detection method. This low overlap between the methods may be due to a high rate of false positives, or to difficulties in detecting certain types of interactions by specific methods. Research bias is another potential explanation of the low overlap; research groups usually have interest focused on finding interactions between certain types of proteins. Further, each PPI identification technique produces a unique distribution of interactions with respect to functional categories of interacting proteins [187]. Assessing the quality of interaction data produced a list of 78,390 yeast PPIs ordered by the level of confidence (high, medium, and low) with the highest confidence being assigned to interactions confirmed by multiple methods [187]. The resulting list of PPIs currently represents the largest publicly available collection of PPIs for *S. Cerevisiae*, and also the largest PPI collection for any organism.

There are other approaches to improve the quality of PPI data sets, such as correlating transcriptome and interactome data [77, 73], computational use of statistical and topological descriptors combined with transcriptional information and other annotations [17], probabilistic modeling of interactions based on the available evidence [13], combining information about evolutionary conserved and essential proteins [196], integrating large screens with information about chromosomal proximity, and phylogenetic profiling and domain fusion [131].

Clearly, false negatives will dominate error in PPI datasets. Due to diverse biases, non-overlap among available databases does not always imply false positives. Careful integration of multiple information sources [46] and hypothesis-driven design of HTP experiments, similar to [114], will enable effective increases in PPI databases, while improving data quality.

4.3 Biological Structures Within PPI Networks and Their Extraction

Biochemical studies used to identify biological structures, such as complexes and pathways, are expensive, time consuming, and of low accuracy. One approach to reduce the time and cost, and increase accuracy of these studies is to computationally detect biological structures from PPI networks. The hope is that with the emergence of high confidence PPI networks, such as the one constructed by Mering et al. [187], computational approaches will

become inexpensive and reliable tools for extraction of known and prediction of still unknown members of these structures.

Despite a large body of literature involving purely theoretical aspects of networks, such as finding clusters in graphs (see Section 5), only a few such methods have been developed specifically for biological applications and applied to PPI networks.

4.3.1 Protein Complexes

Cellular processes are usually carried out by groups of proteins acting together to perform a certain function. These groups of proteins are called *protein complexes*. They are not necessarily of invariable composition, i.e., a complex may have several core proteins, which are always present in the complex, as well as more dynamic, perhaps regulatory proteins, which are only present in a complex from time to time. Also, the same protein may participate in several different complexes during different cellular activities. One of the most challenging aspects of PPI data analysis is determining which of the myriad of interactions in a PPI network comprise true protein complexes [13, 16, 60, 72, 86, 108, 183].

Several mass spectrometry studies have been used to identify protein complexes in yeast *S. cerevisiae*, and report improved data quality compared to yeast 2-hybrid method. Ho *et al.* used HMS-PCI to extract complexes from the *S. cerevisiae* proteome [86]. They reported an approximately threefold higher success rate in detection of known complexes when compared to large-scale yeast 2-hybrid studies [185, 91]. Gavin *et al.* have performed a mass-spec analysis of the *S. cerevisiae* proteome to identify protein complexes [72], with about 70% probability of detecting the same protein in two different purifications. Amongst $1,739$ yeast genes, including $1,143$ human orthologues, they purified 589 protein assemblies, out of which 98 corresponded to protein complexes in the Yeast Protein Database (YPD), 134 were new complexes, and the remaining ones showed no detectable association with other proteins. This led to proposing a new cellular function for 344 proteins, including 231 proteins with no previous functional annotation. They attempted investigating relationships between complexes in order to understand the integration and coordination of cellular functions by representing each complex as a node and having an edge between two nodes if the corresponding complexes share proteins. By color-coding complexes according to their cellular roles they noticed grouping of the same colored complexes, suggesting that sharing of components reflects functional relationships. No graph theoretic analysis of this protein complex network has

been done so far. Comparing human and yeast complexes showed that orthologous proteins preferentially interact with complexes enriched with other orthologues, supporting the existence of an "orthologous proteome", which may represent core functions for all eukaryotic cells [72]. This leads to stronger evolution-based interaction conservation [70, 181, 196], which will increasingly play a role as we move to more complex organisms [37].

Diverse computational approaches have been proposed to identify protein complexes from PPI networks. They have involved measuring connectedness (e.g., k-core concept [15]), node neighborhood "cliquishness" [192] (e.g., MCODE method [16]), partitioning the network's node set into clusters based on a cost function that is assigned to each partitioning [108], or the reliance on reciprocal bait-hit interactions as a measure of complex involvement. The challenge in this analysis is complexity and scalability of the algorithm, and most importantly its specificity and sensitivity. Evaluating these performance characteristics is not trivial, due to the following reasons:

- Since existing databases of protein complexes are incomplete, computationally identified complexes may incorrectly appear as false positives. A few biological validations also does not conclusively prove quality of the algorithm, as only a small fraction of many tests may succeed overall. Thus, multiple computational, comparative, and experimental combinations have to be used for performance evaluation.

- Increasing sensitivity usually decreases specificity, and thus multi-level algorithms that use additional filters to remove potential false positives are necessary. However, depending on the overall goal of the analysis one can change the tradeoff between sensitivity and specificity, and a given algorithm can be tuned for a specific task. Thus, when comparing algorithms one must consider this selection.

The Molecular Complex Detection (MCODE) algorithm [16] exploits the notion of a clustering coefficient [192] (described in Section 3). Bader and Hogue used the notion of a *k-core*, a graph of minimum degree k, and a notion of the "highest k-core of a graph", the most densely connected k-core of a graph, to weight PPI network nodes in the following way. A core-clustering coefficient of a node v is defined as a density of the highest k-core of $N[v]$, where density is the number of edges of a graph divided by the maximum possible number of edges of the graph. The weight of a node v is the product of the node core-clustering coefficient and the highest k-core level, k_{max}, of the $N(v)$. A complex is seeded in the weighted graph with the highest weighted node, and nodes with weights above a given threshold

are recursively included in the complex. The process is repeated for the next highest weighted unexplored node. This generation phase is followed by post-processing, which discards complexes that do not contain a k-core with $k \geq 2$.

The following two options are also included in the algorithm: an option to "fluff" the complexes by adding to them their neighbors unexplored by the algorithm of weight bigger than the "fluff" parameter, and an option to "haircut" the complex by removing nodes of degree 1 from the complex. The resulting complexes are scored according to the product of the complex node set size and the complex density, and they are ranked according to the scoring function. The MCODE algorithm also offers an option to specify a seed node.

MCODE was evaluated against known complexes in the MIPS database [132] and 221 complexes from [72]. In an attempt to maximize the overlap between predicted and known complexes, all combinations of the parameters (true/false for haircut and fluff, and node weight percentage in 0.05 increments) have been varied. However, only 88 out of the 221 complexes from [72] matched, and only 52 of 166 predicted complexes matched MIPS complexes. MCODE identified complexes of high density, which were highly likely to match real complexes. Thus, high sensitivity was achieved at the cost of low specificity.

These results suggest that a different approach of finding efficient graph clustering algorithms to identify highly connected subgraphs should be used to identify protein complexes in PPI networks. We explored this approach and used Hartuv and Shamir's Highly Connected Subgraph (HCS) algorithm [83, 84] (described in section 5) to identify protein complexes from a yeast PPI network with $11,000$ interactions amongst $2,401$ proteins [160]. The algorithm detected a number of known protein complexes (an illustration is presented in Figure 6). Also, 27 out of 31 clusters identified in this way had high overlaps with protein complexes documented in MIPS database. The remaining 4 clusters that did not overlap MIPS contained a functionally homogeneous 6-protein cluster Rib 1-5 and a cluster Vps20, 25, 36, which are likely to correspond to protein complexes. In addition, the clusters identified in this way had a statistically significant functional homogeneity.

A similar approach explored three different methods for identification of highly connected subgraphs in a PPI network constructed on the MIPS database [173]. The first method involves identifying complete subgraphs of the PPI graph. The second method is the Super-Paramagnetic Clustering algorithm [33] to cluster objects in a non-metric space of an arbitrary dimension. The third method comprises a novel Monte-Carlo optimization

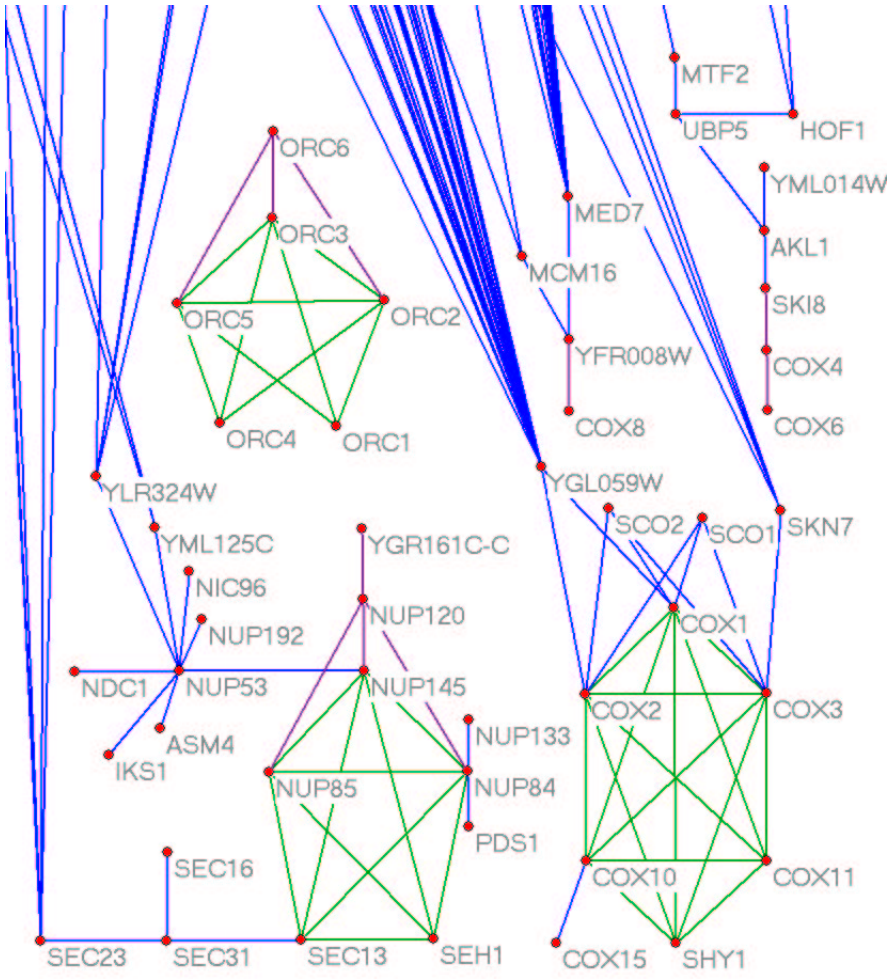

Figure 6: A subnetwork of a yeast PPI network [160] showing some of the identified complexes (green). Violet lines represent PPIs to proteins not identified as biological complex members due to stringent criteria about their connectivity in the algorithm, or due to absence of protein interactions that would connect them to the identified complex (from more details see [160]).

technique to identify highly connected subgraphs in a network (the details of this algorithm have not been published yet). Most of the dense subnetworks that were identified in this way had consistent functional annotation revealing the function of the whole complex [173]. Also, dense subgraphs had a good agreement with the known protein complexes from MIPS, BIND, and the data from [86]. Also, several novel complexes and pathways were predicted, but no further details about these are given [173].

A similar approach has been used to predict functions of uncharacterized proteins [43]. Spectral graph theory methods, previously used for analyzing the World Wide Web [74, 109], have been applied to the yeast PPI network constructed on high and medium confidence interactions from [187]. This method identified "quasi-cliques" and "quasi-bipartites" in the PPI network. Since proteins participating in quasi-cliques usually shared common functions, this method was used to assign function to 76 uncharacterized proteins.

4.3.2 Molecular Pathways

Molecular *pathways* are chains of cascading molecular reactions involved in maintaining life. Different processes involve different pathways. Some examples include metabolic, apoptosis, and signaling pathways for cellular responses. An example of a signaling pathway transmitting information from the cell surface to the nucleus where it causes transcriptional changes, is presented in Figure 7.

Disruption in a pathway function may cause severe diseases, such as cancer. Thus, understanding molecular pathways is an important step in understanding cellular processes and the effects of drugs on cellular processes. As a consequence, modeling and computational pathway prediction from PPI networks has become an active research area.

The Biopathways Consortium[19] was founded to catalyze the emergence and development of computational pathways biology. One of their main goals is to coordinate the development and use of open technologies, standards, and resources for representing, handling, accessing, and analyzing pathways information. Numerous papers addressing these topics have been presented at the Biopathways Consortium Meetings. Many of them used classical graph algorithms in order to integrate genome-wide data on regulatory proteins and their targets with PPI data in yeast [198], reconstruct microbial metabolic pathways [129], determine parts of structure and evolution of metabolic networks [117], etc.

[19] http://www.biopathways.org/

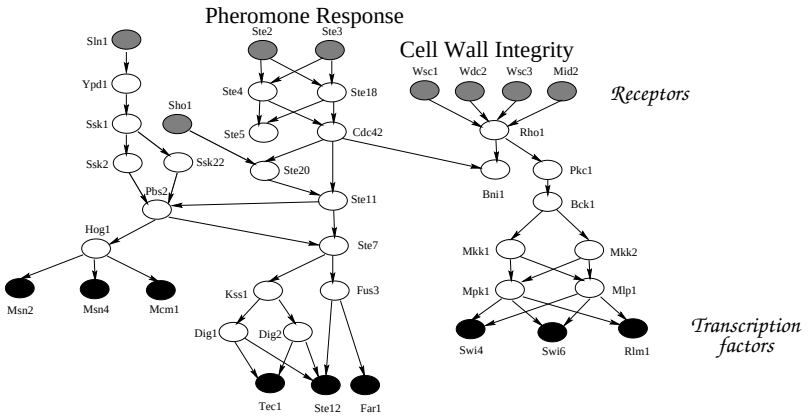

Figure 7: Examples of MAPK (mitogen-activated protein kinases) signal transduction pathways in yeast. Gray nodes represent membrane proteins, black nodes represent transcription factors, and white nodes represent intermediate proteins.

Using simple graph theory and a yeast 2-hybrid data from [91, 167, 185], a model of a *S. cerevisiae* signal transduction network has been constructed [175]. In order to reduce the number of candidate signaling pathways from around 17 million to around 4.4 million, the most highly connected nodes have been first deleted, and then shortest paths of length at most eight between every membrane protein and every DNA-binding protein has been identified in the modified network. To maximize high-scoring pathways in the real PPI network and minimize those in the randomized networks, several parameters have been optimized: the number of clusters in which genes were grouped, the microarray expression datasets used in clustering, the maximum path length of their pathways, and the scoring metric. The resulting putative pathways have been compared with the ones obtained in the same way in three randomized PPI networks. Only paths of length at most 8 have been identified, because the average shortest path length between any two proteins in the given PPI graph was 7.4, and because a fraction of pathways with high microarray clustering ratios over various shortest path lengths peaked at 8. This method reproduced many essential elements of the pheromone response, cell wall integrity, and filamentation MAPK pathways, but it failed to model the High Osmolarity (HOG) MAPK pathway due to missing interactions (false-negatives) in the PPI network.

Although in general pathways are complex, one can identify and model linear components of some pathways. Using this model, it is possible to

predict novel linear pathways [160]. We focused on finding and exploiting the basic structure that these pathways have in PPI networks. We used MAPK as our model pathway, because they are among the most thoroughly studied networks in yeast and because they exhibit linearity in structure [164]. Initial analysis showed that these pathways have source and sink nodes of low degree and internal nodes of high degree in the yeast PPI network. This information was used to create a linear pathway model, and to extract such putative pathways from a PPI network. The approach is based on the following steps [160]:

- Construct a shortest path from a transmembrane or sensing protein of low degree to a transcription factor protein of low degree, such that the internal nodes on the shortest path are of high degree.
- Increase sensitivity by including high degree first and second neighbors of internal nodes of such a shortest path into these predicted pathways (for more details see [160]).

Using this approach, we extracted 399 putative pathways (an example of a predicted pathway is presented in Figure 8).

Other theoretical approaches have been proposed to model pathways. They involve system stoichiometry, thermodynamics, etc. (for example, see [166]). Also, methods for extraction of pathway information from on-line literature are being developed [71, 113, 150].

4.4 Properties of PPI Networks

Systematically analyzing PPI network properties may be used to assess functional meaning of individual proteins and subgraphs within these networks. In addition, comparing properties of multiple PPI networks can be used to evaluate evolutionary characteristics, robustness of the organism, and quality of a given data set.

4.4.1 Scale-Free Network Topology

One of the first approaches to modeling biological networks focused on metabolic pathway networks of different organisms [97, 99] from the WIT database [151]. This database contains predicted pathways based on the annotated genome of an organism and data established from the biochemical literature. This analysis showed that metabolic networks of 43 organisms from the WIT database, containing 6 archaea, 32 bacteria, and 5 eukaryota,

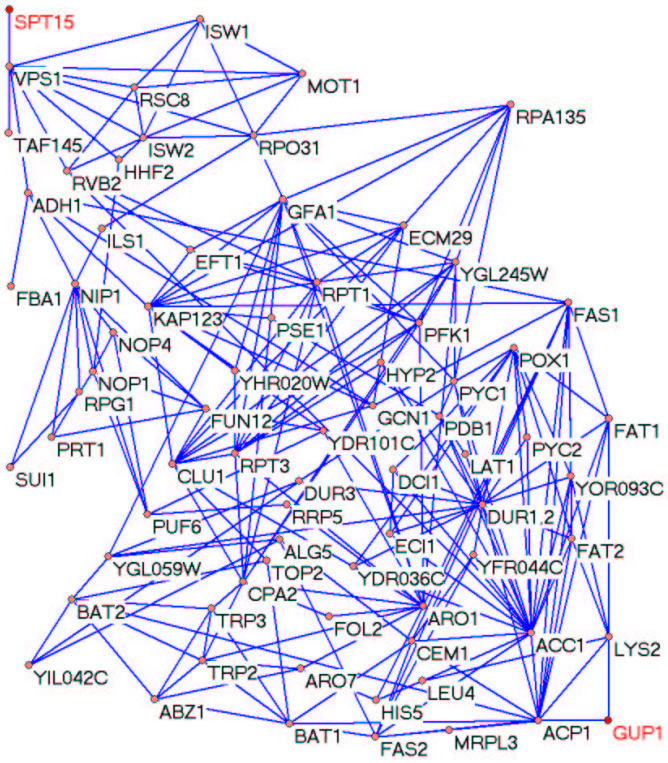

Figure 8: An example of a predicted pathway [160]. Note that this predicted pathway is presented as a subgraph of the PPI graph, and thus some of its internal nodes appear to be of low degree, even though they have many more interactions with proteins outside of this predicted pathway in the PPI graph.

all have scale-free topology with $P(k) \approx k^{-2.2}$ for both in- and out-degrees [99]. The diameter of the metabolic networks was the same for all 43 organisms, indicating that with increasing organism complexity, nodes are increasingly connected. A few hubs dominated these networks, and upon the sequential removal of the most connected nodes, the diameter of the network rose sharply. Only around 4% of the nodes were present in all species, and these were the ones that were most highly connected in any individual organism; species-specific differences among organisms emerged for less connected nodes. A potential bias introduced by a high interest and research being done on some and a lack of interest and research being done on other proteins may also have contributed to this effect. In addition,

randomly removing nodes from these networks, the average shortest path lengths did not change, indicating insensitivity to random errors in these networks.

Subjecting the *S. cerevisiae* PPI network constructed on $1,870$ proteins and $2,240$ interactions derived from the yeast 2-hybrid study [185] and the DIP database [197] to the same analysis established that the yeast PPI network and the PPI network of the human gastric pathogen *Helicobacter pylori* [161] also have heterogeneous scale-free network topology with a few highly connected proteins and numerous less connected proteins. To study robustness of a PPI network, one can correlate removal of a protein of a certain degree and lethality. Interestingly, the same tolerance to random errors, coupled with fragility against the removal of high-degree nodes was observed as in the metabolic networks:

- even though about 93% of proteins had degree at most 5, only about 21% of them were essential;
- only 0.7% of the yeast proteins with known phenotypic profiles had degree at least 15, but 62% of them were essential.

These results suggest that there is evolutionary selection of a common large-scale structure of biological networks and that future systematic PPI network studies in other organisms should uncover an identical PPI network topology [97]. Our results on a larger yeast PPI network confirm this hypothesis [160].

A *genetic regulatory network* of a cell is formed by all pairs of proteins in which one protein directly regulates the abundance of the other. In most of these networks regulation happens at the transcriptional level, where a transcription factor regulates the RNA transcription of the controlled protein. These networks are naturally directed. The analysis of a regulatory network from the YPD database with 682 proteins and $1,289$ edges [52] as well as of the PPI network from 2-hybrid screens with $2,378$ proteins and $4,549$ interactions [91] revealed that both networks had a small number of high-degree nodes (hubs) [121]. Both of these networks had edges between hubs systematically suppressed, while those between a hub and a low-connected protein were favored [121]. Furthermore, hubs tended to share few neighbors with other hubs. This led to the hypothesis that these effects decrease the likelihood of "cross talk" between different functional modules of the cell and increase the overall robustness of a network by localizing effects of harmful perturbations [121]. This may explain why the correlation between the con-

nectivity of a given protein and the lethality of the mutant cell lacking this protein was not strong [97].

However, an alternative explanation of this phenomenon suggests that hubs whose removal disconnects the PPI graph are likely to cause lethality [160]. Analyzing the top 11,000 interactions among 2,401 proteins from [187], which utilizes high confidence interactions detected by diverse experimental methods [160], we confirmed the previously noted result on smaller networks [97], demonstrating that:

- *viable* proteins, whose disruption is non-lethal, have a degree that is half that of *lethal* proteins, whose mutation causes lethality (see Figure 9);
- proteins participating in *genetic interaction pairs* in the PPI network, i.e., combinations of non-lethal mutations which together lead to lethality or dosage lethality, appeared to have a degree closer to that of viable proteins;
- *lethal* proteins are more frequent in the top 3% of high degree nodes compared to viable ones, while viable mutations were more frequent amongst the nodes of degree 1.

Interestingly, lethal mutations were not only highly connected nodes within the network, but were nodes whose removal caused a disruption in network structure – it disconnected one part of the network from the other [160].

The obvious interpretation of these observations in the context of cellular wiring is that lethality can be conceptualized as a point of disconnection in the PPI network. A contrasting property to hubs, which are at the same time points of disconnection, is the existence of alternative connections, called *siblings*, i.e., nodes in a graph with the same neighborhood. The analysis shows that viable mutations have an increased frequency in the group of proteins that could be described as siblings within the network, compared to lethal mutations or genetic interactions [160]. This suggests the existence of alternate paths bypassing viable nodes in PPI networks, and offers an explanation why null mutation of these proteins is less likely to be lethal [160].

However, it is possible that the reason why lethal mutations in current PPI networks have large degrees is due to the extensiveness of research being done in the neighborhood of these nodes. It is possible that the observed low-degrees of viable nodes are due to the high rage of false negatives present in PPI networks, i.e., the lack of interest and research being done in these

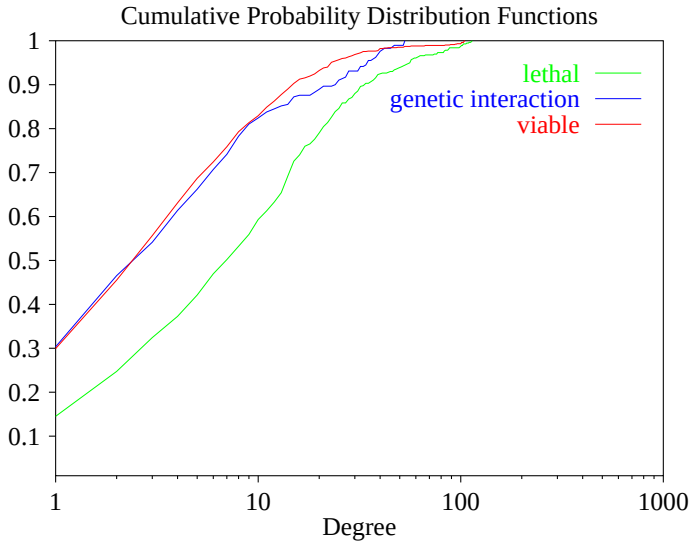

Figure 9: Cumulative distribution functions of degrees of lethal, genetic interaction, and viable protein groups in a yeast PPI network constructed on 11,000 interactions amongst 2,401 proteins [160].

parts of PPI networks. This is further supported by the recently observed weak correspondence between mutational robustness and the topology of PPI networks [53].

4.4.2 Hierarchical Network Topology

Further analysis of metabolic networks of 43 organisms from the WIT database [151] presented a dichotomy between the two phenomena found in metabolic networks [162]:

1. these networks are scale-free with the observed power law degree distribution [20, 99, 188];

2. these networks include hubs, which integrate all nodes into a single, integrated network;

3. these networks have high clustering coefficients [188], which imply modular topologies.

Determining the average clustering coefficients of metabolic networks of 43 different organisms established that all were an order of magnitude

larger than expected for a scale-free network of similar size [162]. This suggested high modularity of these networks. Also, the clustering coefficients of metabolic networks were independent of their sizes, contrasting the scale-free model, for which the clustering coefficient decreases as $n^{-0.75}$. It is possible to integrate the seemingly contradicting phenomena of modularity and integration using a heuristic model of metabolic organization, called a "hierarchical" network [162]. Thus, metabolic networks appear to be organized into many small, highly connected modules, which are combined in a hierarchical manner into larger units [162].

The "hierarchical" network construction is similar to the one described in [23] (see Section 3 for the discussion), but a starting point in this network is a K_4 as a hypothetical module (rather than a P_3 [23]). Nodes of this starting module are connected with nodes of three additional copies of K_4 so that the "central node" of the initial K_4 is connected to the three "external nodes" of new K_4s, as presented in Figure 10 (b), which generates a 16-node module. This process is repeated by making three additional copies of this 16-node module and connecting the "peripheral nodes" of the three new 16-node modules with the "central node" of the initial 16-node module (see Figure 10 (c)). This process can be repeated indefinitely. The architecture of this network has the following characteristics:

- it integrates a scale-free topology with a modular structure;
- it has a power law degree distribution with $P(k) = k^{-2.26}$, which is in agreement with the observed $P(k) \approx k^{-2.2}$ [99];
- it has a clustering coefficient $C \approx 0.6$, which is comparable to coefficients observed for metabolic networks;
- it has a clustering coefficient independent of the network size.

The hierarchical structure of this model is a feature that is not shared by the scale-free, or modular network models. It was also demonstrated that for each of the 43 organisms, the clustering coefficient $C(k)$ of a degree k node is well approximated by $C(k) \approx k^{-1}$. This is in agreement with the theoretical result establishing that the clustering coefficient of a degree k node of a scale-free network follows the scaling law $C(k) \approx k^{-1}$ [55] (further discussed in Section 3.5)). Thus, this hierarchical network model includes all the observed properties of metabolic networks: the scale-free topology, the high, system size independent clustering coefficient, and the power law scaling of $C(k)$.

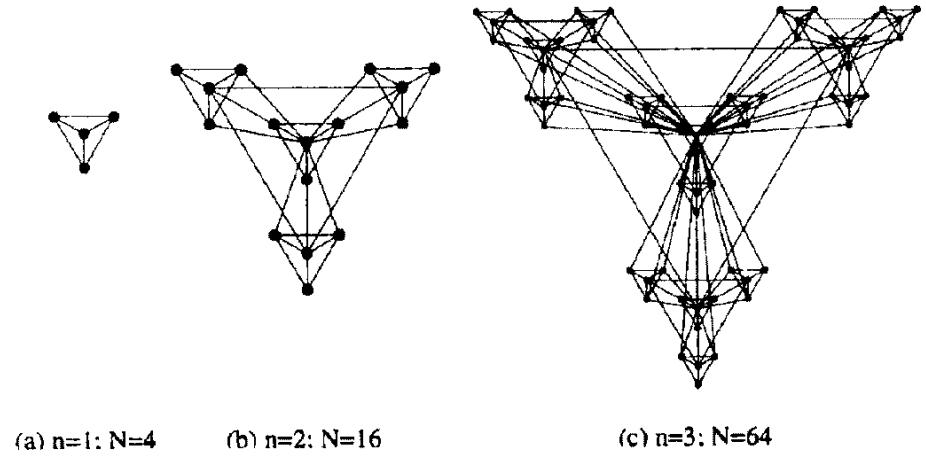

Figure 10: Three steps in the construction of a hierarchical model network. Taken from [162].

To inspect whether this model reflects the true functional organization of cellular metabolism, the focus was turned on the extensively studied metabolic network of *E. coli* [162]. It was established that the model closely overlaps with *E. coli*'s known metabolic network. It was hypothesized that this network architecture may be generic to cellular organization [162]. The existence of small, highly frequent subgraphs in these networks, called "network motifs" [135, 171] makes this hypothesis even more plausible.

These results have been further extended [96] by applying the same analysis on the complete biochemical reaction network of the 43 organisms from the WIT database [151]. The networks were constructed by combining all pathways deposited in the WIT database for each organism into a single network. These networks are naturally directed, which made it possible to examine their in- and out-degree distributions. All of the 43 networks obtained in this way exhibited a power-law distribution for both in- and out-degrees, which suggested that scale-free topology is a generic structural organization of the total biochemical reaction networks in all organisms in all three domains of life [96].

However, the largest portion of the WIT database contains data on core metabolism pathways, followed by the data on information transfer pathways. Thus, these results may have largely been influenced by the domination of metabolic pathways.

To resolve this issue, the same analysis has been performed on the information transfer pathways alone, since apart from the metabolic pathways, these were the only ones present in high enough quantities for doing statistical analyses. The analysis of the information transfer pathways of 39 organisms (four of the 43 organisms had their information pathways of too small size for doing statistics) revealed the same power-law degree distribution both for in- and out-degree as seen for metabolic and complete biochemical reaction networks [96]. Similarly, it was confirmed that the network diameter (which they defined as the average of shortest path lengths between each pair of nodes) remained constant and around 3 for biochemical reaction networks, metabolic networks, and information transfer networks of all 43 organisms, irrespective of the network sizes. Thus, in these networks, the average degree of a node increases with the network size. This is contrary to the results on real non-biological networks, in which the average degree of a node is fixed, so the diameter of the network increases logarithmically with the network size [20, 26, 192]. Further, only about 5% of all nodes were common to the biochemical reaction networks of all 43 species, and these were the highest degree nodes. The same result was observed when repeated analysis was conducted for metabolic and information transfer networks alone.

Applying the same approach, with minor variation of the hierarchical network model (presented in Figure 11) to four independent yeast PPI networks derived from the DIP database [197], data from [91], the MIPS database [132], and data from [185], showed that all networks had hierarchical structures with $C(k)$ scaling as k^{-1} [22].

4.4.3 Network Motifs

"Network motifs" can be defined as patterns of interconnections that recur in many different parts of a network at frequencies much higher than those found in randomized networks [171]. The *transcriptional regulation network* can be represented as a directed graph in which each node represents an *operon*, a group of contiguous genes that are transcribed into a single m-RNA molecule, and each edge is directed from an operon that encodes a transcription factor to an operon that is regulated by that transcription factor [171]. Using network motifs to analyze the transcriptional regulation network of *Escherichia coli* showed that much of the network is composed of repeated appearances of three highly significant motifs, each of which had a specific function in determining gene expression [171]:

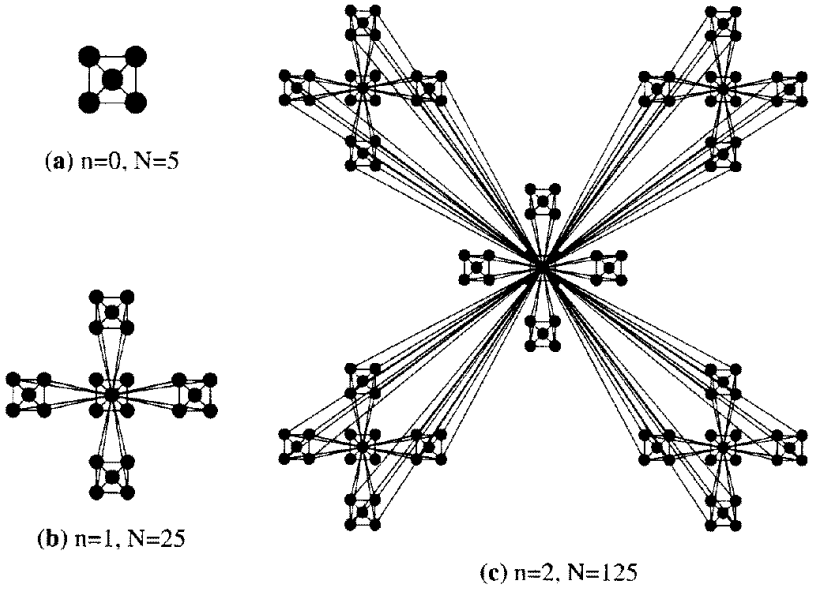

Figure 11: Three steps in the construction of a hierarchical model network. Taken from [22].

1. "feedforward loop";
2. "single-input module" (SIM);
3. "dense overlapping regulons" (DOR) (a *regulon* stands for a group of coordinately regulated operons).

An illustration of these three motifs is presented in Figure 12. Network motifs on 3 and 4 nodes and SIMs have been detected using straightforward adjacency matrix manipulation algorithms. DORs have been identified applying a standard average-linkage clustering algorithm [57], and using a non-metric distance measure between operons [171].

Individual motifs work differently. Feedforward loops can act as circuits that reject transient activation signals and respond only to persistent signals, while allowing a rapid system shutdown: X and Y act in an AND-gate-like manner to control operon Z [171]. SIMs allow temporal ordering of activation of different genes with different activation thresholds, which is useful for processes that require several stages to complete, such as amino-acid biosynthesis processes.

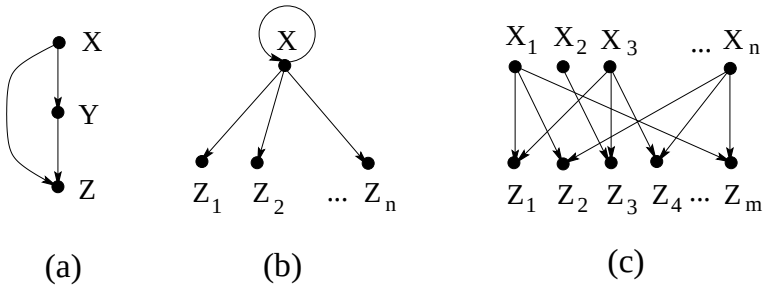

Figure 12: Motifs from [171]: (a) feedforward loop, (b) single input module (SIM), (c) dense overlapping regulons (DOR).

In addition to explaining functionality, motifs can be used to represent the transcriptional network in a compact, modular form [171]. This suggests that performing similar analysis would enable us to identify diverse motifs describing different functional protein groups in large, undirected PPI networks. This in turn would provide mathematical description of cell processes, predict new processes, and aid in determining function of uncharacterized proteins (see section 6).

Further network motif analysis of multiple large networks (*E. coli* and *S. cerevisiae* gene regulation networks (transcription), the neuron connectivity network of *C. elegans*, seven food web networks, the ISCAS89 benchmark set of sequential logic electronic circuits, and a network of directed hyperlinks between World Wide Web pages within a single domain) showed that different networks have different motifs [135]. All possible 3- and 4-node directed subnetworks in these real large networks were found and the frequencies of occurrences of each of these small subnetworks in a real network were compared with the frequencies of their occurrences in randomized networks that have the same connectivity properties and the same number of $(n-1)$-node subgraphs as the real networks, where n is the size of the motif being detected.

The challenge of this analysis is to generate appropriate random networks. During the analysis, one must account for patterns that appear only because of the single-node characteristics of the network, such as the presence of nodes with a large degree, and also to ensure that a high significance is not assigned to a pattern only because it has a highly significant sub-pattern. Following this principle, "network motifs" are defined as those patterns for which the probability of appearing in a randomized network an equal of greater number of times than in the real network is lower than 0.01 [135]. This again brings the tradeoff of higher specificity for lower sensitiv-

ity, as this approach could miss functionally important but not statistically significant patterns.

It was observed that the number of appearances of each motif in the real networks appears to grow linearly with the system size, while it drops in the corresponding random networks [135]. This drop is in accordance with an exact result on Erdös-Rényi random graphs in which the concentration C of a subgraph with n nodes and m edges (i.e., the fraction of times a given n-node subgraph occurs among the total number of occurrences of all possible n-node subgraphs) scales with network size S, as $C \approx S^{n-m-1}$ [34], which in the study of Milo *et al.* is equal to $\frac{1}{S}$, since all but one of their motifs have $n = m$.

In addition, it was established that the identified motifs were insensitive to data errors, since they do not change after addition, deletion, or rearrangement of 20% of the edges at random.

This approach was also applied to an undirected yeast PPI network on $1,843$ nodes and $2,203$ edges [97], which identified one 3-node motif, one 4-node motif, and two 4-node "anti-motifs". Anti-motifs are defined as patterns whose probability of appearing in randomized networks fewer times than in the real network is less than 0.01, and $N_{rand} - N_{real} > 0.1 N_{rand}$, where N_{rand} and N_{real} are the number or subgraph appearances in a real and in randomized networks respectively [135].

An approach to study similarity in the local structure of networks was proposed [134]. A real network was compared with a set of randomized networks with the same degree sequence in the following way. For each 3- and 4-node subgraph i, the statistical significance was expressed by a Z score, $Z_i = (Nreal_i - < Nrand_i >)/std(Nrand_i)$, where $Nreal_i$ is the number of times (i.e., the frequency) the subgraph appears in the network, and $< Nrand_i >$ and $std(Nrand_i)$ are the mean and standard deviation of its appearances in the randomized networks. Then the *significance profile (SP)* for the network is the vector of normalized Z scores, $SP_i = Z_i/(\sum Z_i^2)^{1/2}$. Superfamilies of previously unrelated networks were found based on the similarity of their SPs [134]. For example, protein signaling, developmental genetic networks, and neuronal wiring formed a distinct superfamily. Also, power grids, protein-structure networks and geometric networks formed a network superfamily. Structure and function of some network motifs in transcription networks has been suggested [118, 119].

4.4.4 Geometric Network Topology

We introduced another bottom-up approach of measuring local network structure and showed that the local structure of PPI networks corresponds to the local structure of geometric random graphs [156]. In the *geometric random graph* model, nodes correspond to independently and uniformly randomly distributed points in a metric space, and two nodes are linked by an edge if the distance between them is smaller than or equal to some radius r, where distance is an arbitrary distance norm in the metric space (more details about geometric random graphs can be found in [154]).

We use the term *graphlet* to denote a connected network with a small number of nodes. All of the 29 different 3-, 4-, and 5-node graphlets are presented in Figure 13. We exhaustively searched and found all occurrences of every one of these 29 graphlets in four PPI networks and the corresponding Erdös-Rényi random graphs, generalized random graphs, scale-free networks, and geometric random graphs.

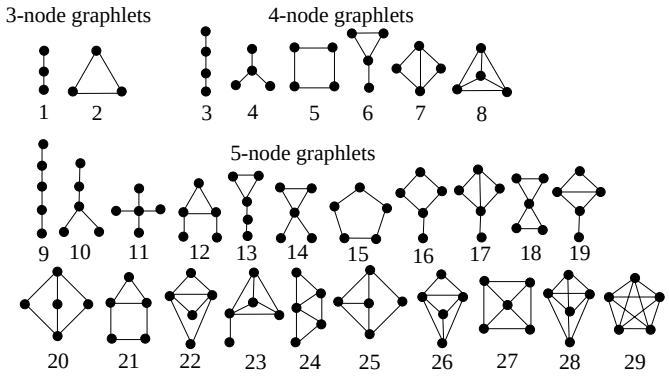

Figure 13: All 3-node, 4-node, and 5-node connected networks (graphlets), ordered within groups from the least to the most dense with respect to the number of edges when compared to the maximum possible number of edges in the graphlet; they are numbered from 1 to 29.

Graphlet counts quantify the local structural properties of a network. Currently, our knowledge of the connections in PPI networks is incomplete. The edges we *do* know are dominated by experiments focused around proteins that are currently considered "important". However, we hypothesize that the local structural properties of the full PPI network, once all connections are made, are similar to the local structural properties of the currently known, highly studied parts of the network. Thus, we expect that the *relative* frequency of graphlets among the currently known connections is similar

to the relative frequency of graphlets in the full PPI network, which is as yet unknown. Therefore, we use the *relative frequency of graphlets* $N_i(G)/T(G)$ to characterize PPI networks and the networks we use to model them, where $N_i(G)$ is the number of graphlets of type i ($i \in \{1, \ldots, 29\}$) in a network G, and $T(G) = \sum_{i=1}^{29} N_i(G)$ is the total number of graphlets of G. In this model, then, the "similarity" between two graphs should be independent of the total number of nodes or edges, and should depend only upon the differences between relative frequencies of graphlets. Thus, we define the *relative graphlet frequency distance* $D(G, H)$, or *distance* for brevity, between two graphs G and H as

$$D(G, H) = \sum_{i=1}^{29} |F_i(G) - F_i(H)|,$$

where $F_i(G) = -\log(N_i(G)/T(G))$.

Using this distance measure, the graphlet frequency distributions of high-confidence PPI networks were closest to the graphlet frequency distributions of geometric random networks [156]. In addition, all global network parameters of high-confidence PPI networks, except for the degree distribution, were closest to the parameters of geometric random networks. However, with added noise in PPI networks, their graphlet distribution and global network parameters become closer to these parameters of scale-free networks. Thus, we expect that the structure of noise-free PPI networks is close to the structure of geometric random networks.

Exhaustively searching for all instances of a graphlet in a large network is computationally intensive. Thus, heuristic techniques for finding approximate frequencies of small subgraphs in large networks have been developed [102]. Analytical solutions for the numbers of 3- and 4-node subgraphs in directed and undirected geometric networks have recently been obtained [93]. Also, topological generalizations of network motifs have been suggested [103].

4.4.5 Function-Structure Relationship in PPI Networks

As discussed above, network organization is not random, and network properties characterize its function. It has been established that complex networks comprise simple building blocks [135, 171], and that distinct functional classes of proteins have differing network properties [160].

Since different building blocks and modules have different properties, it can be expected that they serve different functions. Assuming the functional

classifications in the MIPS database [132], it is possible to statistically determine simple graph properties for each functional group [160]. This analysis shows that proteins involved in translation appear to have the highest average degree, while transport and sensing proteins have the lowest average degree. Figures 14 A and 14 B support this result as half of the nodes with degrees in the top 3% of all node degrees are translation proteins, while none belong to amino-acid metabolism, energy production, stress and defense, transcriptional control, or transport and sensing proteins. This is further supported by the observation that metabolic networks across 43 organisms tested have an average degree of < 4 [99].

Intersecting each of the lethal, genetic interaction, and viable protein sets with each of the functional groups shows that amino-acid metabolism, energy production, stress and defense, transport and sensing proteins are less likely to be lethal mutations (see Figure 14 C). Of all functional groups, transcription proteins have the largest presence in the set of lethal nodes on the PPI graph; approximately 27% of lethals on the PPI graph are transcription proteins, as illustrated in Figure 14 C. Notably, amongst all functional groups, cellular organization proteins have the largest presence in hub nodes whose removal disconnects the network (the nodes whose removal disconnects the network we called *articulation points*; see Figure 14 D).

This strong network structure-function relationship can be exploited for computational prediction. For example, we constructed a simple model for predicting new genetic interaction pairs in the yeast PPI network [160]. The predictive model is based on the distribution of shortest path lengths between known genetic interaction pairs in the PPI network.

New approaches integrating different HTP methods in order to describe known and to predict new biological phenomena continue to appear. Since all HTP techniques contain noise, but often noise caused by different phenomena, resulting data sets not only complement each other, but their integration may also reduce the noise. One approach in this direction is integrating graph theoretic PPI analysis with the results of microarray experiments (for example, see [98]). We discuss some novel approaches in Chapter **??**.

5 Detection of Dense Subnetworks

There is growing evidence that although currently available PPI networks contain a high degree of false positives and false negatives, they do have structure. One of the main goals is to discover more PPI network structure and ultimately exploit it for designing efficient, robust, and reliable algo-

rithms for extracting graph substructures embedded in these networks that have biological meaning and represent biological processes.

Our previous discussion suggests that one example of such substructures may be dense subgraphs of these networks, representing core proteins of protein complexes. Thus, next we describe some useful graph theoretic techniques that can be used as a first step towards addressing extraction of these dense subgraphs in PPI graphs. However, it is possible that protein complexes (and pathways, too) have a distinct graph theoretic structure requiring novel graph theoretic approaches for their detection in PPI networks.

Clustering is an important problem in many disciplines, including computational biology. Its goal is to partition a set of elements into subsets called *clusters*, so that the elements of the same cluster are similar to each other (this property is called *homogeneity*) and elements from different clusters are not similar to each other (this property is called *separation*). Homogeneity and separation can be defined in many different ways leading to different optimization problems. Elements belonging to the same cluster are usually called *mates* and the elements belonging to different clusters are called *non-mates*.

Clustering problems and algorithms can be expressed in graph-theoretic terms. For example, a *similarity graph* can be constructed so that nodes represent elements and edge weights represent similarity values of the corresponding elements. In the analysis of PPI networks, weights on edges have not yet been incorporated in the model, but it may be useful to incorporate them to represent the confidence that the two proteins interact (as in [187]), the strength of the interaction, or possibly the timing of an interaction to distinguish those that are transient.

Several graph theoretic techniques have been developed to cluster microarray gene expression profiles. An overview of these and other gene expression clustering methods can be found in [168]. Identifying protein complexes using graph theoretic methods requires recognizing "dense" subgraphs of PPI networks. Existing algorithms can be characterized as exact algorithms when they have proven properties in terms of solution quality and time complexity, and approximate when heuristics are used to make them more efficient [84, 170]. Some of the algorithms have a probabilistic nature [29, 61, 186].

The early work on graph theoretic clustering determined that highly connected regions of similarity graphs are useful in cluster analysis [122, 123, 124, 125]. This work introduced the *cohesiveness function* for every node and edge in a graph as the maximum edge-connectivity of any subgraph containing that node/edge. By deleting all elements of a graph of cohesiveness less

than k, a maximal k-edge-connected subgraphs of the graph are obtained. First, clusters are identified by using a constant k [123]. This approach was modified to obtain, for any k, clusters which are maximal k-edge-connected subgraphs that do not contain a subgraph with higher connectivity [124]. Several important graph cluster concepts have been introduced in [125]:

- k-bond – a maximal subgraph S such that every node in S has degree at least k in S;
- k-component – a maximal subgraph S such that every pair of nodes in S is joined by k edge-disjoint paths in S;
- k-block – a maximal subgraph S such that every pair of nodes in S is joined by k node-disjoint paths in S.

These notions imply cluster methods with successive refinements going from bond to component to block. These algorithms require solving a minimum cut network flow problem and their time complexities are at least cubic in the input graph node set size for connected graphs. To cope with the complexity, several faster exact graph clustering algorithms have been proposed, some using heuristics for further speed up [83, 84, 170]. A good survey of other graph theoretic clustering techniques, including the probabilistic ones can be found in [8].

The Highly Connected Subgraph (HCS) [83, 84] and CLuster Identification via Connectivity Kernels (CLICK) [170] algorithms operate on a similar principle. The input is a similarity graph, and the algorithm first considers if the graph satisfies a stopping criterion, in which case it is declared a "kernel". Otherwise, the graph is partitioned into two subgraphs, separated by a minimum weight edge cut, and the algorithm recursively proceeds on the two subgraphs, outputting in the end a list of kernels that represent a basis for the possible clusters. The overview of this general algorithm scheme is presented in Algorithm 1 (adapted from [168]). HCS and CLICK have several distinguishing properties, they construct similarity graphs differently and have different stopping criteria.

Algorithm 1: FORM-KERNELS(G)

```
if V(G) = {v} then
    move v to the singleton set
end
else
    if G is a kernel then
        output V(G)
    end
end
else
    (H, H̄) ← MinWeightEdgeCut(G);
    Form-Kernels(H);
    Form-Kernels(H̄);
end
```

The input into the HCS is an unweighted similarity graph G. A *highly connected subgraph (HCS)* is defined to be an induced subgraph H of G, such that the number of edges in a minimum edge cut of H is bigger than $\frac{|V(H)|}{2}$. Thus, if any $\lfloor\frac{|V(H)|}{2}\rfloor$ of edges of H are removed, H remains connected. The algorithm uses these highly connected subgraphs as kernels. Clusters satisfy the two properties [83]:

- clusters are homogeneous, since the diameter of each cluster is at most 2 and each cluster is at least half as dense as a clique;
- clusters are well separated, since any non-trivial split by the algorithm happens on subgraphs that are likely to be of diameter at least 3.

The running time of the HCS algorithm is bounded by $2N \times f(n,m)$, where N is the number of clusters found (often $N \ll n$) and $f(n,m)$ is the time complexity of computing a minimum edge cut of a graph with n nodes and m edges. Currently the fastest deterministic minimum edge cut algorithms for unweighted graphs are of time complexity $O(nm)$ [126, 139]. The fastest simple deterministic minimum edge cut algorithm for weighted graphs is of time complexity $O(nm + n^2 \log n)$ [178] (it is implemented by Mehlhorn and is part of the Leda library [130]).

Several heuristics can be used to speed up the HCS algorithm. Since HCS arbitrarily chooses a minimum edge cut when the same minimum cut value is obtained by several different cuts, this process will often break small clusters into singletons. To avoid this, *Iterated HCS* runs several iterations of HCS,

until no new cluster is found. Theoretically, this would add another $O(n)$ factor to the running time, but in practice only between 1 and 5 iterations are usually needed. *Singletons Adoption* heuristics is based on the principle that singleton nodes get "adopted" by clusters based on their similarity to the clusters. For each singleton node x, the number of neighbors of x in each cluster as well as in the set of all singletons $\mathcal{S}$ is computed, and x is added to a cluster (never to $\mathcal{S}$) with the maximum number of neighbors $\mathcal{N}$, if $\mathcal{N}$ is sufficiently large. This process is repeated a specified number of times to account for changes in clusters resulting from previous adoptions. The last HCS algorithm heuristic is based on *removing low degree nodes*. If the input graph contains many low degree nodes, one iteration of the minimum edge cut algorithm may only separate a low degree node from the rest of the graph contributing to increased computational cost at a low informative value in terms of clustering. This is especially expensive for large graphs with many low degree nodes. For example, around 28% of the nodes of the PPI graph constructed on the top $11,000$ interactions (and $2,401$ proteins) using data from [187], and around 13% of the nodes of the PPI graph constructed on all $\approx$ 78K of the yeast PPIs (and $5,321$ proteins) [187] are of degree 1. Thus, this heuristic may significantly speed up the HCS algorithm applied to these data sets.

We implemented the HCS algorithm without any heuristics and applied it to several PPI graphs constructed on the data set from [187], as described in section 4.3.1. Our results show that clusters identified this way have high overlap with known MIPS protein complexes and a much higher functional group homogeneity than expected at random [160]. Thus, high specificity is favored by this method of protein complex identification. In contrast, Bader and Hogue's approach improves sensitivity at the expense of specificity [16].

The CLICK algorithm builds on the HCS algorithm [83] by incorporating statistical techniques to identify kernels [170]. Similar to HCS, a weighted similarity input graph is recursively partitioned into components using minimum weight edge cut computations. The edge weights and the stopping criterion of the recursion have probabilistic meaning. Pairwise similarity values between mates are assumed to be normally distributed with mean μ_T and variance σ_T, and pairwise similarity values between non-mates are assumed to be normally distributed with mean μ_F and variance σ_F, where $\mu_T > \mu_F$ (this is observed on real data and can also be theoretically justified [170]). The probability p_{mates} of two randomly chosen elements being mates is taken into account when computing edge weights.

If the input similarity matrix between elements is denoted by $S = (S_{ij})$,

the weight of an edge (i,j) in the similarity graph is computed as:

$$w_{ij} = \ln \frac{Prob(i,j \text{ are mates}|S_{ij})}{Prob(i,j \text{ are non-mates}|S_{ij})} = \ln \frac{p_{mates} f(S_{ij}|i,j \text{ are mates})}{(1-p_{mates}) f(S_{ij}|i,j \text{ are non-mates})},$$

where

$$f(S_{ij}|i,j \text{ are mates}) = \frac{1}{\sqrt{2\pi}\sigma_T} e^{-\frac{(s_{ij}-\mu_T)^2}{2\sigma_T^2}}$$

and

$$f(S_{ij}|i,j \text{ are non-mates}) = \frac{1}{\sqrt{2\pi}\sigma_F} e^{-\frac{(s_{ij}-\mu_F)^2}{2\sigma_F^2}},$$

and thus

$$w_{ij} = \ln \frac{p_{mates}\sigma_F}{(1-p_{mates})\sigma_T} + \frac{(S_{ij}-\mu_F)^2}{2\sigma_F^2} - \frac{(S_{ij}-\mu_T)^2}{2\sigma_T^2}.$$

To increase efficiency, edges whose weight is below a predefined non-negative threshold are removed from a graph. A connected subgraph G is called *pure*, if $V(G)$ contains only elements of some cluster. For each cut C of a connected graph, the following hypotheses are tested:

- H_0^C : C contains only edges between non-mates.
- H_1^C : C contains only edges between mates.

G is called a kernel if it is pure, which it is if and only if H_1^C is accepted for every cut C of the graph G. If G is not a kernel, then it is partitioned along a cut C for which the ratio $Pr(H_1^C|C)/Pr(H_0^C|C)$ is minimum. Kernels obtained this way are expanded to obtain clusters, first by singleton adoptions, then by merging "similar" clusters, and finally by performing another round of singleton adoptions. For more details, see Algorithm 2 and [170].

Algorithm 2: CLICK(G)

```
Singletons S ← complete set of elements N;
while some change occurs do
    Execute FORM-KERNELS(G(S));
    Let K be the list of produced kernels;
    Let S be the set of singletons produced;
    Adoption(K, S)
end
Merge(K);
Adoption(K, S)
```

CLICK performance can be improved by using the following two heuristics. Similar to removing low degree nodes for HCS, low weight nodes[20] are filtered from large components in the following way. The average node weight W is computed for the component and multiplied by a factor proportional to the logarithm of the component size; the result is denoted by W^*. Nodes with weight less than W^* are removed repeatedly, updating the weight of the remaining nodes each time a node is removed, until the updated weight of all remaining nodes is greater than W^*. The removed nodes are added to the singleton set and handled later. Although CLICK uses the fastest minimum weight edge cut algorithm [45], with the complexity $O(n^2\sqrt{m})$ [82], the second heuristic exploits a different approach to computing a minimum edge cut. Instead of finding computationally expensive minimum weight edge cuts, they computed a minimum $s-t$ cut of the underlying unweighted graph using $O(nm^{2/3})$ time algorithm [67], with s and t chosen to be nodes that achieve the diameter d of the graph, when $d \geq 4$ (the $O(n+m)$ time breadth first search algorithm is used to find the diameter of the graph).

A polynomial time algorithm for finding the clustering with high probability under the stochastic model of the data has been introduced in [29]. It assumes that the correct structure of the input graph is a disjoint union of cliques (cliques represent clusters), but that errors were introduced to it by independently adding or removing edges with probability $\alpha < \frac{1}{2}$. This heuristic Cluster Affinity Search Technique (CAST) algorithm is built on the theoretical Parallel Classification with Cores (PCC) algorithm, which solves the problem to a desired accuracy with high probability in time $O(n^2(\log n)^c)$ [29].

The input to CAST is the similarity matrix S. CAST uses the notion of the *affinity* of an element v to a putative cluster C, $a(v) = \sum_{i \in C} S(i,v)$, and the affinity threshold parameter t. It generates clusters sequentially by starting with a single element and adding or removing elements from a cluster if their affinity is larger or lower than t, respectively. This process is repeated until it stabilizes. The details are shown in Algorithm 3. In the end, an additional heuristic tries to ensure that each element has the affinity to its assigned cluster higher than to any other cluster by moving elements until the process converges, or some maximum number of iterations is completed.

[20] The weight of a node v is the sum of weights of the edges incident on v.

Algorithm 3: CAST(S)

```
while there are unclustered elements do
    Pick an unclustered element to start a new cluster C;
    repeat
        add an unclustered element v with maximum affinity to C, if
        a(v) > t|C|;
        remove an element u from C with minimum affinity, if a(u) ≤
        t|C|;
    until no changes occur;
    Add C to the list of final clusters;
end
```

A Markov Cluster (MCL) algorithm has been introduced to cluster undirected unweighted and weighted graphs [186]. MCL simulates flow within a graph, promoting flow where the current is strong and demoting flow where the current is weak until the current across borders between different groups of nodes withers away, revealing a cluster structure of the graph (an illustration is presented in Figure 15).

The MCL algorithm deterministically computes the probabilities of random walks through the graph and uses two operators, expansion and inflation, to transform one set of probabilities into another. It uses stochastic matrices (also called Markov matrices) that capture the mathematical concept of random walks on a graph.

Following the notation from [186], for a weighted directed graph $G = (V, E)$, with $|V| = n$, its *associated matrix* M_G is an $n \times n$ matrix with entries $(M_G)_{pq} (1 \leq p, q \leq n)$ being equal to weights of edges between nodes p and q (clearly, weights of all edges of an unweighted graph are equal to 1). Similarly, every square matrix M can be assigned an *associated graph* G_M. For a graph G on n nodes and its associated matrix $M = M_G$, the *Markov matrix* associated with G, denoted by $\mathcal{T}_G$, is obtained by normalizing each column of M so that it sums to 1, i.e., if D is a diagonal matrix with $D_{kk} = \sum_i M_{ik}$ and $D_{ij} = 0$ for $i \neq j$, then $\mathcal{T}_G$ is defined as $\mathcal{T}_G = M_G D^{-1}$. A column j of $\mathcal{T}_G$ corresponds with node j of the stochastic graph associated with $\mathcal{T}_G$, and the matrix entry $(\mathcal{T}_G)_{ij}$ corresponds to the probability of going from node j to node i.

Given such a matrix $\mathcal{T}_G \in \mathbb{R}^{n \times n}, \mathcal{T}_G \geq 0$, and a real number $r > 1$, let $\Gamma_r : \mathbb{R}^{k \times k} \to \mathbb{R}^{k \times k}$ be defined as:

$$(\Gamma_r \mathcal{T}_G)_{pq} = ((\mathcal{T}_G)_{pq})^r / \sum_{i=1}^{n} ((\mathcal{T}_G)_{iq})^r.$$

Γ_r is called the *inflation* operator with power coefficient r and the Markov matrix resulting from inflating each of the columns of $\mathcal{T}_G$ with power coefficient r is written as $\Gamma_r \mathcal{T}_G$. For $r > 1$, inflation changes the probabilities associated with the collection of random walks departing from a node (corresponding to a matrix column) by favoring more probable walks over less probable ones. Inflation can be altered by changing r: larger r makes inflation stronger and produces "tighter" clusters.

Expansion corresponds to computing "longer" random walks. It associates new probabilities with all pairs of nodes with one node being the point of departure and the other being the destination. It relies on the observation that longer paths are more common within clusters than between different clusters, and thus the probabilities associated with node pairs, which are within the same cluster will, in general, be relatively large, since there are many ways of going from one node to the other. Expansion is achieved via matrix multiplication. The MCL algorithm iterates the process of expanding information flow via matrix multiplication and contracting it via inflation. The basics of the MCL algorithm are presented in Algorithm 4.

Algorithm 4: MCL$(G, \Delta, e_{(i)}, r_{(i)})$

G is a graph with every node of degree ≥ 1;
$e_{(i)}$ is a sequence of $e_i \in \mathbb{N}, e_i > 1, i = 1, \ldots$;
$r_{(i)}$ is a sequence of $r_i \in \mathbb{R}, r_i > 0, i = 1, \ldots$;
$G = G + \Delta$; # Possibly add (weighted) self-loops;
$T_1 = \mathcal{T}_G$;
$k = 0$;
repeat
  $k = k + 1$;
  $T_{2k} = (T_{2k-1})^{e_k}$; # Expansion;
  $T_{2k+1} = \Gamma_{r_k}(T_{2k})$; # Inflation;
until T_{2k+1} *is (near-) idempotent*;
Interpret T_{2k+1} as a clustering.

Iterating expansion and inflation results in the matrix that is idempotent under both matrix multiplication and the inflation (such a matrix is called *doubly idempotent*), that is, an equilibrium state is reached when a matrix does not change with further expansion and inflation. The graph associated with such a matrix consists of different directed connected star-like components with an attractor in the centre (see bottom right picture in Figure 15). Each of these components is interpreted as a cluster.

Theoretically, there may be nodes connected to different stars, which is interpreted as cluster overlap [186]. The algorithm iterants converge nearly always to the doubly idempotent matrix. In practice they start noticeably converging after 3 to 10 iterations. Van Dongen proved quadratic convergence of the MCL process in the neighborhood of doubly idempotent matrices [186]. The row of expansion powers, $e_{(i)}$, and the row of inflation powers, $r_{(i)}$, in Algorithm 4 influence the granularity of the resulting clustering.

MCL was successfully applied to cluster protein sequences into families [61, 63]. For this purpose, nodes of the graph represent proteins, edges represent sequence similarities between the corresponding proteins, and edge weights corresponded to sequence similarity scores obtained from an algorithm such as BLAST [9, 10]. The overview of this algorithm, called Tribe-MCL, is presented in Algorithm 5. Tribe-MCL allowed hundreds of thousands of sequences to be accurately classified in a matter of minutes [61].

Algorithm 5: TRIBE-MCL(SET OF PROTEIN SEQUENCES S)

All versus all BLAST(S);
Parse results and symmetrify similarity scores;
Produce similarity matrix M;
Transform M to normalize similarity scores ($-\log$(E-value)) and generate transition probabilities;
MCL(G_M);
Post process and correct domains of resulting protein clusters (families).

The Restricted Neighborhood Search Clustering Algorithm (RNSC) has recently been used to extract identify protein complexes in PPI networks [108]. This algorithm partitions the node set of a network $G(V, E)$ by searching the space of partitions of V, each of which has an assigned cost, for a clustering with low cost. The initial clustering is random, or user-input. The RNSC searches for a low-cost clustering starting from an initial random clustering, by iteratively moving one node from one cluster to another in a randomized fashion to improve the cost of the clustering. A general move is one that reduced the clustering cost by a near-optimal amount. To avoid the tendency to settle in poor local minima, diversification moves shuffle the clustering by occasionally dispersing the contents of a cluster at random. Also, RNSC maintains a list of tabu (forbidden) moves to prevent cycling back to the previously explored partitioning.

Since the RNSC is randomized, different runs on the same input data will result in different clusterings. Thus, to achieve high accuracy in predicting true protein complexes in PPI networks, the RNSC output is filtered for functionally homogeneous, dense, and large clusters. This resulted in an accurate and scalable method for detecting and predicting protein complexes within a PPI network [108].

6 Conclusions

There is growing evidence that the analysis of PPI networks is useful, but multidisciplinary approaches need to be taken. Existing algorithms provide encouraging results, but novel methods have to be designed in order to improve both accuracy and scalability of these algorithms, and biological relevance of the models and hypothesis they generate. We emphasize here those open problems that we consider the most interesting and most pressing.

Understanding interactions between proteins in a cell may benefit from a better model of a PPI network. A full description of protein interaction networks requires a model that would encompass the undirected physical PPIs, other types of interactions, interaction confidence level, source (or method) and multiplicity of an interaction, directional pathway information, temporal information on the presence or absence of a PPI, information on the strength of the interactions, and possibly protein complex information. This may be achieved by designing a weighting function and assigning weights to nodes and edges of a PPI network to incorporate temporal and other interaction specific information, adding directionality to the network subgraphs, and building a hypergraph [21] structure on the top of the network to incorporate information about complexes, or pathways in which proteins take part.

Another interesting research topic is finding an efficient and robust graph clustering algorithm that would reliably identify protein complexes, separate stable from transient complexes [95], or detect pathways despite the noise present in PPI networks. Identifying graph theoretic structural properties that are common to protein complexes or certain pathway types in PPI networks may be crucial to designing such an algorithm. Similarly, modeling signaling pathways and finding efficient algorithms for their identification in PPI networks is another interesting topic. These algorithms would likely

[21] a generalization of graph in which edges may be any subset of the nodes

have to be stochastic, massively parallel, and use local search techniques, due to the presence of noise and large network sizes.

The existence of a "core proteome" has been hypothesized. It has been proposed that approximately 40% of yeast proteins are conserved through eukaryotic evolution [47]. We are approaching the moment when enough information would be available to verify the existence of such a proteome and discover its structural properties within PPI networks. It is already possible to take the first steps towards this goal with the currently available data. We propose to construct putative PPI networks for a number of eukaryotic organisms with mapped genomes by combining protein sequence similarities between different organisms with the known PPI networks of model organisms. With the set of putative PPI networks constructed in this way, it may be possible to do PPI network structural comparisons over different organisms. Preferential attachment to high degree nodes in real world networks has been suggested, implying that the core proteome would consist of high-degree nodes in PPI networks. It is interesting to observe the discrepancy between the high degree of a supposed "core proteome" proteins (hubs) and the separation of hubs by low-degree nodes [121]. Exploring the structural properties of this discrepancy may give an insight not only in the processes of evolution, but also in the properties that a better PPI network model should have. Research in this direction may result in construction of a stochastic, or deterministic large network model (similar to the model in [162] and the model in [96]), which would provide a better framework for understanding PPI networks. Also, the evolutionary mechanisms that produce the recently observed geometric properties of PPI networks [156] need to be explored.

Other interesting topics for future research include distinguishing different graph theoretic properties of proteins belonging to different functional groups. Our results suggest that such differences exist [160]. One way to approach this problem would be to identify different network motifs [171] in the neighborhood of proteins belonging to the same functional group, and compare the enrichment of these motifs over the functionally different sets of proteins. Along the same lines, it may be interesting to compare graph structures of the "same-function modules" over putative (or real, when they become available) PPI networks of different species and possibly infer common and differing elements in the structures of these modules. This could lead to construction of new models, which could be used for identification of false positives and false negatives in PPI networks.

Integration of microarray data with PPI data may be beneficial for finding solutions to many of the above mentioned open problems, as we further

discuss in Chapter **??**. Also, combining PPI data and orthology information based on protein sequence comparison can be used to find conserved protein complexes [169].

Complex biological and artificial networks show graph-theoretic properties that reflect the function these networks carry [76, 135, 59, 176, 184, 200, 195]. A similar analysis could be applied to call graphs of large software [158, 159]. A comparison between PPI networks, software call graphs, and other biological and artificial networks may give further insight into the properties of large, evolving networks. Comparing these networks to intentionally designed and optimized networks, such as circuit designs, could provide further information about the true nature of biological networks – as being truly scale-free, or following other, perhaps geometric properties.

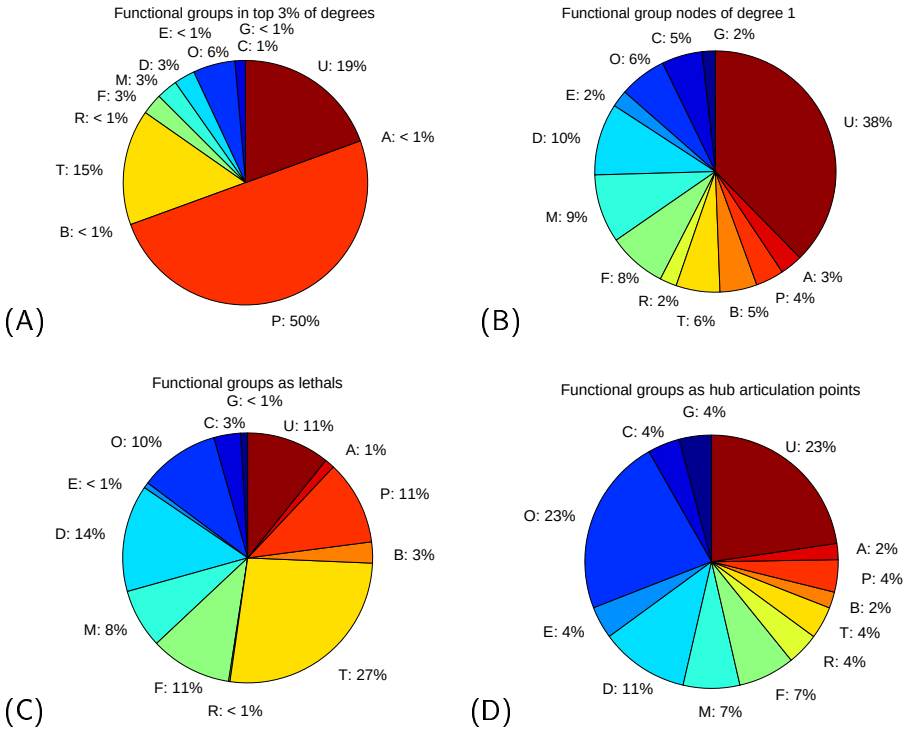

Figure 14: Pie charts for functional groups in the PPI graph: G – amino acid metabolism, C – cellular fate/organization, O – cellular organization, E – energy production, D – genome maintenance, M – other metabolism, F – protein fate, R – stress and defense, T – transcription, B – transcriptional control, P – translation, A – transport and sensing, U – uncharacterized. **A.** Division of the group of nodes with degrees in the top 3% of all node degrees. **B.** Division of nodes of degree 1. Compared with Figure 14 A, translation proteins are about 12 times less frequent, transcription about 2 times, while cellular fate/organization are 5 times more frequent, and genome maintenance, protein fate, and other metabolism are about 3 times more frequent; also, there are twice as many uncharacterized proteins. **C.** Division of lethal nodes. **D.** Division of articulation points which are hubs.

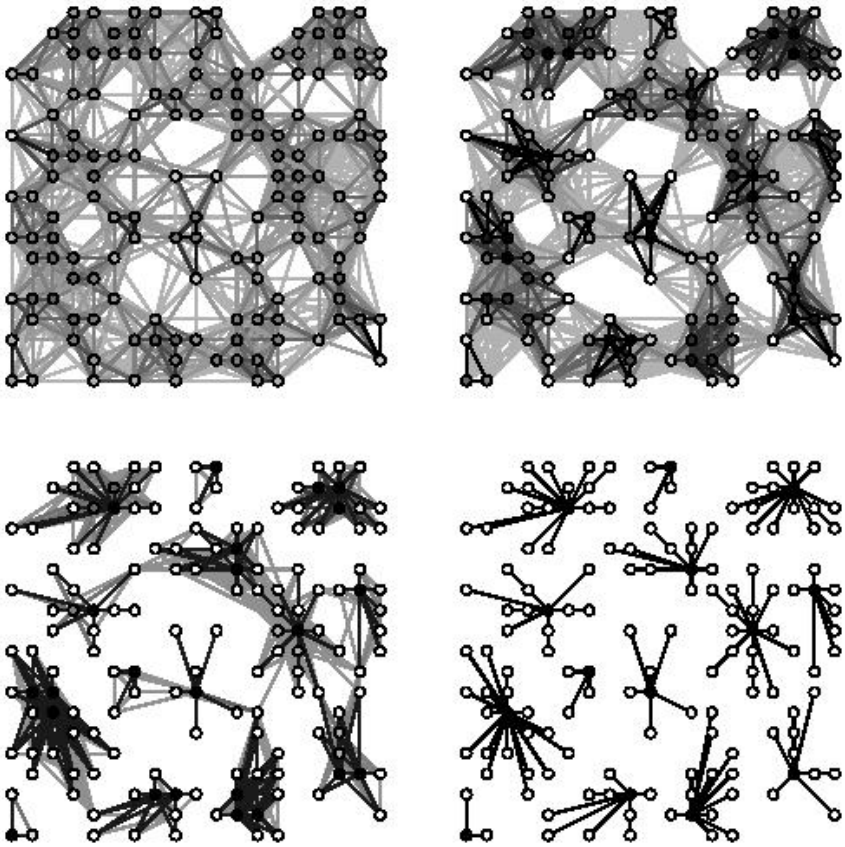

Figure 15: Stages of flow simulation by the MCL process. Taken from [186].

References

[1] J. Abello, A. Buchsbaum, and J. Westbrook. A functional approach to external graph algorithms. *Lecture Notes in Computer Science*, 1461:332–343, 1998.

[2] L. A. Adamic. The small world web. *Lecture Notes in Computer Science*, 1696:443–454, 1999.

[3] W. Aiello, F. Chung, and L. Lu. A random graph model for power law graphs. *Experimental Mathematics*, 10:53–66, 2001.

[4] R. Albert and A. L. Barabasi. Topology of evolving networks: local events and universality. *Phys Rev Lett*, 85(24):5234–7, 2000.

[5] R. Albert and A. L. Barabasi. Statistical mechanics of complex networks. *Reviews of Modern Physics*, 74:47–97, 2002.

[6] R. Albert, H. Jeong, and A. L. Barabasi. Diameter of the world-wide web. *Nature*, 401:387–392, 1999.

[7] R. Albert, H. Jeong, and A. L. Barabasi. Error and attack tolerance of complex networks. *Nature*, 406:378–382, 2000.

[8] C. J. Alpert and A. B. Kahng. Recent directions in netlist partitioning: a survey. *Integration: the VLSI Journal*, 19:1–81, 1995.

[9] S. F. Altschul, W. Gish, W. Miller, and D. J. Lipman. Basic local alignment search tool. *Journal of Molecular Biology*, 215:403–410, 1990.

[10] S. F. Altschul, T. L. Madden, A. A. Schffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, 25:3389–3402, 1997.

[11] L. A. N. Amaral, A. Scala, M. Barthelemy, and H. E. Stanley. Classes of behavior of small-world networks. *Proc Natl Acad Sci U S A*, 97:11149–11152, 2000.

[12] M. Ashburner. FlyBase. *Genome News*, 13:19–20, 1993.

[13] S. Asthana, O. D. King, F. D. Gibbons, and F. P. Roth. Predicting protein complex membership using probabilistic network reliability. *Genome Res*, 14(6):1170–5, 2004. 1088-9051 Journal Article.

[14] G. D. Bader, D. Betel, and C. W. V. Hogue. BIND: the biomolecular interaction network database. *Nucleic Acids Research*, 31(1):248–250, 2003.

[15] G. D. Bader and C. W. V. Hogue. Analyzing yeast protein-protein interaction data obtained from different sources. *Nature Biotechnology*, 20:991–997, 2002.

[16] G. D. Bader and C. W. V. Hogue. An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinformatics*, 4:2, 2003.

[17] J. S. Bader, A. Chaudhuri, J. M. Rothberg, and J. Chant. Gaining confidence in high-throughput protein interaction networks. *Nat Biotechnol*, 22(1):78–85, 2004. 1087-0156 Journal Article.

[18] A. Bairoch and R. Apweiler. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Research*, 28:45–48, 2000.

[19] F. Ball, J. Mollison, and G. Scalio-Tomba. Epidemics with two levels of mixing. *The Annals of Applied Probability*, 7:46–89, 1997.

[20] A. L. Barabasi and R. Albert. Emergence of scaling in random networks. *Science*, 286(5439):509–12, 1999.

[21] A. L. Barabasi, R. Albert, and H. Jeong. Mean-field theory for scale-free random networks. *Physica A*, 272:173–197, 1999.

[22] A. L. Barabasi, Z. Dezso, E. Ravasz, Z.-H. Yook, and Z. N. Oltvai. Scale-free and hierarchical structures in complex networks. *Sitges Proceedings on Complex Networks*, 2004. to appear.

[23] A. L. Barabasi, E. Ravasz, and T. Vicsek. Deterministic scale-free networks. *Physica A*, 299:559–564, 2001.

[24] A. D. Barbour and G. Reinert. Small worlds. *Random Structures and Algorithms*, 19:54–74, 2001.

[25] A. Barrat and M. Weigt. On the properties of small-world network models. *European Physical Journal B*, 13:547–560, 2000.

[26] M. Barthelemy and L. A. N. Amaral. Small-world networks: evidence for crossover picture. *Physical Review Letters*, 82:3180–3183, 1999.

[27] V. Batagelj and A. Mrvar. Pajek – program for large network analysis. *Connections*, 2:47–57, 1998.

[28] A. D. Baxevanis. The molecular biology database collection: 2003 update. *Nucleic Acids Research*, 31(1):1–12, 2003.

[29] A. Ben-Dor, R. Shamir, and Z. Yakhini. Clustering gene expression patterns. *Journal of Computational Biology*, 6(3/4):281–297, 1999.

[30] E. A. Bender and E. R. Canfield. The asymptotic number of labeled graphs with given degree sequences. *Journal of Combinatorial Theory A*, 24:296–307, 1978.

[31] D. A. Benson, I. Karsch-Mizrachi, D. J. Lipman, J. Ostell, B. A. Rapp, and D. L. Wheeler. GenBank. *Nucleic Acids Research*, 30:17–20, 2002.

[32] C. Bettstetter. On the minimum node degree and connectivity of a wireless multihop network. In *Proceedings of the 3rd ACM international symposium on mobile ad hoc networking and computing*, pages 80–01, 2002.

[33] M. Blatt, S. Wiseman, and E. Domany. Superparamagnetic clustering of data. *Physical Review Letters*, 76(18):3251–3254, 1996.

[34] B. Bollobas. *Random Graphs*. Academic, London, 1985.

[35] B. Bollobas and O. Riordan. The diameter of a scale-free random graph. *Combinatorica*, 2001. to appear.

[36] M. Boots and A. Sasaki. 'Small worlds' and the evolution of virulence: infection occurs locally and at a distance. *Proceedings of The Royal Society of London. Series B, Biological Sciences*, 266:1933–1938, 1999.

[37] P. Bork, L. J. Jensen, C. von Mering, A. K. Ramani, I. Lee, and E. M. Marcotte. Protein interaction networks from yeast to human. *Curr Opin Struct Biol*, 14(3):292–9, 2004. 0959-440x Journal Article.

[38] S. Bornholdt and H. Ebel. World-wide web scaling exponent from simon's 1955 model. *Physical Review E*, 64:046401, 2001.

[39] B.-J. Breitkreutz, C. Stark, and M. Tyers. The GRID: The general repository for interaction datasets. *Genome Biology*, 4:R23:R23.1–R23.3, 2003.

[40] B.-J. Breitkreutz, C. Stark, and M. Tyers. Osprey: a network visualization system. *Genome Biology*, 4:R22:R22.1–R22.4, 2003.

[41] A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, and J. Wiener. Graph structure of the web. *Computer Networks*, 33:309–320, 2000.

[42] K. Brown and I. Jurisica. Online predicted human interaction database: (OPHID). 2004.

[43] D. Bu, Y. Zhao, L. Cai, H. Xue, X. Zhu, H. Lu, J. Zhang, S. Sun, L. Ling, N. Zhang, G. Li, and R. Chen. Topological structure analysis of the protein-protein interaction network in budding yeast. *Nucleic Acids Research*, 31(9):2443–2450, 2003.

[44] D. S. Callaway, M. E. J. Newman, S. H. Strogatz, and D. J. Watts. Network robustness and fragility: percolation on random graphs. *Physical Review Letters*, 85:5468–5471, 2000.

[45] C. Chekuri, A. Goldberg, D. Karger, M. Levine, and C. Stein. Experimental study of minimum cut algorithms. In *ACM-SIAM Symposium on Discrete Algorithms (SODA 97)*, pages 324–333, 1997.

[46] Y. Chen and D. Xu. Computational analyses of high-throughput protein-protein interaction data. *Curr Protein Pept Sci*, 4(3):159–81, 2003. 1389-2037 Journal Article Review Review, Academic.

[47] S. A. Chervitz. Comparison of the complete protein sets of worm and yeast: orthology and divergence. *Science*, 282:2022–2028, 1998.

[48] F. Chung and L. Lu. The diameter of sparse random graphs. *Advances in Applied Mathematics*, 26:257–279, 2001.

[49] B. N. Clark, J. C. Colbourn, and D. S. Johnson. Unit disk graphs. *Discrete Mathematics*, 86:165–177, 1991.

[50] R. Cohen, K. Erez, D. ben Avraham, and S. Havlin. Resilience of the internet to random breakdowns. *Physical Review Letters*, 85:4626–4628, 2000.

[51] The Gene Ontology Consortium. Gene ontology: tool for the unification of biology. *Nature Genetics*, 25:25–29, 2000.

[52] M. C. Costanzo, M. E. Crawford, J. E. Hirschman, J. E. Kranz, P. Olsen, L. S. Robertson, M. S. Skrzypek, B. R. Braun, K. L. Hopkins, P. Kondu, C. Lengieza, J. E. Lew-Smith, M. Tillberg, and J. I. Garrels. YPD, PombelPD and WorkPD: model organism volumes of the BioKnowledge library, and integrated resource for protein information. *Nucleic Acids Research*, 29:75–79, 2001.

[53] S. Coulomb, M. Bauer, D. Bernard, and M.-C. Marsolier-Kergoat. Mutational robustness is only wakly related to the topology of protein interaction networks, 2004.

[54] T. Dandekar, B. Snel, M. Huynen, and P. Bork. Conservation of gene order: a fingerprint of proteins that physically interact. *Trends Biochem. Sci.*, 23:324–328, 1998.

[55] S. N. Dorogovtsev, A. V. Goltsev, and J. F. F. Mendes. Pseudofractal scale-free web. *Physical Review E*, 65:066122, 2002.

[56] S. N. Dorogovtsev and J. F. F. Mendes. Evolution of networks with aging of sites. *Physical Review E*, 62:1842–1845, 2000.

[57] R. O. Duda and P. E. Hart. *Pattern Classification and Scene Analysis.* Wiley, New York, 1973.

[58] S. S. Dwight, M. A. Harris, K. Dolinski, C. A. Ball, G. Binkley, K. R. Christie, D. G. Fisk, L. Issel-Tarver, M. Schroeder, G. Sherlok, A. Sethuraman, S. Weng, D. Botstein, and J. M. Cherry. Saccharomyces genome database (SGD) provides secondary gene annotation using the gene ontology (GO). *Nucleic Acids Research*, 30:69–72, 2002.

[59] J. P. Eckmann and E. Moses. Curvature of co-links uncovers hidden thematic layers in the world wide web. *Proc Natl Acad Sci U S A*, 99(9):5825–9, 2002.

[60] A. M. Edwards, B. Kus, R. Jansen, D. Greenbaum, J. Greenblatt, and M. Gerstein. Bridging structural biology and genomics: assessing protein interaction data with known complexes. *Trends Genet*, 18:529–36, 2002.

[61] A. J. Enright. *Computational Analysis of Protein Function within Complete Genomes.* PhD thesis, University of Cambridge, United Kingdom, 2002.

[62] A. J. Enright, I. Iliopoulos, N. C. Kyrpides, and C. A. Ouzounis. Protein interaction maps for complete genomes based on gene fusion events. *Nature*, 402:86–90, 1999.

[63] A. J. Enright, S. Van Dongen, and C. A. Ouzounis. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Research*, 30(7):1575–1584, 2002.

[64] P. Erdos and A. Renyi. On random graphs. *Publicationes Mathematicae*, 6:290–297, 1959.

[65] P. Erdos and A. Renyi. On the evolution of random graphs. *Publ. Math. Inst. Hung. Acad. Sci.*, 5:17–61, 1960.

[66] P. Erdos and A. Renyi. On the strength of connectedness of a random graph. *Acta Mathematica Scientia Hungary*, 12:261–267, 1961.

[67] S. Even. *Graph Algorithms*. Computer Science Press, Rockville, Maryland., 1979.

[68] M. Faloutsos, P. Faloutsos, and C. Faloutsos. On power-law relationships of the internet topology. *Computer Communications Review*, 29:251–262, 1999.

[69] D. Fell and A. Wagner. The small world of metabolism. *Nature Biotechnology*, 19:1121–1122, 2000.

[70] H. B. Fraser and A. E. Hirsh. Evolutionary rate depends on number of protein-protein interactions independently of gene expression level. *BMC Evol Biol*, 4(1):13, 2004. 1471-2148 Journal Article.

[71] C. Friedman, P. Kra, H. Yu, M. Krauthammer, and A. Rzhetsky. Genies: a natural-language processing system for the extraction of molecular pathways from journal articles. *Bioinformatics*, 17 Suppl. 1:74–82, 2001.

[72] A. C. Gavin, M. Bosche, R. Krause, P. Grandi, M. Marzioch, A. Bauer, J. Schultz, J. M. Rick, A. M. Michon, C. M. Cruciat, M. Remor, C. Hofert, M. Schelder, M. Brajenovic, H. Ruffner, A. Merino, K. Klein, M. Hudak, D. Dickson, T. Rudi, V. Gnau, A. Bauch, S. Bastuck, B. Huhse, C. Leutwein, M. A. Heurtier, R. R. Copley, A. Edelmann, E. Querfurth, V. Rybin, G. Drewes, M. Raida, T. Bouwmeester, P. Bork, B. Seraphin, B. Kuster, G. Neubauer, and G. Superti-Furga.

Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature*, 415(6868):141–7, 2002.

[73] H. Ge, Z. Liu, G. M. Church, and M. Vidal. Correlation between transcriptome and interactome mapping data from saccharomyces cerevisiae. *Nat Genet*, 29(4):482–6, 2001.

[74] D. Gibson, J. Kleinberg, and P. Raghavan. Inferring web communities from link toplogy. *Proceedings of the 9th ACM Conference on Hypertext and Hypermedia*, 1998. ACM Press, New York, NY.

[75] E. N. Gilbert. Random plane networks. *SIAM J.*, 9:533–543, 1961.

[76] M. Girvan and M. E. Newman. Community structure in social and biological networks. *Proc Natl Acad Sci U S A*, 99(12):7821–6, 2002.

[77] A. Grigoriev. A relationship between gene expression and protein interactions on the proteome scale: analysis of the bacteriophage t7 and the yeast saccharomyces cerevisiae. *Nucleic Acids Res*, 29(17):3513–9, 2001.

[78] N. Guelzim, S. Bottani, P. Bourgine, and F. Kepes. Topological and causal structure of the yeast transcriptional regulatory network. *Nat Genet*, 31(1):60–3, 2002.

[79] P. Gupta and P. Kumar. Critical power for asymptotic connectivity in wireless networks, 1998.

[80] R. Hafner. The asymptotic distribution of random clumps. *Computing*, 10:335–351, 1972.

[81] K. Han, B. Park, H. Kim, J. Hong, and J. Park. HPID: The human protein interaction database. *Bioinformatics*, 2004. Advance access published April 29, 2004.

[82] J. Hao and J. Orlin. A faster algorithm for finding the minimum cut in a directed graph. *Journal of Algorithms*, 17(3):424–446, 1994.

[83] E. Hartuv and R. Shamir. An algorithm for clustering cdna fingerprints. *Genomics*, 66(3):249–256, 2000. A preliminary version appeared in Proc. RECOMB '99, pp. 188-197.

[84] E. Hartuv and R. Shamir. A clustering algorithm based on graph connectivity. *Information Processing Letters*, 76(4-6):175–181, 2000.

[85] H. Hermjakob, L. Montecchi-Palazzi, G. Bader, J. Wojcik, L. Salwinski, A. Ceol, S. Moore, S. Orchard, U. Sarkans, C. von Mering, B. Roechert, S. Poux, E. Jung, H. Mersch, P. Kersey, M. Lappe, Y. Li, R. Zeng, D. Rana, M. Nikolski, H. Husi, C. Brun, K. Shanker, S. G. Grant, C. Sander, P. Bork, W. Zhu, A. Pandey, A. Brazma, B. Jacq, M. Vidal, D. Sherman, P. Legrain, G. Cesareni, I. Xenarios, D. Eisenberg, B. Steipe, C. Hogue, and R. Apweiler. The hupo psi's molecular interaction format–a community standard for the representation of protein interaction data. *Nat Biotechnol*, 22(2):177–83, 2004. 1087-0156 Journal Article.

[86] Y. Ho, A. Gruhler, A. Heilbut, G. D. Bader, L. Moore, S. L. Adams, A. Millar, P. Taylor, K. Bennett, K. Boutilier, L. Yang, C. Wolting, I. Donaldson, S. Schandorff, J. Shewnarane, M. Vo, J. Taggart, M. Goudreault, B. Muskat, C. Alfarano, D. Dewar, Z. Lin, K. Michalickova, A. R. Willems, H. Sassi, P. A. Nielsen, K. J. Rasmussen, J. R. Andersen, L. E. Johansen, L. H. Hansen, H. Jespersen, A. Podtelejnikov, E. Nielsen, J. Crawford, V. Poulsen, B. D. Sorensen, J. Matthiesen, R. C. Hendrickson, F. Gleeson, T. Pawson, M. F. Moran, D. Durocher, M. Mann, C. W. Hogue, D. Figeys, and M. Tyers. Systematic identification of protein complexes in saccharomyces cerevisiae by mass spectrometry. *Nature*, 415(6868):180–3, 2002.

[87] T. Hubbard, D. Barker, E. Birney, G. Cameron, Y. Chen, L. Clark, T. Cox, J. Cuff, V. Curwen, T. Down, R. Durbin, E. Eyras, J. Gilbert, M. Hammond, L. Huminiecki, A. Kasprzyk, H. Lehvaslaiho, P. Lijnzaad, C. Melsopp, E. Mongin, R. Pettett, M. Pocock, S. Potter, A. Rust, E. Schmidt, S. Searle, G. Slater, J. Smith, W. Spooner, A. Stabenau, J. Stalker, E. Stupka, A. Ureta-Vidal, I. Vastrik, and M. Clamp. The ensembl genome database project. *Nucleic Acids Research*, 30:38–41, 2002.

[88] T. R. Hughes, M. J. Marton, A. R. Jones, C. J. Roberts, R. Stoughton, C. D. Armour, H. A. Bennett, E. Coffey, H. Dai, Y. D. He, M. J. Kidd, A. M. King, M. R. Meyer, D. Slade, P. Y. Lum, S. B. Stepaniants, D. D. Shoemaker, D. Gachotte, K. Chakraburtty, J. Simon, M. Bard, and S. H. Friend. Functional discovery via a compendium of expression profiles. *Cell*, 102(1):109–26, 2000.

[89] M. Huynen and P. Bork. Measuring genome evolution. *Proc Natl Acad Sci U S A*, 95:5849–5856, 1998.

[90] R. F. i Chancho and R. V. Sole. The small world of human language. *Proceedings of The Royal Society of London. Series B, Biological Sciences*, 268(1482):2261–2265, 2001.

[91] T. Ito, , T. Chiba, R. Ozawa, M. Yoshida, M. Hattori, and Y. Sakaki. A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc Natl Acad Sci U S A*, 98(8):4569–4574, 2001.

[92] T. Ito, K. Tashiro, S. Muta, R. Ozawa, T. Chiba, M. Nishizawa, K. Yamamoto, S. Kuhara, and Y. Sakaki. 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 U S A*, 97(3):1143–7, 2000.

[93] S. Itzkovitz and U. Alon. Subgraphs and network motifs in geometric networks.

[94] S. Itzkovitz, R. Milo, N. Kashtan, G. Ziv, and U. Alon. Subgraphs in random networks. *Physical Review E*, 68, 2003.

[95] R. Jansen, D. Greenbaum, and M. Gerstein. Relating whole-genome expression data with protein-protein interactions. *Genome Res*, 12(1):37–46, 2002.

[96] H. Jeong, A. L. Barabasi, B. Tombor, and Z. N. Oltvai. The global organization of cellular networks. *submitted*, 2003. http://www.nd.edu/ networks/cell/.

[97] H. Jeong, S. P. Mason, A. L. Barabasi, and Z. N. Oltvai. Lethality and centrality in protein networks. *Nature*, 411(6833):41–2, 2001.

[98] H. Jeong, Z. N. Oltvai, and A. L. Barabasi. Prediction of protein essentiality based on genomic data. *ComPlexUs*, 1:19–28, 2003.

[99] H. Jeong, B. Tombor, R. Albert, Z. N. Oltvai, and A. L. Barabasi. The large-scale organization of metabolic networks. *Nature*, 407(6804):651–4, 2000.

[100] S. Jones and J. M. Thornton. Protein-protein interactions: a review of protein dimer structures. *Prog. Biophys. Mol. Biol.*, 63:31–65, 1995.

[101] S. Jones and J. M. Thornton. Principles of protein-protein interactions. *Proc Natl Acad Sci U S A*, 93:13–20, 1996.

[102] N. Kashtan, S. Itzkovitz, R. Milo, and U. Alon. Efficient sampling algorithm for estimating subgraph concentrations and detecting network motifs. *Bioinformatics*, 20:1746–1758, 2004.

[103] N. Kashtan, S. Itzkovitz, R. Milo, and U. Alon. Topological generalizations of network motifs. *Phys Rev E*, 70:031909, 2004.

[104] S. Kauffman. *At Home in the Universe.* Oxford, New York, 1995.

[105] S. A. Kauffman. Metabolic stability and epigenesis in randomly constructed genetic nets. *Journal of Theoretical Biology*, 22:437–467, 1969.

[106] J. J. Keeling. The effects of local spatial structure on epidemiological invasions. *Proceedings of The Royal Society of London. Series B, Biological Sciences*, 266:859–867, 1999.

[107] J. O. Kephart and S. R. White. Directed-graph epidemiological models of computer viruses. *Proc. 1991 IEEE Comput. Soc. Symp. Res. Security Privacy*, pages 343–359, 1991.

[108] A. D. King, N. Przulj, and I. Jurisica. Protein complex prediction via cost-based clustering. *Bioinformatics*, 20(17):3013–3020, 2004.

[109] J. Kleinberg. Authoritative sources in a hyper-linked environment. *Proceedings of the 9th ACM Conference on Hypertext and Hypermedia*, 1998. ACM Press, New York, NY.

[110] J. M. Kleinberg. Navigation in a small world. *Nature*, 406:845, 2000.

[111] P. L Krapivsky and S. Redner. Organization of growing random networks. *Physical Review E*, 63:066123–1, 2001.

[112] P. L Krapivsky, S. Redner, and F. Leyvraz. Connectivity of growing random networks. *Physical Review Letters*, 85:4629–4632, 2000.

[113] M. Krauthammer, P. Kra, I. Iossifov, S. M. Gomez, G. Hripcsak, V. Hatzivassiloglou, C. Friedman, and A. Rzhetsky. Of truth and pathways: chasing bits of information through myriads of articles. *Bioinformatics*, 18 Suppl. 1:249–257, 2002.

[114] M. Lappe and L. Holm. Unraveling protein interaction networks with near-optimal efficiency. *Nat Biotechnol*, 22(1):98–103, 2004. 1087-0156 Journal Article.

[115] L. F. Largo-Fernandez, R. Huerta, F. Corbancho, and J. Siguenza. Fast response and temporal coherent oscillations in small-world networks. *Physical Review Letters*, 84:2758–2761, 2000.

[116] T. Luczak. Component behavior near the critical point of the random graph process. *Random Structures and Algorithms*, 1:287, 1990.

[117] H. Ma and A.-P. Zheng. Structure and evolution analysis of metabolic networks based on genomic data. In *4th Biopathways Consortium Meeting*, 2002. Edmonton, Canada, August 1-2.

[118] S. Mangan and U. Alon. Structure and function of the feed-forward loop network motif. *PNAS*, 100:11980–11985, 2003.

[119] S. Mangan, A. Zaslaver, and U. Alon. The coherent feedforward loop serves as a sign-sensitive delay element in transcription networks. *JMB*, 334/2:197–204, 2003.

[120] M. Mann, R. C. Hendrickson, and A. Pandey. Analysis of proteins and proteomes by mass spectrometry. *Annu. Rev. Biochem.*, 70:437–473, 2001.

[121] S. Maslov and K. Sneppen. Specificity and stability in topology of protein networks. *Science*, 296(5569):910–3, 2002.

[122] D. W. Matula. The cohesive strength of graphs. In G. Chartrand and S. F. Kapoor, editors, *The Many Facets of Graph Theory*, pages 215–221. Lecture Notes in Math., Vol. 110, Springer, Berlin, 1969.

[123] D. W. Matula. Cluster analysis via graph theoretic techniques. In R. C. Mullin, K. B. Reid, and D. P. Roselle, editors, *Proc. Louisiana Conference on Combinatorics, Graph Theory and Computing*, pages 199–212. University of Manitoba, Winnipeg, 1970.

[124] D. W. Matula. k-components, clusters and slicing in graphs. *SIAM J. Applied Math*, 22(3):459–480, 1972.

[125] D. W. Matula. Graph theoretic techniques for cluster analysis algorithms. In J. van Ryzin, editor, *Classification and Clustering*, pages 95–129. Academic Press, New York, 1977.

[126] D. W. Matula. Determining edge connectivity in O(nm) time. In *28th IEEE Symposium on Foundations of Computer Science*, pages 249–251, 1987.

[127] R. M. May. *Stability and Complexity in Model Ecosystems.* Princeton Univ. Press, Princeton, 1973.

[128] P. B. McGarvey, H. Huang, W. C. Barker, B. C. Orcutt, J. S. Garavelli, G. Y. Srinivasarao, L. S. Yeh, C. Xiao, and C. H. Wu. PIR: a new resource for bioinformatics. *Bioinformatics*, 16:290–291, 2000.

[129] D. McShan, S. Rao, and I. Shah. Microbial metabolic pathway inference by heuristic search. In *4th Biopathways Consortium Meeting*, 2002. Edmonton, Canada, August 1-2.

[130] K. Mehlhorn and S. Naher. *Leda: A platform for combinatorial and geometric computing.* Cambridge University Press, 1999.

[131] J. C. Mellor, I. Yanai, K. H. Clodfelter, J. Mintseris, and C. DeLisi. Predictome: a database of putative functional links between proteins. *Nucleic Acids Res*, 30(1):306–9, 2002. 21624842 1362-4962 Journal Article.

[132] H. W. Mewes, D. Frishman, U. Guldener, G. Mannhaupt, K. Mayer, M. Mokrejs, B. Morgenstern, M. Munsterkotter, S. Rudd, and B. Weil. MIPS: a database for genomes and protein sequences. *Nucleic Acids Res*, 30(1):31–4, 2002.

[133] S. Milgram. The small world problem. *Psychology Today*, 2:60–67, 1967.

[134] R. Milo, S. Itzkovitz, N. Kashtan, R. Levitt, S. Shen-Orr, I. Ayzenshtat, M. Sheffer, and U. Alon. Superfamilies of evolved and designed networks. *Science*, 303:1538–1542, 2004.

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

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

[137] M. Molloy and B. Reed. The size of the largest component of a random graph on a fixed degree sequence. *Combinatorics, Probability and Computing*, 7:295–306, 1998.

[138] J. M. Montoya and R. V. Sole. Small world patterns of food webs. *Working paper 00-10-059, Santa Fe Institute*, 2001.

[139] H. Nagamochi and T. Ibaraki. Computing edge connectivity in multigraphs and capacitated graphs. *SIAM J. Discrete Math*, 5:54–66, 1992.

[140] M. E. Newman. Scientific collaboration networks: I. network construction and fundamental results. *Physical Review E*, 64:016131, 2001.

[141] M. E. Newman. Ego-centered networks and the ripple effect. *Social Networks*, 25:83–95, 2003.

[142] M. E. Newman and D. J. Watts. Renormalization group analysis in the small-world network model. *Physics Letters A*, 263:341–346, 1999.

[143] M. E. Newman and D. J. Watts. Scaling and percolation in the small-world network model. *Physical Review E*, 60:7332–7342, 1999.

[144] M. E. J. Newman. Models of the small world: a review. *Journal of Statistical Physics*, 101:819–841, 2000.

[145] M. E. J. Newman. The structure and function of networks. *Computer Physics Communications*, 147:44–45, 2001.

[146] M. E. J. Newman. Random graphs as models of networks. In S. Bornholdt and H. G. Schuster, editors, *Handbook of Graphs and Networks*. Wiley-VHC, Berlin, 2002.

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

[148] M. E. J. Newman, C. Mooire, and D. J. Watts. Mean-field solution of the small-world network model. *Physical Review Letters*, 84:3201–3204, 2000.

[149] M. E. J. Newman, S. H. Strogatz, and D. J. Watts. Random graphs with arbitrary degree distributions and their applications. *Physical Review E*, 64:026118–1, 2001.

[150] S. Ng and M. Wong. Toward routine automatic pathway discovery from on-line scientific text abstracts. *Genome Informatics*, 10:104–112, 1999.

[151] R. Overbeek, N. Larsen, G. D. Pusch, M. D'Souza, E. Selkov Jr, N. Kyrpides, M Fonstein, N. Maltsev, and E. Selkov. WIT: integrated system for high-throughput genome sequence analysis and metabolic reconstruction. *Nucleic Acids Research*, 28(1):123–125, 2000.

[152] A. Pandey and M. Mann. Proteomics to study genes and genomes. *Nature*, 405:837–846, 2000.

[153] R. Pastor-Satorras and A. Vespignani. Epidemic spreading in scale-free networks. *Physical Review Letters*, 86:3200–3203, 2001.

[154] M. Penrose. *Geometric Random Graphs.* Oxford Univeristy Press, 2003.

[155] S. Peri, J. D. Navarro, T. Z. Kristiansen, R. Amanchy, V. Surendranath, B. Muthusamy, T. K. Gandhi, K. N. Chandrika, N. Deshpande, S. Suresh, B. P. Rashmi, K. Shanker, N. Padma, V. Niranjan, H. C. Harsha, N. Talreja, B. M. Vrushabendra, M. A. Ramya, A. J. Yatish, M. Joy, H. N. Shivashankar, M. P. Kavitha, M. Menezes, D. R. Choudhury, N. Ghosh, R. Saravana, S. Chandran, S. Mohan, C. K. Jonnalagadda, C. K. Prasad, C. Kumar-Sinha, K. S. Deshpande, and A. Pandey. Human protein reference database as a discovery resource for proteomics. *Nucleic Acids Res*, 32 Database issue:D497–501, 2004. 1362-4962 Journal Article.

[156] N. Przulj, D. G. Corneil, and I. Jurisica. Modeling interactome: Scale-free or geometric? *Bioinformatics*, 20(18):3508–3515, 2004.

[157] N. Przulj, D. G. Corneil, and I. Jurisica. Modeling interactome: Scale-free or geometric? *Bioinformatics*, 20(18):3508–3515, 2004.

[158] N. Przulj and I. Jurisica. A call graph analysis. *CASCON*, 2003. IBM Toronto Lab, Marknam, Ontario, Canada, October 6-9.

[159] N. Przulj, G. Lee, and I. Jurisica. Functional analysis of large software networks. *IBM Academy: Proactive Problem Prediction, Avoidance and Diagnosis*, 2003. IBM T. J. Watson Research Center, NY.

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

[161] J.-D. Rain, L. Selig, H. De Reuse, V. Battaglia, C. Reverdy, S. Simon, G. Lenzen, F. Petel, J. Wojcik, V. Schachter, Y. Chemama, A. Labigne, and P. Legrain. The protein-protein interaction map of helicobacter pylori. *Nature*, 409:211–215, 2001.

[162] E. Ravasz, A. L. Somera, D. A. Mongru, Z. N. Oltvai, and A. L. Barabasi. Hierarchical organization of modularity in metabolic networks. *Science*, 297:1551–5, 2002.

[163] G. Rigaut, A. Shevchenko, B. Rutz, M. Wilm, M. Mann, and B. Seraphin. A generic protein purification method for protein complex characterization and proteome exploration. *Nature Biotechnol.*, 17:1030–1032, 1999.

[164] C. J. Roberts, B. Nelson, M. J. Marton, R. Stoughton, M. R. Meyer, H. A. Bennett, Y. D. He, H. Dai, W. L. Walker, T. R. Hughes, M. Tyers, C. Boone, and S. H. Friend. Signaling and circuitry of multiple mapk pathways revealed by a matrix of global gene expression profiles. *Science*, 287(5454):873–80, 2000.

[165] G. M. Rubin, M. D. Yandell, J. R. Wortman, G. L. G. Miklos, C. R. Nelson, I. K. Hariharan, M. E. Fortini, P. W. Li, R. Apweiler, W. Fleischmann, J. M. Cherry, S. Henikoff, M. P. Skupski, S. Misra, M. Ashburner, E. Birney, M. S. Boguski, T. Brody, P. Brokstein, S. E. Celniker, S. A. Chervitz, D. Coates, A. Cravchik, A. Gabrielian, R. F. Galle, W. M. Gelbart, R. A. George, L. S. B. Goldstein, F. Gong, P. Guan, N. L. Harris, B. A. Hay, R. A. Hoskins, J. Li, Z. Li, R. O. Hynes, S. J. M. Jones, P. M. Kuehl, B. Lemaitre, J. T. Littleton, D. K. Morrison, C. Mungall, P. H. O'Farrell, O. K. Pickeral, C. Shue, L. B. Vosshall, J. Zhang, Q. Zhao, X. H. Zheng, F. Zhong, W. Zhong, R. Gibbs, J. C. Venter, M. D. Adams, and S. Lewis. Comparative genomics of the eukaryotes. *Science*, 287:2204–2215, 2000.

[166] C. H. Schilling, D. Letscher, and B. O. Palsson. Theory for the systematic definition of metabolic pathways and their use in interpreting metabolic function from a pathway-oriented perspective. *Journal of Theoretical Biology*, 203:229–248, 2000.

[167] B. Schwikowski, P. Uetz, and A. Fields. A network of protein-protein interactions in yeast. *Nature Biotechnology*, 18:1257–1261, 2000.

[168] R. Shamir and R. Sharan. Algorithmic approaches to clustering gene expression data. In T. Jiang, T. Smith, Y. Xu, and M. Zhang, editors, *Current Topics in Computational Biology*. MIT Press, 2001.

[169] R. Sharan, T. Ideker, B. P. Kelley, R. Shamir, and R. M. Karp. Identification of protein complexes by comparative analysis of yeast and bacterial protein interaction data. In *Proceedings of the eighth annual international conference on Computational molecular biology (RECOMB'04)*, 2004.

[170] R. Sharan and R. Shamir. CLICK: A clustering algorithm with applications to gene expression analysis. In *International Conference on Intelligent Systems for Molecular Biology (ISMB)*, pages 307–316, 2000.

[171] S. S. Shen-Orr, R. Milo, S. Mangan, and U. Alon. Network motifs in the transcriptional regulation network of escherichia coli. *Nature Genetics*, 31:64–68, 2002.

[172] H. A. Simon. On a class of skew distribution functions. *Biometrika*, 42:425–440, 1955.

[173] V. Spirin, D. Zhao, and L. A. Mirny. Discovery of protein complexes in the network of protein interactions. In *3rd International Conference on Systems Biology (ICSB)*, 2002. Karolinska Institutet, Stockholm, Sweden, Dec. 13 - 15.

[174] O. Sporns, G. Tononi, and G. M. Edelman. Theoretical neuroanatomy: relating anatomical and functional connectivity in graphs and cortical connection matrices. *Cereb. Cortex*, 10:127–141, 2000.

[175] M. Steffen, A. Petti, J. Aach, P. D'haeseleer, and G. Church. Automated modeling of signal transduction networks. *BMC Bioinformatics*, 2002.

[176] J. Stelling, S. Klamt, K. Bettenbrock, S. Schuster, and E. D. Gilles. Metabolic network structure determines key aspects of functionality and regulation. *Nature*, 420:190–3, 2002.

[177] K. E Stephan. Computational analysis of functional connectivity between areas of primate visual cortex. *Phil. Trans. R. Soc. Lond. B*, 355:111–126, 2000.

[178] M. Stoer and F. Wagner. A simple min-cut algorithm. *Journal of the ACM*, 44(4):585–591, 1997.

[179] G. Stoesser, W. Baker, A. van den Broek, E. Camon, M. Garia-Pastor, C. Kanz, T. Kulikova, R. Leinonen, Q. Lin, V. Lombard, R. Leopez, N. Redaschi, P. Stoehr, M. A. Tuli, K. Tzouvara, and R. Vaughan. The EMBL nucleotide sequence database. *Nucleic Acids Research*, 30:21–26, 2002.

[180] S. H. Strogatz. Exploring complex networks. *Nature*, 410:268–276, 2001.

[181] J. M. Stuart, E. Segal, D. Koller, and S. K. Kim. A gene-coexpression network for global discovery of conserved genetic modules. *Science*, 302(5643):249–55, 2003. 1095-9203 Journal Article.

[182] Y. Tateno, T. Imanishi, S. Miyazaki, K. Fukami-Kobayashi, N. Saitou, H. Sugawara, and T. Gojobori. DAN data bank of japan (DDBJ). *Nucleic Acids Research*, 30:27–30, 2002.

[183] A. H. Tong, B. Drees, G. Nardelli, G. D. Bader, B. Brannetti, L. Castagnoli, M. Evangelista, S. Ferracuti, B. Nelson, S. Paoluzi, M. Quondam, A. Zucconi, C. W. Hogue, S. Fields, C. Boone, and G. Cesareni. A combined experimental and computational strategy to define protein interaction networks for peptide recognition modules. *Science*, 295(5553):321–4, 2002.

[184] Y. Tu. How robust is the internet? *Nature*, 406:353–4, 2000.

[185] Peter Uetz, Loic Giot, Gerard Cagney, Traci A. Mansfield, Richard S. Judson, James R. Knight, Daniel Lockshon, Vaibhav Narayan, Maithreyan Srinivasan, Pascale Pochart, Alia Qureshi-Emili, Ying Li, Brian Godwin, Diana Conover, Theodore Kalbfleish, Govindan Vijayadamodar, Meijia Yang, Mark Johnston, Stanley Fields, and Jonathan M. Rothberg. A comprehensive analysis of protein-protein interactions in saccharomyces cerevisiae. *Nature*, 403:623–627, 2000.

[186] S. M. van Dongen. *Graph Clustering by Flow Simulation.* PhD thesis, University of Utrecht, The Netherlands, 2000.

[187] C. von Mering, R. Krause, B. Snel, M. Cornell, S. G. Oliver, S. Fields, and P. Bork. Comparative assessment of large-scale data sets of protein-protein interactions. *Nature*, 417(6887):399–403, 2002.

[188] A. Wagner and D. Fell. The small world inside large metabolic networks. *Proc. Roy. Soc. London Series B*, 268:1803–1810, 2001.

[189] J. Wallinga, K. J. Edmunds, and M. Kretzschmar. Perspective: human contact patterns and the spread of airborne infectious diseases. *Trends in Microbiology*, 7:372–377, 1999.

[190] T. Walsh. Search in small world. *Proc. 16th Int. Joint Conf. Artif. Intell.*, pages 1172–1177, 1999.

[191] D. J. Watts. *Small Worlds.* Princeton University Press, Princeton, 1999.

[192] D. J. Watts and S. H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393:440–442, 1998.

[193] D. B. West. *Introduction to Graph Theory.* Prentice Hall, Upper Saddle River, NJ., 1996.

[194] H. S. Wilf. *Generating Functionology.* Academic, Boston, 1990.

[195] R. J. Williams, E. L. Berlow, J. A. Dunne, A. L. Barabasi, and N. D. Martinez. Two degrees of separation in complex food webs. *Proc Natl Acad Sci U S A*, 99:12913–6, 2002.

[196] S. Wuchty. Evolution and topology in the yeast protein interaction network. *Genome Res*, 14(7):1310–4, 2004. 1088-9051 Journal Article.

[197] I. Xenarios, L. Salwinski, X. J. Duan, P. Higney, S. M. Kim, and D. Eisenberg. Dip, the database of interacting proteins: a research tool for studying cellular networks of protein interactions. *Nucleic Acids Res*, 30(1):303–5, 2002. 1362-4962 Journal Article.

[198] E. Yeger-Lotem and H. Margalit. Detection of regulatory circuits by integration of protein-protein and protein-dna interaction data. In *4th Biopathways Consortium Meeting*, 2002. Edmonton, Canada, August 1-2, 2002.

[199] E Yeger-Lotem, S Sattath, N Kashtan, S Itzkovitz, R Milo, RY Pinter, U Alon, and H Margalit. Network motifs in integrated cellular networks of transcription-regulation and protein-protein interaction. *Proc Natl Acad Sci U S A*, 101(16):5934–5939, April 2004.

[200] S.-H. Yook, H. Jeong, and A. L. Barabasi. Modeling the internet's large-scale topology. *Proc Natl Acad Sci U S A*, 99:13382–6, 2002.

[201] A. Zanzoni, L. Montecchi-Palazzi, M. Quondam, G. Ausiello, Helmer-Citterich M., and G. Cesareni. Mint: A molecular interaction database. *FEBS Letters*, 513(1):135–140, 2002.