Review of uses of network and graph theory concepts within proteomics

Peter Grindrod† and Milla Kibble

The size and the nature of data collected on gene and protein interactions has led to a rapid growth of interest in graph theory and modern techniques for describing, characterizing and comparing networks. Simultaneously, this is a field of growth within mathematics and theoretical physics, where the global properties, and emergent behavior of networks, as a function of the local properties has long been studied. In this review, a number of approaches for exploiting modern network theory to help describe and analyze different data sets and problems associated with proteomic data are considered. This review aims to help biologists find their way towards useful ideas and references, yet may also help scientists from a mathematics and physics background to understand where they may apply their expertise.

It is hoped that by classifying techniques and problems, and making the objectives transparent, the literature will be more accessible to those outside the field. For those, like the authors, focussed on a particular subset of problems, it is useful to gain some perspective from a wider look at current research and understand different areas of progress within the larger context.

Categorizing research fields

Research into large-scale complex networks appears to fall into the following categories, with papers frequently spanning research from several categories.

Proposing models for large-scale complex networks

Traditionally, complex networks have been described with the random graph theory of Erdos and Renyi, for which each pair of nodes...
is connected at random with probability \( p \). These networks are statistically homogeneous with a Poisson connectivity distribution \( P(k) \) that peaks at the average number of links \( \langle k \rangle \). Furthermore, random graphs have relatively short distances between all vertices, whilst triangles are relatively uncommon. A detailed review of random graphs can be found in [2].

Recently, the shortfalls of using random graphs to model large-scale complex networks have become apparent, with the observation that some such networks are extremely heterogeneous with connectivity distributions that decay as a power-law, \( P(k) \sim k^{-\gamma} \), where \( \gamma \) is some constant characteristic of the network. In such networks, termed scale-free networks, nodes have widely different connectivities and the overall topology is dominated by a few highly connected nodes.

Some large-scale complex networks have also been found to have short characteristic path lengths, like random graphs, but a high degree of clustering.

Examples of biological networks found to possess these properties are later discussed. However, first efforts to address the need for more realistic models are cited. An excellent introduction to the following approaches can be found in [3].

Scale-free random graphs

One approach has been to generalize random graphs by constructing models that have a degree distribution which follows a power law, but in all other respects are random (i.e., the edges connect randomly selected nodes). For such models, known as scale-free random graphs, questions similar to those asked by Erdos and Renyi have been studied, for example: is there a threshold at which a giant cluster appears; when does the graph become connected? The characteristic path length and clustering coefficient have also been considered. An estimate of the former can be derived using a general approach to random graphs with a given degree distribution. This approach was developed by Newman, Strogatz and Watts using a generating function formalism [4]. When comparing the characteristic path length of real scale-free networks, \( L_{\text{real}} \), with those predicted under a power law distribution with exponential cut-off, \( L_{\text{pow}} \), Albert and Barabasi observe a general trend that \( L_{\text{real}} \) is larger than both \( L_{\text{pow}} \) and estimates for \( L \) derived from random graph theory. The clustering coefficient, \( C \) (see appendix), does not appear to have been calculated for a scale-free random graph, though it is known to converge to zero as the network size increases. However, a nonvanishing \( C \) can be calculated, again using the generating function method, for networks that are induced by projections of bipartite graphs [4].

Small world networks

The first successful attempt to generate graphs with a short characteristic path length but high clustering coefficient was made by Watts and Strogatz [5]. They introduce networks derived from regular graphs that share this local property of regular graphs and global property of random graphs. Using a ring lattice with \( n \) vertices in which every node is connected to its first \( K \) neighbors as an initial regular graph, they rewire each edge at random with probability \( p \) (to keep the network sparse yet connected, they restrict to \( n > K > \ln(n) > 1 \)). The clustering coefficient \( C(p) \), a local property, and characteristic path length \( L(p) \), a global property, are then observed for different values of rewiring probability, \( p \). The regular lattice at \( p = 0 \) is highly clustered and \( L \) grows linearly with \( n \), whereas the random network at \( p = 1 \) is poorly clustered and \( L \) grows logarithmically with \( n \). However, there is a broad interval of \( p \) over which \( L(p) \) is close to \( L(1) \) (i.e., there is a relatively short distance between all vertices) but \( C(p) > C(1) \) (i.e., there is a high degree of clustering). Networks satisfying both of these conditions are termed small world networks.

Hence \( C(p) \) remains practically unchanged for small \( p \) even though \( L(p) \) drops rapidly. The authors highlight two points: at a local level, transition to a small world network from a regular graph is almost undetectable and that, since only a tiny fraction of shortcuts is necessary in regular graphs to transform to a small world network, small-world behavior may be common in sparse networks.

This article spurred a flood of research into the properties of small world networks, the Watts–Strogatz (WS) model and its variants (with different choices of initial regular graphs and rewiring algorithms). The question of whether the onset of small-world behavior is dependent on system size has been studied and attempts have been made to calculate exactly the distribution of path lengths and the characteristic path length. For the WS model, the dependence of the clustering coefficient on \( p \) can be derived using a slightly different but equivalent definition of \( C \) and the degree distribution has been found to be similar to that of a random graph with a pronounced peak at the average degree and exponential decay.

Barabasi–Albert network

Barabasi and Albert shifted away from modeling network topology to modeling network assembly and evolution. They introduce a theoretical model that, in the limit of infinite time, generates graphs demonstrating a connectivity distribution that decays as a power-law with fixed exponent, \( \gamma = 3 \) [6]. The model is formed from two generic mechanisms: networks are allowed to expand continuously by the addition of new vertices; and these newly added nodes attach preferentially to sites that are already well connected (arguably, both of these key prerequisites are met for protein interaction networks, [7]; however, the process of node loss is perhaps just as important to biological networks).

Comparing a Barabasi–Albert (BA) network with average degree \( \langle k \rangle = 4 \) with that of the appropriate random graph for different network sizes, \( L \) is found to be smaller in the BA network than in a random graph for any network size and \( C \) is found to be about five-times larger than that for a random graph, [3]. However, unlike for small world models, the clustering coefficient of the BA model decays with network size.

A variety of models have been proposed in which addition and deletion of edges can guarantee a power law distribution and these are reviewed in [3].
Wagner considers the processes that influence the structure of the yeast protein interaction network [7]. By estimating their rates from empirical data, Wagner concludes that gene duplications do not alter global network structure drastically since duplicated protein–protein interactions diverge so rapidly and thoroughly. However, interaction turnover without gene duplication is sufficiently rapid to influence network structure drastically since the rates of interaction gain and loss being approximately equal; a fraction of new edge additions also affects preferentially highly connected proteins.

Grindrod introduces a class of sparse graphs that can satisfy the conditions of small-world networks and which are characterized by two simple global parameters [9,10]. These range-dependent random graphs are based on a 1D enumeration of their vertices, $v_k$ for $k = ...,-1,0,1,...$. The range of an edge connecting vertices $v_i$ and $v_j$ is defined to be $|j-i|$. These graphs can be thought of as the superposition of many subgraphs, with each subgraph containing only edges of a certain range, $r = 1,2,...$, which are present with probability $\alpha \lambda^{r-1}$ (where $\alpha$ and $\lambda$ are in $[0,1]$). Notice that choosing $\alpha = 1$ forces neighbors to be connected and $\lambda$ controls the probability of long range edges.

For graphs with infinite vertices and edges, but finite average vertex degree, global characteristics such as expected vertex degree, $z$, expected number of second neighbors, $z_2$, and the WS clustering coefficient, $C$, can be written explicitly in terms of the global parameters, $\alpha$ and $\lambda$. The author shows that for fixed $z$, $C$ is relatively high for middle values of $\lambda$, whereas estimates of $L$ fall to the value for random graphs at much smaller values of $\lambda$, thus giving a range of $\lambda$ where small-world networks exist. Grindrod states that this model is a valid approximation to large finite graphs.

The advantage of simple global characteristics explicitly defined in terms of simple global parameters is that one can calibrate the global parameters based on matching global properties. This will in turn help in answering the inverse problem: given a list of vertices in arbitrary order and a list of edges, assumed to have originated from this class of graphs, can one produce an ordering of the vertices from which the data are likely to have been generated?

Another further recent reference to network modeling and characterization specific to protein interactions is [11].

### Identifying the large-scale structure of networks derived from data

Although we would like to be able to represent a given set of interactions as a graph from an appropriate class, research has mainly concentrated on characterizing macroscopic network properties. In particular, networks derived from data have been characterized by their small-world network properties, $C$ and $L$, and the connectivity distribution $P(k)$ of their nodes. Ensuing sections reveal how such global properties add insights, as well as confirming the need for a model to replace the random graph. Properties that were observed for biological networks are described here. For example, a variety of nonbiological scale-free and/or small-world networks are discussed in [3].

Jeong and coworkers demonstrate that the connectivity distribution of the *Saccharomyces cerevisiae* protein–protein interaction network derived from data in [8,13] follows a power law with exponential cut-off [12]. In this regard, nodes are proteins that are connected if it has been experimentally demonstrated that they bind together. They comment that this topology is shared by *Helicobacter pylori*.

Wagner [14] builds a protein–protein interaction network from data in [8]. Visual inspection demonstrates a feature typical of random graphs: the network has many subsets involving few proteins and one giant component with many proteins. However, the network's degree distribution appears to follow a power law distribution. Furthermore, when considering the giant component, the clustering coefficient is much larger than and the characteristic path length smaller than the corresponding values for an Erdos and Renyi random graph.

Ito introduces another genome-wide two-hybrid protein interaction network for budding yeast *S. cerevisiae* and records a single huge network: 417 proteins linked by 544 interactions and 132 smaller networks with two to 14 proteins [15].

Information gained from the techniques in [16] can be represented as a fully connected graph whose nodes correspond to proteins and whose edges have weights corresponding to the probability that two proteins are part of the same complex. This can be converted to an unweighted graph of high-confidence protein association by including only edges with weights above a chosen threshold. The authors find that the degree distribution for such unweighted graphs resulting from two high-throughput complex purification experiments for the yeast *S. cerevisiae* can be approximated by a power function with a tail that falls off faster than expected under a power law [17,18]. The clustering coefficients in the largest graph components were orders of magnitude greater and the characteristic path lengths were considerably smaller than those observed in [14]. These differences are not surprising, as complex level associations are being studied which include direct and indirect interactions, rather than just direct interactions.

Jeong and coworkers describe a metabolic network has substrates as nodes that are connected to one another if reactions occur between them. [19] They find that scale-free networks describe the metabolic networks in all 43 organisms in all three domains of life (the paper distinguishes between incoming and outgoing edges and works separately for each). These networks are based on data in the WIT database [20].

The genome-wide gene disruption network in [21] is a directed graph with nodes representing genes and edges connecting nodes if the disruption (deletion) of the source gene significantly alters the expression of the target gene. The edges are labeled as upregulating or downregulating depending on whether the expression level has increased or decreased above...
some threshold. The authors find that total degree is distributed roughly according to a power-law distribution. The network is built from gene expression profile data in [22].

Wuchty describes a protein domain network with a vertex set consisting of domains found in proteins (regardless of the protein’s origin) [23]. Two domains are adjacent if they occur together at least once in a single protein. Three databases, collecting protein domain information in completely different ways, were used. The networks are sparse and the frequency distributions (one for each database) of domain connections follow a distribution comparable to a power-law distribution. They each also partially satisfy the structural properties of small-world graphs. Most proteins just contain one domain and there are a large number of unconnected vertices. When major components are studied, they are found to exhibit small-world and scale-free behavior.

Protein domain networks were also built separately for six species and each was found to have a power-law distribution.

In a neural network, nodes are neurons and an edge joins two neurons if they are connected either by a synapse or a gap junction. Watts and Strogatz demonstrated that the neural network of the nematode worm Caenorhabditis elegans, the sole example of a completely mapped neural network (in 1998), is a small-world network [5].

Using properties of the network to determine functional properties of the nodes & gain an understanding of the molecular system

This section includes research aimed at showing how simple properties of a network can lead to useful insights. Thus we gain our first glimpse of the value of taking a network-based viewpoint in combination with knowledge of the network’s individual constituents.

Scale-free networks have a high degree of robustness against random errors. Upon removal of randomly chosen nodes from the network, the diameter (which characterizes the ability of two nodes to communicate with each other) in a scale-free network is less affected than when perturbing an exponential network (a network whose degree distribution decays exponentially rather than as some inverse power). However, scale-free networks are highly vulnerable to perturbations of highly connected nodes; the diameter increases rapidly and the network breaks into many isolated fragments [24]. This led Jeong and coworkers to check whether a correlation between lethality and connectivity exists in their protein interaction network [12]. They correlated the number of links with the phenotypic effect of removing an individual from the yeast proteome. The findings imply that highly connected proteins with a central role in the network’s architecture are three-times more likely to be essential than proteins with only a small number of links to other proteins. Hence the authors argue that despite the importance of individual biochemical function and genetic redundancy, the robustness against mutation is also derived from the organization of interactions and the topological positions of individual proteins.

Similarly Jeong and colleagues also tested tolerance to random error in metabolic networks [19]. Computer simulations were performed on the metabolic network of Escherichia coli. Upon removal of the most connected substrates the network diameter increases rapidly. However, when randomly chosen substrates are removed, the average distance between remaining nodes is not affected, indicating a striking insensitivity to random errors. Similar results were obtained for all 43 organisms.

Ito observed that omitting the most highly connected proteins in the two-hybrid protein interaction network substantially reduced the size of the largest cluster [15].

Based on empirical observations to explain the persistence of the power-law distribution, Wagner argues that natural selection on the global network structure is not essential to sustain the degree distribution [7].

Wuchty considers whether the observed topologies are a direct consequence of domain evolution [23]. The BA model generates scale-free networks by preferential attachment suggesting that highly connected domains may have originated very early [6]. However, this is not the case. The majority of highly connected domains appear to have arisen late in eukaryotes of larger proteome size.

Rung and coworkers correlated cellular role with (in-/out-) degree in the gene disruption network [21]. They ranked Yeast Protein Database (YPD)-annotated genes by their in- and out-degree separately and found that genes with the highest out-degrees are mostly regulators while genes with a high in-degree are mostly involved in metabolism. Genes were also grouped by their cellular role and a median in- and out-degree calculated for each group. The results confirmed their intuition that general regulators are influencing many genes whereas some metabolic genes are being regulated by many other genes.

Whether the gene networks are modular or not is an open question. Rung and colleagues investigated the number of connected components at different thresholds for observed expression levels [21]. Regardless of the cut-off threshold, the disruption networks have only one clearly dominant connected component, with small components of one to three genes being separated for higher significance thresholds. Generally, the same property stands upon the removal of most highly connected genes for different values of cut-off threshold. The author suggests that one may be able to find such modules by examining the network in more detail, for instance by looking for the smallest cuts in the network graph which lead to disconnection and examining whether the resulting components have a biological meaning.

Looking for subnetworks to give testable hypotheses & useful hints for nodes with a similar functional role or to highlight specific properties of the network as a whole

The list has been roughly ordered by increasing mathematical sophistication, but not necessarily by the level of biological insights used or gained.
Rung and coworkers focused on a subnetwork containing genes of particular interest [21]. They filtered the data for a core set of 20 genes known to be involved in pheromone response and looked at their next neighbors in the gene disruption networks.

Ito also looked at the interacting partners of proteins with a known function to predict gene function [15].

Schwikowski and colleagues predicted functional annotations for proteins based on their interaction partners [25]. They combine all published direct protein–protein interaction data available at the time to form a single large network and many small networks of protein interactions. For a particular protein, the method takes the three most frequent functions of interacting proteins as predictions for the function of the original protein. The methodology is validated by finding that 72% of the proteins of characterized function (meaning that the protein has been assigned a cellular role category as in the YPD) with characterized interaction partners could be assigned to a correct functional category by the method. The authors also note that indirect interactions (i.e., second neighbors) may also strengthen functional predictions.

Proteins of known function and cellular location are shown to cluster together.

Gavin observed 232 distinct multiprotein complexes using tandem affinity purification and mass spectrometry and assigned cellular roles to these complexes by computing functional assignments of individual components according to the YPD and by literature mining [17]. In this way, the author proposed new cellular roles for 344 proteins.

At a higher level, Gavin looked at relationships between complexes to understand integration and coordination of cellular functions [17]. Links were established between complexes sharing at least one protein. Nodes (complexes) were double sized if they contained a large number of proteins that are orthologous to each other, and colored according to function. Several complexes belonging to the same class appear to group, suggesting that the sharing of components reflects functional relationships.

This study demonstrated that orthologous proteins for yeast preferentially interact with complexes enriched with other orthologues suggesting the existence of an orthologous proteome that may represent core functions for the eukaryotic cell [17]. Orthologous gene products are thought to be responsible for essential cellular activities. Overlaps with essential gene products are discussed.

Schlitt aimed to study the similarity of genes or proteins by comparing their neighborhoods in gene networks [26]. Six different networks of three types were considered:

- A mutant network, where an arc from gene A to gene B is present if the deletion of A in the mutant significantly alters the expression level of gene B
- An in silico network, where an arc form gene A to B exists if a transcription factor and its binding site is predicted in the putative promoter of B
- Four different chromatin immunoprecipitation (ChIP) networks, where an arc from A to B is present if transcription factor A was empirically found to bind to the putative promoter region of B

All of these networks are directed and so each source gene has a set of target genes in each network. The authors looked at all pairs of source genes within and between networks and tested whether their target sets intersected more than expected by chance using the hypergeometric distribution and Holm's correction for multiple testing.

A total of 816 functional relationships were identified and these were shown to correspond to protein–protein interactions, co-occurrence in the same protein complexes and/or co-occurrence in abstracts of scientific articles. In particular, biological process annotation was assigned to seven previously uncharacterized genes.

Wagner concentrated on a subset of proteins in the protein–protein interaction network, namely products of duplicate genes [14]. By considering, for example, the fraction of duplicates with shared interaction partners as a function of time elapsed since duplication, the author concludes that within approximately 200 Myr after a duplication, the products of duplicate genes become almost equally likely to have common interaction partners and be part of the same subnetwork as two proteins chosen at random from the network. (These criteria might be taken as a crude measure of the overlap in the function of two genes). Furthermore, the author is able to give estimates of the rate at which interactions are lost and new interactions evolve.

Ideker showed that the expression of many genes is measured over multiple conditions [27]. A network is formed from genes as nodes and edges arising if a protein encoded by one gene interacts, or interferes, with the transcription of the second gene. An algorithm based on simulated annealing finds M high-scoring subnetworks simultaneously, where a high-scoring network is one in which all genes have significant expression change over some subset of conditions.

Bader and Hogue describe a graph theoretic clustering algorithm (MCODE) that detects densely connected regions in large protein–protein interaction networks that may represent molecular complexes [28]. The method is based on vertex weighting by local neighborhood density and outward traversal from a locally dense seed protein to isolate the dense regions according to given parameters. The algorithm has the advantage over other graph clustering methods of having a directed mode that allows fine tuning of clusters of interest without considering the rest of the network and allows examination of cluster interconnectivity, which is relevant for protein networks.

Wagner details an algorithm to reconstruct direct regulatory interactions in gene networks from results of gene perturbation experiments [29]. Ideas for improvements are given, for example, using quantitative rather than qualitative information, using several measures of gene activity and developing new ways to deal with cyclic networks. The performance of the algorithm is checked for missing and flawed data.

Grindrod considers the following inverse problem [9,10]: given a list of vertices in arbitrary order and a list of edges that are suspected to be well modeled by the class of range-dependent graphs, can the vertices be reordered to reveal the
range-dependent connectivity? Grindrod introduces an algorithm (specifically designed for this question) that takes the raw interaction data, estimates the global parameters by matching global properties and then uses maximum likelihood modeling to produce an ordering of the vertices. In the last of these steps, the algorithm produces improved orderings by swapping vertices and whole blocks in the ordering, the chosen ordering being the one which maximizes the product of the odds over all edges that exist.

Such an algorithm was validated on test data sets and also a large yeast proteome data set (providing evidence that such networks capture the connectivity structure seen in proteome data). The algorithm was also shown to be robust to such data errors as additions and deletions.

Vertices close together in any ordering have a high probability of being linked by an edge and therefore may highlight clusters of vertices with similar functional roles. In future work, the author will use the range as a measure of distance between proteins within the proteome as co-ordinates in probabilistic models of co-function.

Higham also considers the above inverse problem [30]. He notices that when the graph is viewed in terms of its adjacency matrix, the problem becomes to find a symmetric row/column permutation that places as many nonzeros as possible close to the diagonal. Higham applies three existing algorithms to six instances of Grindrod’s model ($\alpha = 1$, $n = 600$ and $\lambda = 0.8, 0.9$ or 0.975 for both symmetric and unsymmetric matrices). Higham found that the spectral reordering algorithm is particularly accurate in reordering in these single instances. This algorithm uses a two-sum function as an objective function to be minimized:

$$\sum_{i,j, (i,j \neq 0)} (i-j)^2$$

This can be explained by the observation that the average reciprocal of the distance between the nodes. Spectral clustering is then applied using the Laplacian matrix of the graph.

Gardiner and coworkers consider the protein-docking problem: given two proteins, will they interact to form a stable complex and, if so, how [33]? The docking procedure in this study is based upon two main aspects of protein–protein interaction: shape-based complementarity and hydrogen bonding complementarity. The authors represent each of the two proteins as a set of potential hydrogen bond donors and acceptors and use a clique-detection algorithm developed by Bron and Kerbosch to find maximally complementary sets of donor/acceptor pairs.

Hendrickson considers the molecule problem (also called the distance geometry problem): determining the relative locations of a set of objects in Euclidean space relying only upon a sparse set of pair wise distance measurements [34]. Distances between pairs of atoms in a protein can be estimated via nuclear magnetic resonance spectroscopy experiments and so this NP-hard problem has applications to determining 3D protein structure.

The data in an instance of the molecule problem can be represented by a distance graph where vertices correspond to atoms with an edge joining two vertices, $i$ and $j$, if the distance, $d_{ij}$, between the corresponding atoms is known. This distance can then be associated with each edge. A solution to the problem is a realization of the distance graph (a mapping $p$ that takes each vertex, $i$, to a point, $p_i$, in Euclidean space). The problem is then naturally phrased as a nonlinear global optimization problem with a cost function such as:

$$\sum (p_i - p_j)^2 - d_{ij}^2$$

Hendrickson introduces a divide-and-conquer algorithm in which this large global optimization problem is replaced by a sequence of smaller ones. He finds solutions for uniquely realizable subgraphs and then combines them into a solution for the whole graph.

**Inferring evolutionary network properties/universal structural information from numerous different instances of the same network phenomenon**

Jeong and coworkers comment that the simultaneous emergence of an inhomogeneous structure in both protein interaction and metabolic networks suggests an evolutionary selection of a common large-scale structure of biological networks [12]. On the other hand, such simultaneous emergence may be the result of some external, possibly environmental or physical, constraint, preventing or at least inhibiting possible alternatives.

Jeong and colleagues argue that the metabolic network being scale-free in all 43 organisms indicates a generic nature in this structural organization [19]. However, it should be remembered that these were a restricted set of networks.

The investigators found that the network diameter in the metabolic networks for all 43 organisms is the same, irrespective of the number of substrates found in a given species [19]. This can be explained by the observation that the average network diameter is the same

---

Kannan and Vishveshwara describe a method for detecting side chain clusters in protein structures using a graph spectral method [32]. A graph for a protein structure is constructed by considering particular atoms of the side chains ($\gamma$ atoms) as nodes which are connected with an edge of weight equal to the similarity of these spheres. Graph clustering is then applied using the spectral clustering algorithm of Grindrod (for the single objective function) with a cost function such as:

$$\sum_{i,j} (p_i - p_j)^2 - d_{ij}^2$$

Gardiner and coworkers consider the protein-docking problem: given two proteins, will they interact to form a stable complex and, if so, how [33]? The docking procedure in this study is based upon two main aspects of protein–protein interaction: shape-based complementarity and hydrogen bonding complementarity. The authors represent each of the two proteins as a set of potential hydrogen bond donors and acceptors and use a clique-detection algorithm developed by Bron and Kerbosch to find maximally complementary sets of donor/acceptor pairs.

Hendrickson considers the molecule problem (also called the distance geometry problem): determining the relative locations of a set of objects in Euclidean space relying only upon a sparse set of pair wise distance measurements [34]. Distances between pairs of atoms in a protein can be estimated via nuclear magnetic resonance spectroscopy experiments and so this NP-hard problem has applications to determining 3D protein structure.

The data in an instance of the molecule problem can be represented by a distance graph where vertices correspond to atoms with an edge joining two vertices, $i$ and $j$, if the distance, $d_{ij}$, between the corresponding atoms is known. This distance can then be associated with each edge. A solution to the problem is a realization of the distance graph (a mapping $p$ that takes each vertex, $i$, to a point, $p_i$, in Euclidean space). The problem is then naturally phrased as a nonlinear global optimization problem with a cost function such as:

$$\sum (p_i - p_j)^2 - d_{ij}^2$$

Hendrickson introduces a divide-and-conquer algorithm in which this large global optimization problem is replaced by a sequence of smaller ones. He finds solutions for uniquely realizable subgraphs and then combines them into a solution for the whole graph.

**Inferring evolutionary network properties/universal structural information from numerous different instances of the same network phenomenon**

Jeong and coworkers comment that the simultaneous emergence of an inhomogeneous structure in both protein interaction and metabolic networks suggests an evolutionary selection of a common large-scale structure of biological networks [12]. On the other hand, such simultaneous emergence may be the result of some external, possibly environmental or physical, constraint, preventing or at least inhibiting possible alternatives.

Jeong and colleagues argue that the metabolic network being scale-free in all 43 organisms indicates a generic nature in this structural organization [19]. However, it should be remembered that these were a restricted set of networks.

The investigators found that the network diameter in the metabolic networks for all 43 organisms is the same, irrespective of the number of substrates found in a given species [19]. This can be explained by the observation that the average
number of reactions in which a certain substrate participate increases with the number of substrates found within a given organism. This apparent conservation of network diameter may represent an additional survival and growth advantage. A particular instance of a scale-free model has fixed average connectivity as the number of nodes increases. This implies that the diameter of the network grows logarithmically with the addition of new nodes.

The importance of the most highly connected substrates to maintaining network diameter was previously discussed. This led Jeong and coworkers to ask whether these hubs in metabolic networks are organism specific [19]. They find that the ranking (on the basis of the number of links) of the most connected substrates is practically identical for all 43 organisms. Only approximately 4% of all substrates were found in all species and these represented the most highly connected substrates found in any individual substrate. This may indicate a generic utilization of the same substrates by each species.

Wuchty built protein domain networks for six individual species that developed differently during the course of evolution [23]. The slopes of the regression lines of the frequency distribution of links increase with organism complexity, suggesting a trend that guides multicellular organisms to higher domain connectivity.

**Validation of biological information**

This includes comparing data- and literature-driven networks to check the amount of biological information in the data-derived network.

The gene disruption network described by Rung and colleagues demonstrated significantly higher overlap with a literature network (where there is an undirected edge between genes x and y if gene x is mentioned in the description of gene y) than a randomly generated network with the same overall topology [21].

To check that the emergence of a single huge network is not inherent to the systematic two-hybrid analysis, Ito extracted protein interactions identified in conventional studies from the YPD and found that these also formed a big cluster [15].

Posterior probabilities that pairs of proteins are associated within the same protein complex were calculated by Gilchrist and coworkers [16]. These probabilities were found to be remarkably similar to the observed association probabilities for proteins in the MIPS Complex Catalogue, a reference set of manually curated protein complexes.

**Combining results & dealing with errors**

A protein–protein interaction network mediated by peptide recognition modules is derived by combining two networks obtained from two very different methodologies [35].

Two main sources of database errors could affect the analysis of metabolic networks [19]: erroneous annotation of biological reactions and missing reactions and pathways. The results were found to be robust to these errors.

Data from two independent large-scale two-hybrid projects failed to largely overlap, the reason for this being unclear [8,15]. This suggests that it would be better to have several independent interaction sequence tag (IST) projects using different constructs and to combine their results. Such a combination should represent multiple crossvalidation of interactions that is some form of intersection, not a simple union or averaging of results.

Bader and Hogue investigated the biases, overlaps and complementarities among published interaction information on *S. cerevisiae* [36]. In particular, they carried out analysis on two high-throughput mass spectrometry-based protein interaction data sets from budding yeast, comparing them with each other and to other interaction data sets [17,18]. Their analysis also leads to the conclusion that integrating many different experimental data sets would yield a clearer biological view than any single method alone.

The two-hybrid system is generally claimed to show many false positives or biologically meaningless signals. Ito checked false positives using the YPD and known functions [15]. False negatives were also shown to be very common, which may have been due to the use of full-length proteins.

Gilchrist and colleagues offered a statistical framework to interpret high-throughput proteomic data sets (which often have high rates of error) and to combine information from different high-throughput experiments [16]. The approach was illustrated for the type of data resulting from complex purification experiments where prey proteins, which are in the same complex as a bait protein, were identified. In particular, they applied their methods to data sets [17,18]. Each pair of proteins is represented by a pair of values (*t*, *s*), where *t* indicates the number of opportunities to observe an association between the proteins and *s* indicates the number of observed associations. The authors use this information, together with a binomial model of protein sampling, Bayes' law and a maximum-likelihood estimation of error probabilities, to calculate a posterior probability that any two proteins are associated within the same protein complex. They show that their approach permits combining information from multiple experiments within and between different high-throughput data sets.

The maximum-likelihood estimation of experimental error rates has the advantage of allowing the evaluation of data quality without the use of a reference set of verified protein interactions. For this estimation, one needs the approximate number of different prey proteins that a particular experiment can detect. The authors accordingly provide a technique to estimate this number.

**Expert opinion**

The authors have categorized quite a wide selection of network-based, distinct research into different themes, each with distinct objectives.

The increasing awareness of networks and network properties, from a variety of data sources in the biosciences, has led to work developing suitable mathematical models of networks. It is essential for workers to have a common vocabulary and to be able to compare and contrast different networks (large interaction databases). The common theme here is of very distinctive
global connectivity properties being derived from actual networks with scale-free and/or small world properties. Surely all researchers with access to large interaction databases should publish and characterize the degree distribution and other parameters such as the clustering coefficient. This is the first step to standardizing insight and making comparable statements about data from different sources. The very fact that the emergence of proteome and genome data are likely to spur research into the development of more suitable classes of networks and concepts should illustrate to those developing the data how new and sophisticated this field is. It is clearly essential that analysis of such interaction data can be automated and that algorithms are available to deal with very large networks. The days when researchers had to select a small, manageable subset of their favorite proteins to focus on are, or must soon be, surely over.

The extraction of comparative structure is key from the point of view of contrasting different data sources; distinct techniques, distinct laboratories, and distinct researchers. Moreover, it is the global organization of the proteome interactions that will hopefully give the clue to how organisms and functions emerge, in an organized way, during embryo growth.

Clearly the program of life, which runs from conception through embryogenesis, building organisms, must be written and operate through interactions of basic units (in this case individual genes and proteins); but how is it organized? It has multiple properties which together are unlike any computer program: re-use of units, substitution for missing faulty units, self organization, recursion, sensitivity to environmental external fields, and so on. More importantly, it is compact and efficient (there are only tens of thousands of basic units); and functionally robust to perturbations. The recognition of the common structure elements and properties of interacting networks will provide us with the first important clues as to how this program is organized.

Research where the imposition of network concepts allows us to gain understanding of the possible hierarchical organization (or relative importance) of proteins; and research that causes us to identify subsets of connected nodes that may share common functional roles has also been considered. Of course, it is possible for proteins to be in more than one such sub network: hence the sets correspond to higher order functionality.

The extraction of comparative structure is fascinating (what makes us like ourselves?): how is it organized? Clearly the program of life, which runs from conception to death, must be written and operate through interactions of basic units (in this case individual genes and proteins); but how is it organized? It has multiple properties which together are unlike any computer program: re-use of units, substitution for missing faulty units, self organization, recursion, sensitivity to environmental external fields, and so on. More importantly, it is compact and efficient (there are only tens of thousands of basic units); and functionally robust to perturbations. The recognition of the common structure elements and properties of interacting networks will provide us with the first important clues as to how this program is organized.

Research where the imposition of network concepts allows us to gain understanding of the possible hierarchical organization (or relative importance) of proteins; and research that causes us to identify subsets of connected nodes that may share common functional roles has also been considered. Of course, it is possible for proteins to be in more than one such sub network: hence the sets correspond to higher order functionality.

The extraction of comparative structure is fascinating (what makes us like ourselves?): how is it organized? Clearly the program of life, which runs from conception to death, must be written and operate through interactions of basic units (in this case individual genes and proteins); but how is it organized? It has multiple properties which together are unlike any computer program: re-use of units, substitution for missing faulty units, self organization, recursion, sensitivity to environmental external fields, and so on. More importantly, it is compact and efficient (there are only tens of thousands of basic units); and functionally robust to perturbations. The recognition of the common structure elements and properties of interacting networks will provide us with the first important clues as to how this program is organized.
Network and graph theory concepts in proteomics

weight analysis of super large data resources is still a thing for the future. There is a further problem with data from the perspective of the network analysts. Whilst they accept the data are noisy with false negatives as well as positives (and algorithms need to be robust to these errors, and can be designed to tackle this problem), it can often be difficult to obtain raw data for which the provenance is fully understandable. Undoubtedly the network analysis cannot be bolted on after the fact and the best work will be done by groups containing both expertise and intimate knowledge of the nature and sensitivity of the particular data sources and the theory of networks.

Clearly there are obvious and queer uses for networks and graph theory emerging at a great pace due to the sheer volume of interaction data and size of the proteomes. We hope that this snapshot review will encourage researchers to develop their own ideas in these areas and to rise to some of the challenges and questions that this early work has raised.

Appendix

All the data sets considered here consist of some set of objects supplemented with information about relationships between the objects. For example, protein–protein interaction data derived from two-hybrid assays tells us which pairs of proteins have been experimentally found to interact directly [8,15]. Complex composition data derived from mass spectroscopy of purified protein complexes tells us whether a pair of proteins are found in the same complex [17,18]. Here the relationship between the pair of proteins may be a direct or indirect interaction, complex formation being more than the sum of binary interactions. Gene deletion data derived from a cDNA microarray hybridization assay tells us how much the deletion of a particular gene alters the expression of another gene [22].

Such data are naturally represented in a graph, a mathematical object consisting of nodes (also known as vertices) and edges. If two nodes are joined by an edge, they are said to be adjacent. For protein interaction data, each protein forms a node in the graph, with two nodes (proteins) being adjacent if they bind together. Depending on the type of information available about the relationships between the objects in the data, representative graphs can be made more complicated for example by giving the edges a direction or by assigning weights to the edges or the nodes. In a gene disruption network, nodes represent genes and a directed edge from node \(i\) to node \(j\) is drawn if the deletion of gene \(i\) alters the expression of gene \(j\). Furthermore, the amount of change in expression could be appended to each edge as an edge weight.

Several measures on graphs are available for their characterization and comparison:

A path between two nodes, \(i\) and \(j\), is a sequence of adjacent proteins leading from \(i\) to \(j\). The distance (or shortest path length) between two nodes is, as one would expect, the number of edges along the shortest path connecting them. Small-world behavior manifests itself in a graph when there is a relatively short distance between any two nodes of the graph. This can be quantified by the network diameter (also known as the characteristic path length and often denoted by \(L\)), that is, the shortest path length averaged over all pairs of nodes.

Another way to characterize the structure of a graph is to measure the tendency of nodes to form clusters of highly connected nodes. One such measure, often denoted by \(C\), is the WS clustering coefficient. Consider a vertex \(i\) with \(n_i\) neighbors. If all \(n_i\) neighbors were to form a clique, there would be \(\frac{n_i(n_i-1)}{2}\) edges between them. The proportion of these edges that actually exist is denoted by \(C_i\). The clustering coefficient of the whole network is then the average of all individual \(C_i\)’s. Graphs where there is a high degree of clustering but a relatively short distance between all nodes have been termed small world networks.

The simplest measure on a graph is that of the number of edges connecting a node to other nodes, that is, the degree of a node, and its distribution in the graph, that is, the degree (or connectivity) distribution. This \(P(k)\) gives the probability that a randomly selected node has exactly \(k\) edges. Graphs characterized by a connectivity distribution that peaks at the average number of links \(<k>\) and decays exponentially are known as exponential networks. Prominent protagonists of this type are the random graph model and the WS small-world model. Both lead to fairly homogeneous networks with nodes comprising approximately the same number of links \(k \sim <k>\). Scale-free networks are a class of networks for which the connectivity distribution decays as a power-law \(P(k) \sim k^{-\gamma}\). Compared with exponential networks, the probability that a node is highly connected \((k >> <k>)\) is statistically significant in scale-free networks.

A good introduction to graph theory is given in [37].

---

**Key issues**

- It should become common practice to present and publish network (graph theoretical) properties of protein interaction networks, such as the degree distribution, the clustering coefficient, and so on, calculated according to their standard definitions.
- There are a number of alternative model candidates for characterizing large interaction networks and, since this is an ongoing area of theoretical research, it would be useful for data sets to be made easily available to the mathematical and physics communities.
- Experience of network analysis shows that global and emergent properties of networks are dependent on small-scale structure.
- Biologists should attempt to analyze large-scale networks themselves and avoid premature reduction to subsets of proteins: mathematicians and physicists should be encouraged further to find opportunities to collaborate.
- The many-to-many nature of the proteins to functions relationship requires a network approach not a clustering type of methodology. The availability of protein–interaction network data are likely to challenge and force the pace of research within network theory in the next few years.
References


Affiliations

- Peter Grindrod, PhD
  Lawson Software, Magdalen Centre, The Oxford Science Park, Oxford OX4 4GA, UK
  Tel.: +44 1865 784 264
  Fax: +44 1865 784 265
  peter.grindrod@lawson.com
- Miila Kibble, PhD
  Department of Mathematics, University of Turku, FIN-20014 Turku, Finland
  milanr@utu.fi