

# Protein Interaction Networks and Their Statistical Analysis

Waqar Ali, Charlotte Deane, and Gesine Reinert, University of Oxford,  
Department of Statistics



# 1

## Protein Interaction Networks and Their Statistical Analysis

### 1.1 Introduction

A major aim of post-genomic biology is to provide a complete systems-level snapshot of the cells in an organism. This requires not only a detailed description of the different components of a cellular system, but also a deep understanding of how those components interact with each other. During the last few years large scale genome sequencing projects and advances in protein analysis technologies have gathered a huge data set of the components of living systems. The focus now is on developing successful models of the interactions of these components that can explain real living systems.

Protein-protein interactions (PPI) are the corner-stone of most biological processes taking place in cells. Recent advances in high throughput interaction detection techniques have led to the elucidation of substantial parts of the entire protein interaction set for several species. These large datasets of interactions can be conveniently represented in the form of networks, where the nodes represent proteins and edges represent interactions.

Given this data, some pertinent research questions are:

1. How do these networks work? How can a network be manipulated in order to prevent, say, tumor growth?
2. How did these biological networks evolve? Could mutation affect whole parts of the network at once?
3. How similar are these networks? How much can we infer from the PPI network of one organism to those of other organisms?
4. How are these networks linked with other networks, such as gene interaction networks?
5. What are the building principles of these networks? How are resilience and flexibility achieved?

In order to attack these questions we first focus on

1. How to best describe networks?
2. How to compare networks from related organisms?
3. How to model network evolution?

4. How to find relevant sub-structures of a network?
5. How to predict functions from networks?
6. How to infer and validate edges?

In the following sections, we first introduce proteins and their interactions (Section 1.2), experimental techniques to detect them and the high rate of false-positives and false-negatives. Section 1.3 covers network descriptions and model fitting. Approaches for the comparison of networks across species are discussed in Section 1.4, and Section 1.5 discusses evolution in networks. Community detection, and the identification of relevant sub-structures in a network, is found in Section 1.6, and Section 1.7 contains prediction using network structure as well as validation and inferring of edges. We conclude with a brief overview of current and future research directions.

## 1.2 Proteins and their interactions

Proteins are the most versatile macromolecules in living systems and serve crucial functions in most biological processes. For example, they function as catalysts, transport and store other molecules such as oxygen, provide mechanical support and immune protection and control growth differentiation. Proteins are linear polymers built of monomer units called amino acids. They fold up into three-dimensional structures that are thought to be determined by the sequence of amino acids in the protein polymer (Berg et al. 2006).

### 1.2.1 Protein structure and function

Due to interactions between the chemical groups on amino acids, a few characteristic patterns occur frequently within folded proteins. These recurring shapes are called secondary structure, and they occur repeatedly as they are particularly stable (Brändén and Tooze 1991). The two most commonly occurring secondary structures are the alpha-helix and the beta strand. These are both highly regular local sub-structures (Figure 1.1). The term tertiary structure is used to refer to the three-dimensional structure of a single protein molecule. This final shape is determined by a variety of bonding interactions between the amino acids. The tertiary structure of a protein is thought to determine its functionality. Some proteins also possess quaternary structure which involves the association of two or more polypeptide chains into a multisubunit or oligomeric protein. The polypeptide chains of an oligomeric protein may be identical or different.

**Figure 1.1** Protein structure : (A) Secondary structure elements and (B) Tertiary structure.

From the biological perspective, the function of a protein is the most important characteristic, which in turn is determined to a large extent by its structure. Although proteins can often be classified into functional groups, many proteins can carry out multiple functions dependent on the cellular context. Some major classifications include enzymes, antibodies,

transport proteins, hormones, signalling proteins and structural proteins (Berg et al. 2006). Proteins can interact with each other and with other macromolecules to form complex assemblies. The proteins within these assemblies often act synergistically to generate capabilities not afforded by the individual component proteins. These assemblies include macromolecular machines that carry out the accurate replication of DNA, the transmission of signals within cells and many other essential processes.

### *1.2.2 Protein-protein interactions*

Most proteins function through interaction with other molecules, and often these are other proteins. There is an important distinction between transient and obligate protein interactions. Many proteins exist as parts of permanent obligate complexes such as multi-subunit enzymes, which may often fold and bind simultaneously. Other interactions are fleeting encounters between single proteins or larger complexes. These include enzyme-inhibitor, hormone-receptor, and signaling-effector types of interactions. This distinction is not always well understood, and the classification is sometimes difficult (Mintseris and Weng 2005).

The interactions between proteins are important for many biological functions and operate at almost every level of cell function including in the structure of sub-cellular organelles, the transport machinery across the various biological membranes, packaging of chromatin, the network of sub-membrane filaments, muscle contraction, signal transduction and regulation of gene expression (Huthmacher et al. 2008). Thus, the elucidation of protein interactions is a central problem in biology today. Unless we understand the complex interaction patterns of the tens of thousands of proteins that constitute our proteome, we cannot hope to even attempt to efficiently combat some of the most important diseases, let alone gain an integrated understanding of the living cell.

### *1.2.3 Experimental techniques for interaction detection*

Given their importance, there has been a surge in studies of protein interactions during the last decade. Some of the initial experiments focused on small and specific sets of interactions of interest to a particular research group, and were characterised by repeated observations. However the sheer scale of the number of possible interactions that proteins in a cell may undergo soon made researchers worldwide realise that there are probably more different possible interactions than there are researchers in the field. Thus, high throughput approaches for the elucidation of protein-protein interactions have rapidly gained appreciation.

A few of the most popular and widely used experimental techniques are summarised below. These approaches differ widely in the quality and quantity of interaction data reported. Moreover, large scale studies using these methods show little overlap with each other.

#### **Yeast two hybrid**

The two-hybrid system is a genetic method that uses transcriptional activity as a measure of protein-protein interaction (Chien et al. 1991). Two hybrid proteins are created: one is a bait protein of interest fused to a DNA-binding domain and the other is a prey protein fused to a transcription activation domain. These two hybrids are then expressed in a cell containing one or more reporter genes. If the bait and prey proteins interact, this can be detected by

expression of the reporter genes. While the assay has been generally performed in yeast cells, it works similarly in mammalian cells. If all proteins in a genome are treated as prey and bait in pairwise tests, all possible interactions can be probed. The main criticism applied to the yeast two-hybrid screen of protein-protein interactions is the possibility of a high number of false positive (and false negative) identifications. The exact rate of false positive results is not known, but estimates are between 35 and 70% (Hart et al. 2006).

### **Tandem affinity purification**

The two-hybrid system uses binary combinations to explore the interaction space of a set of proteins. A different strategy to solve this problem is to purify all protein complexes from a living cell, subsequently characterizing their constituent parts. This is the strategy that lies at the heart of tandem affinity purification (TAP)-tagging approaches. First the nucleotide sequence encoding the TAP tag is inserted at the end of the open reading frame to be investigated. A column with immunoglobulin beads would retain the TAP-tagged protein and associated complexed proteins. The complex is then purified and separated to its constituent protein parts and analysed on a mass spectrometer (Puig et al. 2001). With the help of software, peptide sequences and protein identities are obtained from mass spectrometry. Compared to the yeast two hybrid system, TAP is thought to have lower false negative rates (15%), and a false positive rate of 35% (Hart et al. 2006).

### **Co-immunoprecipitation**

One of the most common and rigorous demonstrations of protein-protein interaction is the co-immunoprecipitation (Co-IP) of suspected complexes from cell extracts. Co-IP confirms interactions utilising a whole cell extract where proteins are present in their native conformation in a complex mixture of cellular components that may be required for successful interactions. An antibody specific to the bait protein is used to extract the complex of interest. This complex is purified and then evaluated using SDS-PAGE followed by Western blotting with specific antibodies (Phizicky and Fields 1995). Although very accurate, Co-IP can only determine the interaction between one pair of proteins at a time.

#### *1.2.4 Computationally predicted data-sets*

Parallel to experimental efforts, a number of computational methods have been developed for the prediction of protein interactions. Complete genome sequencing projects provide the vast amount of information needed for these analyses. The methods utilize the genomic and biological context of genes in complete genomes to predict functional linkages between proteins. Given that experimental techniques remain expensive, time-consuming, and labour-intensive, these methods represent an important advance in proteomics.

One of the first methods for predicting protein-protein interactions from the genomic context of genes utilizes the idea of co-localisation, or gene neighbourhood. Such methods exploit the notion that genes which physically interact (or are functionally associated) will be kept in close physical proximity to each other on the genome (Bowers et al. 2004; Overbeek et al. 1999; Tamames et al. 1997). This method has been successfully used to identify new members of metabolic pathways (Dandekar et al. 1998).

Another method exploits the co-occurrence of homologous pairs of genes across multiple genomes. The fact that a pair of genes remains together across many disparate species represents a concerted evolutionary effort that suggests that these genes are functionally associated or physically interacting. The analysis of phylogenetic context in this fashion has been termed phylogenetic profiling (Pellegrini et al. 1999). This method has been used not only to infer physical interaction, but also to predict the cellular localisation of gene products (Bowers et al. 2005; Marcotte et al. 2000).

Methods using the analysis of gene fusion across complete genomes have also been proposed (Enright et al. 1999; Marcotte et al. 1999). A gene fusion event represents the physical fusion of two separate parent genes into a single multi-functional gene. This is the ultimate form of gene co-localisation as interacting genes are not just kept in close proximity on the genome, but are also physically joined into a single entity (Skrabanek et al. 2008). These events are detected by cross-species sequence comparison and provide a way to computationally detect functional and physical interactions between proteins. Although the method is not generally applicable to all genes, it has been shown to have an accuracy as high as 90% and has been successfully applied to a large number of genomes, including eukaryotes (Enright and Ouzounis 2001).

It must be noted that all of these methods use some experimental data sources and as a result, they all suffer from the limitations of experimental approaches and incompleteness of observed data. Moreover, many of these techniques detect functional associations between proteins that do not necessarily indicate physical interactions.

### 1.2.5 Protein interaction databases

As a consequence of the experimental and computational approaches providing data about interacting proteins on a genome- and proteome-wide scale, several research groups have made an important effort in designing and setting up databases. The interaction data in these databases usually results from the integration of diverse data sets. Public databases of protein interactions include:

- Biomolecular Interaction Network Database - BIND (Bader et al. 2001);
- Database of Interacting Proteins - DIP (Xenarios et al. 2002);
- General Repository for Interaction Datasets - GRID (Breitkreutz et al. 2003);
- Molecular Interactions Database - MINT (Zanzoni et al. 2002);
- Search Tool for the Retrieval of Interacting Genes/Proteins - STRING (Mering et al. 2003);
- Human Protein Reference Database - HPRD (Keshava Prasad et al. 2009).

The structure and type of data that these databases contain is similar, but not identical. Most of these databases contain protein-protein interaction data only, though MINT and BIND also feature interactions involving non-protein entities such as promoter regions and mRNA transcripts. DIP is probably the most highly curated database of protein interactions. Curation in DIP is carried out manually by experts and also automatically using computational approaches.

The sheer volume of interaction data available in these databases poses many challenges along with opportunities. On the one hand, such large scale data can enable one to infer

large scale properties of cellular systems. On the other hand, the data has to be presented and analyzed in a manageable framework.

### 1.2.6 Error in PPI data

Recent estimates suggest that the full yeast protein-protein interaction network contains 37,800-75,500 interactions and the human network 154,000-369,000 (Hart et al. 2006), but owing to a high false negative rate, current experimental data sets are roughly only 10 to 50 percent complete. Analysis of yeast, worm, and fly data indicates that 25 to 45 percent of the reported interactions are likely false positives (Huang et al. 2007). Membrane proteins have higher false-discovery rates on average, and signal transduction proteins have lower rates. The overall false-negative rate ranges from 75 percent for worm to 90 percent for fly, which arises from a roughly 50 percent false-negative rate due to statistical under-sampling and a 55 to 85 percent false-negative rate due to proteins that appear to be systematically lost from the assays (Huang et al. 2007).

Error rates for large-scale PPI datasets can be estimated computationally using methods like the expression profile reliability (EPR) index and paralogous verification method (PVM) (Deane et al. 2002). The EPR index estimates the biologically relevant fraction of protein interactions detected in a high throughput screen. It does so by comparing the RNA expression profiles for the proteins whose interactions are found in the screen with expression profiles for known interacting and non-interacting pairs of proteins. PVM judges an interaction likely if the putatively interacting pair has paralogs that also interact. More recent methods such as IG1, IG2 (Saito et al. 2002, 2003) and IRAP (Chen et al. 2005) use network structure in assessing individual interaction reliabilities.

Current PPI networks are, therefore, a sample of the complete network. Biases in sampling could lead to even more drastic differences between the complete network and the sub-sample that we observe. Even data derived from high-throughput studies are not an unbiased sample of the complete network; rather, they are biased toward proteins from particular cellular environments, toward more ancient, conserved proteins and toward highly expressed proteins (von Mering et al. 2002). Current interaction maps represent the first steps on the way to accurate networks, and should continue to improve in accuracy and sensitivity with time.

### 1.2.7 The interactome concept and protein interaction networks

The compendium of all molecular interactions present in cells is called the *interactome*. When spoken in terms of proteomics, interactome refers to the entire set of protein-protein interactions for a species. Due to limitations of current knowledge, the experimentally and computationally determined set of protein interactions available in databases is a subset of the real interactome. Still, the sheer number of known protein interactions makes even the simplest analysis a difficult task. It has therefore become routine to represent this data in the form of protein interaction networks. A protein interaction network can summarize large amounts of interaction data in the form of graphs, with proteins as nodes and interactions as edges. The networks are undirected, and may be weighted. The weights of the edges could represent the confidence level for the interaction (typically based on the experimental or computational method used to detect that particular interaction). A distinct advantage of such a representation is the visual and computational ease in detecting higher level structures

in interaction data. For instance, many biological processes are a result of more than two proteins acting in sequential pathways or simultaneously forming multi-protein complexes, which can be identified relatively easily in a network.

### 1.3 Network analysis

Computational analysis of PPI networks makes extensive use of graph theoretical techniques and concepts from literature on random networks. In the following sections, we introduce some basic terminology and present several methods as well as models for network analysis.

#### 1.3.1 Graphs

A *graph* consists of *nodes* (also called *vertices*) and *edges* (also called *links*). Nodes may possess characteristics which are of interest (such as protein structure or function). Edges may possess different weights, depending on for example, the strength of the interaction or its reliability. Mathematically, we abbreviate a graph as  $G = (V, E)$ , where  $V$  is the set of nodes and  $E$  is the set of edges. We use the notation  $|S|$  to denote the number of elements in the set  $S$ . Then  $|V|$  is the number of nodes, and  $|E|$  is the number of edges in the graph  $G$ . If  $u$  and  $v$  are two nodes and there is an edge from  $u$  to  $v$ , then we write that  $(u, v) \in E$ , and we say that  $v$  is a *neighbour* of  $u$ ; and  $u$  and  $v$  are *adjacent*. If both endpoints of an edge are the same, then the edge is a *loop*. In general in PPI networks, we exclude self-loops as well as multiple edges between two nodes. Edges may be *directed* or *undirected*; here we shall mainly deal with undirected edges. While some authors assume that, in contrast to graphs, *networks* are connected, here we make no such assumption and use the terms graph and network interchangeably.

#### 1.3.2 Network summary statistics

Although large networks are typically high dimensional and complex objects, many of their important properties can be captured by calculating relatively simple summary statistics. One of the most is the average degree. The *degree*  $deg(v)$  of a single node  $v$  is the number of edges which are adjacent to  $v$ . The *average degree* of a graph is then the average of its node degrees. In protein interaction networks it has been found that the vast majority of nodes have low degrees, whereas a few nodes are highly connected (Figure 1.2). This apparent similarity to the power law distribution prompted the popular classification of PPI networks as *scale-free* (Barabasi and Oltvai 2004; Jeong et al. 2001), although subsequent studies have challenged this view (de Silva et al. 2006; Lima-Mendez and van Helden 2009; Tanaka et al. 2005).

The *clustering coefficient* for a graph measures the tendency of the formation of tightly connected groups of nodes. Two versions of the clustering coefficient are in use: The global clustering coefficient is defined as the number of closed triplets of nodes in the network divided by the total number of triplets. The local clustering coefficient is defined for single nodes and is defined as the number of links existing between the neighbours of node divided by the total number of possible links; for node  $i$  with  $k_i$  neighbours in its set  $N_i$  of neighbours, the clustering coefficient  $C_i$  is

$$C_i = \frac{2|\{(v_j, v_k) \in E : v_j, v_k \in N_i\}|}{k_i(k_i - 1)}.$$

The *shortest path length* and *average shortest path length* of a graph are also commonly used summary statistics, where path length is defined as the number of edges traversed to reach a target node from a source node. Other popular summaries include the *betweenness of an edge (or node)* which counts the proportion of all the shortest paths in the network which pass through this edge (or node). For a comprehensive review of graph summary statistics, see for example Luciano et al. (2007).

**Figure 1.2** Frequency of node degrees ( $k$ ) in the yeast DIP network (accessed July 2010)

A comparison of yeast and human interaction networks indicates very similar clustering and path-length statistics, despite the difference in size (Table 1.1). To judge whether these difference are significant, network models are needed (see Section 1.3.4).

**Table 1.1** Summary statistics for yeast DIP and human HPRD protein interaction networks. Data downloaded from DIP and HPRD on 15 July 2010.

Summary Statistic	Yeast DIP	Human HPRD
Nodes	4823	12937
Edges	17471	43496
Avg. Degree	6.10	6.72
Avg. Clustering Coefficient	0.1283	0.1419
Avg. Shortest Path	4.14	4.40

### 1.3.3 Network Motifs

In addition to considering general graph summary statistics, it has proven fruitful to describe the smaller-scale structure of networks in terms of *subgraphs* and *motifs*. Given a graph  $G = (V, E)$ , a subgraph  $G_S = (V_S, E_S)$  consists of a subset of nodes  $V_S \subseteq V$  and a subset of edges  $E_S \subseteq E$  connecting the nodes of  $V_S$  in the original graph. The subgraph induced by  $V_S$  is the subgraph  $G_S$  that includes all the edges of  $G$  which connect the vertices of  $V_S$ . A motif is commonly defined as a subgraph with a fixed number of nodes and a given topology that appears more often in a graph than expected by chance. The over-representation of a subgraph is established on the basis of its frequency compared to the average frequency of the same subgraph in a set of random networks (either based on a suitable model or generated by shuffling the edges of the original network while keeping the same degree distribution). A motif of size  $k$ , i.e. containing  $k$  nodes, is called a  $k$ -motif. As the number of possible  $k$ -motifs grows very fast with  $k$ , only small size  $k$ -motifs have been studied in PPI networks. The two most commonly studied motifs in the context of PPI networks are *cliques*, i.e. complete subgraphs, and *k-cores*, i.e. graphs where every node has the degree at least  $k$ .

The enumeration of cliques and  $k$ -cores in particular has been used as a method of detecting protein complexes and functionally related proteins in protein interaction networks. Apart from PPI networks, motifs have also been found to be present in gene regulatory, metabolic and transcription networks (Alon 2006).

**Figure 1.3** Some examples of motifs: (A) Line, (B) Triangle, (C) Square and (D) 5-node clique

There is a large body of literature on network models in different fields from physics to sociology to the internet and biology (Alm and Arkin 2003; Dorogovtsev and Mendes 2003; Wasserman and Faust 1995). In the next section, we present some random network models used to model protein interaction networks. Unfortunately, for some research questions none of these models provide a good fit for the data (Rito et al. 2010). Yet there may be research areas or new data which make these models relevant.

#### 1.3.4 Models of random networks

In order to judge whether a network summary is "unusual" or whether a motif is "frequent", there is an underlying assumption of randomness in the network. To understand mechanisms which could explain the formation of networks, mathematical models have been suggested. The main models, discussed elsewhere in this volume, are firstly *Bernoulli or Erdős-Renyi (ER) random graphs* (Erdos and Renyi 1960), with a finite node set and independent identically distributed edges; a variant is the random graph model  $G(n, m)$  with  $n$  nodes and  $m$  edges chosen uniformly at random from all  $\binom{n}{2} = \frac{1}{2}n(n-1)$  possible edges. *Barabási-Albert (BA) models* (Barabasi and Albert 1999) start with a small complete graph; new nodes attach to existing nodes with probability proportional to (a power of) the degree of the existing node, resulting in an asymptotic power-law degree distribution. *Erdős-Renyi Mixture Graphs*, also known as *latent block models* in social science (Nowicki and Snijders 2001) assume that nodes are of different types, edges are independent, and the probability for an edge varies depending on the type of the nodes at its endpoints. Another set of models are *exponential random graph ( $p^*$ ) models* where all edges of the network are modelled simultaneously, making it easy to incorporate dependence. A variation of the ER model is an ER graph with fixed degree distribution, abbreviated *ER-DD*. For a given real graph as input, an ER-DD graph is constructed to have not only the same number of nodes and edges as the input graph, but also the same degree distribution. Finally, *geometric random graph models (GEOdD)* have also been proposed (Penrose 2003), which are constructed by dropping  $n$  nodes randomly uniformly into the unit square (or more generally according to some arbitrary specified density function on  $d$ -dimensional Euclidean space) and adding edges to connect any two nodes distant at most  $r$  from each other.

The above models were initially proposed in non-biological contexts. While studies suggest that they are able to reproduce some coarse properties of biological networks, it is difficult to relate their growth mechanisms to real biological systems. This has led to the proposal of models specifically aimed at protein interaction networks. For instance, although

the Barabási and Albert class of models proposes a preferential attachment rule resulting in a power-law degree distribution observed in protein interaction networks, the underlying reason for preferential attachment is unclear. A more biologically plausible mechanism is gene *duplication and divergence* (DD) (Ispolatov et al. 2005a), where nodes are randomly selected and copied along with their links (Figure 1.4). In the DD model, the degree of a node increases mainly by having duplicate genes as its neighbours. Therefore, the preferential attachment rule is achieved implicitly, with highly connected nodes having more chance to have duplicate genes as their neighbours. DD models have been shown to closely model the degree distribution observed in real protein networks (Evlampiev and Isambert 2007). The DD model is also shown to generate hierarchically modular networks under certain conditions. If self-interactions (homo-oligomers) are taken into consideration, the DD model gives rise to networks with patterns of clustering and abundance of cliques similar to those found in natural networks (Ispolatov et al. 2005b).

**Figure 1.4** Duplication divergence model. The duplicate node (blue) loses some of the original links and creates new links.

While the basic duplication divergence model remains by far the most widely accepted one in protein interaction network literature, some recent studies have proposed enhancements such as mixture models combining DD and preferential attachment (Ratmann et al. 2007). Alternatives to the DD model have also been investigated, including a crystal growth model that captures the age-dependency of interaction density in the yeast interaction network along with hierarchical modularity (Kim and Marcotte 2008).

### 1.3.5 Parameter estimation for network models

In most of these models it is necessary to estimate parameters. In ER graphs, where the unknown parameter is the edge probability, this probability can be estimated using standard maximum likelihood, yielding the *graph density* as an estimate. The graph density is the number of edges that are present in the network, divided by the total possible number of edges in the network. In the  $G(n, m)$  version, once the number of nodes and the number of edges are observed, no parameters are to be estimated. In Barabási-Albert models, the parameters include the power exponent for the node degree, as occurring in the probability for an incoming node to connect to some node already in the network, and the size of the initial complete graph. Estimation depends on the precise model formulation - the general Barabási-Albert model does not specify the joint distribution of edges. In exponential random graphs, unless the network is very small, maximum-likelihood estimation quickly becomes numerically unfeasible. Instead, Markov chain Monte Carlo estimation is employed. Unfortunately in exponential random graph models it is known that in some small parameter regions the stationary distribution of the Markov chain is not unique.

To assess the model fit, the distribution of a network statistic under the model of choice can be used to see whether the observed value of the network statistic is unusual. This could be carried out by establishing the (asymptotic) distribution of the network statistic of choice and

finding the  $p$ -value of the observed statistic, or by using Monte Carlo tests. For example, Lin and Reinert (2009) showed that in ER graphs and generalisations which include Erdős-Renyi mixture graphs, the number of triangles and the number of squares are asymptotically Poisson distributed if the edge probabilities are small, and asymptotically normally distributed when the edge probabilities are moderate, with non-trivial covariance matrix.

As full-likelihood based parameter estimation is often computationally intractable for even moderate-sized networks and relatively simple evolutionary models, studies of protein network growth models have mostly been restricted to comparing the observed degree distribution to a probability model for the degree distribution. Ratmann et al. (2007) developed a novel, model-based approach for Bayesian inference on biological network data that centres on *Approximate Bayesian Computation* (ABC; see Section 1.3.6). Instead of computing the intractable likelihood of the protein network topology, their method summarizes key features of the network and then uses a MCMC algorithm to approximate the posterior distribution of the model parameters. This was used to fit a mixture model that captures network evolution by combining preferential attachment and duplication divergence with attachment, to data from *Helicobacter pylori* and *Plasmodium falciparum*.

Fitting this above model using ABC indicated that gene duplication has played a larger part in the PPI network evolution of the eukaryote than in the prokaryote.

### 1.3.6 Approximate Bayesian Computation

In standard Bayesian inference the posterior distribution for a parameter  $\theta$  is given by

$$P(\theta|D) \propto P(D|\theta)\pi(\theta).$$

Here  $\pi$  is the prior distribution for  $\theta$ ,  $D$  are the data, and  $P(D|\theta)$  is the likelihood of the data  $D$  given the parameter  $\theta$ . When simulating from sufficiently complex models and large data sets, the probability of happening upon a simulation run that yields precisely the same dataset as the one observed will be very small, often unacceptably so. This is especially true in the case of network data, where it is nearly impossible to simulate a network with exactly the same topology as the data-set. The explicit evaluation of the likelihood  $P(D|\theta)$  is avoided in *Approximate Bayesian Computation* (ABC) approaches by considering distances between observations and data simulated from a model with parameter  $\theta$ . Rather than considering the data itself, we consider a summary statistic of the data,  $S(D)$ , and use a distance  $\Delta(S(D), S(X))$  between the summary statistics of real and simulated data,  $D$  and  $X$ , respectively. The generic ABC approach to infer the posterior probability of a parameter  $\theta$  is as follows:

1. Sample a candidate parameter vector  $\theta^*$  from some proposal distribution.
2. Simulate a dataset  $X$  from the model with parameter  $\theta^*$ .
3. If  $\Delta(S(D), S(X)) < \varepsilon$  then accept  $\theta^*$  as a sample from the posterior.

For  $\varepsilon$  sufficiently small, the ABC procedure should deliver a good approximation to the true posterior, in particular if the summary statistic  $S$  is a sufficient statistic of the probability model. If sufficient statistics do not exist or are hard to obtain, setting up a satisfying and efficient ABC approach can be challenging.

The generic procedure outlined above can be computationally inefficient but ABC procedures can be combined with standard computational approaches used in Bayesian inference such as Markov chain Monte Carlo and sequential Monte Carlo. In these frameworks ABC can be used to tackle otherwise computationally intractable problems. For a review of ABC see for example Beaumont (2010).

### 1.3.7 Threshold behaviour in graphs

It is widely believed that there is a correspondence between topological motifs or subgraphs in PPI networks and biologically relevant functional modules (Bachman and Liu 2009; Hartwell et al. 1999; Spirin and Mirny 2003). Thus rigorous theoretical studies of the conditions under which certain subgraphs might arise are of great interest. For a graph  $G(n, m)$  with  $n$  nodes and  $m$  edges, many theoretical properties change dramatically in a narrow range of  $m$ , which lead to the concept of threshold functions (Erdos and Renyi 1960). Let  $G_{n, f(n)}$  be a family of random graphs induced by  $n$  number of nodes and  $f(n)$  a function that gives edges according to the specific model. If  $Q$  is a graph property,  $P(Q)$  denotes the probability that  $G_{n, f(n)}$  has or belongs to  $Q$ . We say that almost every graph in  $G_{n, f(n)}$  has the property  $Q$  if  $P(Q) \rightarrow 1$  as  $n \rightarrow \infty$ . For a given monotone increasing property  $Q$  (such as the appearance of a certain subgraph), we define a threshold function  $t(n)$  for  $Q$  as any function which satisfies  $P(Q) \rightarrow 0$  if  $\frac{f(n)}{t(n)} \rightarrow 0$  and  $P(Q) \rightarrow 1$  if  $\frac{f(n)}{t(n)} \rightarrow \infty$ .

Threshold functions for the ER model are not unique although they are so within certain factors (Bollobas 2001). For the ER model  $G(n, M(n))$ , with  $f(n) = M(n)$  it is possible to show that the threshold function for the property of containing a fixed, non-empty graph  $F$  is  $n^{2-2/m}$ , where  $m = m(F)$  is the maximum average degree of  $F$  (Bollobas 2001).

For the ER model  $G(n, M(n))$  it is possible and more informative to calculate the graph density such that the expected number of copies of a given subgraph  $F$  is approximately 1. For a subgraph on  $v$  vertices with  $e$  edges, the approximate expected count for the subgraph under the ER model  $G(n, M(n))$  with  $\rho = \rho(n) = \frac{M(n)}{\binom{n}{2}}$  is

$$E(\text{number of occurrences}) = \lambda := \binom{n}{v} \rho^e (1 - \rho)^{\binom{v}{2} - e} \sim n^v \rho^e / v!,$$

for small  $\rho$ . When the number of occurrences is well approximated by a Poisson process, as in the case for balanced graphs,  $P(\text{no occurrence of subgraph}) \sim 1 - e^{-\lambda} \sim \lambda$  and hence the threshold function and the expectation formula coincide.

Threshold functions for other models are not so well understood, but Goel et al. (2005) have shown that every monotone graph property has a threshold in geometric random graphs, generalising a similar result by Friedgut and Kalai (1996) for ER graphs. One can, nonetheless, calculate approximate threshold values for the appearance of induced graphlets with  $k$  vertices. This is based on the fact that for a random geometric graph placed in  $\mathbb{R}^d$  with  $n$  vertices and a radius  $r$ , the  $k$ -vertices subgraph count satisfies a Poisson limit when the product  $n^k r^{d(k-1)}$  tends to a finite constant (Penrose 2003). Choosing  $r$  such that  $n^k r^{d(k-1)} = 1$  then gives an approximate threshold value. To translate this value  $r$  into a graph density, we use that the radius  $r$  can be related to the expected average degree  $\alpha$  by

using the gamma function  $\Gamma(x)$  (Dall and Christensen 2002),

$$r = \frac{1}{\sqrt{\pi}} \left[ \frac{\alpha}{n} \Gamma\left(\frac{d+2}{2}\right) \right]^{1/d}. \quad (1.1)$$

Using  $r$  such that  $n^k r^{d(k-1)} = 1$  and solving for  $\alpha$  in (1.1) approximates the threshold graph density  $\rho$  as  $\rho = \alpha n / 2 \binom{n}{2}$ .

While threshold behaviour has been almost exclusively studied in theoretical models so far, in a recent paper Rito et al. (2010) show that PPI networks are situated in a region of graph density close to the threshold behaviour in ER and GEO3D models. Yeast has about 6600 protein-coding genes and is predicted to have about 25000-35000 interactions (Stumpf et al. 2008); such a network would have a graph density between 0.0011 and 0.0016. For human, estimates of about 25000 genes (Human Genome Project) and 650000 interactions (Stumpf et al. 2008) would also lead to graph densities around 0.002. Both these networks would be placed in the threshold region for the appearance of specific types of motifs under the ER model. As the authors subsequently show, GDDA (see Section 1.4.1) which is a widely used measure of fit between interaction datasets and theoretical models, is unstable in this very region. It is conjectured that this instability in model fitting could be a consequence of the threshold for specific motifs appearing in the networks.

## 1.4 Comparison of protein interaction networks

So far we have discussed analysis techniques for single networks. The availability of interaction data for multiple species also opens up the opportunity for comparative techniques. Current research in the comparison of networks follows two generally separate streams: (1) Comparing experimental networks to theoretical models in order to assess the fit, and, (2) comparison of experimental networks across multiple species to identify conservation at systems level. Here we concentrate on using subgraph counts for the first problem, and network alignment for the second problem. Indeed the second problem itself is also referred to as *network alignment* as it is essentially a graph-matching problem. While there has been much work done and many algorithms proposed recently for network alignment, some of which we discuss later in this section, not many measures exist to measure the agreement between experimental data and theoretical network models.

### 1.4.1 Network comparison based on subgraph counts

One possible way to compare empirical and model generated networks is by quantifying their similarity in terms of abundance of specific classes of subgraphs. GraphCrunch (Milenkovic et al. 2008) is an open source software tool that compares large real-world networks with random graph models. These are automatically generated to have the same number of nodes and edges (to within 1%) as those of the real-world network being compared. (This is approximate; with a 12-star as input, GraphCrunch generates ER-DD graphs with 10, 11 and 12 edges.) As well as many global standard properties, the software supports the local statistics *RGF-distance* (Przulj et al. 2004) and *GDDA* (Przulj 2007). RGF-distance compares the frequencies of the appearance of all 3 to 5-node subgraphs in two networks. The networks being compared by GraphCrunch always have the same number of nodes as well as edges,

and thus the frequencies of occurrence of the only 1-node subgraph, a node, and the only 2-node subgraph, an edge, are also taken into account by this measure. GDDA uses orbit degree distributions, which are based on the automorphism orbits of the 29 subgraphs on 2 to 5 vertices, as follows. Automorphisms are edge-preserving bijections from a graph to itself, and together they form a permutation group. An automorphism orbit is a node that represents this group (Figure 1.5). Within the 29 subgraphs, 73 different orbits can be found and each one will have an associated orbit degree distribution.

**Figure 1.5** Some subgraphs and their automorphism orbits (Rito et al. 2010).

An orbit  $i$  from subgraph  $G_j$  has orbit degree  $k$  in the graph  $G$  if there are  $k$  copies of  $G_j$  in  $G$  which involve orbit  $i$ . Let  $d_G^j(k)$  be the sample distribution of the node counts for a given orbit degree  $k$  in a graph  $G$  and for a particular automorphism orbit  $j$ . This sample distribution is then scaled by  $1/k$  in order that large degrees do not dominate the score, and normalise  $d$  to give a total sum of 1,

$$N_G^j(k) = \frac{d_G^j(k)/k}{\sum_{l=1}^{\infty} d_G^j(l)/l}.$$

The comparison  $D^j(G, H)$  of two graphs  $G$  and  $H$  with respect to orbit  $j$  is simply the Euclidean distance between the two scaled and normalise  $d$  vectors  $N$ , which is scaled by  $1/\sqrt{2}$  to be between 0 and 1, as pointed out in Przulj (2010); the resulting expression is

$$D^j(G, H) = \frac{1}{\sqrt{2}} \left( \sum_{k=1}^{\infty} [N_G^j(k) - N_H^j(k)]^2 \right)^{1/2}.$$

This is then turned into an agreement by subtracting from 1, and the agreements are combined into a single value by taking the arithmetic mean over all  $j$ , yielding the GDDA,

$$GDDA = \frac{1}{73} \sum_{j=0}^{72} (1 - D^j(G, H)).$$

According to Przulj (2007), a perfect score can be achieved when comparing networks of the same random model type; for example generate a network from an ER model with given number of edges and nodes, and compare it to other randomly generated ER models with the same parameters. Przulj (2007) found the mean GDDA of comparing ER versus ER, ER-DD versus ER-DD or GEO3D versus GEO3D to be  $0.84 \pm 0.07$ , where 0.07 denotes one standard error. This was updated in Przulj (2010) where they found the highest score for two GEO3D networks to be  $0.95 \pm 0.002$ .

However, Rito et al. (2010) found that GDDA values have not only striking differences amongst different model types but also a pronounced dependency on the number of vertices of the network. For a specific graph, drawn from one model type and with a fixed number

of vertices, they also observed a strong dependency of the GDDA score with graph density when comparing to graphs of the same type and with the same number of vertices (Figure 1.6). Furthermore, these dependencies are not monotone. They propose a new protocol where several same model versus model comparisons with roughly the same number of vertices and edges should be carried out in order to assess the best obtainable score for this specific case. GDDA should then be calculated between the query network and graphs from the model network. Model fit can be evaluated by gauging the differences between the distributions of agreement scores resulting from query network versus model and model versus model comparisons, for example by using a Monte Carlo test.

**Figure 1.6** Dependency of GDDA for model versus model comparisons on the number of vertices and edges of a network. GDDA of ER versus ER (A) and GEO3D versus GEO3D (B) graphs with 500, 1000 and 2000 vertices are plotted against graph density (Rito et al. 2010). The arrows indicate thresholds for the appearance of the subgraphs  $G_1$  up to  $G_{29}$  for ER, and for the appearance of triangles in GEO3D.

#### 1.4.2 Network alignment

Research in cross-species network comparison, or network alignment has been spurred on by the introduction of the *interolog* concept. An interolog is a conserved interaction between a pair of proteins which have interacting orthologs in another organism, where orthologs are proteins descended from a common ancestor. The evidence for the existence of such protein interactions that are conserved across species is increasing. Proteins in the same pathway have been found to be present or absent in a genome as a group (Kelley et al. 2003; Pellegrini et al. 1999), and many protein interactions in the yeast network have also been identified for the corresponding protein orthologs in *C. elegans* (worm), see Matthews et al. (2001). These discoveries have led to research directed at identifying conserved complexes and functional modules through network alignment, analogous to traditional sequence alignment (Dandekar et al. 1999; Kelley et al. 2004; Ogata et al. 2000). Given two or more networks the aim of network alignment algorithms is to identify sets of interactions that are conserved across the networks. This alignment is achieved by first identifying a mapping between the nodes of two or more networks based on some biological information, usually sequence similarity. This step is followed by the actual alignment process, incorporating concepts from graph matching where the goal is to maximise the overlap in the interaction patterns of mappable nodes (Figure 1.7). The premise is that patterns of interactions which are conserved across species have biological significance and hence are more likely to correspond to real protein complexes or functional modules. The large and ever-increasing size of interaction datasets (typically >5000 nodes and >25000 edges) combined with the fact that graph matching is an NP-hard problem, makes network alignment a computationally challenging problem.

One of the earliest network alignment algorithms, NetworkBlast (Sharan and Ideker 2006) carries out alignment by first defining an alignment graph where each node represents a set of orthologous proteins based on sequence similarity. The edges in the alignment

**Figure 1.7** Pair-wise network alignment of two graphs  $G$  and  $G'$ . Dotted red lines indicate homology.

graph represent conserved interactions. A search is then usually carried out over this alignment graph for high scoring subgraphs. NetworkBlast has been used to perform three-way comparisons of yeast, worm and fly which yielded conserved modules displaying good overlap with MIPS (Mewes et al. 2004) complexes.

Graemlin (Flannick et al. 2006) uses progressive pair-wise alignments to compare multiple networks. Graemlin's probabilistic formulation of the topology-matching problem eliminates restrictions on the possible architecture of conserved modules such as those imposed by NetworkBlast. However it requires parameter learning through a training set of known alignments. The sensitivity of the method was assessed by counting the number of KEGG (Kanehisa and Goto 2000) pathways in the alignments. The KEGG coverage of the alignment results was between 21 and 39%. In terms of speed, it far outperforms NetworkBlast with a running time approximately linear to the number of networks.

Other alignment algorithms have tried to take into account the evolutionary forces shaping the interaction networks. For example, MaWISH (Koyuturk et al. 2006), which implements a duplication divergence model to carry out pair-wise network alignment. More recently an evolutionary based multiple network alignment algorithm CAPPI (Dutkowski and Tiuryn 2007) was developed which tries to reconstruct the ancestral network for the input species and maps it back onto the extant networks to identify common modules. Graemlin 2.0 (Flannick et al. 2008) is also a multiple network aligner, with a scoring function that can use evolutionary events.

Some recent network alignment methods proposed recently include IsoRank (Singh et al. 2008) and IsoRankN (Liao et al. 2009), GNA and PATH (Zaslavskiy et al. 2009) and DOMAIN (Guo and Hartemink 2009). DOMAIN is the first algorithm to introduce protein domains into the network alignment problem and uses a novel direct-edge-alignment paradigm to directly detect equivalent interaction pairs across species. It should be noted that global network alignment methods such as IsoRank and GNA do not directly address the conserved module detection problem, and focus on finding the best node-to-node match across the entire networks.

### *1.4.3 Using functional annotation for network alignment*

A common theme of the studies described above for protein interaction network alignment is the use of protein sequence similarity to map orthologous proteins across different species. However, this does not necessarily provide a complete picture of orthologous relationships in the context of interaction networks. When aligning networks from species that are very distant in evolutionary terms, the proteins may not display enough sequence similarity to achieve a reasonable degree of mapping. This would result in a severely restricted alignment graph that may miss biologically conserved regions in the networks. Ali and Deane (2009) explored the possibility of using a different measure of protein similarity. Since the goal of alignment is to extract modules that correspond to specific biological processes, functional similarity of proteins across networks was employed to aid alignment.

To be useful for network alignment, a subjective concept like functional similarity must be expressed in a quantitative form that reflects the closeness in the biological functions of the proteins being compared. Functional annotation of proteins is an ongoing scientific activity and one of the most widely used resources is Gene Ontology or GO (Ashburner et al. 2000). GO offers substantial coverage of major protein databases and provides a species-specific, structured set of terms describing gene products. A simple measure of functional similarity was used which is based on the most specific and hence most informative GO annotation of each protein. For simplicity they focussed only on the Biological Process category of GO, the method being identical for the other top categories of Molecular Function and Cellular Component. Let there be a total of  $N$  proteins in the dataset under consideration and the GO functional annotation of each protein be defined as a set of terms  $S_A$ . Define a multi-set of size  $n$  as a pair  $(S, \sigma)$  where  $\sigma : S \rightarrow \mathbb{N}$ , with the conditions:

$$S = \bigcup_{A \in N} S_A \quad , \quad \sum_{y \in S} \sigma(y) = n.$$

Here,  $\sigma$  is a function that maps a GO term to the number of times it occurs in the dataset. Terms having fewer proteins annotated to them occur less frequently in the dataset and are thus classified as more specific. For any two proteins  $A$  and  $B$  with annotation sets  $S_A$  and  $S_B$ , the functional similarity score (*funsim*) was then calculated as follows:

$$funsim(A, B) = \max_{t \in S_A \cap S_B} \left( 1 - \frac{\sigma(t)}{n} \right).$$

The above scoring scheme assigns higher functional similarity to protein pairs that share more specific GO annotations. It should be noted that other, more sophisticated scoring schemes for functional similarity based on GO are possible. Several measures of functional similarity have been proposed in recent years making use of the information content of GO terms as well as the semantics (is a, part of) of the GO relationships (Resnik 1995; Schlicker et al. 2006).

In the study by Ali and Deane (2009), function-based alignment using the above similarity score was successful in uncovering a larger number of proteins participating in conserved interactions than sequence-based alignment. The human network used for the study contained 9305 proteins and 35458 interactions while the yeast network contained 4941 proteins and 17387 interactions. As shown in Table 1.2, the number of conserved interactions discovered in the human network (aligned to yeast) increased from 612 (between 457 unique proteins) to 1034 (between 727 unique proteins). Moreover, the two sets share only 58 proteins (<15%), indicating that the interactions targeted by the two methods are nearly disjoint. Sets of conserved interactions detected using function based alignment also displayed higher overlap with experimentally detected complexes in the MIPS database.

As can be seen, despite the development of increasingly sophisticated network alignment methods, conserved interactions detected between model organisms constitute only a tiny fraction of existing data-sets. Taking into account that some of the detected conserved interactions will be due purely to chance, such low evolutionary conservation at the interactome level is quite surprising and warrants more attention.

**Table 1.2** Comparison of sequence and function based alignment of the yeast and human PPI networks.

Alignment Method	No. Interactions	No. Proteins	MIPS Coverage	MIPS Accuracy	Functional Coherence
Seq. based	612	457	96	0.18	0.36
Func. based	1034	727	126	0.24	0.51
MaWISH	596	543	83	0.1	0.32

Results are also compared to sequence based alignment using the MaWISH algorithm.

## 1.5 Evolution and the protein interaction network

Related intimately to the problem of cross-species network alignment is the study of likely factors underlying network evolution. Many models have been proposed for the growth of the protein interaction network, some of which (namely duplication-divergence and geometric evolution) are discussed in Section 1.3.4. In terms of underlying mechanisms, evolution at the network level is thought to be a consequence of protein evolution. Errors in replication can result in a change in copy number of proteins, from individual genes being duplicated or lost (Zhang June 2003), to the whole genome being duplicated (Kasahara 2007; Scannell et al. 2007). After a gene duplication event, divergence of function is possible. There are two main competing models for such divergence: sub-functionalisation (partitioning of ancestral function between gene duplicates) and neo-functionalisation (the de novo acquisition of function by one duplicate). Whichever model is chosen, this functional divergence at protein level can manifest itself in the form of diverging interaction patterns at the network level. While evolution in PPI networks opens up several research questions including the co-evolution of interacting proteins and constraints imposed by interactions on protein evolution (Lewis et al. 2010b), here we focus on one issue: whether network divergence after speciation can explain the low conservation observed in species-level network alignments.

### 1.5.1 How evolutionary models affect network alignment

Alignments of protein interaction networks have found little conservation among several species (Ali and Deane 2009). This is in sharp contrast to the genome level sequence alignments, where even evolutionarily distant organisms share significant portions of their gene repertoire. While this could be a consequence of the incompleteness of interaction data-sets and presence of error, an intriguing prospect is that the process of network evolution is sufficient to erase any evidence of conservation. Ali and Deane (2010) tested this hypothesis using models of network evolution and also investigated the role of error in the results of network alignment. Using the duplication divergence and geometric evolution models, pairs of networks were grown from a single ancestor to the size of current interaction data-sets for model species. Pair-wise network alignment was then carried out using several different network alignment algorithms, and a distance metric based on summary statistics was used to assess the fit between experimental and simulated network alignments (Figure 1.8). Results indicated that network evolution alone is unlikely to account for the poor quality alignments given by real data. Alignments of simulated networks undergoing evolution are considerably

**Figure 1.8** Measuring the effect of error and evolution on network alignment (Ali and Deane 2010).

(4 to 5 times) larger than real alignments. The authors also compared several error models in their ability to explain this discrepancy. For a given error model with a single rate parameter  $\theta$ , they estimate the posterior density for  $\theta$  using the following ABC algorithm:

1. Draw  $(\theta^1, \theta^2) \sim Uniform[0, 1]^2$ .
2. Simulate error in networks  $(NW_1, NW_2)$  using error models  $\mathcal{M}(\theta^1), \mathcal{M}(\theta^2)$ .
3. Align  $(NW_1, NW_2)$  and compute summary vector  $\mathcal{S}$  from alignment.
4. Calculate the distance  $d(\mathcal{S}, \mathcal{D})$  where  $\mathcal{D}$  is the summary vector for the real alignment.
5. Accept  $\theta$  if  $d(\mathcal{S}, \mathcal{D}) \leq \delta$ .

In the algorithm  $d$  is a scaled Euclidean distance between two summary vectors defined as

$$d(\vec{x}, \vec{y}) = \sqrt{\sum_{i=1}^N \frac{(x_i - y_i)^2}{\sigma_i^2}}.$$

Using this setup, posterior estimates of false negative rates for the yeast and human protein networks vary from 20 to 60% dependent on whether current incomplete proteome sampling is taken into account or not (Figure 1.9). It was also found that false positives appear to affect network alignments little compared to false negatives indicating that incompleteness, not spurious links, is the major challenge for interactome-level comparisons.

**Figure 1.9** Posterior density for error parameters in *H.sapiens* (HN) and *S.cerevisiae* (YN) PPI networks (Ali and Deane 2010).

## 1.6 Community detection in PPI networks

In parallel with network comparisons and model fitting, another area with obvious potential benefits is the automatic detection of functional modules and complexes in PPI data. Discovering these structures can not only help us better understand the hierarchical organisation of cellular systems but may also assist in better targeting/design of drugs. Thus a very active area of research in the computational analysis of PPI networks is related to graph clustering approaches.

Within the broader networks literature, a great many algorithms have been proposed that locate dense regions in a network, called communities or clusters (reviewed in Fortunato 2010; Porter et al. 2009). A *community* is loosely defined as a group of nodes that are more closely associated with themselves than with the rest of the network. In the context

of biological networks, such communities are potentially good candidates for functional modules, and many studies report running one of the myriad algorithms for detecting community structure on PPI networks (Bu et al. 2003; Dunn et al. 2005; Li et al. 2008; Luo et al. 2007; Pereira-Leal et al. 2004). Having located communities, such studies then attempt to assess their functional homogeneity by searching for terms in a structured vocabulary—usually GO or MIPS—that are significantly over-represented within communities. If such terms exist, the identified communities are said to be ‘enriched’ for biological function. In many studies such enriched communities are found, and hence are plausible candidates for biological modules.

### 1.6.1 Community detection methods

A much used approach is the algorithm by Newman and Girvan (2004). It involves simply calculating the betweenness of all edges in the network and removing the one with highest betweenness, and repeating this process until no edges remain. If two or more edges tie for highest betweenness then one can either choose one at random to remove, or simultaneously remove all of them. As a guide to how many communities a network should be split into, they use the *modularity*. For a division with  $g$  groups, define a  $g \times g$  matrix  $e$  whose component  $e_{ij}$  is the fraction of edges in the original network that connect nodes in group  $i$  to those in group  $j$ . Then the modularity is defined to be

$$Q = \sum_i e_{i,i} - \sum_{i,j,k} e_{i,j} e_{k,i},$$

that is, the fraction of all edges that lie within communities minus the expected value of the same quantity in a graph where the nodes have the same degrees but edges are placed at random. A value of  $Q = 0$  indicates that the community is no stronger than would be expected by random shuffling.

The Potts method (Reichardt and Bornholdt 2006) partitions the proteins into communities at many different values of a resolution parameter, thus finding communities at different scales within the network. The method seeks a partition of nodes into communities that minimises a quality function (‘energy’):

$$H = - \sum_{ij} J_{ij}(\lambda) \delta(s_i, s_j),$$

where  $s_i$  is the community of node  $i$ ,  $\delta$  is the Kronecker delta,  $\lambda$  is the resolution parameter, and the interaction matrix  $J_{ij}(\lambda)$  gives an indication of how much more connected two nodes are than one would expect at random (i.e., in comparison to some null hypothesis). The energy  $H$  is thus given by a sum of elements of  $J$  for which the two nodes are in the same community.

Other methods include the Markov clustering algorithm or MCL (van Dongen 2000) which simulates random walks on networks to isolate dense regions and MCODE (Bader and Hogue 2003) which uses high density  $k$ -cores to search for potential complexes.

Here we have only mentioned community detection techniques based purely upon network structure, though there are several studies in literature where network data is supplemented by additional biological information such as gene expression (Segal et al. 2003) and phenotypic sensitivity (Tanay et al. 2004) to achieve potentially more meaningful graph partitions.

### 1.6.2 Evaluation of results

Community detection algorithms are generally evaluated in terms of some quality function defined over the output clusters. A commonly used measure of 'goodness' is functional homogeneity, measuring how similar the proteins in a cluster are in terms of their biological function. Given a measure of functional similarity between pairs of proteins, one way to express the homogeneity of a cluster is

$$H(C) = \frac{\sum_{i,j \in C} \text{Similarity}(i, j)}{|C|(|C| - 1)};$$

here,  $H(C)$  represents the homogeneity of a cluster  $C$  by the average pairwise protein similarity within  $C$ . Modules can also be statistically evaluated using the  $p$ -value from the hypergeometric distribution, which is defined as

$$p = 1 - \sum_{i=0}^{k-1} \frac{\binom{|X|}{i} \binom{|N|-|X|}{n-i}}{\binom{|N|}{n}},$$

where  $|N|$  is the total number of proteins,  $|X|$  is the number of proteins in a reference function,  $n$  is the number of proteins in an identified module, and  $k$  is the number of proteins in common between the function and the module. This is the probability that at least  $k$  proteins in a module of size  $n$  are included in a reference function of size  $|X|$  assuming that all proteins are investigated independently and have the same probability to be included in  $|X|$ . A low  $p$ -value indicates evidence for the hypothesis that the module corresponds to the function.

Lewis et al. (2010a) used the Potts method to study the biological relevance of, and the relationship between, communities detected in the yeast protein network at different resolutions. They found that the large communities present at small values of the resolution parameter  $\lambda$  are not judged to be functionally homogeneous. As  $\lambda$  is increased, larger numbers of proteins occur in functionally homogeneous communities, peaking in the range  $1.5 < \log(\lambda) < 2$ . At  $\log(\lambda) = 1.5$ , the mean community size is 73 proteins, and the majority of proteins (3071 out of 4980) are in functionally homogeneous communities (Figure 1.10).

**Figure 1.10** Communities identified in the yeast protein interaction network (Lewis et al. 2010b). When the resolution parameter  $\lambda$  is very small, all nodes are assigned to the same community. As  $\lambda$  is increased (viewing the network at progressively closer distances), more structure is revealed. The figures on the left hand side show visualisations of the networks' partition into communities at four different values of  $\lambda$ . Each circle represents a community, with size proportional to the number of proteins in that community, positioned at the mean position of its constituent nodes. The shade of the connecting lines is proportional to the number of links between two communities.

## 1.7 Predicting function using PPI networks

Given the relative sparsity of protein functional annotations for even well-studied organisms, predicting protein function using PPI networks has received a lot of attention in recent years. Protein functions may be predicted on the basis of module detection algorithms. If an unknown protein is included in a functional module, it is expected to contribute toward the function that the module represents. Several topology-based approaches that predict protein function on the basis of PPI networks have also been introduced. The simplest method (majority vote) assigns the function of an unknown protein as the function of the majority of its neighbours (Schwikowski et al. 2000). The assumption that the function of a protein can be assumed to be independent of all other proteins, given the functions of its immediate neighbours, has led to several methods using Markov random fields for function prediction (Deng et al. 2003; Letovsky and Kasif 2003). The number of common neighbours of the known protein and of the unknown protein has also been taken as the basis for the prediction of function (Lin et al. 2006). Here we describe the method of Chen et al. (2007) to consider how such prediction of characteristics can be achieved using network structure. As characteristics they considered structure (7 categories) and function (24 categories).

From the PPI network, they built an upcast set of category-category interactions. A category-category interaction is constructed by two characteristic categories from two interacting proteins. Denote the set of all categories the characteristic can assume by  $S$ . For a protein  $x$ ,  $S(x)$  is the set of categories that protein  $x$  is classified into. If two proteins  $x$  and  $y$  interact, the category-category interactions it generates are the edges between any two characteristic categories,  $a$  and  $b$  ( $a \in S(x)$ ,  $b \in S(y)$ ), from each of the two proteins (denoted by  $a \sim b$ ). The upcast set of category-category interactions is a collection of all category-category interactions extracted from the protein-protein interaction network (Figure 1.11).

**Figure 1.11** Upcast sets of characteristic pairs. In this example, we consider only a single characteristic (e.g., protein function), so that the characteristic vector for a protein is a 1-vector. There are three single-category proteins and one two-category protein in the protein interaction network (left), which result in an upcast set of six characteristic pairs (Chen et al. 2007).

Let  $f(a \sim b)$  denote the relative frequency of category-category interaction  $\{a \sim b\}$  among all category-category interactions. The score  $F(a, x)$  for the query protein  $x$  with annotated neighbours  $B(x)$ , to be in a specific category  $a$  is proportional to the product  $C(a, x)$  of the relative frequencies  $f$  of observing category  $a$  for all category-category interactions of  $x$ 's neighbours in the prior data base;

$$C(a, x) = \prod_{\substack{b \in S(n) \\ n \in B(x)}} f(a \sim b),$$

and is defined by

$$F(a, x) := \frac{C(a, x)}{\sum_{k \in S} C(k, x)}.$$

The protein  $x$  is then predicted to possess the characteristic category, or categories, with the highest score. This *frequency method* can be extended to include two or more protein characteristics in the prediction of a specific protein characteristic; it is then called the *enhanced frequency method*.

**Table 1.3** The accuracies of function prediction using Majority Vote (M.V), Chen’s method (F.) and Chen’s enhanced method (E.F.). The enhanced method combines structure and function into a category vector for category-category interactions. The accuracy is calculated as the ratio between the number of correctly predicted proteins and all predicted proteins.

Organism (DIP)	Predicted proteins	M.V.	F.	E. F.
D.melanogaster	1275	0.53	0.67	<u>0.69</u>
C.elegans	85	0.38	0.55	<u>0.71</u>
S.cerevisiae	1618	0.67	0.61	0.67
E.coli	154	0.69	0.69	0.70
M.musculus	32	0.59	0.88	<u>0.81</u>
H.sapiens	274	0.79	0.90	<u>0.89</u>

<sup>a</sup> Predicting methods are Majority Vote (M.V.), Chen’s method (F.), Chen’s enhanced method (E.F.)

<sup>c</sup> A function prediction is counted as correct if one of the best three predicted categories is correct.

Underline: where the result outperforms M.V. with statistical significance, based on a  $z$ -test.

One can use the above model to see how well one network would predict characteristics in another network. The model was implemented to use pattern frequencies from pooled interactions in other organisms as well. Protein interactions were grouped into prokaryotes including *E.coli* and *H.pylori* and eukaryotes including *C.elegans*, *S.cerevisiae*, *D.melanogaster*, *M.musculus*, *H.sapiens*, and a final global pooled dataset including all interactions. As shown in Figure 1.12, in the prediction of structure, higher accuracy was gained when predicting eukaryotes using a prior database from eukaryotes only. On the other hand, the use of pooled interactions does not significantly improve predictions. While predicting function, the predictions significantly deteriorated when using prokaryotes as prior data. The prediction results thus clearly differentiated between using prior databases of eukaryotes and prokaryotes.

**Figure 1.12** Structure prediction using different priors (Chen et al. 2007).

These results indicate that the type of prior data may be more important than the data quantity and that there might be network similarity within kingdoms not observed across kingdoms. We conclude that while there are network similarities within kingdoms, interaction networks in prokaryotes and eukaryotes may be different.

## 1.8 Predicting interactions using PPI networks

As discussed earlier, current PPI networks are highly incomplete even for model organisms. Existing PPI networks from experimental data-sets can be useful resources on which to base the prediction of new interactions or the identification of reliable interactions. Deng et al. (2002) used evolutionarily conserved domains defined in the Pfam database (Finn et al. 2010) and applied a maximum likelihood estimation method to infer interacting domains that are consistent with observed protein interactions. They estimated the probabilities of interactions between every pair of domains and measured the accuracies of their predictions at the protein level. Using the inferred domain-domain interactions, they predict interactions between proteins. Liu et al. (2005) extended this approach by integrating large-scale PPI data from three organisms to estimate the probabilities of domain-domain interactions. They found that the integrated analysis provides more reliable inference of protein interactions than the analysis from a single organism. Jonsson et al. (2006) predicted interactions by integrating experimental PPI data from many species and translating it into the reference frame of the rat. The putative rat protein interactions were given confidence scores based on their homology to proteins that were experimentally observed to interact.

In continuation to the previous section, we describe a method by Chen et al. (2008) who used ideas from logistic regression to develop a score to predict and to validate protein interactions, based on the protein characteristics and the PPI network.

### 1.8.1 Tendency to form triangles

The score is based on the following observation from exploratory data analysis. Let  $a, b, c \in S$  be three vectors, and let  $N$  be the set of proteins in the protein interaction network. Assume that all of  $a, b$  and  $c$  are indeed observed in the proteins, so that using the notation from Section 1.7,

$$\sum_{x,y,z \in N} \mathbf{1}(a \in S(x), b \in S(y), z \in S(z)) > 0,$$

and

$$\sum_{x,y \in N} \mathbf{1}(a \in S(x), b \in S(y)) > 0.$$

Here  $\mathbf{1}(A)$  denotes the indicator function; it takes on the value 1 if  $A$  is satisfied, and is 0 otherwise.

For each type of category-category pair  $\{a, c\}$  with a fixed category  $b$ , the ratio of probabilities  $r_{abc} = \frac{P(a \sim c | a \sim b \sim c)}{P(a \sim c)}$  is estimated by

$$\hat{r}_{abc} = \frac{\hat{P}(a \sim c | a \sim b \sim c)}{\hat{P}(a \sim c)} = \frac{\hat{P}(a \sim c, a \sim b \sim c)}{\hat{P}(a \sim b \sim c \sim a) + \hat{P}(a \sim b \sim c \not\sim a)},$$

where  $\hat{P}(a \sim c)$  is the proportion of pairs of proteins  $x, y$  in  $N$ , with characteristics such that  $a \in S(x), b \in S(y)$ , which interact, relative to all pairs of proteins with such characteristics. Similarly,  $\hat{P}(a \sim b \sim c \sim a)$  is the proportion of protein triplets, with given characteristics, which form a triangle, and  $\hat{P}(a \sim b \sim c \not\sim a)$  is the proportion of protein triplets, with given characteristics, which form a line (but not a triangle).

For each organism (protein interaction network),  $\bar{r}$  is the average of  $\hat{r}_{abc}$  for all  $a, b, c \in S$ ,

$$\bar{r} = \frac{\sum_{a,b,c \in S} \hat{r}_{abc}}{\frac{1}{2}|S|^2(|S| + 1)}.$$

If  $\bar{r} \ll 1$ , the existence of the interacting partner tends to decrease the chance of interaction. If  $\bar{r} \gg 1$ , the interaction is more likely if two protein have an common interacting partner. The average ratios of conditional probabilities from different organisms are estimated in Table 1.4.

**Table 1.4** Estimates of  $\bar{r}$  from triplets: function

Organisms	#obs. pairs <sup>†</sup>	#obs. triples <sup>‡</sup>	$\bar{r}$	S.E.	$\bar{r} > 1$ <sup>§</sup>
D.Melanogaster	110	534	48.3	67.6	
S.Cerevisiae	214	1850	55.9	91.36	
E.Coli	76	494	16.1	9.91	*
H.Sapiens	60	350	76.8	125.25	

<sup>†</sup> number of different pairs  $\{a, c\}$  forming triples  $\{a \sim b \sim c\}$

<sup>‡</sup> total number of different triples  $\{a \sim b \sim c\}$

<sup>§</sup> 5% level of significance based on Z-tests

\* organism showing tendency of formation of triangles

From this table we conclude that there is some evidence, albeit weak, for a tendency to form triangles in the upcast networks. This tendency can be exploited for prediction, as follows.

### 1.8.2 Using triangles for predicting interactions

Within the triplet interactions, the odds are assessed to observe triangles versus lines around the query protein pair. The *triangle rate score*,  $tri(x, y)$  for the protein pair  $\{x, y\}$  is defined as the odds of observing triangles versus lines among triangles and lines in its neighbourhood.

This scoring method was compared to the Deng, Liu and Jonsson scores using ROC curves (Figure 1.13) and, using a subsampling scheme, the areas under the ROC curves were tested for significant difference. The results show that the triangle rate score outperforms both the domain-based and homology-based scores. The success of this method provides a good argument for representing PPI data as networks as the triangle information is crucial.

**Figure 1.13** ROC curves, 1 minus specificity vs. sensitivity, for predicting yeast protein interactions using domain interaction based approaches (Deng's score and Liu's score), a homology-based approach (Jonsson's score plus paralogs) and Chen's network-based approach (the triangle rate score)

In the above two sections, we have presented in detail two related methods for predicting protein characteristics and protein interactions based on network data. It should be noted that there is a vast number of similar as well as very different computational prediction techniques reported in the literature (see Browne et al. 2010; Rentzsch and Orengo 2009, for recent reviews).

## 1.9 Current trends and future directions

We conclude this chapter with a brief overview of current research trends in the analysis of protein networks and some areas which we believe will generate significant activity in the near future.

### 1.9.1 Dynamics

Studies of large-scale biological networks are gradually shifting from the analysis of their organisational principles and guilt-by-association predictions of the function of individual network components towards examining cell dynamics. In such studies, experimentally determined static networks are often used as scaffolds for modeling of dynamical changes in the system. Information about dynamics can be provided, for example, by measurements of gene expression at different time points or in different conditions. Han et al. (2004) examined the extent to which hubs in the yeast interactome are co-expressed with their interaction partners. They defined hubs as proteins with degree at least 5. Based on the averaged Pearson correlation coefficient (avPCC) of expression over all partners, they concluded that hubs fall into two distinct classes: those with a low avPCC (which they called date hubs) and those with a high avPCC (so-called party hubs). They inferred that these two types of hubs play different roles in the modular organisation of the network: Party hubs are thought to coordinate single functions performed by a group of proteins that are all expressed at the same time, whereas date hubs are described as higher-level connectors between groups that perform varying functions and are active at different times or under different conditions. The validity of the date/party hub distinction has since been debated in some recent papers (Batada et al. 2006, 2007; Wilkins and Kummerfeld 2008). Agarwal et al. (2010) used an interaction data set from the Online Predicted Human Interaction Database - OPHID (Brown and Jurisica 2005) and found that the form of the distribution of hub avPCC does not support bi-modality and is not robust to methodological changes (Figure 1.14).

**Figure 1.14** Probability density plots of the distribution of hub avPCC values for human interaction data from OPHID. Gene expression data from GeneAtlas (Su et al. 2004), normalised using (a) MAS5 (Hubbell et al. 2002) and (b) GCRMA (Wu et al. n.d.). From Agarwal et al. (2010).

Expression data can also be utilised to infer causality as well as information flow within cellular networks. A particularly illuminating source of dynamic data comes from knock-out experiments, where a gene is perturbed or removed from a genetic background and the expression levels of all other genes are measured. Yeang et al. (2004) developed a

probabilistic approach for explaining observed gene expression changes due to a knock-out by inferring molecular cascades of flow through the interaction network. These molecular cascades correspond to paths beginning from the knock-out gene and ending at the gene whose expression has changed. RNA interference (RNAi) screens are a powerful technique for obtaining perturbation data in higher organisms. Here, a known gene from a pathway of interest is chosen as a reporter gene, other genes in the genome are systematically knocked-down using RNAi, and the effect on the reporter is measured. As RNA-based loss of function screens are increasingly being applied with automated image analysis to detect effects on specific processes or phenotypes (Jones et al. 2009; Moffat et al. 2006), these types of network analyses are likely to be relevant to study a broad range of interesting biological questions.

### *1.9.2 Integration with other networks*

Until recently, the large-scale modelling of network dynamics has been focused on individual network types. However, within a cell, all network types are interrelated and dynamics of any individual network has an impact on the behaviour of other networks. Several recent studies have begun to address the challenge of coupling large-scale dynamical models for different network types to obtain one consistent dynamical network. Such methods have been spearheaded by approaches to combine metabolic and regulatory networks. For example, to obtain a combined model of metabolic and regulatory networks, Covert et al. (2001) used flux-balance analysis to model the metabolic network component while the transcriptional regulatory network was modelled as a Boolean network. The genes in the transcriptional regulatory network were assigned Boolean (binary) values indicating whether or not a given gene is being expressed. An interactive procedure was applied to ensure that the combined model satisfies both the metabolic and the regulatory constraints. A subsequent study used mixed integer linear programming (a general optimisation framework for capturing problems with both discrete and continuous variables) to couple such metabolic and regulatory models (Shlomi et al. 2007).

Wang and Chen (2010) propose an approach for integrating transcription regulation and protein-protein interactions using dynamic gene-expression data. They start with candidate gene regulatory and signaling networks obtained from genome-scale data. These candidate networks are then pruned and combined, utilizing gene-expression data at multiple time points, to obtain an integrated and focused network under a specific condition of interest.

This area is in its infancy; integrated networks may shed new lights on all the research questions tackled above.

### *1.9.3 Limitations of models, prediction and alignment methods*

While there is an expanding literature on methods for predicting protein characteristics and interactions using network information, there are several shortcomings which could hamper the practical application of such techniques. Many methods for protein function prediction using network neighbours fail to cope with the relatively sparse level of annotation currently available (Freschi 2009). Even given sufficient annotations, the network structure itself can be quite heterogeneous and too few links in some regions can make prediction either impossible or highly unreliable. For instance, performance of methods such as (Chen et al. 2008) which

use a triangle score for interaction prediction, can be negatively affected in networks or network regions with low density.

Similarly, network alignment methods are now sophisticated enough to cope with the computationally hard problem of matching multiple large networks, but such algorithms tend to be purely graph-theory based. Unlike sequence alignment, where rigorous models based on evolutionary theory are present to explain the results, network alignment is very often based on heuristic arguments. This is perhaps partly a consequence of the fact that evolution at the network level is poorly understood at the moment and we have no universally applicable models for PPI networks. Future research should be able to shed more light on the mechanisms shaping interaction networks through evolutionary time-scales and make it possible to incorporate these principles in comparative and predictive studies. A related research question could be the classification of existing as well as more refined network models according to the type of problem they best tackle.

#### *1.9.4 Biases, error and weighting*

Apart from techniques employing diverse biological information, some serious issues related to the nature of interaction data itself also need to be resolved. Current experimental techniques, such as yeast two-hybrid and coaffinity purification, sample subsets of the interaction data space (Stumpf et al. 2005). As mentioned earlier, these subsets show very limited overlap. Moreover, interaction data are non-binary by nature for any multi-component complex; their conversion to binary pair-interactions is nontrivial and relies on processing protocols that may introduce further biases in the final screening output (Wodak et al. 2009; Yu et al. 2009). It is vital that we understand to what extent observed discrepancies between different networks reflect sampling biases of their experimental methods, as opposed to topological features due to biological functionality. While some studies have focused on the possible effect of biases, uncertainty and incompleteness on network inference and comparison (Fernandes et al. 2010; Stumpf and Wiuf 2005), we feel there is great need for directly modelling in these factors as part of predictive studies.

#### *1.9.5 New experimental sources of PPI data*

Finally, it is expected that new sources of PPI data will help relieve some of the issues discussed above. Sanderson (2009) pointed out that 20 to 30% of human genes encode membrane proteins and currently very little PPI interaction data exists for these proteins. This is true for both intracellular and extracellular membrane protein interactions. Nonetheless, exciting new experimental techniques are now being developed to probe these so-called dark regions of the interactome. These include the membrane yeast two hybrid (MYTH) assays (Stagljar et al. 1998), the AViditybase EXtracellular Interaction Screen (AVEXIS) system (Bushell et al. 2008) and the yeast adapted version of the DyHydroFolate-Reductase (DHFR) Protein fragment Complementation Assays (PCA) (Tarassov et al. 2008). It will be interesting to see how network topology differs amongst the different techniques and whether these new datasets, when combined with existing data, have an influence on the overall network topology. It is important to bear in mind that most of these assays still rely on protein over-expression and, since PPIs are dependent on both relative affinity and protein concentration, validating PPI data is still a major challenge. Diverent techniques

might however help to build a system of increasing confidence, by assigning higher scores to interactions verified by several independent techniques.

## References

- Agarwal S, Deane CM, Porter MA and Jones NS 2010 Revisiting date and party hubs: Novel approaches to role assignment in protein interaction networks. *PLoS Comput Biol* **6**(6), e1000817.
- Ali W and Deane CM 2009 Functionally guided alignment of protein interaction networks for module detection. *Bioinformatics* **25**(23), 3166–3173.
- Ali W and Deane CM 2010 Evolutionary analysis reveals low coverage as the major challenge for protein interaction network alignment. *Mol. BioSyst.* **6**, 2296–2304.
- Alm E and Arkin AP 2003 Biological networks. *Current Opinion in Structural Biology* **13**(2), 193 – 202.
- Alon U 2006 *An Introduction to Systems Biology: Design Principles of Biological Circuits* (Chapman & Hall/CRC *Mathematical & Computational Biology*) 1 edn. Chapman and Hall/CRC.
- Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM and Sherlock G 2000 Gene Ontology: tool for the unification of biology. *Nature Genetics* **25**(1), 25–29.
- Bachman P and Liu Y 2009 Structure discovery in PPI networks using pattern-based network decomposition. *Bioinformatics* **25**(14), 1814–1821.
- Bader G and Hogue C 2003 An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinformatics* **4**(1), 2.
- Bader GD, Donaldson I, Wolting C, Ouellette BFF, Pawson T and Hogue CWV 2001 BINDThe Biomolecular Interaction Network Database. *Nucleic Acids Research* **29**(1), 242–245.
- Barabasi AL and Oltvai ZN 2004 Network biology: understanding the cell's functional organization. *Nature Reviews Genetics* **5**(2), 101–113.
- Barabasi AL and Albert R 1999 Emergence of Scaling in Random Networks. *Science* **286**(5439), 509–512.
- Batada NN, Reguly T, Breitkreutz A, Boucher L, Breitkreutz BJ, Hurst LD and Tyers M 2006 Stratus not altocumulus: A new view of the yeast protein interaction network. *PLoS Biol* **4**(10), e317.
- Batada NN, Reguly T, Breitkreutz A, Boucher L, Breitkreutz BJ, Hurst LD and Tyers M 2007 Still stratus not altocumulus: Further evidence against the date/party hub distinction. *PLoS Biol* **5**(6), e154.
- Beaumont M 2010 Approximate Bayesian Computation in Evolution and Ecology. *Annual Review of Ecology, Evolution, and Systematics*.
- Berg JM, Tymoczko JL and Stryer L 2006 *Biochemistry (Biochemistry (Berg))* sixth edition edn. W. H. Freeman.
- Bollobas B 2001 *Random Graphs*. Cambridge University Press.
- Bowers P, Pellegrini M, Thompson M, Fierro J, Yeates T and Eisenberg D 2004 Prolinks: a database of protein functional linkages derived from coevolution. *Genome Biology* **5**(5), R35.
- Bowers PM, O'Connor BD, Cokus SJ, Sprinzak E, Yeates TO and Eisenberg D 2005 Utilizing logical relationships in genomic data to decipher cellular processes.. *FEBS J* **272**(20), 5110–5118.
- Brändén C and Tooze J 1991 *Introduction to protein structure*. Garland Publishing, New York.
- Breitkreutz BJ, Stark C and Tyers M 2003 The grid: The general repository for interaction datasets. *Genome Biology* **4**(3), R23.
- Brown KR and Jurisica I 2005 Online Predicted Human Interaction Database. *Bioinformatics* **21**(9), 2076–2082.
- Browne F, Zheng H, Wang H and Azuaje F 2010 From experimental approaches to computational techniques: a review on the prediction of protein-protein interactions. *Adv. in Artif. Intell.* **2010**, 7:5–7:5.
- Bu D, Zhao Y, Cai L, Xue H, Zhu X, Lu H, Zhang J, Sun S, Ling L, Zhang N, Li G and Chen R 2003 Topological structure analysis of the protein-protein interaction network in budding yeast. *Nucleic Acids Research* **31**(9), 2443–2450.
- Bushell KM, Sllner C, Schuster-Boeckler B, Bateman A and Wright GJ 2008 Large-scale screening for novel low-affinity extracellular protein interactions. *Genome Research* **18**(4), 622–630.
- Chen J, Hsu W, Lee ML and Ng SK 2005 Discovering reliable protein interactions from high-throughput experimental data using network topology. *Artificial Intelligence in Medicine* **35**(1-2), 37 – 47. Computational Intelligence Techniques in Bioinformatics.
- Chen PY, Deane CM and Reinert G 2007 A statistical approach using network structure in the prediction of protein characteristics. *Bioinformatics* **23**(17), 2314–2321.
- Chen PY, Deane CM and Reinert G 2008 Predicting and validating protein interactions using network structure. *PLoS Comput Biol* **4**(7), e1000118.
- Chien CT, Bartel PL, Sternglanz R and Fields S 1991 The two-hybrid system: a method to identify and clone genes for proteins that interact with a protein of interest. *Proceedings of the National Academy of Sciences of the United States of America* **88**(21), 9578–9582.
- Covert MW, Schilling CH and Palsson B 2001 Regulation of gene expression in flux balance models of metabolism. *Journal of Theoretical Biology* **213**(1), 73 – 88.

- Dall J and Christensen M 2002 Random geometric graphs. *Phys. Rev. E* **66**(1), 016121.
- Dandekar T, Schuster S, Snel B, Huynen M and Bork P 1999 Pathway alignment: application to the comparative analysis of glycolytic enzymes. *Biochem. J.* **343**(1), 115–124.
- Dandekar T, Snel B, Huynen M and Bork P 1998 Conservation of gene order: a fingerprint of proteins that physically interact. *Trends in biochemical sciences* **23**(9), 324–328.
- de Silva E, Thorne T, Ingram P, Agrafioti I, Swire J, Wiuf C and Stumpf M 2006 The effects of incomplete protein interaction data on structural and evolutionary inferences. *BMC Biology* **4**(1), 39.
- Deane CM, Salwiski u, Xenarios I and Eisenberg D 2002 Protein Interactions. *Molecular & Cellular Proteomics* **1**(5), 349–356.
- Deng M, Mehta S, Sun F and Chen T 2002 Inferring Domain-Domain Interactions From Protein-Protein Interactions. *Genome Research* **12**(10), 1540–1548.
- Deng M, Zhang K, Mehta S, Chen T and Sun F 2003 Prediction of protein function using protein-protein interaction data. *Journal of Computational Biology* **10**(6), 947–960.
- Dorogovtsev SN and Mendes JFF 2003 *Evolution of Networks: From Biological Nets to the Internet and WWW*. Oxford University Press.
- Dunn R, Dudbridge F and Sanderson C 2005 The use of edge-betweenness clustering to investigate biological function in protein interaction networks. *BMC Bioinformatics* **6**(1), 39.
- Dutkowski J and Tiurnyn J 2007 Identification of functional modules from conserved ancestral protein-protein interactions. *Bioinformatics* **23**(13), 1149–158.
- Enright A and Ouzounis C 2001 Functional associations of proteins in entire genomes by means of exhaustive detection of gene fusions. *Genome Biology* **2**(9), research0034.1–research0034.7.
- Enright AJ, Iliopoulos I, Kyrpides NC and Ouzounis CA 1999 Protein interaction maps for complete genomes based on gene fusion events. *Nature* **402**(6757), 86–90.
- Erdos P and Renyi A 1960 On the evolution of random graphs. *Publ. Math. Inst. Hung. Acad. Sci* **5**, 17–61.
- Evlampiev K and Isambert H 2007 Modeling protein network evolution under genome duplication and domain shuffling. *BMC Systems Biology* **1**(1), 49.
- Fernandes LP, Annibale A, Kleinjung J, Coolen ACC and Fraternali F 2010 Protein networks reveal detection bias and species consistency when analysed by information-theoretic methods. *PLoS ONE* **5**(8), e12083.
- Finn RD, Mistry J, Tate J, Coghill P, Heger A, Pollington JE, Gavin OL, Gunasekaran P, Ceric G, Forslund K, Holm L, Sonnhammer ELL, Eddy SR and Bateman A 2010 The Pfam protein families database. *Nucleic Acids Research* **38**(suppl 1), D211–D222.
- Flannick J, Novak A, Srinivasan BS, McAdams HH and Batzoglou S 2006 Grmlin: General and robust alignment of multiple large interaction networks. *Genome Research* **16**(9), 1169–1181.
- Flannick J, Novak AF, Do CB, Srinivasan BS and Batzoglou S 2008 Automatic parameter learning for multiple network alignment *RECOMB*, pp. 214–231.
- Fortunato S 2010 Community detection in graphs. *Physics Reports* **486**(3-5), 75 – 174.
- Freschi V 2009 A graph-based semi-supervised algorithm for protein function prediction from interaction maps In *Learning and Intelligent Optimization* (ed. Sttze T) vol. 5851 of *Lecture Notes in Computer Science* Springer Berlin / Heidelberg pp. 249–258.
- Friedgut E and Kalai G 1996 Every monotone graph property has a sharp threshold. *Proceedings of the American Mathematical Society* **124**(10), pp. 2993–3002.
- Goel A, Rai S and Krishnamachari B 2005 B.: Monotone properties of random geometric graphs have sharp thresholds. *ANNALS OF APPLIED PROBABILITY*.
- Guo X and Hartemink AJ 2009 Domain-oriented edge-based alignment of protein interaction networks. *Bioinformatics* **25**(12), i240–1246.
- Han JDD, Bertin N, Hao T, Goldberg DS, Berriz GF, Zhang LV, Dupuy D, Walhout AJ, Cusick ME, Roth FP and Vidal M 2004 Evidence for dynamically organized modularity in the yeast protein-protein interaction network. *Nature* **430**(6995), 88–93.
- Hart GT, Ramani A and Marcotte E 2006 How complete are current yeast and human protein-interaction networks?. *Genome Biology* **7**(11), 120.
- Hartwell LH, Hopfield JJ, Leibler S and Murray AW 1999 From molecular to modular cell biology. *Nature* **402**(6761 Suppl), C47–C52.
- Huang H, Jedynak BM and Bader JS 2007 Where have all the interactions gone? estimating the coverage of two-hybrid protein interaction maps. *PLoS Comput Biol* **3**(11), e214.
- Hubbell E, Liu WM and Mei R 2002 Robust estimators for expression analysis. *Bioinformatics* **18**(12), 1585–1592.
- Huthmacher C, Gille C and Holzhtter HG 2008 A computational analysis of protein interactions in metabolic networks reveals novel enzyme pairs potentially involved in metabolic channeling. *Journal of Theoretical Biology* **252**(3), 456 – 464. In Memory of Reinhart Heinrich.
- Ispolatov I, Krapivsky PL and Yuryev A 2005a Duplication-divergence model of protein interaction network. *Phys. Rev. E* **71**(6), 061911.
- Ispolatov I, Krapivsky PL, Mazo I and Yuryev A 2005b Cliques and duplication-divergence network growth. *New Journal of Physics* **7**(1), 145.

- Jeong H, Mason SP, Barabási AL and Oltvai ZN 2001 Lethality and centrality in protein networks. *Nature* **411**(6833), 41–42.
- Jones TR, Carpenter AE, Lamprecht MR, Moffat J, Silver SJ, Grenier JK, Castoreno AB, Eggert US, Root DE, Golland P and Sabatini DM 2009 Scoring diverse cellular morphologies in image-based screens with iterative feedback and machine learning. *Proceedings of the National Academy of Sciences* **106**(6), 1826–1831.
- Jonsson P, Cavanna T, Zicha D and Bates P 2006 Cluster analysis of networks generated through homology: automatic identification of important protein communities involved in cancer metastasis. *BMC Bioinformatics* **7**(1), 2.
- Kanehisa M and Goto S 2000 KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research* **28**(1), 27–30.
- Kasahara M 2007 The 2r hypothesis: an update. *Current Opinion in Immunology* **19**(5), 547 – 552. Hematopoietic cell death/Immunogenetics/Transplantation.
- Kelley BP, Sharan R, Karp RM, Sittler T, Root DE, Stockwell BR and Ideker T 2003 Conserved pathways within bacteria and yeast as revealed by global protein network alignment. *Proceedings of the National Academy of Sciences of the United States of America* **100**(20), 11394–11399.
- Kelley BP, Yuan B, Lewitter F, Sharan R, Stockwell BR and Ideker T 2004 PathBLAST: a tool for alignment of protein interaction networks. *Nucleic Acids Research* **32**(suppl 2), W83–W88.
- Keshava Prasad TS, Goel R, Kandasamy K, Keerthikumar S, Kumar S, Mathivanan S, Telikicherla D, Raju R, Shafreen B, Venugopal A, Balakrishnan L, Marimuthu A, Banerjee S, Somanathan DS, Sebastian A, Rani S, Ray S, Harrys Kishore CJ, Kanth S, Ahmed M, Kashyap MK, Mohmood R, Ramachandra YL, Krishna V, Rahiman BA, Mohan S, Ranganathan P, Ramabadrans S, Chaerkady R and Pandey A 2009 Human Protein Reference Database 2009 update. *Nucleic Acids Research* **37**(suppl 1), D767–D772.
- Kim WK and Marcotte EM 2008 Age-dependent evolution of the yeast protein interaction network suggests a limited role of gene duplication and divergence. *PLoS Comput Biol* **4**(11), e1000232.
- Koyuturk M, Kim Y, Topkara U, Subramaniam S, Szpankowski W and Grama A 2006 Pairwise alignment of protein interaction networks. *Journal of Computational Biology* **13**(2), 182–199. PMID: 16597234.
- Letovsky S and Kasif S 2003 Predicting protein function from protein/protein interaction data: a probabilistic approach. *Bioinformatics* **19**(suppl 1), i197–i204.
- Lewis A, Jones N, Porter M and Deane C 2010a The function of communities in protein interaction networks at multiple scales. *BMC Systems Biology* **4**(1), 100.
- Lewis ACF, Saeed R and Deane CM 2010b Predicting protein[ $\leftrightarrow$ ]protein interactions in the context of protein evolution. *Mol. BioSyst.* **6**, 55–64.
- Li M, Wang J and Chen J 2008 A graph-theoretic method for mining overlapping functional modules in protein interaction networks In *Bioinformatics Research and Applications* (ed. Mandoiu I, Sunderraman R and Zelikovsky A) vol. 4983 of *Lecture Notes in Computer Science* Springer Berlin / Heidelberg pp. 208–219.
- Liao CS, Lu K, Baym M, Singh R and Berger B 2009 IsoRankN: spectral methods for global alignment of multiple protein networks. *Bioinformatics* **25**(12), i253–258.
- Lima-Mendez G and van Helden J 2009 The powerful law of the power law and other myths in network biology. *Mol. BioSyst.* **5**, 1482–1493.
- Lin C, Jiang D and Zhang A 2006 Prediction of protein function using common-neighbors in protein-protein interaction networks *Bioinformatics and BioEngineering, 2006. BIBE 2006. Sixth IEEE Symposium on*, pp. 251–260.
- Lin K and Reinert G 2009 Joint Vertex Degrees in an Inhomogeneous Random Graph Model. *ArXiv e-prints*.
- Liu Y, Liu N and Zhao H 2005 Inferring protein-protein interactions through high-throughput interaction data from diverse organisms. *Bioinformatics* **21**(15), 3279–3285.
- Luciano, Rodrigues FA, Travieso G and Boas VPR 2007 Characterization of complex networks: A survey of measurements. *Advances in Physics* **56**(1), 167–242.
- Luo F, Yang Y, Chen CF, Chang R, Zhou J and Scheuermann RH 2007 Modular organization of protein interaction networks. *Bioinformatics* **23**(2), 207–214.
- Marcotte EM, Pellegrini M, Ng HL, Rice DW, Yeates TO and Eisenberg D 1999 Detecting Protein Function and Protein-Protein Interactions from Genome Sequences. *Science* **285**(5428), 751–753.
- Marcotte EM, Xenarios I, van der Bliik AM and Eisenberg D 2000 Localizing proteins in the cell from their phylogenetic profiles. *Proceedings of the National Academy of Sciences of the United States of America* **97**(22), 12115–12120.
- Matthews LR, Vaglio P, Reboul J, Ge H, Davis BP, Garrels J, Vincent S and Vidal M 2001 Identification of Potential Interaction Networks Using Sequence-Based Searches for Conserved Protein-Protein Interactions or Interologs. *Genome Research* **11**(12), 2120–2126.
- Mering Cv, Huynen M, Jaeggi D, Schmidt S, Bork P and Snel B 2003 STRING: a database of predicted functional associations between proteins. *Nucleic Acids Research* **31**(1), 258–261.
- Mewes HW, Amid C and Arnold R 2004 MIPS: analysis and annotation of proteins from whole genomes. *Nucl. Acids Res.* **32**, D41–44.

- Milenkovic T, Lai J and Przulj N 2008 Graphcrunch: A tool for large network analyses. *BMC Bioinformatics* **9**(1), 70.
- Mintsers J and Weng Z 2005 Structure, function, and evolution of transient and obligate protein-protein interactions. *Proceedings of the National Academy of Sciences of the United States of America* **102**(31), 10930–10935.
- Moffat J, Grueneberg DA, Yang X, Kim SY, Kloepper AM, Hinkle G, Piqani B, Eisenhaure TM, Luo B, Grenier JK, Carpenter AE, Foo SY, Stewart SA, Stockwell BR, Hacohen N, Hahn WC, Lander ES, Sabatini DM and Root DE 2006 A lentiviral RNAi library for human and mouse genes applied to an arrayed viral high-content screen. *Cell* **124**(6), 1283–1298.
- Newman MEJ and Girvan M 2004 Finding and evaluating community structure in networks. *Phys. Rev. E* **69**(2), 026113.
- Nowicki K and Snijders TAB 2001 Estimation and prediction for stochastic block structures. *Journal of the American Statistical Association* **96**(455), pp. 1077–1087.
- Ogata H, Fujibuchi W, Goto S and Kanehisa M 2000 A heuristic graph comparison algorithm and its application to detect functionally related enzyme clusters. *Nucl. Acids Res.* **28**(20), 4021–4028.
- Overbeek R, Fonstein M, D'Souza M, Pusch GD and Maltsev N 1999 The use of gene clusters to infer functional coupling. *Proceedings of the National Academy of Sciences of the United States of America* **96**(6), 2896–2901.
- Pellegrini M, Marcotte EM, Thompson MJ, Eisenberg D and Yeates TO 1999 Assigning protein functions by comparative genome analysis: Protein phylogenetic profiles. *Proceedings of the National Academy of Sciences of the United States of America* **96**(8), 4285–4288.
- Penrose M 2003 *Random Geometric Graphs*. Oxford University Press.
- Pereira-Leal JB, Enright AJ and Ouzounis CA 2004 Detection of functional modules from protein interaction networks. *Proteins: Structure, Function, and Bioinformatics* **54**(1), 49–57.
- Phizicky E and Fields S 1995 Protein-protein interactions: methods for detection and analysis. *Microbiol. Rev.* **59**(1), 94–123.
- Porter MA, Onnela JP and Mucha PJ 2009 Communities in Networks. *Notices of the American Mathematical Society, Vol. 56, No. 9, 2009*.
- Przulj N 2007 Biological network comparison using graphlet degree distribution. *Bioinformatics* **23**(2), e177–e183.
- Przulj N 2010 Biological network comparison using graphlet degree distribution. *Bioinformatics* **26**(6), 853–854.
- Przulj N, Corneil DG and Jurisica I 2004 Modeling interactome: scale-free or geometric?. *Bioinformatics* **20**(18), 3508–3515.
- Puig O, Caspary F, Rigaut G, Rutz B, Bouveret E, Bragado-Nilsson E, Wilm M and Srafin B 2001 The tandem affinity purification (tap) method: A general procedure of protein complex purification. *Methods* **24**(3), 218–229.
- Ratmann O, Jrgensen O, Hinkley T, Stumpf M, Richardson S and Wiuf C 2007 Using likelihood-free inference to compare evolutionary dynamics of the protein networks of *h. pylori* and *p. falciparum*. *PLoS Comput Biol* **3**(11), e230.
- Reichardt J and Bornholdt S 2006 Statistical mechanics of community detection. *Phys. Rev. E* **74**(1), 016110.
- Rentsch R and Orengo CA 2009 Protein function prediction - the power of multiplicity. *Trends in Biotechnology* **27**(4), 210–219.
- Resnik P 1995 Using information content to evaluate semantic similarity in a taxonomy. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, pp. 448–453.
- Rito T, Wang Z, Deane CM and Reinert G 2010 How threshold behaviour affects the use of subgraphs for network comparison. *Bioinformatics* **26**(18), i611–i617.
- Saito R, Suzuki H and Hayashizaki Y 2002 Interaction generality, a measurement to assess the reliability of a protein-protein interaction. *Nucleic Acids Research* **30**(5), 1163–1168.
- Saito R, Suzuki H and Hayashizaki Y 2003 Construction of reliable protein-protein interaction networks with a new interaction generality measure. *Bioinformatics* **19**(6), 756–763.
- Sanderson CM 2009 The Cartographers toolbox: building bigger and better human protein interaction networks. *Briefings in Functional Genomics & Proteomics* **8**(1), 1–11.
- Scannell DR, Butler G and Wolfe KH 2007 Yeast genome evolution: the origin of the species. *Yeast* **24**, 929–942.
- Schlicker A, Domingues F, Rahnenfuhrer J and Lengauer T 2006 A new measure for functional similarity of gene products based on gene ontology. *BMC Bioinformatics* **7**(1), 302.
- Schwikowski B, Uetz P and Fields S 2000 A network of protein-protein interactions in yeast. *Nature biotechnology* **18**(12), 1257–1261.
- Segal E, Wang H and Koller D 2003 Discovering molecular pathways from protein interaction and gene expression data. *Bioinformatics* **19**(suppl 1), i264–i272.
- Sharan R and Ideker T 2006 Modeling cellular machinery through biological network comparison. *Nature Biotechnology* **24**, 427–433.
- Shlomi T, Eisenberg Y, Sharan R and Ruppin E 2007 A genome-scale computational study of the interplay between transcriptional regulation and metabolism. *Molecular Systems Biology*.
- Singh R, Xu J and Berger B 2008 Global alignment of multiple protein interaction networks with application to functional orthology detection. *Proceedings of the National Academy of Sciences* **105**(35), 12763–12768.

- Skrabaneck L, Saini H, Bader G and Enright A 2008 Computational prediction of protein-protein interactions. *Molecular Biotechnology* **38**, 1–17. 10.1007/s12033-007-0069-2.
- Spirin V and Mirny LA 2003 Protein complexes and functional modules in molecular networks. *Proceedings of the National Academy of Sciences of the United States of America* **100**(21), 12123–12128.
- Stagljar I, Korostensky C, Johnsson N and te Heesen S 1998 A genetic system based on split-ubiquitin for the analysis of interactions between membrane proteins in vivo. *Proceedings of the National Academy of Sciences of the United States of America* **95**(9), 5187–5192.
- Stumpf MPH and Wiuf C 2005 Sampling properties of random graphs: The degree distribution. *Phys. Rev. E* **72**(3), 036118.
- Stumpf MPH, Thorne T, de Silva E, Stewart R, An HJ, Lappe M and Wiuf C 2008 Estimating the size of the human interactome. *Proceedings of the National Academy of Sciences* **105**(19), 6959–6964.
- Stumpf MPH, Wiuf C and May RM 2005 Subnets of scale-free networks are not scale-free: Sampling properties of networks. *Proceedings of the National Academy of Sciences of the United States of America* **102**(12), 4221–4224.
- Su AI, Wiltshire T, Batalov S, Lapp H, Ching KA, Block D, Zhang J, Soden R, Hayakawa M, Kreiman G, Cooke MP, Walker JR and Hogenesch JB 2004 A gene atlas of the mouse and human protein-encoding transcriptomes. *Proceedings of the National Academy of Sciences of the United States of America* **101**(16), 6062–6067.
- Tamames J, Casari G, Ouzounis C and Valencia A 1997 Conserved clusters of functionally related genes in two bacterial genomes. *Journal of molecular evolution* **44**(1), 66–73.
- Tanaka R, Yi TM and Doyle J 2005 Some protein interaction data do not exhibit power law statistics. *FEBS Letters* **579**(23), 5140–5144.
- Tanay A, Sharan R, Kupiec M and Shamir R 2004 Revealing modularity and organization in the yeast molecular network by integrated analysis of highly heterogeneous genomewide data. *Proceedings of the National Academy of Sciences of the United States of America* **101**(9), 2981–2986.
- Tarassov K, Messier V, Landry CR, Radinovic S, Molina MMS, Shames I, Malitskaya Y, Vogel J, Bussey H and Michnick SW 2008 An in Vivo Map of the Yeast Protein Interactome. *Science* **320**(5882), 1465–1470.
- van Dongen SM 2000 *Graph Clustering by Flow Simulation* PhD thesis University of Utrecht, The Netherlands.
- von Mering C, Krause R, Snel B, Cornell M, Oliver SG, Fields S and Bork P 2002 Comparative assessment of large-scale data sets of protein-protein interactions. *Nature* **417**(6887), 399–403.
- Wang YC and Chen BS 2010 Integrated cellular network of transcription regulations and protein-protein interactions. *BMC Systems Biology* **4**(1), 20.
- Wasserman S and Faust K 1995 *Social Network Analysis: Methods and Applications (Structural Analysis in the Social Sciences)*. Cambridge University Press.
- Wilkins MR and Kummerfeld SK 2008 Sticking together? falling apart? exploring the dynamics of the interactome. *Trends in Biochemical Sciences* **33**(5), 195–200.
- Wodak SJ, Pu S, Vlasblom J and Srafin B 2009 Challenges and Rewards of Interaction Proteomics. *Molecular & Cellular Proteomics* **8**(1), 3–18.
- Wu Z, Irizarry RA, Gentleman R, Martinez-Murillo F and Spencer F n.d. A Model-Based Background Adjustment for Oligonucleotide Expression Arrays. *Journal of the American Statistical Association* **99**(468), 909+.
- Xenarios I, Salwinski L, Duan XJ, Higney P, Kim SM and Eisenberg D 2002 DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions. *Nucleic Acids Research* **30**(1), 303–305.
- Yeang CH, Ideker T and Jaakkola T 2004 Physical network models. *Journal of Computational Biology* **11**(2-3), 243–262.
- Yu X, Ivanic J, Wallqvist A and Reifman J 2009 A novel scoring approach for protein co-purification data reveals high interaction specificity. *PLoS Comput Biol* **5**(9), e1000515.
- Zanzoni A, Montecchi-Palazzi L, Quondam M, Ausiello G, Helmer-Citterich M and Cesareni G 2002 Mint: a molecular interaction database. *FEBS Letters* **513**(1), 135–140. Protein Domains.
- Zaslavskiy M, Bach F and Vert JP 2009 Global alignment of protein-protein interaction networks by graph matching methods. *Bioinformatics* **25**(12), i259–i267.
- Zhang J June 2003 Evolution by gene duplication: an update. *Trends in Ecology and Evolution* **18**, 292–298(7).