Conservation, error and dynamics in protein interactions networks

Waqar Ali

Department of Statistics

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

March 2011

I would like to dedicate this thesis to my loving parents.

Acknowledgements

I would first and foremost like to thank my supervisor Dr. Charlotte Deane for all her support and advice during my doctoral studies. She constantly provided guidance and encouragement throughout the research projects and read (and re-read) many drafts of my papers and this thesis.

I would also like to thank all members, past and present, of the Oxford Protein Informatics Group for providing a friendly, lively and intellectually stimulating environment for research work: Sebastian and Habib for all the interesting discussions, academic and otherwise, in room 50.14; Anna, Sumeet and Tiago for sharing ideas and being fellow members of the 'network' clique; Rebecca and Mireille for the lunch-time crosswords; and K1, for being himself.

I am thankful to the Department of Statistics at Oxford for providing me a teaching assistant bursary that funded my studies and members of its academic and support staff for their help at key stages of my DPhil.

Last, but not the least, many thanks to Beverley for her delicious cookies which made Tuesday mornings worth looking forward to.

Abstract

The availability of large scale protein interaction networks for several species has motivated many comparative studies in recent years. These studies typically employ network alignment algorithms for the task and use the sequence similarity of proteins to aid the alignment process. In this thesis I use a quantitative measure of protein functional similarity and show that the results are superior to sequence based network alignment. I present a method for module detection that combines results from network alignments with clustering measures to achieve superior results over several existing methods. Next, I address the issue of generally low conservation detected by alignments of interaction networks from model organisms. By explicitly modelling evolutionary mechanisms on pairs of networks I test the hypothesis that divergent evolution alone may be the cause. I use a distance metric based on graph summary statistics to assess the fit between experimental and simulated network alignments. Our results indicate that network evolution alone is unlikely to account for the poor quality alignments given by real data. We also find 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. Finally, I focus on the comparative analysis of a subset of the interaction network related to mitosis in Yeast, Human and Fly. Manual ordering of mitosis-related functional annotations allows the study of temporal aspects of the network. I also use a Markov random field approach to infer temporal labels for unlabelled proteins. Sequence based network alignment of the mitotic networks in the three species finds little conservation despite the proteins being functionally very similar. Further investigation suggests a fuzzy relationship between protein sequence and function that may have implications for future network alignment studies.

Contents

List of Figures

List of Tables

Chapter 1

Introduction and background

The central focus of this thesis is the analysis of conservation in large-scale protein interaction networks from model organisms and an investigation into the extent of the role played by factors such as noisy data. While new computational methods for comparing interaction networks of multiple species have been regularly proposed in recent years, there is a lack of systematic studies that quantify the amount of conservation detected using available methods and the sensitivity of this comparative analysis to issues such as incompleteness and noise inherent in experimental protein interaction data. In this thesis I discuss results of my research aimed at addressing some of these issues. This chapter introduces some of the basic concepts and provides a review of the most relevant research areas of what is now the very rich and diverse field of protein interaction network analysis. I start with a brief introduction to proteins and their interactions and then move on to describe existing techniques for the analysis and modelling of genome-scale interaction datasets.

1.1 Proteins

Proteins are the most versatile macromolecules in living systems and serve crucial functions in most biological processes. 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 primarily by the sequence of amino acids in the protein polymer (Berg et al., 2006).

1.1.1 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 structures, 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.

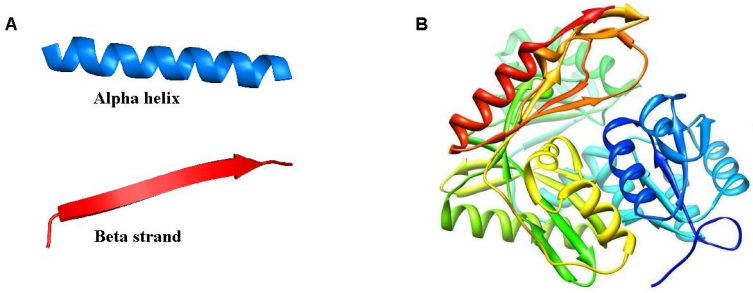


Figure 1.1: Protein structure: (A) Secondary structure elements and (B) Tertiary structure.

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 pro-

teins 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 Protein-protein interactions

Most proteins function through interaction with other molecules, and often these are other proteins. 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.

There is a 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 difference is not always well understood, and the classification is sometimes difficult (Mintseris and Weng, 2005). Moreover, many interactions do not fall into distinct types. Rather, a continuum exists between non-obligate and obligate interactions, and the stability of all complexes very much depends on the physiological conditions and environment. Still, this distinction is an important consideration when collating data from multiple experiments as the different empirical techniques for interaction detection are usually biased towards either binary interactions (which are more likely to be transient) or stable complexes.

1.2.1 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 is a much larger number of 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 (Uetz et al., 2008). Moreover, large scale studies using these methods show little overlap with each other (Bader and Hogue, 2002; von Mering et al., 2002).

1.2.1.1 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 (Figure 1.2). 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). The reason for this high error rate lies in the principle of the screen: The assay investigates the interaction between over-expressed fusion proteins in the yeast nucleus. Each of these points alone can give rise to false results. For example, overexpression can result in non-specific interactions. Moreover, a mammalian protein is sometimes not correctly modified in yeast (e.g., missing

phosphorylation), which can also lead to false results. Finally, some proteins might specifically interact when they are co-expressed in the yeast, although in reality they are never present in the same cell at the same time. Due to the combined effects of all error sources the overall confidence of the yeast two-hybrid assay is rather low.

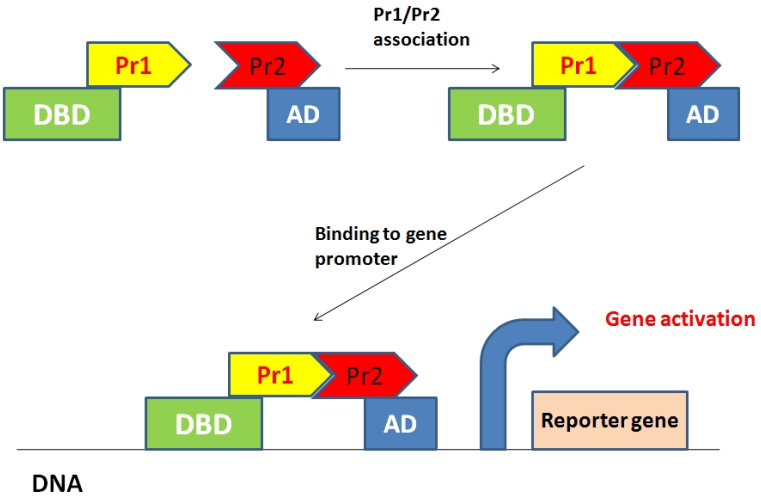


Figure 1.2: The yeast two hybrid method (Lhost, 2011). Pr1 and Pr2 are the two putatively interacting proteins. DBD is the DNA-binding domain and AD is the transcription activation domain.

1.2.1.2 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. The TAP method (Figure 1.3) requires fusion of the TAP tag to the target protein of interest. 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. This purification procedure significantly reduces the possible occurrence of nonspecific protein contaminants, thus decreasing both the unspecific background noise and the possible presence of false positives. Compared to the yeast

two hybrid system, TAP is thought to have lower false negative rates (15%), and a false positive rate of up to 35% (Hart et al., 2006).

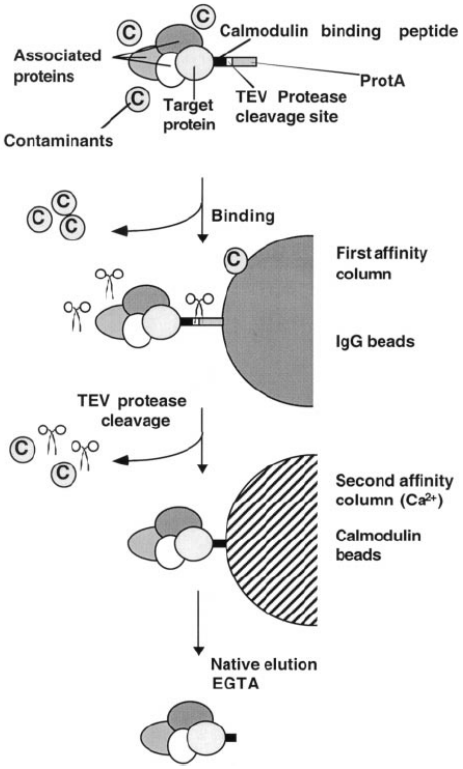


Figure 1.3: The tandem affinity purification (Puig et al., 2001) method. It involves the fusion of the TAP tag to the target protein and the introduction of the construct into the host cell or organism. Cell extracts are prepared and the fusion protein as well as associated partners are recovered by two specific affinity purification/elution steps.

1.2.1.3 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 (Figure 1.4). 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.

Figure 1.4: Coimmunoprecipitation (ThermoScientific, 2011). The procedure includes: 1) An antibody specific to the protein of interest is added to a cell lysate. 2) The antibody-protein complex is then precipitated. If there are any protein/molecules that bind to the first protein, they will also be precipitated. 3) Co-precipitated proteins can then be identified by Western blot analysis.

1.2.2 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 base data 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 experimental data sources to some extent 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 (e.g indicating participation in the same biological process) that do not necessarily imply physical interactions.

1.2.3 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 designed and set up databases to store this information. 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., 2002b);
- 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., 2009b).

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.

1.2.4 Error and incompleteness in PPI datasets

It is essential to appreciate how accurately empirical data reflects the true interactome. For this reason it is of crucial importance to have an understanding of the noise found in protein interaction data when performing any global analyses. Determining whether two proteins interact is hard to achieve. Many issues with experimental data exist, including: biases or systematic errors from experimental techniques (Aloy and Russell, 2002; Chiang et al., 2007); how to use binding affinities to infer interactions (Aloy and Russell, 2006); and basic uncertainties regarding our understanding of the regulation system of the cell. Apart from the sources of error mentioned earlier for the yeast two hybrid system, complex purification methods such as TAP also have potential pitfalls. Complications arise, for example, when proteins belonging to the

same complex are tagged and the resulting complexes are purified. In most cases this leads to conflicting information, because these purifications have slightly different protein compositions, depending on which protein was the tagged one. Different complexes are recovered even when the same tagged protein is purified repeatedly (Goll and Uetz, 2006). In addition, many proteins are part of several different complexes: one bait protein may thus pull down several independent complexes that appear in the experiment to be one large complex. Noise is thus an inherent property of current experimental techniques.

The false-positive set of interactions contains reported protein pairs that are erroneously reported as true interactions. The false-negative set of interactions are those that are tested but incorrectly not reported in an experiment. Recent estimates suggest that the complete 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 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.

1.3 Protein interaction networks as graphs

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 (Figure 1.5). 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. It also makes the interaction data immediately amenable to automated analysis via a very rich set of techniques from the graph theory literature. 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 computationally in a network using clustering algorithms (Section 1.4.1).

1.3.1 Graphs and networks

In the current and the next section, I review some of the fundamental mathematical terminology pertinent to the study of networks. Much of the terminology and methodology for network analysis comes from graph theory. While some authors assume that, in contrast to graphs,

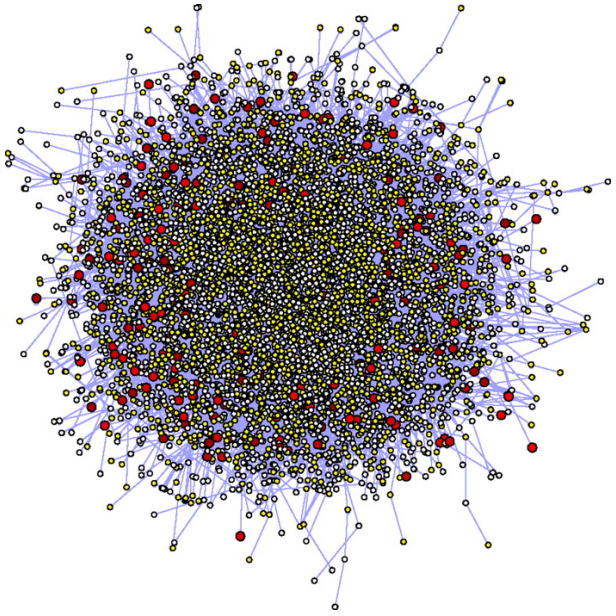


Figure 1.5: The human protein interaction network (Kim et al., 2007).

networks are connected, here we make no such assumption and use the terms graph and network interchangeably. A graph consists of nodes (also called vertices) and edges (also called links). For protein interaction data, we typically think of the nodes as proteins, and an edge as an interaction between the two nodes (proteins). 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. Therefore $|V|$ 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.

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 basic statistic 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 1.6). 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

$$C_i = \frac{2|\{(v_j, v_k) \in E : v_j, v_k \in N_i\}|}{k_i(k_i - 1)}. \tag{1.1}$$

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).

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, theoretical models for these networks are needed (see Section 1.3.5).

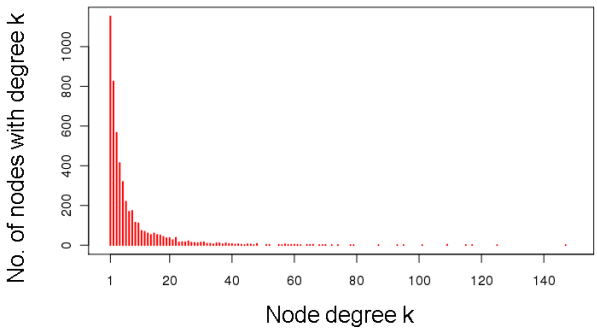


Figure 1.6: Frequency of node degrees (k) in the yeast DIP network (accessed July 2010)

Table 1.1: Summary statistics for yeast DIP and human HPRD protein interaction networks. Data downloaded from DIP and HPRD on 15 Sep 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 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 1.7 illustrates some examples of 2 to 5-node motifs.

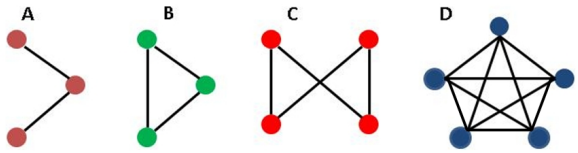


Figure 1.7: Some examples of motifs: (A) Line, (B) Triangle, (C) Square and (D) 5-node clique

1.3.4 Modularity in PPI networks

Protein molecules bind to each other to form stable complexes that often can be purified. At a higher level of structure, proteins and protein complexes can interact with preferred partners weakly, transiently, or conditionally to form a biological module serving a specific collective function (Hartwell et al., 1999). A functional module is thus defined as a group of genes or their products which are related by one or more genetic or cellular interactions, e.g. coregulation, coexpression or membership of a protein complex, of a metabolic or signaling pathway or of a cellular aggregate (e.g. chaperone, ribosome, protein transport facilitator, etc.). An important property of a module is that its function is separable from other modules (Hartwell et al., 1999; Rives and Galitski, 2003) and that its members have more relations among themselves than with members of other modules, which is reflected in the network topology. The separability may stem from, for example, cellular localization or specific interaction of proteins or specific regulation of genes. Functional modules need not be rigid, fixed structures; a given component may belong to different modules at different times. The function of a module can be quantitatively regulated, or switched between qualitatively different functions, by chemical signals from other modules. Higher-level functions can be built by connecting modules together. A picture of the interaction network has started to emerge in which various regions of high connectivity correspond to groups of proteins forming higher level structures to perform specific

biological functions. These protein complexes are linked together by extended networks of weaker, transient interactions, to form interaction networks that integrate pathways mediating the major cellular processes. The protein interaction network of the Yeast cell has been shown to have a nonrandom power-law distribution of node degree and a low frequency of direct connections between high-connectivity nodes (Maslov and Sneppen, 2002). These observations suggest modular organization consistent with the insights of biologists. Various methods of network clustering (Section 1.4.1) have been developed and applied to identify modules in biological systems. Such methods typically use graph theoretic approaches to detect regions of high interaction density in a network known as clusters (or communities). The underlying assumption is that these clusters or communities correspond to biological structures: either protein complexes or functional modules (Figure 1.8).

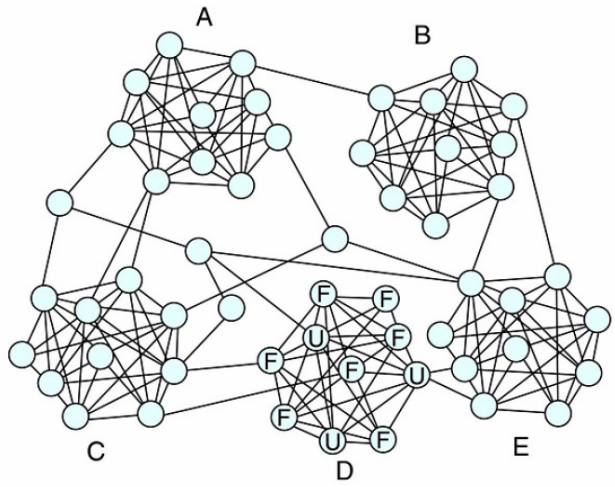


Figure 1.8: A highly clustered network. A-E are distinct clusters in this graph. In a biological context, nodes within a cluster might share properties such as protein function (F,U).

1.3.5 Theoretical models for PPI 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. Some of the main models are discussed here. The Bernoulli or Erdős-Renyi (ER) random graph

model (Erdos and Renyi, 1960) is one of the earliest, 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 (Barabsi 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 PPI networks (Barabasi and Albert, 1999; Barabasi and Oltvai, 2004; Lee et al., 2005), it is difficult to relate their growth mechanisms to real biological systems. This has led to the proposal of models specifically aimed at 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. 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. 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).

1.3.6 Parameter estimation for network models

In most of the above network 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 (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 infeasible. 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.

As full-likelihood based parameter estimation is often computationally intractable for even moderate-sized networks and relatively simple network models, studies in the context of protein interaction networks 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).

Instead of computing the intractable likelihood of the protein network topology, their method summarizes key features of the network and then uses a Markov Chain Monte Carlo 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 a eukaryotic species (*H.pylori*) and a prokryote ($P.falciparum$). Fitting this above model using ABC indicated that gene duplication has played a larger part in the PPI network evolution of the eukaryotic species than in the prokaryote.

1.4 Computational analysis of PPI networks

Although the theoretical models underlying protein interaction networks and their evolution are still not well-understood, research in the computational analysis of these networks has seen a remarkable increase during the last decade. Many studies have used protein interaction networks, sometimes in conjunction with other types of biological data, to better understand cellular systems. The focus of most of the initial work was the prediction and validation of the properties (protein function in particular) of individual proteins. While this is still one of the most active areas in network analysis, much research is also being carried out to glean the higher level organizing principles of these networks. Automated detection of protein complexes and functional modules in large-scale networks is one of these promising areas which has branched out further into cross-species network comparisons to detect evolutionarily conserved units of biological function. Below I review some of the dominant research themes in the analysis of these networks from an applied perspective.

1.4.1 Computational detection of protein complexes and functional modules

Experimentalists study complexes in many ways including identifying their structure using techniques such as X-ray crystallography or Nuclear Magnetic Resonance (NMR). Although experimental techniques like NMR and X-ray crystallography produce accurate and reliable results, they are labour intensive, time-consuming and expensive, so the number of experimentally determined structures for protein complexes is quite small. It is also not easy to detect functional modules using experimental techniques as their components may not assemble at a single particular time (Hartwell et al., 1999). Therefore, computational methods for the discovery of protein complexes and modules are becoming increasingly important. The abundance of interaction data inside publicly available interaction networks for several species has made them a natural choice for this computational analysis. Most of the methods are based on the understanding that proteins in a complex display a high number of interactions with each other so protein complexes would generally correspond to dense clusters in the PPI network. This idea is also lent support by evidence of the modular nature of protein networks discussed earlier in Section 1.3.4. Several studies (Spirin and Mirny, 2003) have shown that clusters in the networks exhibit properties of real complexes and/or functional modules. Unfortunately the situation is far more complicated than the ideal painted in Figure 1.8. Not all real complexes display high connectivity inside currently available interaction networks. Moreover, not all dense clusters in the networks have biological significance (Spirin and Mirny, 2003). These issues combined with the high error rates in interaction datasets, make complete elucidation of protein complexes and functional modules from interaction networks a very challenging task, albeit one with great benefits. For example, proteins that participate in the same complex generally share some functional aspects (Dezs et al., 2003). It is therefore possible to assign functions to un-annotated proteins based on the types of proteins they form complexes with. Elucidation of complexes can also afford valuable insights into the organizational properties of interaction networks, and answer questions such as whether networks of different species share molecular machineries at a higher level than individual proteins. Finally, there is potential for more effective drug discovery, targeting sets of proteins which tend to function in a concerted

manner.

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 $\mathbf{e}$ 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

$$Q = \sum_i e_{i,i} - \sum_{i,j,k} e_{i,j} e_{k,i}, \tag{1.2}$$

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'):

$$H = -\sum_{ij} J_{ij}(\lambda)\delta(s_i, s_j), \tag{1.3}$$

where s_i is the community of node i, δ is the Kronecker delta, λ 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.

The Markov clustering algorithm (MCL) (Dongen, 2000) simulates random walks on the underlying interaction network, by alternating two operations: expansion, and inflation. First, loops are added to the input graph - by default, the loop weight for each node is assigned as

the maximum weight of all edges connected to the node and this graph is then translated into a stochastic "Markov" matrix. This matrix represents the transition probabilities between all pairs of nodes, and the probability of a random walk of length n between any two nodes can be calculated by raising this matrix to the exponent n: a process referred to as expansion. As higher length paths are more common between nodes in the same cluster than nodes within different clusters, the probabilities between nodes in the same complex will typically be higher in expanded matrices. MCL further exaggerates this effect by taking entry wise exponents of the expanded matrix, and then rescaling each column so that it remains stochastic: a process called inflation. Clusters are identified by alternating expansion and inflation until the graph is partitioned into subsets so that there are no longer paths between these subsets.

MCODE (Bader and Hogue, 2003) is another graph clustering algorithm that operates in three stages; vertex weighting, complex prediction and an optional post-processing step. The first stage of MCODE weights all vertices based on their local network density using the highest k-core of the vertex neighbourhood. MCODE defines a term core-clustering coefficient of a vertex, v, as the density of the highest k-core of the immediate neighbourhood of v (vertices connected directly to v) including v. The weight given to a vertex is the product of the vertex core-clustering coefficient and the highest k-core level, $kmax$, of the immediate neighbourhood of the vertex. The second stage takes as input the vertex weighted graph, seeds a cluster with the highest weighted vertex and recursively moves outward from the seed vertex. The third stage involves filtering out clusters that do not contain at least a 2-core.

Affinity Propagation (AP) (Frey and Dueck, 2007) identifies cluster centers, or exemplars, from the graph, which in some sense are a representative member of the cluster. Initially, all nodes are considered as exemplars, though each node is manually assigned a "preference" that it should be chosen as an exemplar. If no prior knowledge is available on which nodes should be favoured as exemplars, then all nodes can be assigned the same preference value - where the magnitude can be used to control cluster granularity. For each node i and each candidate exemplar k, AP computes the "responsibility" $r(i,k)$, which indicates how well suited k is as an exemplar for i, and the "availability" $a(i,k)$ reflecting the evidence that i should choose k

as an exemplar.

$$r(i,k) \leftarrow s(i,k) - \max_{k':k' \neq k} \{a(i,k') + s(i,k')\} \tag{1.4}$$

$$a(i,k) \leftarrow \min \left\{ 0, r(k,k) + \sum_{i':i' \notin \{i,k\}} \max\{0, r(i',k)\} \right\}$$

Where the matrix $s(i,k)$ denotes the similarity (eg. edge weight) between the two nodes i and k, and the diagonal of this matrix contains the preferences for each node. The above two equations are iterated until a good set of exemplars emerges. Each node i can then be assigned to the exemplar k which maximizes the sum $a(i,k)+r(i,k)$, and if $i = k$, then i is an exemplar. A damping factor between 0 and 1 is used to control for numerical oscillations.

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

$$H(C) = \frac{\sum_{i,j \in C} Similarity(i,j)}{|C|(|C|-1)}; \tag{1.5}$$

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

$$p = 1 - \sum_{i=0}^{k-1} \frac{\binom{|X|}{i}\binom{|N|-|X|}{n-i}}{\binom{|N|}{n}}, \tag{1.6}$$

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.

The above-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.

1.4.2 Function prediction using PPI networks

Protein interaction networks are a rich source of information about the context of a protein, i.e., the position of an individual protein in a larger view of the biological processes in a organism. Huynen et al. (2000) suggested that exploiting this context information is more effective for predicting functional associations and specific functions than pairwise comparison-based approaches such as mRNA co-expression.

Approaches that attempt to predict function from a protein interaction network can be broadly categorized into the following categories:

1.4.2.1 Neighbourhood-based approaches

These approaches utilize the neighbourhood of the query protein in the interaction network and the most dominant annotations among these neighbours to predict its function. Given a set of interconnections among a set of entities, the most intuitively straightforward approach for inferring the characteristics of these entities is to extrapolate the characteristics of their neighbours, known as the majority vote method (Figure 1.9).

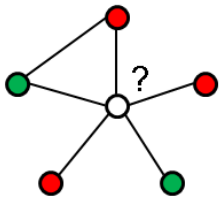


Figure 1.9: Majority vote function prediction. The central node will be labelled red as most of its neighbours share this label.

This idea of guilt-by-association directly addresses the problem of protein function prediction from protein interaction networks and was used by an early paper which addressed this problem (Schwikowski et al., 2000). In this study, a network of 2709 interactions among 2039

yeast proteins were assembled from various sources such as MIPS and DIP. Even though the prediction method was simple: the functions of a protein are assigned as the (at most) three most frequent functions among its neighbours, an accuracy of 72% was achieved for 1393 characterized proteins. A strategy to improve the statistical significance of these predictions was proposed by Hishigaki et al. (2001). First, instead of just the immediate neighbours, a set of n-neighbouring proteins consisting of proteins reached via n links is considered for prediction. Second, the frequencies of all the functions in this neighbourhood is recorded. Finally, the most significant function in this set is assigned to the protein of interest. This significance is tested using a χ^2-test, that compares the frequency of the function in this neighbourhood with that expected according to its occurrence probability across the whole interaction network.

In addition to function, networks can be also be used to predict other features of proteins such as structure. These protein features can normally be categorised, and Chen et al. (2007) proposed two methods for predicting protein properties using network structure that rely on sets of protein feature categories. A protein x is annotated with a set of categories $S(x)$; these categories could relate for example to structure, to function or to subcellular location. The protein interaction network provides a set $B(x)$ of proteins interacting with protein x. The characteristics of these interacting partners, together with the characteristics of x, give an upcast set of triples. For their frequency-based method, they give a category score based on the counts of relative frequencies of pairwise category-category interactions, and predict the category with the highest score, which is the most common category in interaction pairs that the protein x is involved in. The method differs from majority vote in that relative frequencies of all category pairs are taken into account. Their triple method and its variants use the lines and triangles of the category interactions in the prediction of protein characteristics. A triangle is a subnet formed by an interacting protein pair with a common neighbour. A line, by contrast, is a subnet formed by an non-interacting protein pair with a common neighbour. Heuristically, for a protein x they look at all the triples that x is involved in. They then translate these triples into category triples. In the network, the frequencies of different category triples differ considerably. They predict, for x, the category which is most common in the type of triples that x is involved with. The benefit of using this triangle based approach was demonstrated through better prediction of protein function for several experimental datasets.

1.4.2.2 Global optimization-based approaches

In many cases, the neighbourhood of the query protein may not contain enough information, such as annotated proteins, for determining the function of the query protein robustly. Under these conditions, it may be advantageous to consider the structure of the entire network and use the annotations of the proteins indirectly connected to the query protein also. The approaches in this section are based on this idea, and in most cases, are based on the optimization of an objective function based on the annotations of the proteins in the network. One of the first methods that approached the problem from this viewpoint was proposed by Deng et al. (2003) and used the theory of Markov random fields (MRF) to determine the probability of a protein having a certain function. This theory is used to determine the joint probability of the entire network with respect to a certain function. The MRF formulation allows this joint probability to be transformed to that of the conditional probability of a protein having a certain function given the annotations of its interaction partners. The Gibbs sampling technique is used iteratively to determine the stable values of this probability for each protein. This strategy was found to outperform the neighbourhood-based approaches (Hishigaki et al., 2001; Schwikowski et al., 2000) in the functional annotation task for the MIPS interaction data for Yeast. In one of the extensions of this work, the same strategy was applied for the mapping of Gene Ontology terms (see Section 2.2.1) to proteins, with similar results (Deng et al., 2004). In another extension (Lee et al., 2006), the MRF approach was generalized by using a diffusion kernel-based similarity between proteins in the network. This enabled the approach to transfer annotations from farther away proteins, in addition to only the neighbouring proteins, weighted by their diffusion kernel-based similarity with the query protein.

1.4.2.3 Clustering-based approaches

The approaches in this section are based on the hypothesis that dense regions in the interaction network represent functional modules, which are natural units in which proteins perform their function. Thus, these approaches apply graph clustering algorithms (discussed in Section 1.4.1) to networks and then determine the functions of un-annotated proteins in the extracted modules using measures such as majority (Figure 1.10).

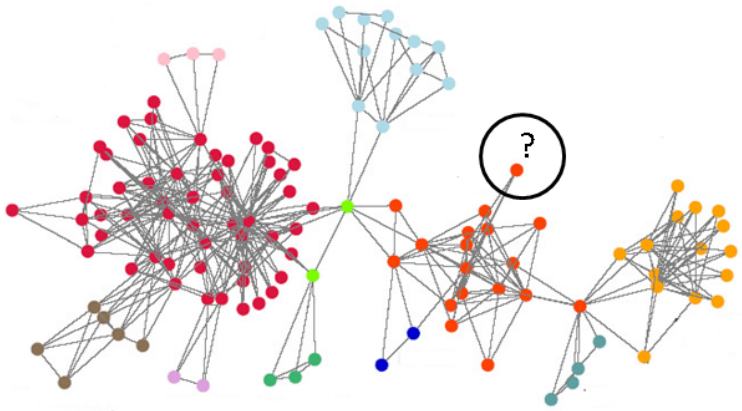


Figure 1.10: Protein function prediction using graph clustering. Nodes with the same colour signify proteins inside the same cluster in this hypothetical interaction network. The function of the unannotated protein can be inferred from its participation in a cluster with known function.

One of the obstacles to systematic evaluation of the different module-assisted methods for functional annotation is the lack of agreed upon technique for function prediction within a module. A systematic quantitative evaluation of module-assisted clustering algorithms has been carried out by Brohee and van Helden (2006). They found that the MCL algorithm is remarkably robust to graph alternations. MCL had the best performance on both simulated and real data sets, whereas MCODE was shown to be clearly inferior under most conditions.

It is worth pointing out here that despite the large number of techniques suggested for functional annotation using networks, systematic annotation is still mostly based on other data sources, such as sequence homology.

1.4.3 Interaction prediction using PPI networks

As discussed earlier, current PPI networks are very 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. Traditionally interaction prediction using already available interaction networks has not received as much attention as the prediction of protein function. Some notable methods include the domain based method by Deng et al. (2002), who 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. Beyond the pair-based approaches, Chen et al. (2008) incorporated higher level network structures like triangles and lines into the interaction prediction framework. They exploit the fact that protein networks display a high level of local clustering by suggesting a score based on triplets of observed protein interactions. Their score utilises both protein characteristics and network properties. Comparing their method to Deng's domain-based approach and the extension by Liu, they found that their method outperforms the others on the subset of interactions in the DIP Yeast data. The success of this method provides a good argument for representing PPI data as networks, as the triangle information appears to be crucial for good prediction.

1.4.4 Studying network 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 modelling 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 (Figure **??**). The validity of the date/party hub distinction has since been debated in some recent papers (Batada et al., 2006b, 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.

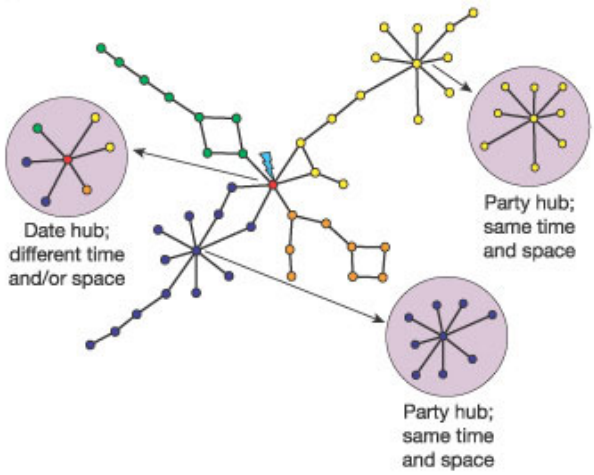


Figure 1.11: Date and party hubs (Han et al., 2004). Proteins are coloured here according to mutual similarity in their mRNA expression patterns.

RNA 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.

1.4.5 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 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 (also referred to as network alignment). While there has been much work done and many algorithms proposed recently for the alignment of empirical networks, not many methods exist to measure the agreement between experimental data and theoretical network models. Network alignment is not of much use for the latter, as the aim of the comparison in this case is to see how well the model replicates the properties (not the exact topology) of a given empirical network. In the following sections I will focus solely on network alignment methods and their application to large scale datasets.

1.4.6 Protein interaction 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 (Figure 1.12).

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

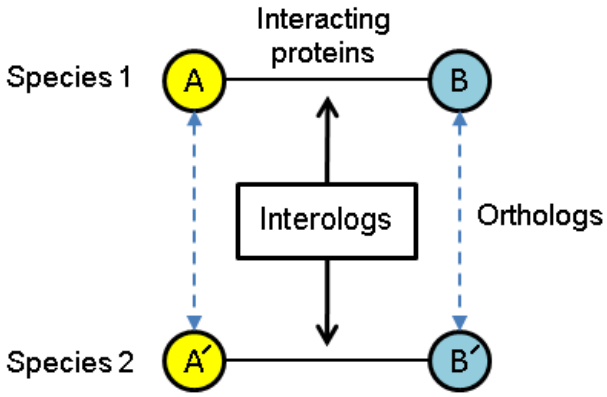


Figure 1.12: An example of an interolog. The two interacting proteins, A and B, in Species 1 have corresponding orthologs A' and B' in Species 2 that also interact with each other. This gives rise to a conserved interaction or interolog.

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 1.13). 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 computationally challenging.

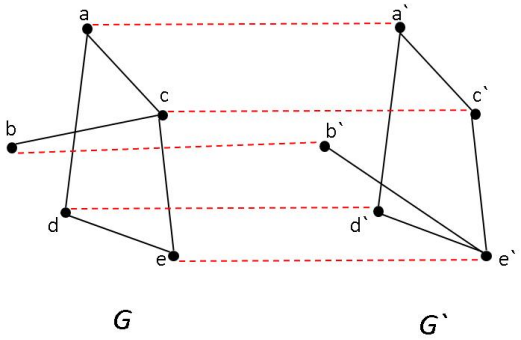


Figure 1.13: Pair-wise network alignment of two graphs G and G'. Dotted red lines indicate homology.

The network alignment problem can be formulated in various ways, depending on the kind of input (pairwise versus multiple alignments) and the scope of node mapping desired. Here,

I draw an analogy from the sequence alignment problem to distinguish between local and global network alignment. The goal in local network alignment is to find multiple, unrelated regions of isomorphism (i.e., same graph structure) between the input networks, each region implying a mapping independently of others. Many independent, high-scoring local alignments are usually possible between two input networks; which need not even be mutually consistent (i.e., a protein might be mapped differently under each alignment). The motivations behind local sequence alignment and local network alignment are similar-the former is often used to search for a conserved motif in the target sequence; the latter could be used to search for a known functional component (e.g., pathways, complexes, etc.) in the network of a new species. The aim in global network alignment is to find the best overall node-to-node mapping between the input networks. The mapping in a global network alignment should cover all of the input nodes: Each node in an input network is either matched to one or more nodes in the other network(s) or explicitly marked as a gap node (i.e., with no match in another network). In contrast, a local network alignment algorithm is essentially intended for finding similar motifs/patterns between two networks, and the mappings corresponding to different motifs may be mutually inconsistent.

1.4.6.1 Local network alignment methods

NetworkBlast (Sharan and Ideker, 2006) is one of the earliest algorithms that makes use of protein network alignment for conserved functional module detection. NetworkBlast carries out the alignment process by defining a network alignment graph. Each node in this alignment graph corresponds to a group of k sequence-similar proteins, one from each species (k is the number of networks being aligned). Two proteins are considered to have sufficient sequence similarity if their BLAST E-value is smaller than 10^{-7}, and each is among the 10 best BLAST matches of the other. Each edge in the alignment graph represents a conserved interaction. A search over the network alignment graph is performed to identify two types of conserved sub-network structures: short linear paths of interacting proteins, which model signal transduction pathways, and dense clusters of interactions, which model protein complexes. The results are filtered by comparison with a random model that assumes no conservation. Only those modules which have p-values < 0.01 and have less than 80% overlap with other modules are reported

(Figure 1.14). NetworkBlast has been used to perform three-way comparisons of yeast, worm and fly (Sharan et al., 2005) which yielded conserved modules displaying containing a total of 649 proteins, more than half of which were also found to be present in experimentally detected protein complexes from MIPS.

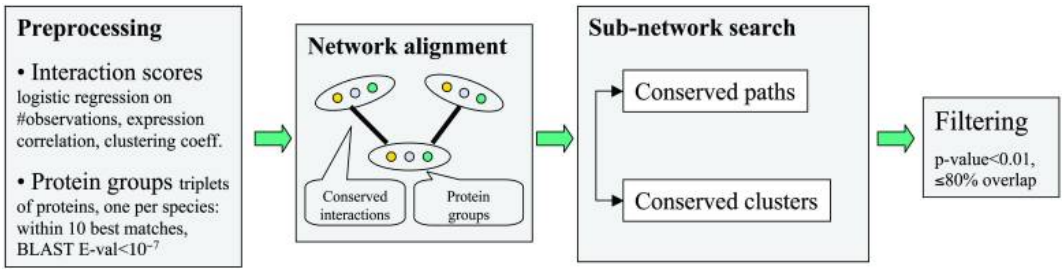


Figure 1.14: NetworkBlast pipeline.

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 scoring system based on the duplication divergence model to carry out pair-wise network alignment. While searching for conserved groups of interactions, it evaluates mismatched interactions in light of the duplication/divergence model. It introduces the concepts of match (interaction conservation), mismatch (interaction emergence or elimination), and duplication to discover alignments that also allow speculation about the structure of the network in the common ancestor. In one test, the yeast and human interaction networks were aligned using MaWish, identifying 151 modules. The identified modules were compared to MIPS complexes of size 3-25, and the reported MIPS coverage was 20%.

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. At the heart of the CAPPI algorithm lies reconstruction of the conserved ancestral PPI network. First, it reconstructs the hypothetical sequence of evolutionary events (duplications, deletions and speciations), by building gene trees for the input proteins. The proteins in the input networks are split up into families using the MCL clustering method. The phylogenetic history of each family is then calculated using the Neighbour Joining method implemented in the PHYLIP package (Felsenstein, 1989) and the gene trees are then reconciled with the species tree to give the sequence of evolutionary events. The posterior probabilities of interaction between proteins are then determined at each stage of evolution up to and including the common ancestor network. The probability of protein interaction is calculated under the proposed stochastic model of network evolution. The topology of the ancestral network (and each network at every stage of evolution) is determined by the most probable interactions. The ancestral network is finally projected back onto the input networks to determine the alignment. Compared to previous approaches the proposed framework provides more insight into the protein network evolution by explicitly modelling the ancestral history of the input networks. Its parameters are also tied to the evolutionary events shaping the network. In a comparison with NetworkBlast, CAPPI identified a lower number of conserved modules when aligning the yeast, worm and fly networks but the results were more functionally pure.

DOMAIN (Guo and Hartemink, 2009) is the first algorithm to introduce protein domains into the network alignment problem and uses a novel direct-edge-alignment paradigm to detect equivalent interaction pairs across species. DOMAIN does not explicitly restrict its attention to putatively homologous proteins. Instead, it directly aligns PPIs across species by decomposing PPIs in terms of their constituent domain-domain interactions (DDIs) and looking for conservation of these DDIs. DOMAIN was used to identify conserved protein complexes in the yeast-fly and yeast-worm protein interaction networks, and was shown to exhibit comparable performance to MaWISH and NetworkBlast on most performance metrics.

1.4.6.2 Global network alignment methods

While the detection of conserved functional modules is better approached as a local network alignment problem, some network alignment methods proposed recently have also started to focus on the global network alignment problem. These include Graemlin 2.0 (Flannick et al., 2008), IsoRank (Singh et al., 2008) and IsoRankN (Liao et al., 2009), and GNA (Zaslavskiy et al., 2009).

Graemlin 2.0 is a new global multiple network aligner with (1) a novel scoring function that can use arbitrary features of a multiple network alignment, such as protein deletions, protein duplications, protein mutations, and interaction losses; (2) a parameter learning algorithm that uses a training set of known network alignments to learn parameters for the scoring function and thereby adapt it to any set of networks; and (3) an algorithm that uses the scoring function to find approximate multiple network alignments in linear time. IsoRankN is also a global multiple network alignment algorithm based on the IsoRank algorithm (for pairwise alignment) that simultaneously uses both PPI network data and sequence similarity data in an eigenvalue-based framework to compute network alignments, the relative weight of the two data sources being a free parameter. IsoRank uses spectral graph theory to first find pairwise alignment scores across all pairs of input networks. These pairwise scores, computed by spectral clustering on the product graph, work well in capturing both the topological similarity as well sequence similarity between nodes of the networks. However, to find multiple network alignments, IsoRank uses these scores in a time-intensive greedy algorithm whereas IsoRankN uses a different method of spectral clustering on the induced graph of pairwise alignment scores. GNA formulates alignment as two different graph matching problems: The constrained GNA formulation corresponds to a situation where one has a strong a priori about which pairs of nodes can be matched. In the balanced GNA problem, the binary constraints on which pairs are allowed are replaced by a more global objective function that balances the matching of similar proteins with the conservation of interactions, with a parameter to smoothly control the trade-off between these two contradictory goals. In both formulations, GNA was found to outperform IsoRank in terms of its ability to disambiguate orthologs between the Yeast and Human networks. Direct comparisons of GNA with IsoRankN and Gramlin 2.0 have not been carried out,

though IsoRankN was shown to provide better (in terms of number of nodes aligned) mappings compared to Graemlin 2.0 on a test set of five eukaryotic networks (Liao et al., 2009). It should be noted that the global network alignment methods discussed above do not directly address the conserved module detection problem. Instead, they focus on finding the best node-to-node match across the entire networks, which can help identify functional orthologs across species.

1.5 Overview

This introductory chapter, provides the background and context regarding protein interaction data and computational analysis of protein interaction networks. The rest of this thesis is organized as below:

Chapter 2

In this chapter I discuss a functional similarity measure based on the Gene Ontology to map similar proteins across species. I show that using this measure instead of the traditional sequence based Blast scores improves network alignment results in terms of the functional coherence and overlap with experimentally verified protein complexes. Exploiting the fact that the results from functional similarity-based network alignment display little overlap (<15%) with sequence similarity-based alignment, I develop a combined approach that integrates sequence and function-based network alignment with graph clustering concepts to substantially increase alignment coverage. Keeping in view the high error rates and incompleteness in interaction datasets I also investigate simultaneous clustering of multiple networks as an alternative to network alternative that relaxes strict topological constraints on graph comparisons.

Chapter 3

In this chapter I address the issue of low conservation detected through the alignment of current interaction networks of model organisms. 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. In this chapter, I test this hypothesis using models of network evolution and also investigate the role of error in the results

of network alignment. I devise a distance metric based on graph summary statistics to assess the fit between experimental and simulated network alignments. The results indicate that network evolution alone is unlikely to account for the poor quality alignments given by real data. I then compare several error models in their ability to explain this discrepancy and estimate error rate parameters using approximate Bayesian computation. My estimates of false negative rates vary from 20 to 60% dependent on whether incomplete proteome sampling is taken into account or not. I also find 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.

Chapter 4

Here I focus on the comparative analysis of a subset of the interaction network related to mitosis between Yeast, Human and Fly. Manual ordering of the GO labels within this category allows me to study temporal aspects of the mitotic interaction network. To alleviate sparse temporal labelling of proteins, I use a Markov random field approach to infer temporal labels for proteins that connect two or more annotated proteins. The resulting interaction networks exhibit strong temporal clustering with most interactions among proteins with similar time labels. I then investigate the conservation of the mitotic network within the three species using sequence similarity based network alignment. Surprisingly, I found that sequence similarity completely fails to detect any conservation between the two model organisms despite the sets of mitotic proteins being functionally very similar. I then investigate the relationship between sequence and function and the implications for network alignment studies.

Chapter 5

This chapter concludes the dissertation and summarizes my findings. I also touch upon some possible future research directions relevant to the application and extension of the work presented here.

Chapter 2

Functionally guided network alignment

Note: Work presented in this chapter has been published:

> Ali W and Deane CM, 2009. Functionally guided alignment of protein interaction networks for module detection. *Bioinformatics* **25**(23), 3166-3173.

2.1 Introduction

The aim of protein interaction network alignment is to identify subgraphs that are conserved across species and hence likely to be functionally important modules. A potential disadvantage of network alignment for functional module detection is that despite its success in identifying conserved modules in multiple species, it offers limited coverage compared to graph clustering methods (as single network clustering detects all potential modules and not just those conserved in multiple species). It is also highly dependent on the graph topology for correct results, thus error rates pose a special challenge. This is a critical issue due to the unusually high percentage of false positive and false negative interactions in current networks. Recent estimates have put these numbers as high as 70 and 90% (Hart et al., 2006; Saeed and Deane, 2008), respectively. A common theme in previous studies of protein interaction network alignment has been 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.

In this chapter, we explore 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, we examine the use of functional similarity of proteins across networks to aid alignment. We use a simple functional similarity-based measure to carry out network alignment that increases by more than 30%, the number of conserved interactions found. The modules found using our measure display 15% higher functional coherence on average compared to sequence-based alignment. Module detection was carried out purely through alignment of functionally similar proteins across species. Specifically, functional similarity of proteins within a species was not used to guide module detection. We also investigate the benefits of combining network alignment with expression data and clustering measures to identify larger modules. The combined method improves the coverage of experimentally verified complex datasets by nearly 200% compared to either sequence or function-based network alignment alone. We also present a novel representation that attempts to perform simultaneous clustering of multiple networks constrained by the similarity links between them. This method accounts for the errors in interaction data by entirely relaxing the restrictions on the conservation of module topology and can therefore identify both conserved and non-conserved modules in multiple networks at the same time.

I conclude the chapter with an analysis of the protein sets participating in conserved interactions that led to the observation that they are significantly enriched for essentiality and overwhelmingly involved in metabolism-related processes. These observations support the view that essential genes ought to be conserved across a wide range of organisms.

2.2 Methods

2.2.1 Functional similarity score

In order to be of use 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 (Ashburner et al., 2000). GO offers substantial coverage of major protein databases and provides a structured set of terms describing gene products. The ontology covers three domains:

- Cellular component: The parts of a cell or its extracellular environment.
- Molecular function: The elemental activities of a gene product at the molecular level, such as binding or catalysis.
- Biological process: The operations or sets of molecular events with a defined beginning and end, pertinent to the functioning of integrated living units: cells, tissues, organs, and organisms.

Each GO term within the ontology has a term name, which may be a word or string of words; a unique alphanumeric identifier; a definition with cited sources; and a namespace indicating the domain to which it belongs. GO is structured as a directed acyclic graph, and each term has defined relationships to one or more other terms in the same domain, and sometimes to other domains (Figure 2.1). The GO vocabulary is designed to be species-neutral, and includes terms applicable to prokaryotes and eukaryotes, single and multicellular organisms.

We devised a simple measure of functional similarity which is based on the most specific and hence most informative GO annotation of each protein. We focus here only on the biological process domain of GO as it is the most densely annotated. The method is equally applicable to the molecular function and cellular component domains. 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. We define a multi-set of size n as a pair (S, σ) where $\sigma : S \to \mathbb{N}$, with the

Figure 2.1: Hierarchical structure of the Gene Ontology under the biological process domain (http://www.yeastgenome.org/help/GO.html).

conditions:

$$S = \bigcup_{N} S_i \qquad , \qquad \sum_{y \in s} \sigma(y) = n \tag{2.1}$$

Here, σ 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 ($funsim$) is then calculated as follows:

$$funsim(A, B) = max\left(1 - \frac{\sigma(t)}{n}\right) \qquad , \qquad t \in \{S_A \cap S_B\} \tag{2.2}$$

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;

Wang et al., 2007). We compared our score to that proposed by Wang et al. (2007) and found the lists of functionally similar proteins generated in both cases to be in good agreement.

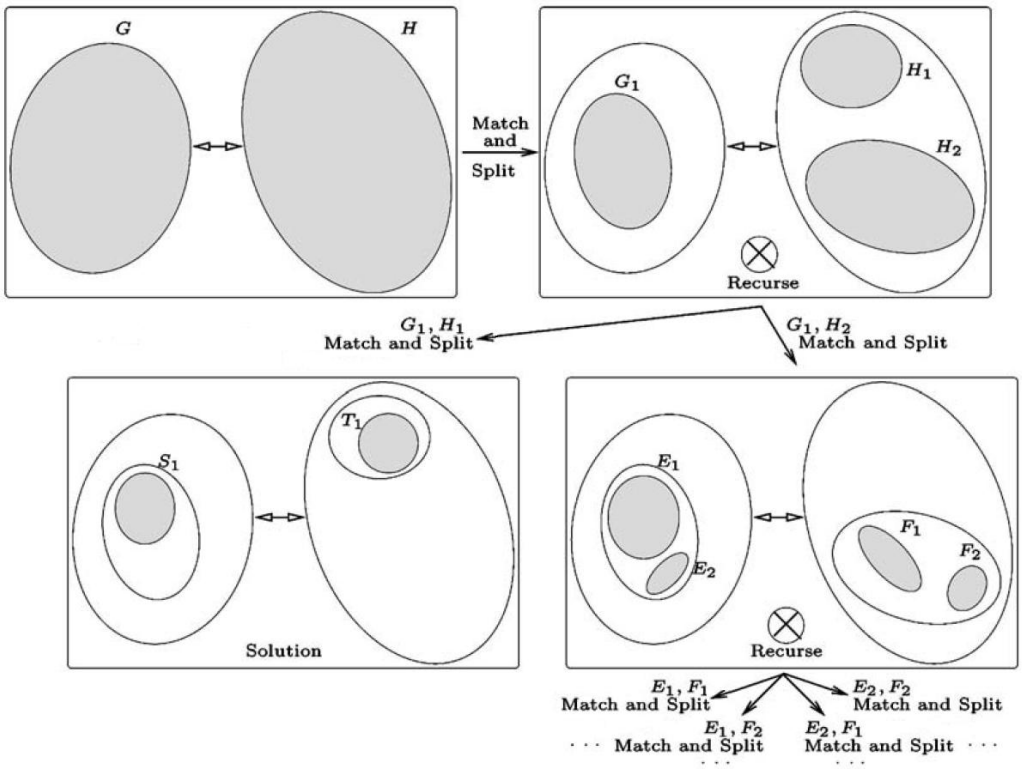
2.2.2 Alignment algorithm

We used the Match-and-Split (MAS) network alignment algorithm (Narayanan and Karp, 2007) to compare alignment results based on sequence similarity to our functional similarity measure (Section 2.2.1). Instead of explicitly creating an alignment graph, MAS uses a recursive process that alternately identifies locally matching nodes across two networks and then splits the matching sub-graphs into connected components (Figure 2.2). Nodes are deemed locally matching if they share sequence similarity as well as network neighbourhood. In the case of multiple orthologs for a node, the orthologs are aligned independently of each other. They can therefore be part of different sub-graphs in the alignment. MAS is relatively fast and uses a flexible node similarity component that uses BLAST (Altschul et al., 1997) E-values in the original implementation. We modified it to use our functional similarity scores. A cutoff score of 0.9 was used to select highly similar proteins. We found that this particular choice of cutoff identifies functional matches for a significant proportion of proteins across two species whilst keeping multiple hits within reasonable limits.

Algorithm 1 Match-and-Split

INPUTS: Graphs G and H
OUTPUTS: Set of maximally matching subgraph pairs

[Match] Compute induced subgraph:
G' of G over the locally matching nodes $lm(G, H)$, and
H' of H over the locally matching nodes $lm(H, G)$.
[Split] Find connected components:
$G_1, ..., G_c$ of G', and
$H_1, ..., H_d$ of H'.
[Recurse]
if $c = 1$, $d = 1$ and $G' = G$,$H' = H$ **then**
    Output the maximal solution G,H. [base case]
else
    for $i = 1$ to c; $j = 1$ to d **do**
        Match-and-Split (G_i, H_j). [recursive case]
    end for
end if



Narayanan M and Karp RM (2007), 'Comparing protein interaction networks via a graph matchand-split algorithm', Journal of computational biology: a journal of computational molecular cell biology 14(7), 892-907. Reprinted with permission from Mary Ann Liebert Inc. Publishers.

Figure 2.2: The match and split algorithm for graph matching (Narayanan and Karp, 2007). The input graphs are recursively split into smaller, locally matching sub-graphs.

2.2.3 Combining function and sequence

As discussed in the results section, even though function-based network alignment substantially improves upon the level of conservation detected by sequence-based alignments, module size and number of detected modules is still quite low. This led us to the development of an extended algorithm which combines high quality sequence and function-based alignment results with common graph clustering measures to identify larger modules in a network (Figure 2.3). Briefly, a given query-species network is pair-wise aligned separately to multiple networks and each edge in the query network is scored based on its conservation in other networks. A seed set of edges is then identified from the weighted interaction network, where the weights are based on the degree of edge conservation and other complementary measures. Modules are then generated from this set in a greedy fashion. The following sections detail the various scores devised to weight each edge in the the network and how their linear combination is used by the module

expansion algorithm.

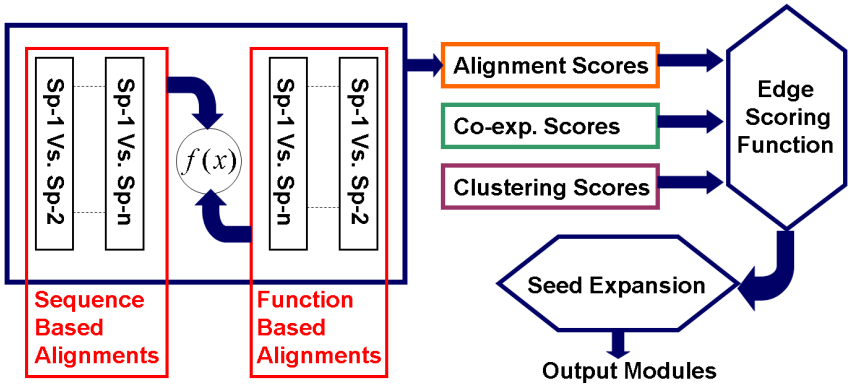


Figure 2.3: Module detection by combining sequence and function based alignment results with expression data and graph clustering measures..

2.2.3.1 Alignment based edge score

Intuitively, an interaction that is conserved across many species should have a high score. We align the query network with each of the other input networks separately using MAS. For each edge in the query network a track is kept of the number of alignments, x, in which it was found to be conserved. The alignment-based score is found through a logistic scoring function,

$$Alignment\ Score\ (AS) = \frac{k}{k + ce^{-tx}} \tag{2.3}$$

This choice is motivated by the initial exponential growth of the function followed by saturation. This models the requirement that the score increases rapidly initially as an edge is found to be conserved in multiple species and then should slowly approach a limiting value with increasing evidence of conservation. Here, the parameters k, c and t can be adjusted to set an upper limit on the score as well as the saturation point. In our implementation these have been set to 1, 20 and 2, respectively. This choice limits the alignment score to a maximum of 1 and also ensures that the score saturates when an edge is found to be conserved in two species (as we carried out our tests on pair-wise alignments).

We align the networks using both sequence and our functional similarity measures. Each edge will therefore have two alignment scores assigned to it, AS_{seq} (for sequence-based align-

ment) and AS_{func} (for function-based alignment).

2.2.3.2 Graph-based edge score

Our graph-based score is constructed from two commonly used graph statistics. The clustering coefficient is a local network measure of how close a vertex and its neighbours are to being a clique. Consider a selected node i in a network, having k_i neighbours. The value of the clustering coefficient of the node i is given by the ratio between the number of edges E_i that actually exist between these k_i nodes and the total number $k_i(k_i-1)/2$ of such edges that could exist in the neighbourhood of i:

$$C_i = \frac{2E_i}{k_i(k_i-1)} \tag{2.4}$$

As the clustering coefficient is a node-based score and our algorithm is based on edge scoring, we assigned to each edge the average clustering coefficient of its endpoints, so that an edge which links two nodes with high clustering coefficients is more likely to be within a dense cluster.

The betweenness of an edge is defined as the number of shortest paths between pairs of nodes that run along it (Girvan and Newman, 2002). If a network contains groups or communities that are only loosely connected by a few inter-group edges then all the shortest paths between different communities must go along one of these edges. Thus, the edges connecting communities will have high betweenness while edges inside the clusters would tend to have lower betweenness.

To favour edges that have both a high clustering coefficient and low betweenness, we use the product of normalized edge betweenness values (NB) and edge clustering coefficients (C) to calculate the graph-based edge score:

$$Graph\ Score\ (GS) = C(1-NB) \tag{2.5}$$

2.2.3.3 Co-expression-based edge score

Co-expression alone is not necessarily an effective measure of co-membership in a module, though it is still a useful indicator of biological coordination between proteins. Combined with the other measures presented above, it may contribute to better module detection. Co-

expression data for Human proteins was obtained from COXPRESdb (Obayashi et al., 2008). We use the Pearson correlation coefficient values (ranging from -1 to 1) as the Co-expression Score (CS). Where no co-expression data is available, a CS of 0 is assigned.

2.2.3.4 Combined edge score and module expansion

The four different scores for each edge in the network are finally integrated through a weighted linear combination:

$$Edge\ Score = \alpha AS_{seq} + \beta AS_{func} + \gamma GS + \delta CS \tag{2.6}$$

The weights $(\alpha, \beta, \gamma, \delta)$ can be adjusted to assign relative importance to the different scores, depending upon the confidence level attributed to the corresponding data sources and the type of results sought. In our implementation we assigned the set of weights $(\alpha = 9, \beta = 7, \gamma = 2, \delta = 1)$ based on performance optimization on a test set of 100 randomly selected experimentally detected Human complexes (performance criteria given in 2.7 and 2.8). This set of complexes was subsequently removed from the validation set. To test the robustness of the combined score, weights were inferred from the Human-Yeast analysis and the same set of values was then used for all subsequent analyses, including Human-Fly, Fly-Yeast and Human-Worm comparisons. We found that using weights optimized for one species can be used for the analysis of other species with little effect on the results. This can of course be better tested as data for more species becomes available. After edge-weighting, modules are extracted from the network using the highest-scoring edges (Algorithm 2). First, a seed set is selected consisting of edges with scores $\geq 3\mu$, μ being the average edge score of the network. These seeds are then expanded stepwise into modules. At each step, the highest scoring neighbour of a seed is added to the module if it does not decrease the average score, S, of the module (averaged over all edge scores in the module) by more than a user defined threshold, t. We carried out all our tests with t set to 0.75. The algorithm terminates when no more edges can be added. At this point, the input network is divided into a set of modules which potentially correspond to real biological complexes.

Algorithm 2 Module expansion

INPUTS: Weighted graph $G = (V, E)$, Expansion threshold t
OUTPUTS: Set of modules M

$\mu \leftarrow \frac{\sum_{e \in E} weight(e)}{|E|}$
$M \leftarrow$ initial set of modules with edge weight $\geq 3\mu$
for each module m in M **do**
  repeat
    for each edge $e \in neighbors(m)$ **do**
      if $\frac{weight(m)+weight(e)}{|m+1|} > \frac{t \cdot weight(m)}{|m|}$ **then**
        $m.add(e)$
        for each $k \in \{M - m\}$ **do**
          if $k \cap m > 0$ **then**
            $merge(k, m)$
          end if
        end for
      end if
    end for
  until no more edges added to m
end for

2.2.4 Simultaneous clustering

An inherent issue with alignment-based methods for module detection is low coverage due to emphasis only on conserved modules along with sensitivity to errors in network topologies. To tackle these problems we introduce the concept of simultaneous clustering. A basic implementation of this concept takes as input multiple interaction networks along with the similarity relationships between proteins from different species. A global graph is then built with all the nodes present in the input networks and two types of links: inter- and intra-species edges. We stress here that the global graph is different from the alignment graph used in network alignment algorithms. In particular, the alignment graph is a product of the input species networks where each node is a merged representation of orthology relationships. In contrast, the global graph does not involve merging of nodes and edges and all orthologs are represented individually. The alignment task can then be reduced to a clustering of this global graph based on edge density. Proteins in a species that are highly connected to each other as well as to a highly connected group of proteins in another species through similarity links would be clustered together and identified as meta-clusters (Figure 2.4). These meta-clusters represent putative conserved modules.

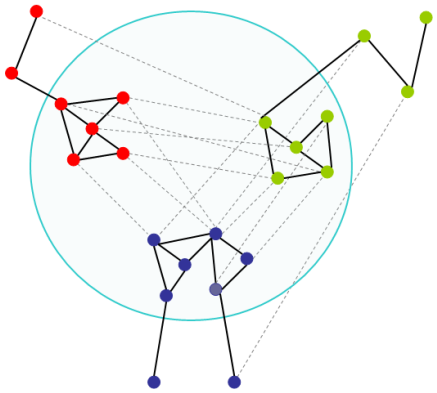


Figure 2.4: A meta-cluster within the global graph composed of networks of multiple species (node colours differentiate between species). The meta-clusters are characterized by relatively higher number of intra-species links (protein interactions, bold lines) as well as high number of inter-species links (orthology relationships, dotted lines).

The crucial difference to network alignment is that to be part of the same meta-cluster, clusters from different species need not be very similar in edge topology: they only need to be well-connected within as well as with each other. Moreover, unlike network alignment, module detection using this technique is not only limited to conserved regions. Dense regions in the interaction network of one species that do not have sufficient similarity links to any other species will still be clustered into modules. In this case the meta-cluster would only contain proteins from that particular species. For the clustering process, any of the myriad of already available algorithms can be used, provided they can deal with the presence of two different types of links in the global graph. We used an implementation of the popular clustering algorithm MCL to carry out our tests, where the inter- and intra-species links are differentiated by their weights (All intra-species links were weighted as 1.00 as no information was available for interaction confidence. The inter-species links were weighted with the functional similarity values ranging from 0.00 to 1.00).

2.2.5 Data sources

The species selected for analysis were *H.sapiens* (Human), *S.cerevisiae* (Yeast), *D.melanogaster* (Fly) and *C.elegans* (Worm). Interaction data for Yeast (4941 proteins, 17 387 interactions), Fly (6701 proteins, 20 092 interactions) and Worm (2328 proteins, 3495 interactions) was down-

loaded from the Database of Interacting Proteins or DIP (Xenarios et al., 2002a), while data for Human (9305 proteins, 35 458 interactions) was taken from the Human Protein Reference Database or HPRD (Keshava Prasad et al., 2009a). All datasets were downloaded in June 2008.

2.2.6 Testing criteria

The modules recovered were analyzed in terms of their functional coherence and compared to experimentally determined complexes to assess their quality. We define the functional homogeneity of a module M, as the average functional similarity of all possible protein pairs (i, j) in the module. Pairs for which functional similarity could not be calculated due to lack of GO annotation for one or both proteins were not included.

$$Functional\ Homogeneity = \frac{1}{|M|} \sum_{i,j \in M} funsim(i, j) \tag{2.7}$$

To check whether the extracted modules corresponded to real complexes we compared them to high quality datasets containing experimentally identified complexes. Experimentally determined complexes for Yeast and Human were downloaded from MIPS CYGD (Güldener et al., 2005) and MIPS CORUM (Ruepp et al., 2008) databases, respectively. For each MIPS complex (A), we identify its best matching module in the solutions (M) as the one having the greatest value of the following comparison score:

$$MIPS\ Comparison\ Score = \frac{|A \cap M|}{|M|} \tag{2.8}$$

In addition to the module-specific MIPS comparison score, we also calculated the MIPS coverage (the proportion of proteins in MIPS complexes that are also in the output modules) and MIPS accuracy (the proportion of proteins in the output modules that are present in MIPS complexes).

We compared the performance of function and sequence-based network alignment by aligning the Human network with Yeast and Fly networks. Function-based alignment was carried out using MAS only, while sequence-based alignment was done using MAS, MaWISH and NetworkBlast. For sequence-based alignment, the orthology file was generated by selecting pairs

of proteins with BLAST E-values $\leq 10^{-7}$ following the cut-off used by Sharan et al. (2005). The combined method was tested by using alignment results from MAS (both sequence and function based) as input. To observe the effect of multiple networks, the combined method was executed using sets of two (Human-Yeast, Human-Fly, Fly-Yeast and Human-Worm) as well as three (Human-Yeast-Fly) networks. In both 2-way and 3-way analysis, one network is chosen as the query network, aligned separately with each of the other networks, and the resulting edge conservation information is then fed into the alignment scoring function and module expansion algorithm as described in Section 2.2.3.4

In addition to alignment-based methods, we also compared our combined method and simultaneous clustering method to the more commonly used technique of single network clustering. This was done by weighting the edges in the network with GO functional similarity of interacting proteins and applying MCL, one of the most popular network clustering algorithms. In the results section, we refer to this method as GO clustering.

2.3 Results

2.3.1 Human: aligned to Yeast

Here we discuss results for the Human network aligned with Yeast. Detailed results including the Human-Fly and other comparisons can be found in subsequent sections.

Function-based alignment using our similarity score was successful in uncovering a larger number of proteins participating in conserved interactions than sequence-based alignment. As shown in Table 2.2, the number of conserved modules discovered in the Human network increased from 74 (spanning 457 unique proteins) to 94 (spanning 727 unique proteins). Moreover, the two sets share only 58 proteins (<15%), indicating that the modules targeted by the two methods are nearly disjoint. We successfully exploit this observation in our combined method.

In addition to greater coverage, modules identified using function-based alignment displayed higher biological coherence than sequence-based alignment using any of the other methods (Figure 2.5a). Almost 50% of the modules identified by our technique scored 0.5 or more compared to only 30% when using sequence-based MAS; MaWISH and NetworkBLAST both

Table 2.1: List of methods discussed in the results section.

Method	Description
Sequence based	Pair-wise alignment of two networks using sequence similarity as the node mapping criteria and MAS as the alignment algorithm
Function based	Pair-wise alignment of two networks using functional similarity as the node mapping criteria and MAS as the alignment algorithm
Combined	Our combined method incorporating network alignment results with expression data and clustering measures. The query species is aligned to one other species and the alignment results are used to extract modules from the query species using 2.6.
Combined 3-way	In this case the query species is aligned separately to two other species and the alignment results are used to extract modules from the query species using 2.6.
MaWISH	Pairwise alignment of two networks using sequence similarity as the node mapping criteria and MaWISH as the alignment algorithm
NWBlast	Pairwise alignment of two networks using sequence similarity as the node mapping criteria and MaWISH as the alignment algorithm
Simultaneous	Simultaneous clustering of two networks via the MCL algorithm using both protein interactions and sequence similarity links
GO clustering	Clustering of a single weighted network using MCL. Interactions are weighted by the functional similarity of interacting proteins.

performed far worse. As illustrated in Figure 2.5b, function-based alignment also identifies modules that correlate better with experimentally verified complexes. Around one-third of the modules correspond well (overlap >50%) with a MIPS complex as opposed to only one-tenth of the modules from MaWISH and NetworkBlast.

Using the combined method, in the Human-Yeast case a total of 303 modules (1479 proteins) were identified, far higher than either sequence or function-based alignment alone. This increased overage does not affect the module quality as both the functional homogeneity and overlap with MIPS complexes is still better than sequence-based alignment methods. An even greater number of modules are found when the Fly network is also added for a three-way analysis (Table 2.2, combined 3-way), accompanied by an increase in MIPS coverage.

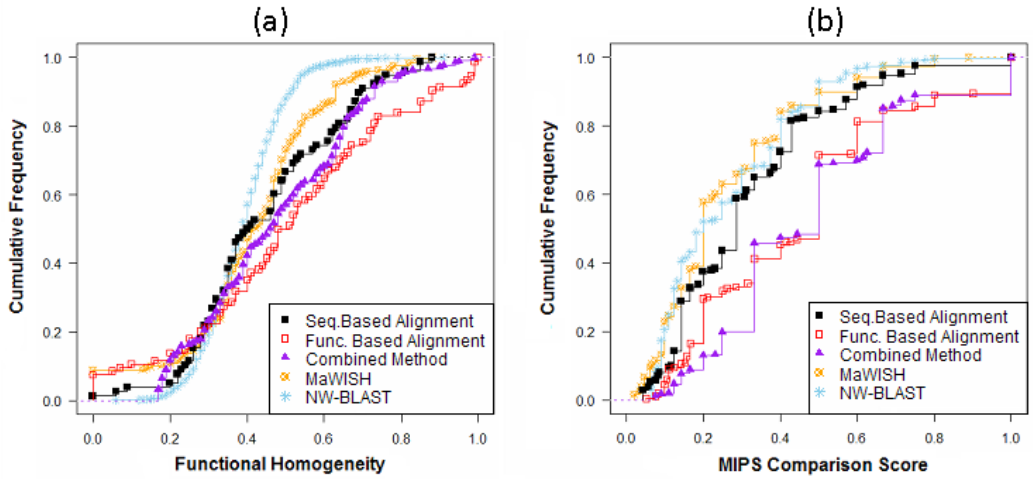


Figure 2.5: Comparison of methods on modules extracted from the Human network (aligned to Yeast): cumulative frequency distribution of (a) functional homogeneity and (b) MIPS comparison scores. (Results for the simultaneous clustering method not included.) Plots shifted towards right (higher values of functional homogeneity and MIPS score) indicate better results.

Figure 2.6 plots the coverage and accuracy of the various methods in terms of the total number of MIPS interactions covered by the modules detected by the various methods. The simultaneous clustering approach exhibits the highest coverage although its accuracy is relatively low. This is probably because this method clusters the entire network, instead of identifying only the conserved regions. This would extract many modules that are not yet present in MIPS and thus drive down the accuracy of the method. Our combined method offers a superior coverage-accuracy trade-off amongst all techniques. The coverage of MIPS interactions using this approach is better than any of the other alignment-based methods, accompanied by a higher accuracy than MaWISH and NetworkBlast. Finally, GO clustering of single networks performs worse than all methods based on multiple networks. Specifically, this method suffers a 5-fold drop in accuracy compared to our combined method (with three networks) for a marginal increase in coverage. Upon detailed inspection of the results we found that while some of the modules extracted using this technique are highly functionally homogeneous (by construction), this comes at the cost of a substantial number of spurious clusters. These results also support the view that module detection based on evidence from multiple species leads to more reliable, though fewer results.

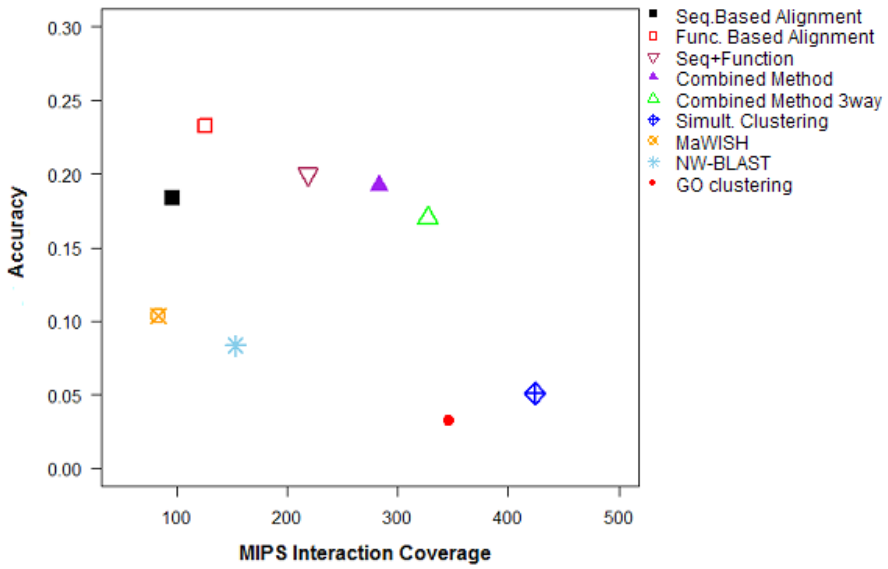


Figure 2.6: Coverage of interactions in MIPS plotted against accuracy for each method (Human network aligned to Yeast). The combined method clearly offers the best coverage-accuracy tradeoff.

Table 2.2 also shows the execution time for each of the methods on a machine with a 1.66 GHz processor with 1 GB RAM. Further analysis of other species indicated that the run times of the alignment algorithms are extremely sensitive to the number of orthologs between two species. The simultaneous clustering approach is an exception to this, as it does not experience an explosion in the number of possible alignments due to multiple possible orthologs, experienced by MAS, NetworkBlast and MaWISH.

2.3.1.1 Effect of the linear combination parameters

The parameters for the linear combination in 2.6 control the relative weighting of each component (sequence based alignment, function based alignment, co-expression and clustering score). As mentioned earlier in Section 2.2.3.4, these parameters were learnt from the Human network and used subsequently in all other comparisons. To better understand the effect of the weighting given to sequence and function based alignment, we also carried out tests by changing the relative weighting of these two components (while keeping the other two constant). The results are visualized in Figure 2.7. In general, assigning higher weights to the function based alignment component improves both the functional homogeneity of output modules and their

Table 2.2: Comparison of alignment methods: Human network (aligned to Yeast).

Method	# Modules	# Proteins	MIPS-c	MIPS-a	FH	Time (s)
Sequence based	74	457	96	0.18	0.36	3061
Function based	94	727	126	0.24	0.51	5432
Combined	303	1479	283	0.21	0.43	21
MaWISH	242	543	83	0.1	0.32	663
NWBlast	2353	894	153	0.08	0.3	68977
Combined 3-way	430	1603	327	0.17	0.40	29
Simultaneous	1197	7371	424	0.05	0.23	369
GO clustering	1093	6663	346	0.03	0.29	112

* MIPS-c is MIPS coverage, MIPS-a is MIPS accuracy and FH is Functional homogeneity.

overlap with experimental complexes.

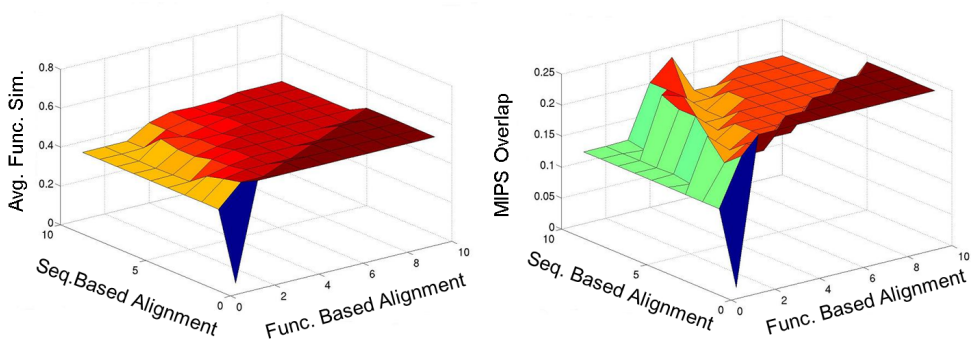


Figure 2.7: Effect of relative weighting of sequence and function based alignment components in our combined method 2.6. On the vertical axes, Avg.Func.Sim. is the average functional similarity of all pairs of interacting proteins in the output modules and MIPS overlap is the average proportion of proteins in each module that overlaps with a MIPS complex. Using either measure, higher quality modules are extracted if higher weights are assigned to the function based alignment component.

2.3.2 An example: the Human DAB complex

The Human DAB (Maldonado et al., 1990) is a multi-protein complex involved in transcription initiation and consists of the TFIID complex, TFIIA complex and TFIIB. It is present in the MIPS database as complex ID 493 which lists 16 proteins as its subunits, 15 of which are transcription initiation factors and 1 is a TATA-box binding protein. It is a typical target of module detection methods both in terms of the number of proteins involved as well as the

tight functional relationships between them. Figure 2.8 shows the performance of our combined approach along with NetworkBlast and MaWish in terms of their ability to correctly identify this complex in the Human network when aligned to the Yeast network.

It must be noted that the Human network from HPRD on which all module detection methods were tested was missing four out of the 16 protein subunits of the DAB complex. We have found this to be a widespread problem with over 30% of MIPS proteins missing from the HPRD network. Circular nodes in the figure represent components of the DAB complex while other nodes represent proteins that were mistakenly classified as part of the same complex by any of the methods. MaWish and NetworkBlast which use only sequence similarity to carry out network alignment manage to capture <50% of the complex. Both these methods capture almost the same set of proteins while missing the rest. This probably indicates an inherent weakness of using just sequence information, rather than the alignment algorithms themselves. Note that the combined method which uses sequence and functional similarity with additional module expansion correctly identifies almost the entire complex except two proteins, one of which is not identified as part of the complex by any of the methods.

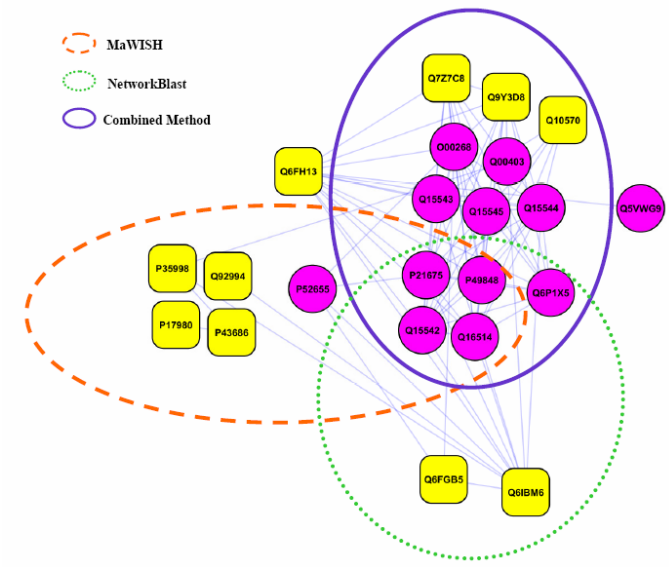


Figure 2.8: Identification of the Human DAB complex (MIPS ID 493) using MaWISH, NetworkBlast and the Combined method. Circular nodes in the figure represent components of the DAB complex while other nodes represent proteins that were mistakenly classified as part of the same complex by any of the methods.

The above example is just one of several cases in which the combined method improves the coverage of real complexes demonstrating the ability of functional mapping to capture conserved interactions independently of sequence similarity. In addition to a comparison of alignment-based methods, this example also demonstrates the benefits of cross-species analysis. When we carried out GO clustering of the single Human network, the DAB complex was detected as part of a much larger cluster consisting of 23 proteins, of which only nine are present in the real complex (results not shown in Figure). Similarly, analysis of the Human Rap1 complex (MIPS complex ID 1204) using the combined method detects all six component proteins, while GO clustering misclassifies three of the components as part of a different cluster. The use of interaction evidence from multiple species can therefore not only identify modules at a higher resolution, but can also detect components missed by single-network clustering.

The above results for the Human network (aligned to Yeast) validate the efficacy of using functional similarity as an alternative to sequence similarity for network alignment. The conserved modules detected using the former are not only larger in size, but also biologically more coherent. Our combined method incorporates alignments using both measures, substantially increasing alignment outputs whilst retaining superior module quality. In the following sections, I discuss results from other species that lend further support to these observations.

2.3.3 Yeast: aligned to Human

Results for the Yeast network (Figures 2.9, 2.10 and Table 2.3) also demonstrate the superiority of function based alignment, which clearly outperforms both MaWISH and NWBlast. This is also the case for overlap with MIPS complexes. Using function and combined methods dramatically improves coverage while slightly improving accuracy at the same time. Overall, modules in the Yeast network exhibit a higher functional coherence for all methods, compared to the Human network. The Yeast modules also match better with experimental complexes, perhaps indicating better quality and more dense annotation for Yeast in the databases.

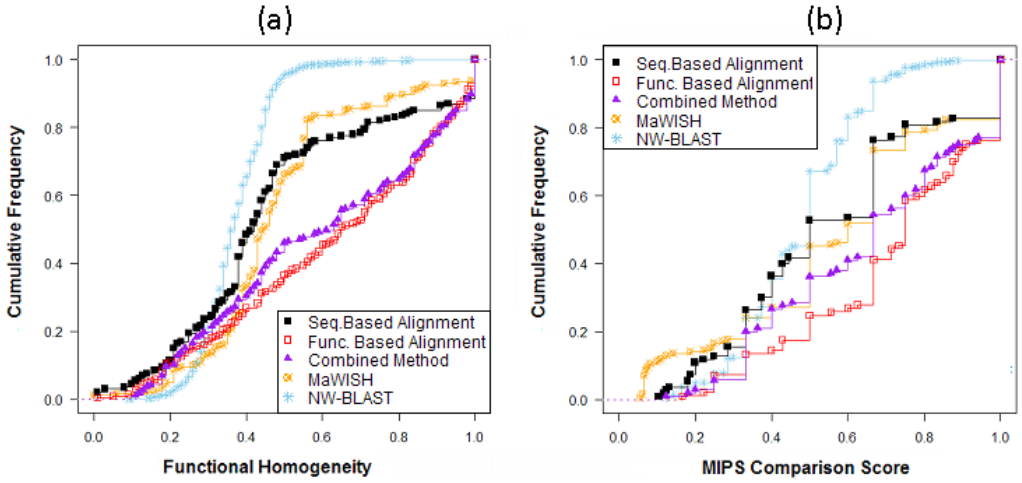


Figure 2.9: Cumulative frequency distribution of (a) functional homogeneity and (b) MIPS comparison scores of modules extracted from the Yeast network (aligned to Human). Modules from the functional similarity based alignment and combined method display far higher functional homogeneity than sequence based alignment methods. Almost 50% of the modules have functional homogeneity scores higher than 0.75 in the former case compared to only around 20% or less in the latter. The results of comparison with MIPS complexes also follow a similar pattern.

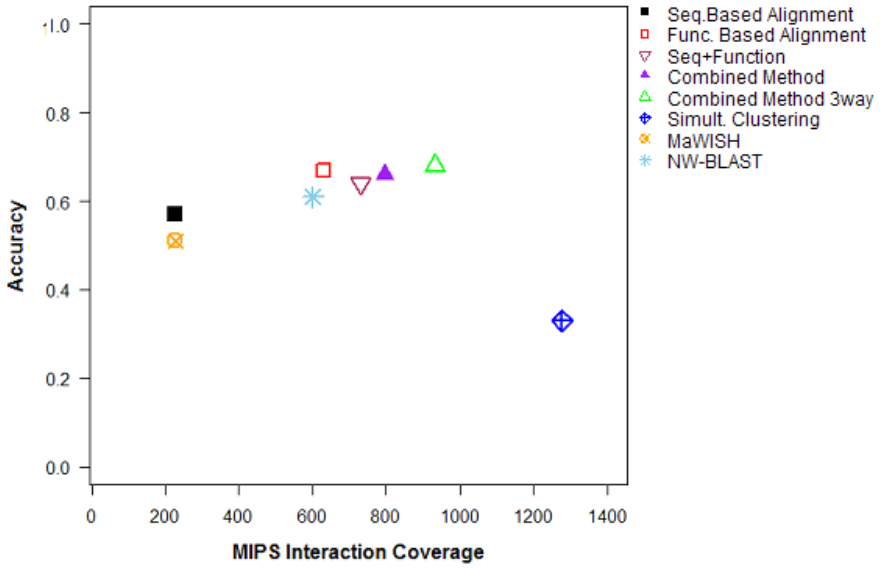


Figure 2.10: Coverage of interactions in MIPS plotted against accuracy for each method (Yeast network aligned to Human). The accuracy of all methods is notably higher in this case compared to the results for the Human network. The combined methods still offer the best coverage than all alignment based approaches.

Table 2.3: Comparison of alignment methods: Yeast network (aligned to Human).

Method	# Modules	# Proteins	MIPS-c	MIPS-a	FH	Time (s)
Sequence based	111	338	226	0.57	0.47	3061
Function based	119	616	629	0.67	0.63	5432
Combined	109	808	795	0.66	0.60	21
MaWISH	242	322	228	0.51	0.48	663
NWBlast	2353	497	601	0.61	0.37	68977
Combined 3-way	143	989	932	0.68	0.55	29
Simultaneous	491	3546	1278	0.33	0.38	369

* MIPS-c is MIPS coverage, MIPS-a is MIPS accuracy and FH is Functional homogeneity.

2.3.4 Human: aligned to Fly

Alignment of the Human network with Fly results in surprisingly little detected conservation (Figures 2.11, 2.12 and Table 2.4), given that the two species are evolutionarily much closer than Human and Yeast. This can be partly explained by the fact that both the Fly and Human interaction datasets are very incomplete relative to the Yeast network. The detected modules in this case also show much lower functional homogeneity than alignment with the Yeast network. Very few of the detected modules in Human overlap well with MIPS complexes and at much lower accuracy values than the Human-Yeast case, strengthening the argument for a bias in database annotations in favour of Yeast.

Table 2.4: Comparison of alignment methods: Human network (aligned to Fly).

Method	# Modules	# Proteins	MIPS-c	MIPS-a	FH	Time (s)
Sequence based	315	638	51	0.08	0.37	2754
Function based	410	708	61	0.12	0.48	4122
Combined	367	1126	194	0.18	0.41	13
MaWISH	412	772	47	0.06	0.36	577
NWBlast	1392	923	73	0.03	0.32	59198
Combined 3-way	430	1603	297	0.17	0.40	17
Simultaneous	1885	4967	117	0.04	0.36	311

* MIPS-c is MIPS coverage, MIPS-a is MIPS accuracy and FH is Functional homogeneity.

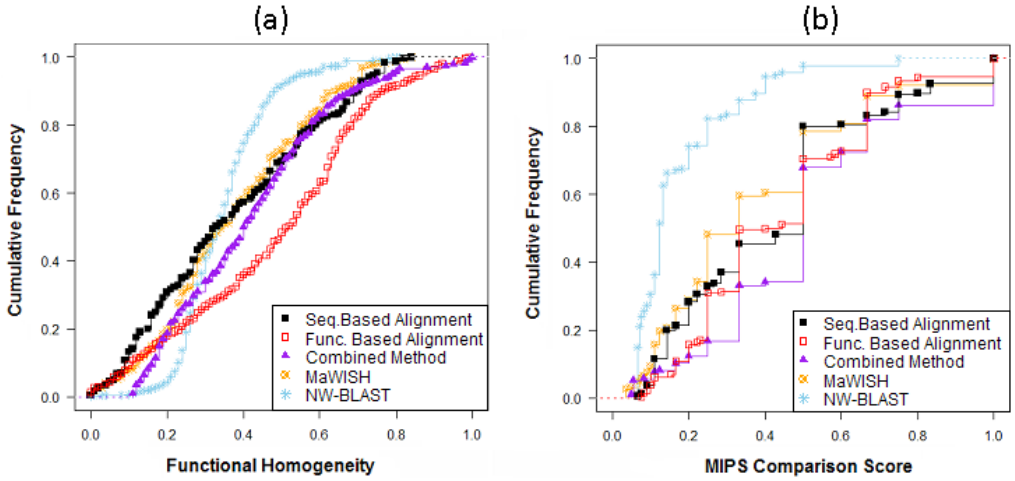


Figure 2.11: Cumulative frequency distribution of (a) functional homogeneity and (b) MIPS comparison scores of modules extracted from the Human network (aligned to Fly). As in the case for Human-Yeast, modules from functional similarity based alignment display the highest functional homogeneity, while the combined method performs best in terms of overlap with MIPS complexes.

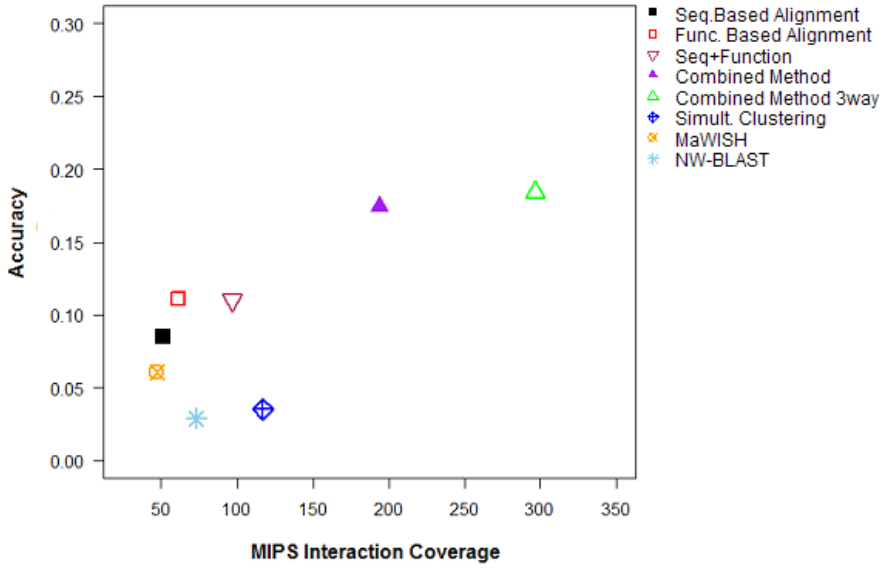


Figure 2.12: Coverage of interactions in MIPS plotted against accuracy for each method (Human network aligned to Fly). The combined method offers almost 300% higher coverage than sequence based alignment, while achieving higher accuracy.

2.3.5 Optimization of parameters

Although our results show that for the combined method, use of the same parameters that were learnt for the Human-Yeast comparison for the analysis of other species leads to good results, we also tested the effect of optimization for specific cases. In Figure 2.13, we present the results for Yeast-Fly comparison, using parameters learnt from 100 randomly selected experimentally verified yeast complexes from the MIPS database. Upon comparison with the Yeast-Fly analysis carried out using same parameters as Human-Yeast, we find that despite slight changes in the learnt parameter values ($\alpha = 8$, $\beta = 8$, $\gamma = 3$, $\delta = 1$), the results largely follow the same pattern.

2.4 What is conserved?

In the previous sections we were considering if functional modules could be extracted from interaction data using network alignment. Here we investigate whether proteins involved in modules that are found conserved across multiple species share any biological properties. One obvious explanation for module conservation could be that the pathways and biological processes carried out by these proteins form the essential backbone of life and thus are expected to be retained through long evolutionary times. This can be explored in two complementary ways: (1) A general GO term enrichment analysis to find out if all the conserved protein sets are highly overrepresented by specific biological process terms, and (2) Given a reasonably well-defined standard set of essential proteins in a species, are the proteins involved in conserved interactions also more likely to be in the essential set?

2.4.1 GO enrichment

We tested the sets of proteins in conserved modules for enrichment of GO terms under the biological process domain. Enrichment analysis was carried out via the GOrilla web tool (Eden et al., 2009) which uses a hypergeometric test with multiple testing correction to identify terms in GO which are found to be significantly overrepresented in a given gene set, relative to a background set (annotation for the entire Yeast proteome in this case). P-values using the

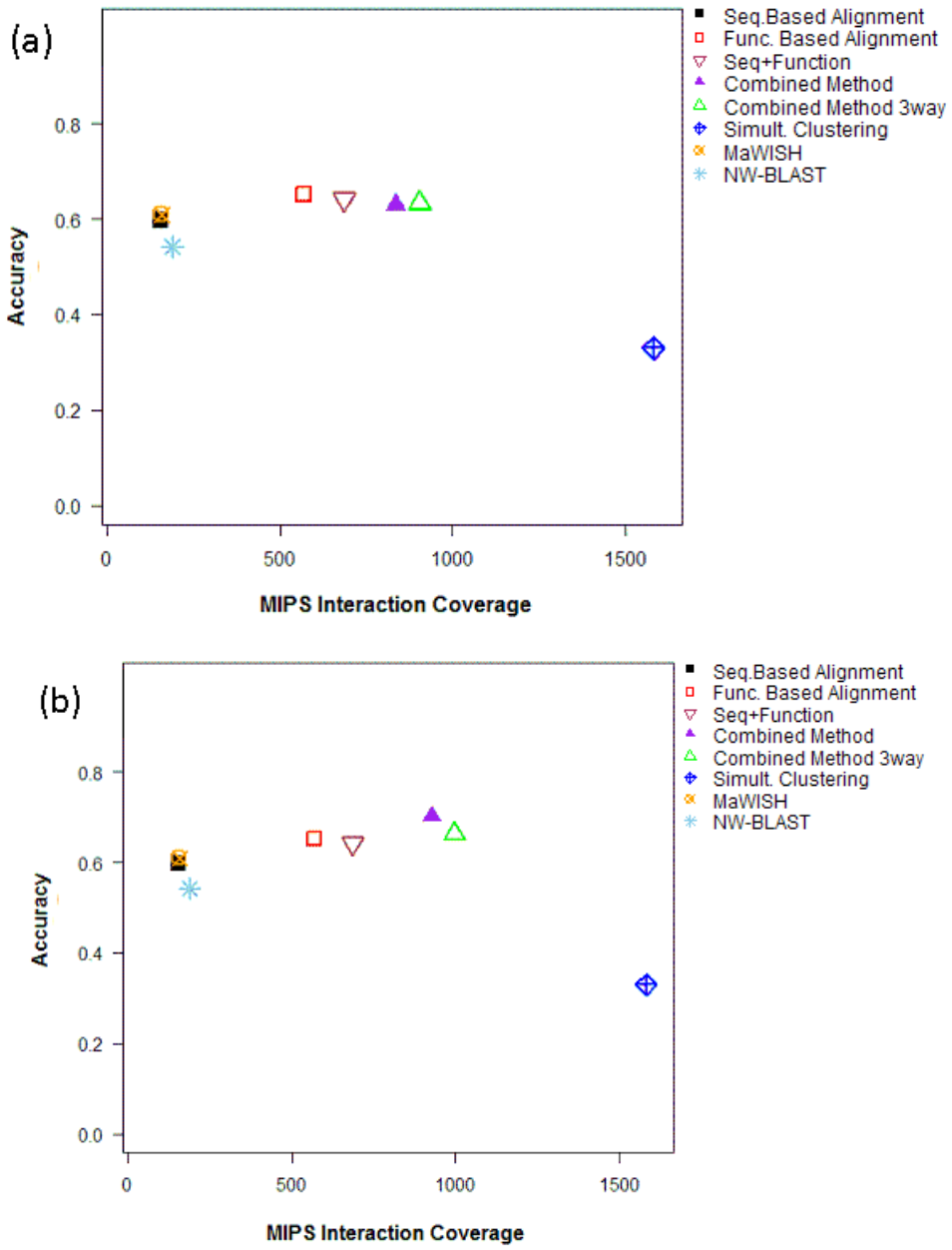


Figure 2.13: Results for Yeast-Fly using (a) non-optimized weights and (b) weights optimized for the Yeast-Fly analysis. The combined method (2-way and 3-way) displays a slight improvement after weight optimization, though the results follow the same overall trend. Note that results for all other methods are the same in both cases as only the combined method uses the weights for the linear combination.

hypergeometric distribution are calculated as follows: If an annotations file contains N genes, a given GO term has M annotated genes, and the user inputs a list of n genes of interest, the probability of seeing k or more genes of interest annotated to a given GO term is,

$$p - value = \sum_{j=k}^{n} \frac{\binom{M}{j}\binom{N-M}{n-j}}{\binom{N}{n}} \tag{2.9}$$

In Figure 2.14, we plot the most significant enriched terms (p-value < 0.01) at GO level 2 for each set. All sets are enriched in key processes such as the cell cycle, cell death and ageing in particular, while there are also categories specific to each set. For example, the sets containing Yeast proteins with conserved interactions in Human are enriched for the reproductive developmental process, which is not the case for the Yeast proteins with conserved interactions in Fly.

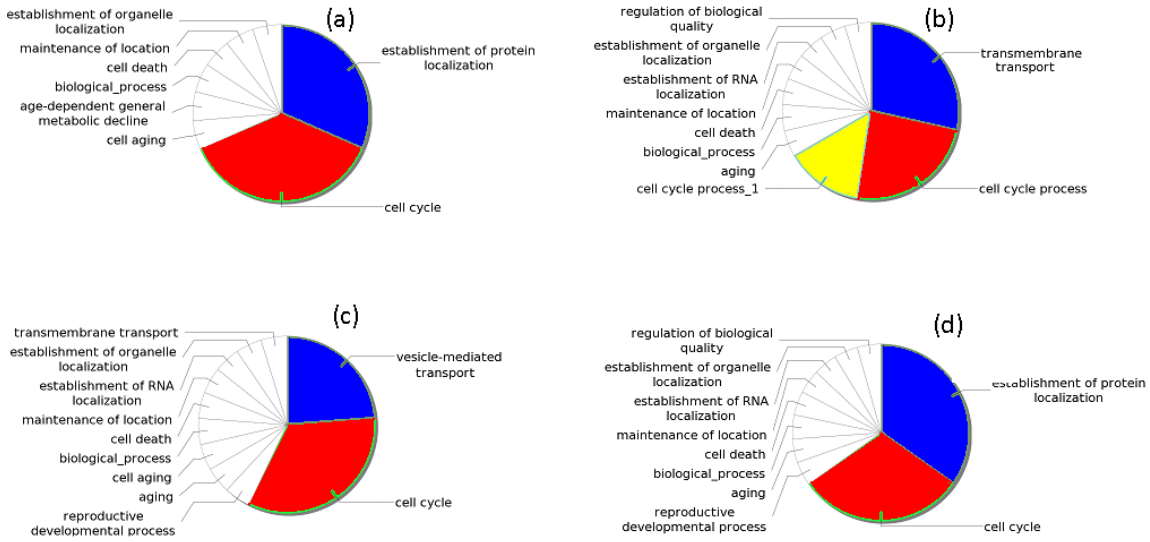


Figure 2.14: Level 2 Gene Ontology terms enriched in the sets of proteins with conserved interactions found by network alignment : (a) Yeast (aligned to Human using sequence similarity), (b) Yeast (aligned to Human using functional similarity), (c) Human (aligned to Yeast using sequence similarity) and (d) Human (aligned to Yeast using functional similarity).

As GO level 2 is perhaps too general to spot detailed differences and commonalities in the sets, we provide more complete lists of enriched terms in the four sets at all GO levels in Appendix A. While the terms span quite a broad spectrum of biological activity in the cell

and are thus not amenable to general conclusions, the over-abundance of metabolism-related proteins is worth commenting on and not altogether surprising. As this is one of the most critical processes responsible for life, significant conservation in model organisms would be expected.

2.4.2 Essentiality

We also tested whether proteins taking part in conserved interactions are more likely to be essential. The list of essential genes was downloaded from the Yeast deletion project at Stanford (Winzeler et al., 1999). Out of a total of 1135 essential proteins, 955 were present in the DIP Yeast network used for our alignment studies. One-sided hypergeometric tests were carried out in R (R Development Core Team, 2011) to check for over-representation of essential proteins in the sets of Yeast proteins taking part in conserved interactions (found by aligning Yeast with the Fly and Human networks). Table 2.5 lists the number of essential genes in each set and the associated p-values. All sets are significantly enriched for essential proteins and the sets from function-based alignments show lower p-values.

Table 2.5: Empirical distribution of essential proteins.

Yeast protein set	# Proteins	# Essential proteins	p-value
H-Seq.based*	338	118	8.93e-13
H-Func.based	616	275	6.7e-55
F-Seq.based	284	119	5.08e-20
F-Func.based	432	206	1.41e-45

* H-Seq.based and H-Func.based are the sets of Yeast proteins found in conserved modules detected from the Yeast-Human network alignment using sequence and function similarity measures respectively. F-Seq.based and F-Func.based are the sets of Yeast proteins found in conserved modules detected from the Yeast-Fly network alignment using sequence and function similarity measures respectively.

Fig 2.15 shows the empirical densities for the number of essential genes in 100,000 randomly selected sub-sets (equal in size to the conserved sets) from the complete Yeast network. The actual number of essential genes in the conserved sets are indicated by red marks for comparison.

This observed enrichment in essential genes fits well with the conservation of their interactions in model species; it is to be expected that genes (and their interactions) that are essential for life would tend to be conserved between even evolutionarily distant organisms. However, it is pertinent to note here that various studies in the literature have explored correlations

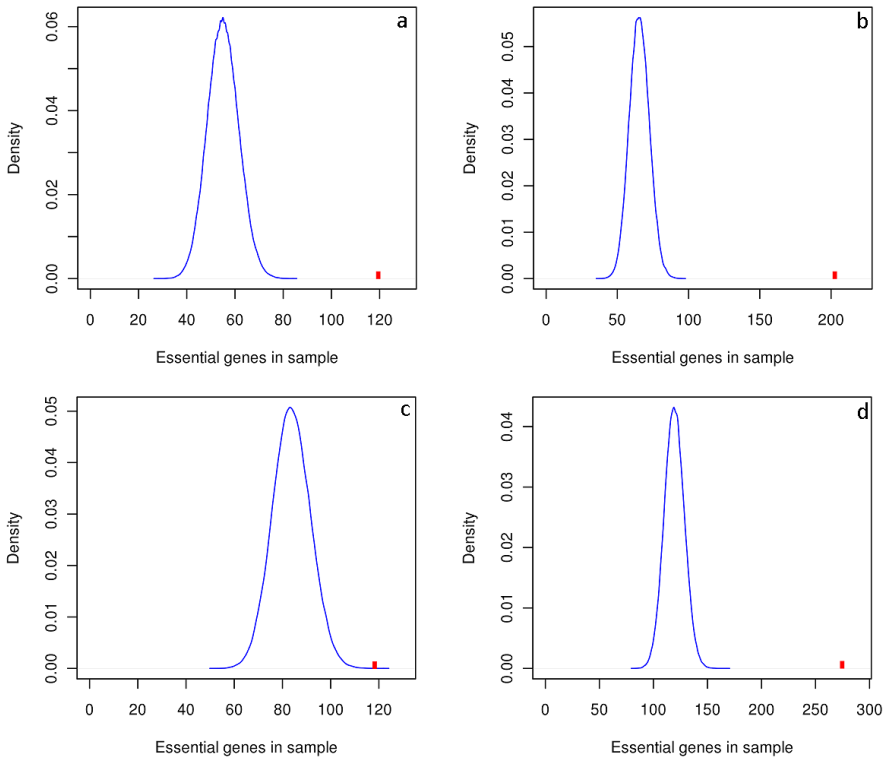


Figure 2.15: Distribution of the number of essential genes in 100,000 samples taken from the complete Yeast proteome. The sample sizes are equal in size to the conserved protein sets in Table 2.5. The actual number of essential proteins in each set is marked in red. All sets contain a significantly higher number of essential proteins than expected at random.

between gene essentiality and network properties of the corresponding nodes. That highly connected nodes in a protein interaction network tend to be more essential is a well-known (albeit contested) relationship between network topology and gene essentiality (Batada et al., 2006a; Jeong et al., 2001; Yu et al., 2007; Zotenko et al., 2008). This centrality-lethality rule states that essential proteins are likely to have a high degree of centrality, where centrality of a node in a network can be defined in a number of ways. The apparent enrichment of conserved modules in essential genes may therefore be caused by other confounding factors. Moreover, the repertoire of known essential genes is almost certainly incomplete. Combined with the substantial gaps in the current interaction datasets that we used for our analysis, this enrichment of essential genes may well be an artefact of noisy data.

2.5 Conclusions

Previous to our work, all protein network alignment studies used sequence similar proteins across species to aid the network comparison process. Here, we demonstrate our method that uses a quantitative measure of functional similarity to align protein interaction networks. Our results indicate that modules found by alignment using functional similarity exhibit higher functional coherence compared to sequence similarity-based alignment. This is encouraging because functionally coherent modules are more likely to be biologically relevant. These observations were further confirmed by the comparison of identified modules to experimentally determined complexes in the MIPS database. The modules found using our functional similarity score also displayed higher levels of overlap with real complexes. Given that <15% of proteins were common in the modules from the functional similarity and sequence similarity-based alignments, a question that arises naturally is whether using both techniques simultaneously can increase the power of computational complex detection.

Our combined method that uses network alignment based on both function and sequence similarity led to several improvements in the module detection results. First, the combined approach produced better results in terms of agreement with experimental datasets. The coverage of MIPS was more than twice that of using sequence-based alignment alone. In terms of the functional coherence of the detected modules, the combined method performs far better than sequence-based alignment. Adding simple clustering measures from graph theoretic methods and gene co-expression information improves the results further by increasing the size of the solution set. While these two measures alone are not powerful enough to produce high quality results, they can be used to expand the solution sets of alignment-based methods and thus increase their coverage. Finally, the weighted combination of different techniques in our method provides a natural way of optimizing the results for a particular measure of goodness. Modules with high functional coherence can be produced by assigning a relatively high weight to the functional similarity-based alignment component while higher weights for the graph-based component will identify larger modules.

Results for our simultaneous clustering-based alignment method are less conclusive. The coverage of MIPS is naturally much higher in this case, though the modules are not of com-

parable quality to the other approaches. This could be a consequence of the larger sample size, more likely to contain highly connected sub-graphs with no biological relevance. Furthermore, not all real modules are expected to be completely functionally homogeneous (Spirin and Mirny, 2003). Still, developments to our method are possible with the potential to improve the results. We differentiated between inter- and intra-species links by assigning them different weights whereas they might need to be treated entirely differently, for instance as a bi-partite graph. Also, currently all networks in the global graph are treated at the same level. One way forward could be to take a more evolutionary realistic approach and assign relative ordering to the protein orthology links based on how evolutionary distant the respective species are.

In conclusion, we have demonstrated that using function as a metric for protein network alignment offers improved performance over traditional sequence-based network comparisons. The two measures manage to identify an almost disjoint set of conserved interactions which indicates that network alignment methods may benefit by exploiting still other ways of mapping similar proteins across species. We have also simultaneously clustered entire networks from several species using both protein similarity and interaction links as constraints. This method offers far greater coverage than any network alignment approach and fewer restrictions on module topology making it more suitable for error-prone data.

Chapter 3

Modelling the effects of evolution on network alignment

Note: Work presented in this chapter has been published:

> 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.

3.1 Introduction

In the previous chapter I discussed the notion of protein network alignment in detail and compared results from several algorithms including joint alignment and clustering. Despite increasingly sophisticated algorithms, including the use of functional annotation to aid alignment, results of such studies indicate low conservation at network level given the relatively high similarity at genome level. Traditionally, this has mostly been attributed to the potentially high rate of error in the interaction data-sets. This is in sharp contrast to the sequence alignment field, where the data is generally of very high quality and the deficiencies, if any, are probably on the algorithmic side.

In the following sections of this chapter, I discuss an indirect method of error rate estimation in interaction datasets that exploits the mismatch between the levels of observed and expected

conservation in the interaction networks of model species. We used commonly accepted network growth models to evolve pairs of networks from a single ancestor and aligned these simulated networks to give explicit estimates of the best possible expected performance (Figure 3.1).

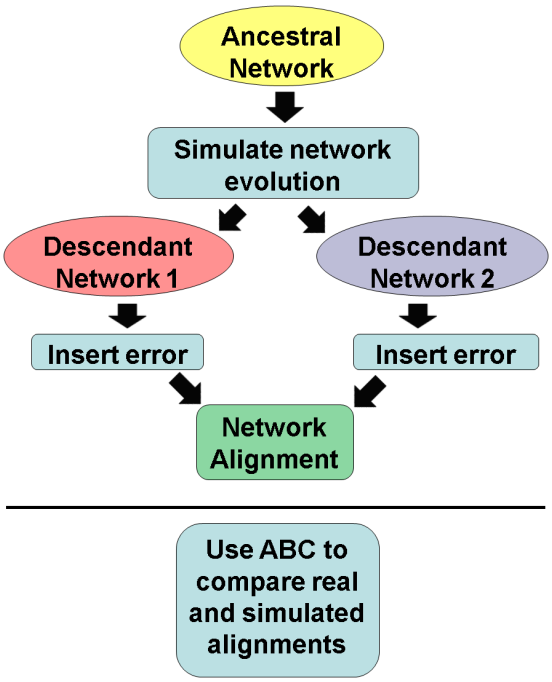


Figure 3.1: Flowchart of our error estimation method. Error estimation using ABC is carried out based on the agreement between alignments obtained from experimental networks and those obtained from networks simulated via network evolution models.

We found that these alignments of simulated networks are on average far larger than those observed in real networks which indicates that network evolution is unlikely to be the sole cause of low conservation. These results are subsequently used in calculating the extent of the role played by error in alignment results. Our use of alignment comparisons based on summary statistics enables us to quantitatively judge the performance of competing error models and to estimate model parameters. Simple random error models seem to perform best with error estimates of around 50-60%. These estimates are substantially lowered to around 15-20% when we account for the fact that current networks contain only a fraction of the proteins encoded by the entire genome. We also observe that the incompleteness of experimental data-sets is the major factor affecting network alignments while the presence of false positives even at very

high rates seems to have little effect.

3.2 Methods

3.2.1 Datasets

The interaction datasets used in this study were downloaded from HPRD for Human (9130 proteins, 33170 interactions), and DIP for Fly (7522 proteins, 22013 interactions) and Yeast (4987 proteins, 19531 interactions). Interaction datasets for two bacterial species *E.coli* (2920 proteins, 13117 interactions) and *C.jejuni* (1344 proteins, 12143 interactions) were downloaded from the IntAct database (Aranda et al., 2009). All interaction data was extracted in June 2010.

Before detailing the components of our network alignment based error estimation method, in Section 3.2.2 I discuss a few existing interaction scoring/verification methods which can be used to assess the level of error in a given dataset.

3.2.2 Interaction network verification methods

One way of quantifying error in existing experimental datasets is by using interaction verification or confidence scoring methods. Such methods typically exploit either the topological properties of the networks or known biological properties of individual proteins to identify putative interactions that have a high probability of being true. Thresholds on some scoring function are usually employed to filter away low-probability interactions. A few of these methods are briefly described below along with a discussion in the results section regarding their performance on the DIP Yeast interaction dataset.

The interaction generality measure (IG) was proposed by Saito et al. (2002) and can be used to computationally assess the reliability of interaction data using only a list of protein-protein interactions. The definition of interaction generality is based on the idea that there are some sticky proteins which seem to interact with many other proteins and that most of these interactions may not be physiologically important. In particular, in Yeast two-hybrid assays some proteins appear to activate transcription of a reporter gene without actually interacting with their partners, a situation that can lead to an excess number of candidate partners (some

of which are erroneous) for a single protein. Therefore interaction generality is defined as the number of proteins that directly interact with the target protein pair. The authors claim that interactions with low generalities are predominantly physiologically meaningful and eliminating high generality interactions may help create more reliable networks. They showed that by eliminating high generality interactions from a combined Yeast interaction dataset extracted from MIPS, Ito et al. (2001) and Uetz et al. (2000), the rate of interactions with common cellular roles increased from 63% to 79% in the refined networks.

IRAP (Chen et al., 2005) is another method for assessing the reliability of protein interactions based on the underlying topology of the network. A candidate PPI is considered to be reliable if it is involved in a closed loop in which the alternative path of interactions between the two interacting proteins is strong (path strength is defined as the product of edge weights along the path). A reliable PPI is accompanied by at least one reliable alternative interaction path in the underlying interaction network. An algorithm called AlternativePathFinder is used to compute the IRAP value for each interaction in a complex PPI network. By testing the method on a Yeast interaction dataset containing experimentally reproducible and non-reproducible interactions, the authors showed that unreliable (non-reproducible) experimental interactions can be filtered away by choosing higher IRAP thresholds.

Deng et al. (2002) approached the problem of protein interaction verification and prediction by studying the large-scale conserved patterns of interactions between protein domains. Using evolutionarily conserved domains defined in the PFAM database (Finn et al., 2010), they applied a maximum likelihood estimation method to infer interacting domains that are consistent with the observed protein-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 also predicted interactions between proteins and tested for overlap with 2575 experimentally detected Yeast protein-protein interactions from MIPS, demonstrating 100 fold better prediction compared to random interaction assignment. This method can also be used to validate existing interactions along with predicting new interactions.

STRING is a database and web resource dedicated to protein-protein interactions, including both physical and functional interactions. The database is a meta-resource that aggregates

most of the available information on protein-protein associations, scores and weights it, and augments it with predicted interactions, as well as with the results of automatic literature-mining searches. The basic interaction unit in STRING is the functional association, which is defined in this database as a specific and meaningful interaction between two proteins that jointly contribute to the same functional process. STRING contains a unique scoring-framework based on benchmarks of the different types of associations against a common reference set, integrated in a single confidence score per prediction. The database currently covers 2,590,259 proteins from 630 organisms.

3.2.3 Network evolution models

Several network evolution models were introduced in Chapter 1 (Section 1.3.5). These models in general aim to mimic the underlying biological mechanisms in real systems and try to reproduce some specific properties of empirical datasets. For protein interaction networks, the duplication divergence model is biologically most plausible as it is rooted in the process of gene duplication and subsequent neofunctionalization. In order to model network growth, we used a duplication-divergence model (DD) including the process of hetero-dimerization. In hetero-dimerization, a protein is initially self-interacting and the propensity for this interaction is preserved after duplication and divergence. This model has been shown to capture both the degree distribution and relatively high clustering values (measured by the clustering coefficient) of experimental interaction networks (Ispolatov et al., 2005b). The DD model we implemented possesses a basic mechanism of gene duplication at each time step followed by asymmetric divergence.

- *Duplication: a node in the network is chosen randomly and its duplicate is introduced and connected to each neighbour of the original node.*
- *Divergence (i): each link emanating from the duplicate is retained with probability* σ_d
- *Divergence (ii): each link emanating from the original node is retained with probability* σ_t
- *Heterodimerization: the original node and its duplicate are linked with probability* δ.

The parameter values were set to: σ_d=0.3, σ_t=0.7 and δ=0.05. These values were chosen in such a manner that the final evolved networks closely match the size of experimental data-sets.

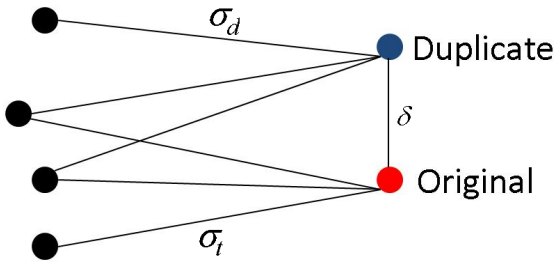


Figure 3.2: Duplication divergence model for protein network evolution with heterodimerization. The original and duplicate nodes retain their links to other nodes with probabilities σ_t and σ_d respectively, while a link between them is inserted with probability δ.

Specifically for Yeast-Human alignment analysis, the simulated networks were grown from a common ancestor (see Section 3.2.4) till the number of nodes match the number of proteins in DIP Yeast and HPRD Human interaction data-sets. The number of links in the simulated networks were then found to be within 5% of the interactions in the data-sets. The fidelity of the model and parameters were further demonstrated by the observation that the average degree and clustering coefficient of the simulated networks also closely match the data-sets (see Table 3.1). We found that addition of the heterodimerization process in the basic duplication divergence model was crucial for the close agreement in the average clustering coefficients of experimental and simulated networks.

For further validation of our results, we also tested another model of network evolution proposed recently by N. Przulj and Hayes (2010). In this model of geometric network evolution, called Geo-GD expansion, the proteins of the initial network are first randomly embedded inside a hypersphere in low-dimensional space. Following this, a duplication is modelled by placing the duplicate node in a uniformly randomly chosen direction and at a uniformly random distance up to a maximum of 2ϵ from the parent node where ϵ is a distance threshold. The value of ϵ is chosen based on the number of links required in the initial network. Once all duplications have been simulated by inserting new nodes, new links are generated between nodes if the distance between them is less than ϵ (Figure 3.3). The crucial parameters of this model are the dimensionality of the metric space and the distribution of nodes in that space. Intuitively, each protein can be described with its biochemical properties and therefore proteins reside in some multidimensional biochemical space. The further the duplicate is moved away from its parent,

the more different their biochemical properties are. This model is found to fit real networks at least as well as the DD and several other competing models. In our implementation, we used geometric expansion in 3-dimensional Euclidean space (Geo3d).

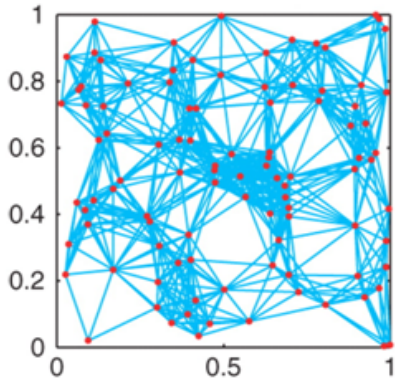


Figure 3.3: Geometric model of network growth (Higham et al., 2008). This figure illustrates a two-dimensional geometric model, with the nodes embedded randomly in the 2-d space. New nodes are embedded close to parent nodes in a random direction. Links between nodes are added based on proximity in terms of Euclidean distance.

Table 3.1: Summary statistics for experimental Yeast network from DIP and simulated Yeast networks using the DD and Geo3d models.

Network	# Nodes	# Edges	Avg. degree	Avg. clust.coeff.
DIP Yeast	4987	19531	6.86	0.113
Simulated-DD	4987	20031	7.02	0.117
Simulated-Geo3d	4987	19114	6.97	0.108

3.2.4 Ancestral network

Network evolution simulation starts with an initial ancestral network. Here we create a representation of the network for the last common ancestor for Yeast and Human. First, all potential orthologs between Human and Yeast were identified using bi-directional Blast hits with E-values less than 10^{-10}. The sub-graph induced by these proteins in the DIP Yeast network was then selected as the putative ancestral network (1836 nodes, 4761 edges). The reciprocal case was also tested by using a sub-graph of the HPRD Human network as the ancestral network (1962 nodes, 4317 edges). An alternative strategy for ancestral network creation could be the use

of conserved interactions (interologs) between Yeast and Human. However a straightforward enumeration of interologs in current data-sets leads to a highly restricted ancestral network (192 nodes, 276 edges). This would be a severe underestimate of the real ancestral network size due to incomplete coverage of the two species and also does not account for divergent evolution after speciation. We therefore proceeded with the former strategy. As the aim of this analysis is only to model the effect of evolutionary mechanisms on network alignment, the exact topology of the ancestral network is not expected to be significant. This was validated by tests indicating that our results are robust to random permutations of the ancestral network (see Section 3.3.6). The method for ancestral network creation described above for the Human-Yeast pair was also used for the analysis of other pairs of species. For the Yeast-Fly analysis, the ancestral network was created from Yeast (sub-graph induced by Yeast proteins with Fly orthologs), while for the Human-Fly analysis the ancestral network was created from Human.

3.2.5 Evolution of orthology

To align two networks, a list of orthology relationships between their nodes (proteins) is also needed. We therefore supplemented our network evolution model with a mechanism of orthology evolution that preserves orthology up to one duplication step (Figure 3.4). Suppose we start with an ancestral network with node set $\{A, B, ...\}$ that undergoes speciation to form two descendant networks nw_1 and nw_2. Just after speciation, there is a one-to-one correspondence between the proteins in the two descendant networks: i-e, if the two networks have node sets $\{A_1, B_1, ...\}$ and $\{A_2, B_2, ...\}$, the orthology relationships are simply $\{A_1 \leftrightarrow A_2, B_1 \leftrightarrow B_2, ...\}$. Now assuming that protein A_1 gets duplicated to A_1d in nw_1 and protein A_2 gets duplicated to A_2d in nw_2. Then two new orthology links $A_1d \leftrightarrow A_2$ and $A_2d \leftrightarrow A_1$ will arise alongside the original link $A_1 \leftrightarrow A_2$. However the two duplicates will not be orthologs of each other and subsequent duplications of these duplicates will not lead to any new orthology links.

Since the number of orthologs between two species can significantly impact network alignment results, it is important that the simulated and real networks possess approximately the same number of orthologs. We find that using our mechanism of orthology evolution mirrors both the actual number and distribution of orthologs between Yeast and Human found through

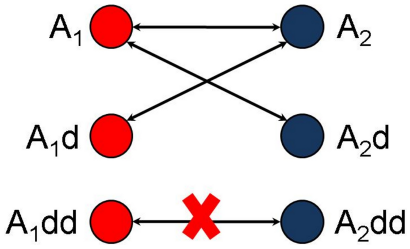


Figure 3.4: Mechanism of orthology evolution implemented in our model. New orthology relationships are created only up to a single duplication step. After initial speciation A_1 and A_2 are orthologs as indicated by the arrow. If A_1 and A_2 are duplicated to A_1d and A_2d, the two duplicates are not orthologs of one another but A_1 is orthologous to A_2d and A_2 is orthologous to A_1d. Further duplications do not lead to new orthologous relationships.

bi-directional Blast hits (Figure 3.5). This ensures that the alignment results are comparable and the simulated networks do not gain an unfair advantage by preserving a much higher number of orthology relationships and consequently, larger alignments.

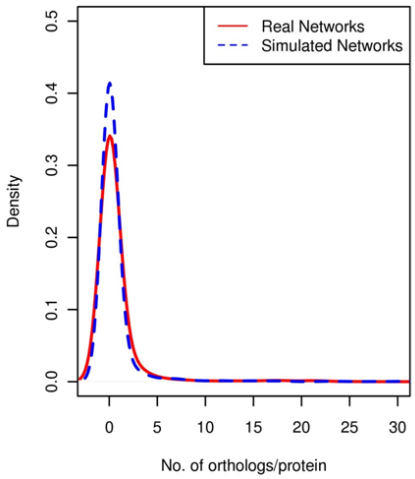


Figure 3.5: Real vs. simulated ortholog distribution: Distribution of orthologs between the real Yeast and Human networks (bidirectional blast E-value $< 10^{-10}$), and simulated Yeast and Human networks using our orthology evolution mechanism.

3.2.6 Alignment method

We chose the MaWISh network alignment protocol due to its superior speed and the fact that it provides an explicit alignment graph as output which we can use for error estimation. Moreover, it requires no additional input except two interaction networks and a list of orthologs. This

was crucial as our analysis is based on aligning simulated networks which contain no other biological information such as protein domains and COG groups (Tatusov et al., 2000) required by some more recent network alignment methods. We also repeated our analysis using NWBlast to avoid any bias due to the alignment protocol used. While the Match-and-Split algorithm discussed in Chapter 2 produces good network alignments and could theoretically be used in place of MaWISH, it was found to be unsuitable for error parameter estimation due to its slow execution speed. Posterior density estimation for error parameters via the approximate Bayesian computation algorithm (see Section 3.2.9) requires a very large number of simulated alignments to be generated which would not have been possible using MAS.

3.2.7 Comparison of real and simulated alignments

In order to quantify how closely the alignments of evolved networks match the alignments of experimental networks, a distance function over alignments is needed. We exploit the fact that the alignments themselves are undirected graphs, and the problem is transformed to one of graph comparison. However, the comparison in this case should not take into account the exact topology of the two alignments, as they will necessarily differ (one of these alignments is from a pair of simulated networks). Instead, we are only interested in relatively coarse statistical properties of the alignments. This observation motivated our choice of a distance function over pairs of alignments, which is essentially a Euclidean distance between vectors in N-dimensional space, where N is the number of graph summary statistics we wish to incorporate in our distance function. Use of summary statistics to compare high dimensional data has been used previously in the literature (Beaumont et al., 2002; Przeworski, 2003). In the context of protein networks for example, Ratmann et al. (2007) used several graph summary statistics including clustering coefficient, fragmentation and within-reach distribution to determine the fit between artificial and real networks while estimating evolutionary parameters of *H.pylori* and $P.falciparum$ networks. The summary statistics used by us are: number of nodes (NN), number of edges (NE), average degree of largest connected component (AD), average clustering coefficient (ACC) and size of largest connected component (LC). We deliberately avoid using any of the more complicated statistics as alignment graphs are relatively small in size and using

advanced measures of graph structure could prove counterproductive. Each alignment is thus represented as a vector of the form (NN, NE, AD, ACC, LC) in 5-dimensional space. Given two vectors $\vec{x}$ and $\vec{y}$ representing two alignments, the distance is defined as,

$$d(\vec{x}, \vec{y}) = \sqrt{\sum_{i=1}^{N} \frac{(x_i - y_i)^2}{\sigma_i^2}} \tag{3.1}$$

where σ_i^2 is the sample variance of $(X_{i1}, ..., X_{ik})$ over k samples. We set $N = 5$ and use $k = 1000$ samples for our estimates. To generate a single sample two descendant networks are evolved independently from a single ancestor and then aligned. Taking a large number of samples provides a view of the average behaviour of the network evolution trajectories.

3.2.8 Uniform error models

We initially assumed a uniform random error model where each positive and negative interaction in the network is equally likely to be mis-reported. Starting with this assumption, we investigate two simple error models that cater for both false positives and false negatives in the data. The edge rewiring error (Figure 3.6a) models the simplest case of a single false negative and single false positive interaction in the data. In this case a randomly selected edge is removed from the network, followed by the addition of an edge between two randomly selected nodes. The model is completely described by a single parameter θ_{rw} which is the fraction of edges rewired in the network.

The node relabelling error (Figure 3.6b) works at the node level instead of the edge level. A pair of proteins is selected and their labels (protein identifiers) are swapped, essentially replacing all interactions of one protein with another. This can potentially introduce multiple false positives/false negatives at each step. Moreover, unlike edge rewiring that maintains exactly the same rate of false positives and false negatives (one at each step), node relabelling leads to unequal rates in general. Like the rewiring error, the model has a single parameter θ_{rl} which is the fraction of nodes relabelled in the network.

The underlying assumption for the single-parameter uniform error models is that existing experimental interaction datasets are of the same size as the (unknown) real networks and are

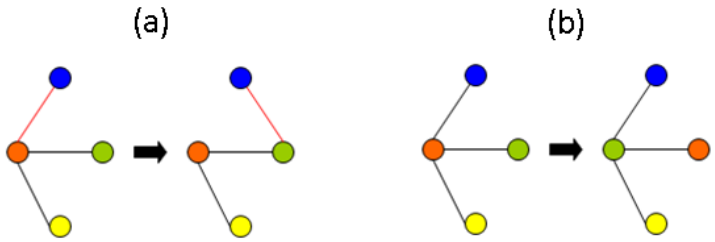


Figure 3.6: Uniform error models: (a) Edge rewiring error and (b) Node relabelling error. The edge rewiring error removes one true interaction (blue-orange) and inserts one spurious interaction (blue-green) at each step. The node relabelling error swaps the identities of two proteins in the network at each step. This can affect multiple edges simultaneously. Swapping green with orange removes two true interactions (blue-orange, yellow-orange) and inserts two spurious interactions (blue-green,yellow-green).

obtained by random edge removal and an equal number of edge insertions. While useful for modeling purposes, it must be pointed out that this assumption may not be very realistic and one possible way of decoupling the false positive and false negative rate would be to introduce two separate parameters in the error model. Additionally, as the false positive and negative rates are not independent of each other, error models with separate parameters for both would perhaps also need to model the relationship between the two rates to be more realistic. In the context of this thesis though, we limit ourselves to the single parameter approach, especially as shown in subsequent sections, false positives do not appear to have a noticeable impact on network alignments.

3.2.9 Estimation of error parameters

Once an error model has been selected, estimation of the parameters for the model proceeds as follows. A pair of networks (NW_1, NW_2) is evolved from a single ancestor until they reach the size of the experimental networks. The error models discussed above are completely specified with a single parameter θ (θ signifies the proportion of edges rewired in the edge rewiring model and the proportion of nodes with swapped label in the relabelling error model) for one network. Since in alignment problems we look at pairs of networks, here we deal with a two-dimensional parameter space (θ^1, θ^2). The problem is reduced to calculating the joint posterior density of the error parameters for the two networks. We calculate the posterior using an approximate Bayesian computation approach (Algorithm 3). In standard Bayesian inference the posterior

distribution for a parameter θ is given by

$$P(\theta|D) \propto P(D|\theta)\pi(\theta). \tag{3.2}$$

Here π is the prior distribution for θ, D are the data, and $P(D|\theta)$ is the likelihood of the data D given the parameter θ. 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 ABC approaches by considering distances between observations and data simulated from a model with parameter θ. Rather than considering the data itself, one considers 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. In our implementation samples are randomly drawn from a 2-D uniform prior and the respective error is put in the evolved networks. These error-filled networks are then aligned using network alignment algorithm and the simulated alignment is compared to the experimental network alignment using the distance function in 3.1. Samples which match experimental alignments within a certain threshold $\delta = 0.001$ are accepted. This small threshold value ensures that only samples with properties very similar to real alignments are accepted. The effect of higher threshold values on the error density estimation is discussed in Section 3.3.5.

Algorithm 3 Error parameters posterior estimation using ABC

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

3.2.10 Non-uniform error models

The node relabelling and rewiring error models described earlier assume a uniform error probability distribution throughout the network. We also implemented several non-uniform error models which attack edges and nodes with high importance for network structure as measured by node degree (deg), clustering coefficient (cc) and edge-betweenness (edgebet). Given a network with N nodes and E edges, nodes were picked for relabelling using the distribution $P(v_i) = \frac{deg(v_i)}{\sum_{j=1}^{N} deg(v_j)}$ for degree based preferential error (pref-relabel). Weighing the probability of a node's selection with its normalized degree ensures that highly connected nodes are more likely to be attacked. Similarly, preferential error based on the clustering co-efficient of nodes (ccremoval) is implemented by the distribution $P(v_i) = \frac{cc(v_i)}{\sum_{j=1}^{N} cc(v_j)}$ to attack highly inter-connected groups of nodes. For preferential edge rewiring, edges are picked based on their importance in maintaining network connectivity. We use edge-betweenness which measures the proportion of all shortest paths passing through a particular edge. Edges are thus picked using the distribution $P(e_i) = \frac{edgebet(e_i)}{\sum_{j=1}^{E} edgebet(e_j)}$.

3.3 Results

3.3.1 Existing verification methods show little agreement

As an investigation into the efficacy of existing interaction verification methods discussed in Section 3.2.2, we start with the Yeast protein interaction network from DIP which is classified into a highly reliable core subset and a larger set of less reliable interactions. Not surprisingly, the STRING weights for interactions in the Yeast DIP-core dataset are much higher than non-core interactions (Figure 3.7).

For each of the interaction datasets (core, non-core and both) we also investigated the relationship between the experimental technique used for interaction detection (Y2H, CoIP or TAP) and the confidence score assigned by STRING to that interaction. Figure 3.8a indicates that for all three datasets, interactions detected using TAP are assigned higher STRING scores. For the core dataset, this difference is not so pronounced and interactions from all three techniques exhibit high scores. Figure 3.8b shows the results broken down by experimental technique and

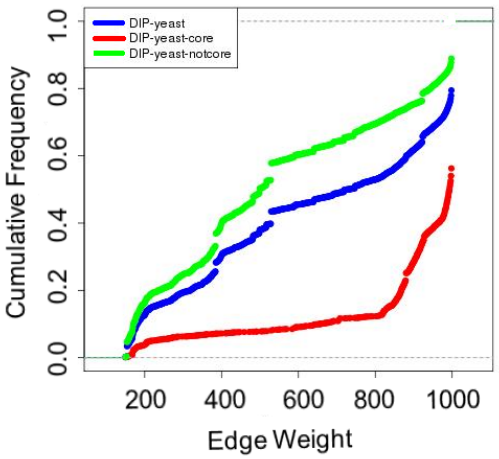


Figure 3.7: Distribution of STRING interaction scores for DIP, DIP-core and DIP-noncore datasets for Yeast.

it is the case that TAP interactions are highly scored, both for core and non-core DIP datasets.

The agreement between STRING scores and the other methods was then tested by verifying the full Yeast DIP dataset using IG, IRAP and Deng's domain-domain interaction method. An interaction was deemed as verified by STRING if it had a score >700, by IG with a generality score < 5, by IRAP with a value > 0.8 and by Deng's method with an interaction probability > 0.8. The thresholds for each method were the same as used in the original papers to filter away unreliable interactions (the threshold for STRING scores was the same as used by von Mering et al. (2005) to identify high confidence interactions). Each interaction was then given an error score based on the number of methods that classified it as not verified (error score is equal to 4 when all four methods flag an interaction as a potential false positive and 0 when all four methods classify it as a true positive). As shown in Figure 3.9, there is little consensus between the different interaction verification methods. Most importantly, very few interactions (< 500) are classified as true positives by all four methods. Any analysis of error detection in interaction datasets using one particular method out of the above is thus fraught with the risk of underestimating the number of false positives whereas using all of them could severely overestimate false positives due to the lack of consensus. Moreover, interaction verification methods such as the above can only detect potential false positives and do not provide any insights into the levels of incompleteness (false negatives) in the interaction datasets.

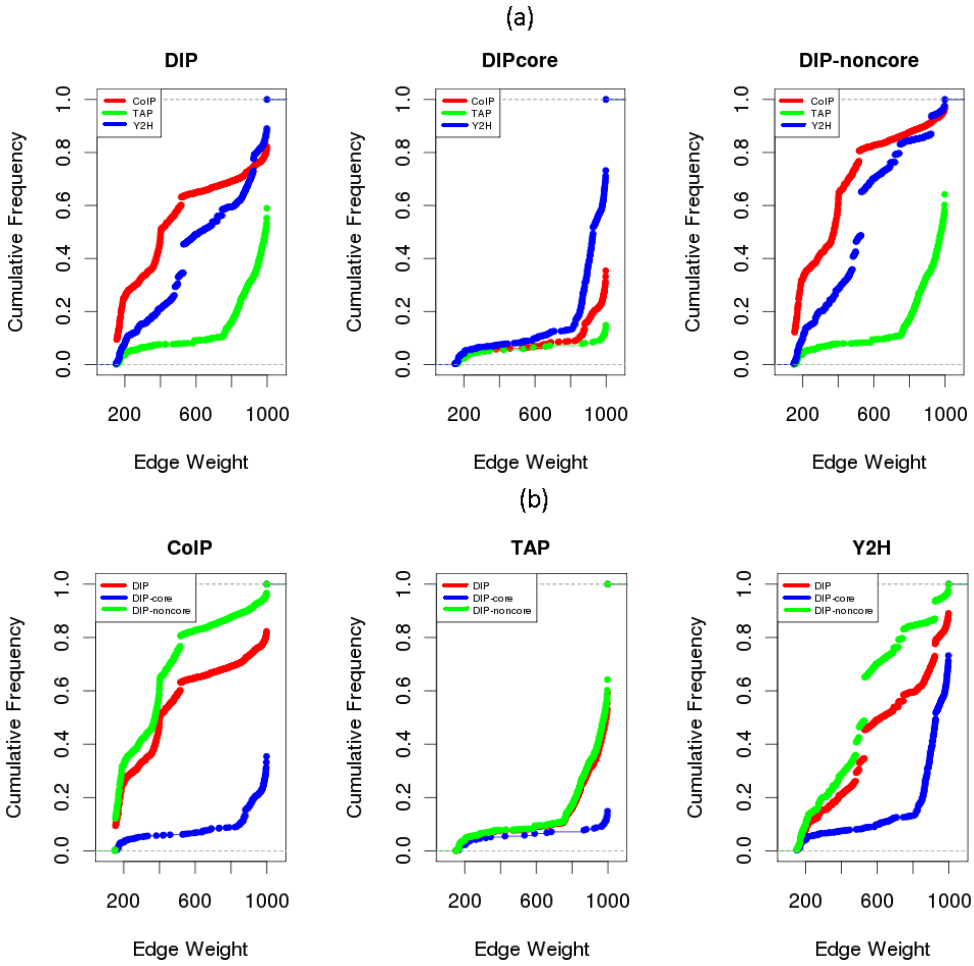


Figure 3.8: Distribution of STRING interaction scores for DIP Yeast datasets classified by (a) core/non-core and (b) experimental technique.

In the following sections, I discuss the results from our network alignment based error estimation method. I mainly focus on the results relevant to the pair-wise analysis of Yeast and Human networks unless stated otherwise. We tested our methods further by carrying out the same analysis for Human-Fly and Yeast-Fly pairs. These tests indicated that our method is applicable to multiple datasets.

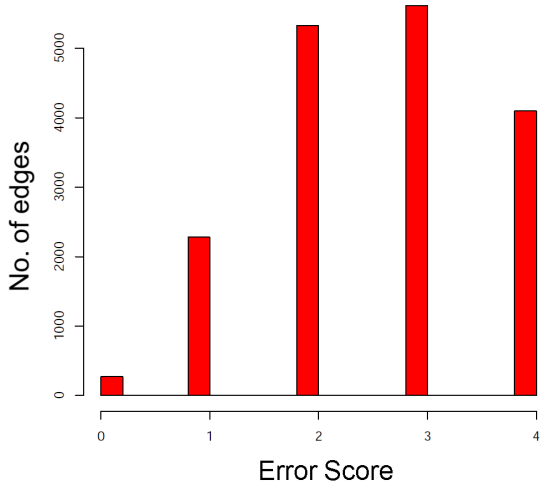


Figure 3.9: Frequency of error scores for Yeast DIP interactions. The error score for each interaction is the number of methods (out of STRING, IRAP IG and Deng) that classify it as low-confidence. Very few interactions are flagged as high confidence by all four methods.

3.3.2 Alignment of experimental networks

Pair-wise alignments of experimental networks for Yeast and Human were initially carried out using the MaWISh algorithm. The resulting alignment contained 580 conserved interactions between 431 proteins (Table 3.2). Carrying out the same alignment through NWBlast resulted in 785 conserved interactions between 601 proteins. For additional insights, we also aligned the prokaryotic networks of *E.coli* and *C.jejuni*. To the best of our knowledge these are the only two prokaryotic species with significant experimental interaction datasets. However, we found the alignment between *E.coli* and *C.jejuni* to be extremely small (22 proteins, 23 interactions). The same pattern was observed when the two prokaryotic networks were aligned to Yeast (in both cases the alignment found less than 20 conserved interactions). Given that the *C.jejuni* and *E.coli* networks are fairly complete (Hu et al., 2009; Parrish et al., 2007), the extremely low conservation detected cannot be explained away by even very high false negative rates. This may point to a fundamental difference between prokaryotic and eukaryotic networks and thus we may need considerably more interaction data for prokaryotes before similar error-focused studies can be replicated. A related observation was previously made by Chen et al. (2007) while predicting protein properties using interaction networks. They found that using prior information from networks of the same kingdom (eukaryotes or prokaryotes) results in better

Table 3.2: Alignment statistics for experimental and simulated pairs of networks.

Model	*Aln.Method*	*NN*	*NE*	*AD*	*ACC*	*LC*	*d*
Experimental	MaWISh	431	580	2.71	0.15	372	0
Networks	NWBlast	601	785	3.03	0.17	564	0
DD Y-ancestor	MaWISh	1712	3425	3.79	0.11	1697	0.15 ± 0.003
(1000 samples)	NWBlast	2101	4217	3.41	0.14	1863	0.13 ± 0.004
DD H-ancestor	MaWISh	1654	3172	3.57	0.098	1611	0.16 ± 0.004
(1000 samples)	NWBlast	1883	3721	3.35	0.11	1792	0.14 ± 0.004
Geo3d Y-ancestor	MaWISh	1625	2987	3.22	0.09	1459	0.21 ± 0.005
(1000 samples)	NWBlast	1959	4006	3.82	0.13	1754	0.20 ± 0.003
Geo3d H-ancestor	MaWISh	1505	2731	3.12	0.10	1342	0.19 ± 0.006
(1000 samples)	NWBlast	1709	3676	3.91	0.15	1366	0.17 ± 0.007
DD-optimal error	MaWISh	447	612	2.69	0.13	401	0.0092 ± 0.003
(1000 samples)	NWBlast	547	801	3.22	0.15	547	0.0091 ± 0.002
Geo3d-optimal error	MaWISh	427	536	2.54	0.11	421	0.0073 ± 0.0027
(1000 samples)	NWBlast	652	913	3.09	0.14	652	0.0051 ± 0.0023

* Number of Nodes (NN), Numder of Edges (NE), Average Degree (AD), Average Clustering Coefficient (ACC), Largest Connected Component (LC) and distance between real and simulated alignments (d). Optimal error refers to the error rate that provides the best fit to real alignments. This was 0.6 (60%) for the DD model and 0.55 (55%) for the Geo3d model.

prediction than the case where the priors from the two kingdoms are combined, indicating that the networks in eukaryotes and prokaryotes may differ.

3.3.3 Alignment of error-free simulated networks

We start with an ancestral network for Yeast and Human, created by preserving only those proteins in the DIP Yeast network which have orthologs in Human. The ancestral network contains 1836 nodes and 4761 edges. Two descendant networks were then grown independently from this common ancestor to the size of DIP Yeast (5000 nodes, 21000 edges) and HPRD Human (9100 nodes, 33000 edges) networks, using both DD and GeoEvol methods. Orthology relationships between the evolving networks were also updated in parallel. The evolved networks were then aligned, and as expected, the alignments achieved from these error-free networks were substantially better than the alignment of experimental networks. The average alignment size (over 1000 independent runs of the network evolution and alignment steps) for evolved networks was more than four times the size of alignment of experimental networks (Table 3.2). Use of a different ancestral network extracted from the HPRD Human data-set (induced by proteins with orthologs in Yeast) produced very similar alignment statistics. Our estimates therefore

appear to be robust to changes in the topology of the ancestral network which is especially important as the true ancestral topology is unknown. To ensure that our estimates are not dependent on the alignment method, we also performed the above analysis using NWBlast. Although the alignment results are better than MaWISh in terms of absolute numbers, the relative sizes of the simulated and experimental alignments are around the same (Table 3.2).

3.3.4 Adding uniform error

Once it was determined that the alignments of simulated networks evolved from a single ancestor are much better than alignments of currently available experimental networks, we then asked the question: What is the most likely error model and the minimum amount of error that explains this discrepancy? Two basic uniform error models were tested: node relabelling and edge rewiring. The models are tested by selecting a value for the error rate (θ) and inserting the error in the evolved networks. The networks are then aligned again and the alignments are compared to experimental Yeast and Human network alignments using the distance function given in 3.1. A discrete 2-D map of the effect of error rates in a pair of networks on the alignment distance function is shown in Figure 3.10.

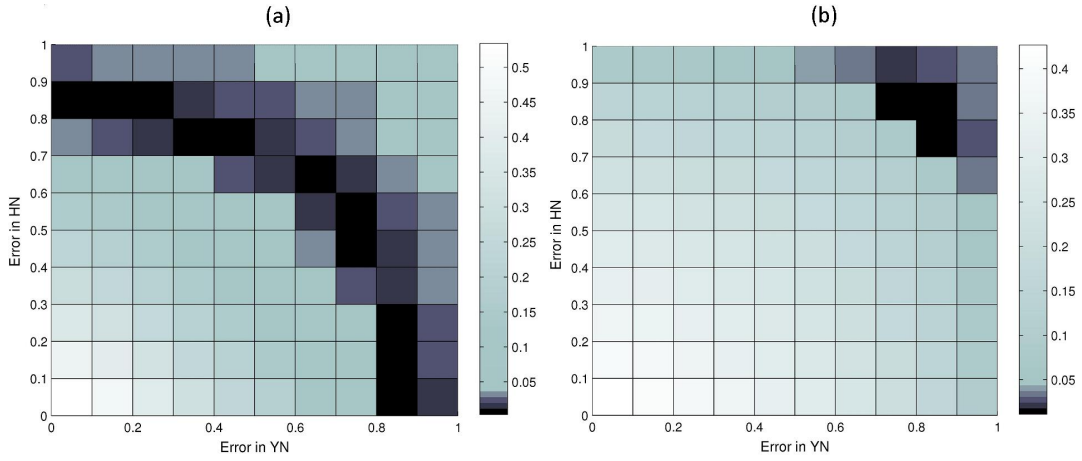


Figure 3.10: Alignment distance versus error rate for (a) Node relabelling error and (b) Edge rewiring error. For a pair of simulated networks (Yeast and Human in this case), the two axes represent the error rate in each, while the colouring stands for the value of the distance function between the alignment of these simulated networks and the alignment of corresponding experimental networks. Lower values of the distance function indicate that the simulated alignment better mimics the real alignment.

For very low error values, the simulated alignments are very good, translating into large values of our distance function between experimental and simulated network alignments. The distance decreases with increasing error, reaches a minimum and then starts increasing again as the simulated alignments become worse. Figure 3.10 also indicates that relabelling error fits the real alignment data with lower error estimates than rewiring error. In terms of biological mechanism, rewiring is easily explained by errors in an experimental protocol, for example the Yeast two-hybrid system missing a true interaction and reporting a spurious interaction. We investigated the relationship between the two error models to explain relabelling in terms of rewiring. This was done by first introducing a certain amount of relabelling error in the network and observing its effect in terms of edges that were rewired. Relabelling is essentially a non-linear version of rewiring with potentially several edges rewired at each step (Figure 3.11). This non-linearity may be a more realistic reflection of the actual error and thus can explain the superior performance of the relabelling model. We therefore discuss all subsequent error estimation based on the relabelling model.

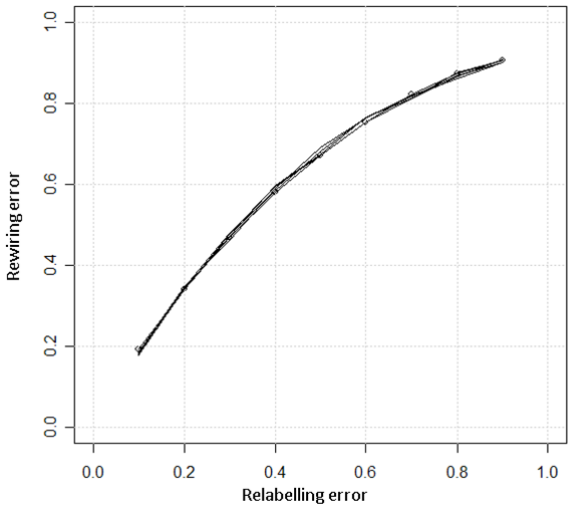


Figure 3.11: Relationship between relabelling and rewiring errors. For a given rate of relabelling error, this figure plots the equivalent rate rewiring error. Non-linearity arises because each relabelling error can perturb multiple links.

Calculation of the posterior density of error rates using approximate Bayesian computation clearly indicates a band of likely error rates in the two networks (Figure 3.12). It is worth

noting that even under the assumption of roughly equal error in the pair of networks, nearly 60% random error is required in both networks to explain the quality of current real alignments. Since we know that the Yeast network has been determined to a reasonable level of completion, our error bands would suggest even higher than 60% error rates for the Human network. Similar error estimates of around 55% were obtained using the NWBlast algorithm for alignment instead of MaWISh.

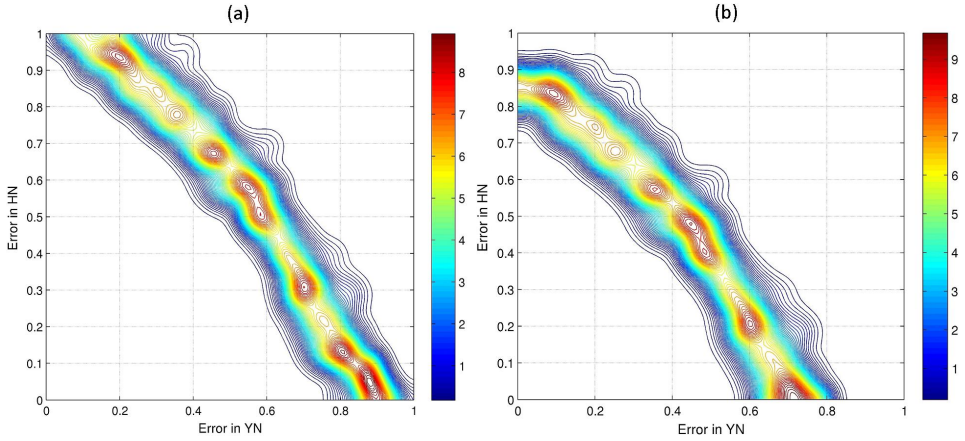


Figure 3.12: Contour map showing the density estimates for error rates in pairs of networks using the (a) DD model and (b) Geo3d model. The error in YN and HN is the amount of error introduced in the simulated Yeast and Human networks respectively. Regions of high density are in red.

A somewhat surprising result concerning the error rate densities in Figure 3.12 is that the assumption of very low error rates in one network lead to extremely high ($> 90\%$) error rates in the other network. Table 3.3 summarizes the number of accepted samples in each region of the parameter space and indicates that a significant number of accepted samples do indeed lie in regions of very high error rates for one of the networks. This implies that the alignments of real datasets are comparable to alignments of simulated pairs in which one of the networks is almost entirely randomized. One factor responsbile for this may be the randomization process itself. There is a small chance that an interaction that is rewired is still found to be preserved in the other network because the newly interacting nodes still have orthologs in the other network that interact. This is due to the possible one-to-may relationships introduced by the orthology evolution process. Thus, not every rewiring step necessarily results in the loss of a conserved interaction, and may in fact lead to a gain (albeit with a very low probability).

Table 3.3: Distribution of 5000 samples from the joint error rate posterior for Yeast (rows) and Human (cloumns) .

Error rate	$0.0-0.2$	$0.2-0.4$	$0.4-0.6$	$0.6-0.8$	$0.8-1.0$
$0.0-0.2$	0	0	0	0	544
$0.2-0.4$	0	0	0	348	496
$0.4-0.6$	0	0	564	612	24
$0.6-0.8$	352	704	300	12	0
$0.8-1.0$	660	48	0	0	0

Almost completely randomizing one of the simulated networks will therefore still result in the detection of some conservation. This is enough for the sample to be accepted given that one of the experimental datasets in this particular example (Human) is highly incomplete and provides very small alignments.

We derive more conservative estimates later for other species and also when we take into account incomplete sampling at node level in the current networks.

3.3.5 Threshold δ for approximate Bayesian computation

In approximate Bayesian computation the choice of the threshold δ for the distance between observed and simulated data can be quite important. For δ sufficiently small the ABC procedure should deliver a good approximation to the true posterior, whereas the the posterior estimate approaches the diffuse prior as δ increases. Ideally, the value of δ should be as small as possible, but this can be highly inefficient as the proportion of accepted samples can drop dramatically. The threshold choice is thus a problem-dependent trade-off between efficiency and accuracy. We used a value of 0.001 for the Yeast-Human error estimation as this resulted in fairly accurate error estimates. Figure 3.13 illustrates how assigning higher values to the threshold leads to very diffuse posterior densities.

3.3.6 Effect of ancestral topology

In our network evolution method, we used the network induced by Yeast proteins which have orthologs in Human. While this is a reasonable estimate of the order of the network for the last common ancestor of Yeast and Human, it is unlikely that the exact topology of the network

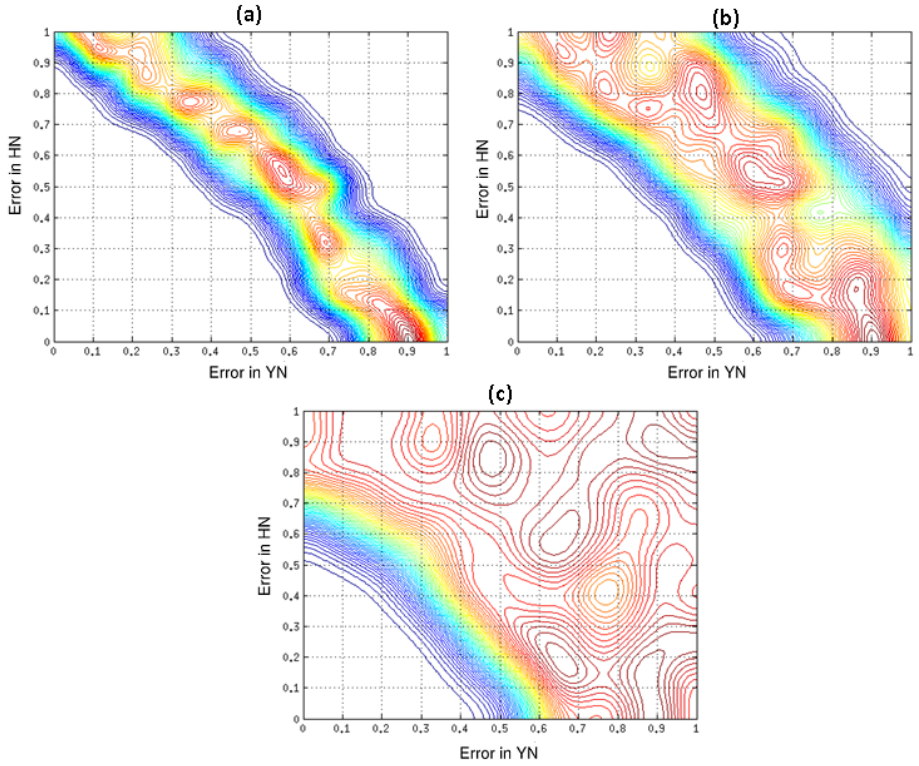


Figure 3.13: Effect of threshold δ on error rate density estimation: The axes and colour coding are the same as in Figure 3.12. This figure illustrates how using higher values of the threshold δ in Algorithm 3 negatively affects the error rate density estimates. Plots a-c show error estimates with the threshold set to 0.005, 0.01 and and 0.02 respectively. Note that lower threshold values require a much higher number of samples to be generated, as the proportion of samples accepted is very small. The density estimates in Figure 3.12 were calculated by generating 5000 random samples from the two dimensional error rate parameter space and accepting only those with distance less than 0.001 to real alignments.

is well approximated by this method. We therefore investigated the extent to which our error estimation method is sensitive to changes in ancestral network topologies.

Let n_0 be the original ancestral network as defined in Section 3.2.4 and n_x be the same ancestral network with $x\%$ of its links randomly rewired. Also, let A_{n_0} and A_{n_x} be the alignments of pairs of networks evolved from these ancestors and A_E be the alignment of a pair of experimental networks (Human and Yeast in this case). We define a measure RCAP (Relative Change in Alignment Properties) as,

$$RCAP = \frac{d(A_{n_0}, A_{n_x})}{d(A_{n_0}, A_E)} \tag{3.3}$$

where the distance d between the alignments is calculated using 3.1. RCAP essentially measures the effect of changes in the ancestral network topology on the alignment of its descendants. This is normalized by the distance between the simulated alignments (using n_0) and alignment of experimental networks. As our error estimation method using ABC is dependant on the normalizing term, the ratio in 3.3 quantifies the relative effect of different ancestral topologies. As seen in the Figure 3.14, even for 100% randomization, the change in alignment statistics is very small relative to the distance between real and simulated alignments. The effect of ancestral network topology on our error estimates is thus negligible.

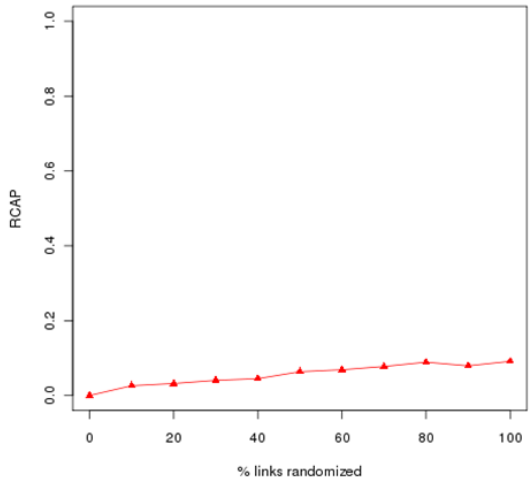


Figure 3.14: Effect of changes in ancestral topology on alignments of pairs of descended networks: On the x-axis is the percentage of links that are randomly shuffled in the original ancestral network topology (created as in Section 3.2.4). On the y-axis is the RCAP score from 3.3, measuring the effect of this shuffling. The RCAP score can have a minimum value of 0.0 and is unbounded from above, although our results indicate that is is well below 1.0 even for 100% randomization in our test-case.

3.3.7 Effect of false positives and negatives

An implicit assumption underlying the edge rewiring error model is that the number of false interactions added is equal to the number of true edges removed. This may well be unrealistic and we tested the effect of unequal rates of false positives and false negatives on our estimates. Given the widely held view that the real protein networks are at least as dense as current experimental data-sets (Stumpf et al., 2008), the number of spurious interactions is expected to be less than the number of undetected interactions. The false positive rate (fp) is therefore

unlikely to be higher than false negative rate (fn). Here we investigated two extreme cases: (a) $fn = 10 * fp$ and (b) $fp = 10 * fn$. In the first case, starting with a pair of error-free simulated networks , a false edge is inserted for every ten true edges removed (the starting networks were made denser than before to ensure that they mimic properties of real networks even after such a high false negative rate). Missing links therefore account for most of the error in the network. As shown in Figure 3.15, under this extreme assumption the error estimates increase by around 18% percent. The second situation involves the insertion of ten spurious links with the removal of a single true link. We observe that this has a negligible effect on the error estimate indicating that false positives do not significantly affect network alignment results.

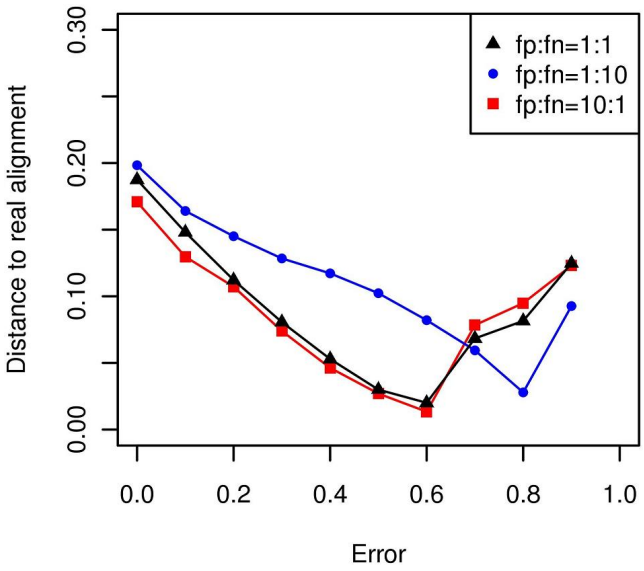


Figure 3.15: Effect of using unequal false positive (fp) and false negative (fn) rates on the optimal error estimate. While using equal fp and fn provides the best match to real alignments at around 62% error , this increases to nearly 80% under the extreme assumption of $fn = 10fp$. On the other hand, assuming $fp = 10fn$ leads to only a negligible difference from the original estimate.

Intuitively, inserting random links into a pair of networks is not expected to alter their alignments significantly as the number of possible links is vast, whereas removing true links can rapidly deteriorate alignments.

3.3.8 Adding non-uniform error

We also tested preferential or non-uniform versions of the relabelling and rewiring error models. In our tests however, all of the non-uniform error models failed to perform as well as the uniform models in terms of agreement with experimental network alignments. Figure 3.16 plots the distance function against error (equal rate in both networks) for the various non-uniform models along with uniform rewiring and relabelling. Although preferential rewiring and relabelling reach a minimum at much lower values of error, the resulting alignments are quite different in their summary statistics as indicated by the high values of distance function. Attacking nodes with high clustering coefficient preferentially does produce a minimum value comparable to random relabelling, but at a prohibitively high error rate. The simplest random error models appear to provide the best fit to real data. However the actual error in the networks may well be a complex mixture of various models.

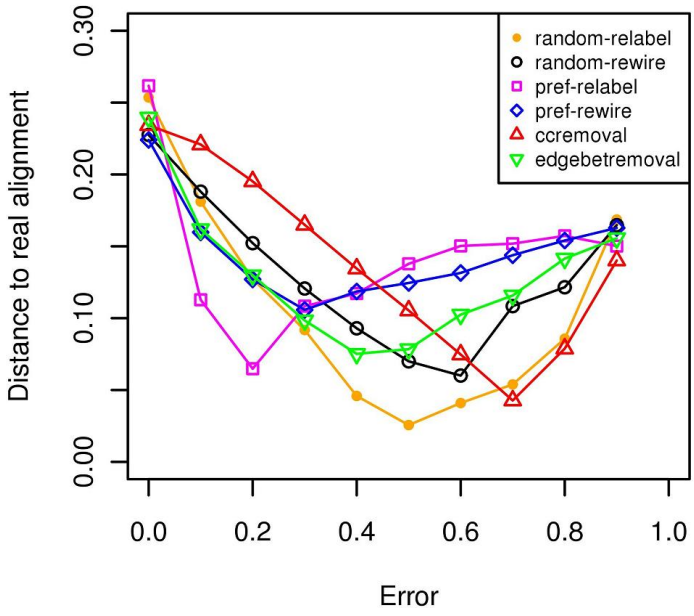


Figure 3.16: Rate of error versus distance to real alignment for several uniform and non-uniform error models (described in methods). The node relabelling error achieves the best match to real alignment, at an error estimate of 50%.

3.3.9 Sampling from complete proteomes

An issue that can further complicate current alignment studies and provide misleading results is the fact that existing interaction networks are essentially samples extracted from the complete proteomes in terms of nodes as well as edges. Poor alignments could therefore be a consequence of the possibility that the samples come from unrelated regions of two complete networks. For instance, the current Yeast interaction network may be better aligned to as yet unobserved parts of the Human network. To investigate the effect of this hypothesis on our error estimates we supplemented our evolution-alignment strategy with a sampling step. In this case instead of evolving a pair of networks to the size of existing Human (9000 nodes) and Yeast (5000 nodes) networks, the networks were grown to the approximate order of gene-coding ORF's in Human (25,000) and Yeast (6000). The network evolution process resulted in a total of 26,340 and 172,500 links respectively for the two networks. These numbers are in reasonable agreement with recent estimates of complete interactomes which range from 25-35,000 for Yeast and 160-600,000 for Human (Stumpf et al., 2008; Venkatesan et al., 2009). From these simulated complete interactomes, samples equalling experimental networks in order were drawn and aligned.

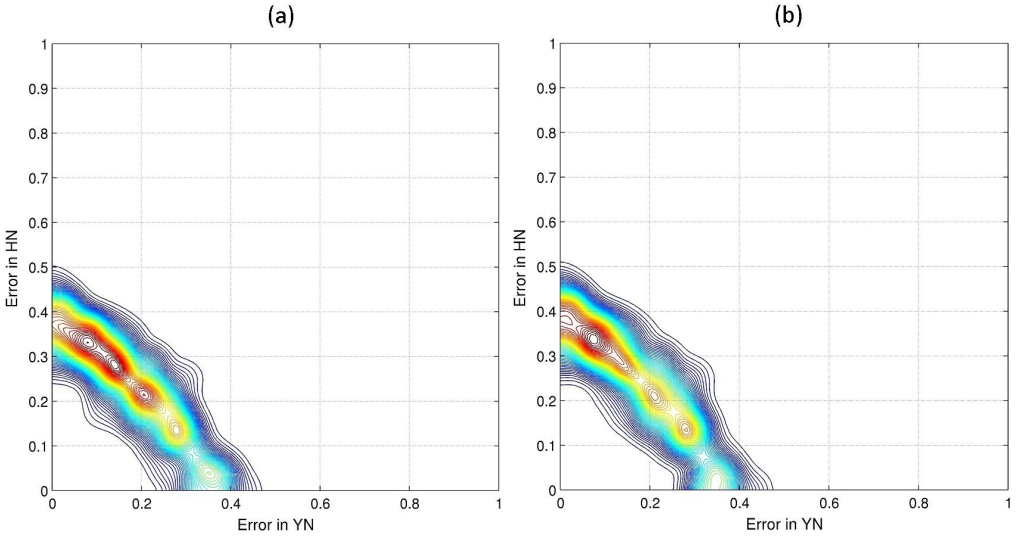


Figure 3.17: Density estimates for error rates in the Yeast and Human networks - with proteome sampling for (a) DD model and (b) Geo3d model. In this case two networks were grown to the complete proteome sizes of Human and Yeast and sub-samples of current network sizes were extracted and then aligned. These alignments were compared to real alignments to estimate error rates in Yeast and Human networks. The colour scheme for the density contours is the same as in Figure 3.12.

The simulated alignments in this case are substantially worse than the previous analysis. Consequently, this evolution-sampling-alignment strategy gives far lower estimates of error than before. The error rate density (Figure 3.17) indicates a much lower curve of possible values, centered around 20-25% error. An important underlying assumption in this analysis is that the samples drawn from the two networks are independent. In reality, this might not be totally accurate as it is expected that exploration of the Yeast interactome affects or drives research in the corresponding proteins in Human and vice versa.

3.3.10 Fly and Human networks have high error rates

The results presented in the previous sections for the Human-Yeast pair indicated rates ranging from 20-50%. We carried out the same analysis on Human-Fly and Human-Yeast and the error posteriors are shown in Figures 3.18 and 3.19. The estimated error rates are higher (35-65%) for the Human-Fly case than for Human-Yeast whereas they are lower for Yeast-Fly. Taken together, these results reinforce the prevailing view that the Yeast network has been characterized to a much larger extent compared to the Human and Fly networks, which are still mostly incomplete.

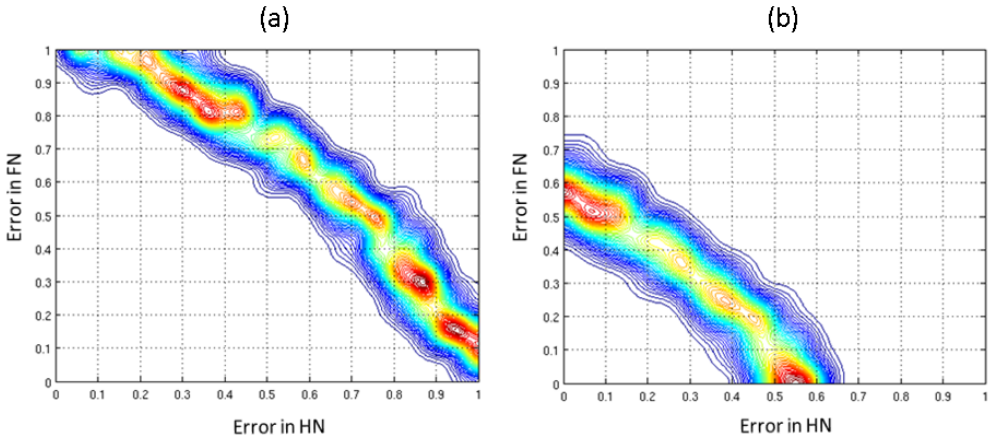


Figure 3.18: Error estimates for the Human-Fly pair using the DD model: (a) Without accounting for proteome sampling and (b) accounting for proteome sampling. The colour scheme for the density contours is the same as in Figure 3.12. Estimates are higher in this case than Human-Yeast, indicating that the Yeast network is better characterized than the Fly network.

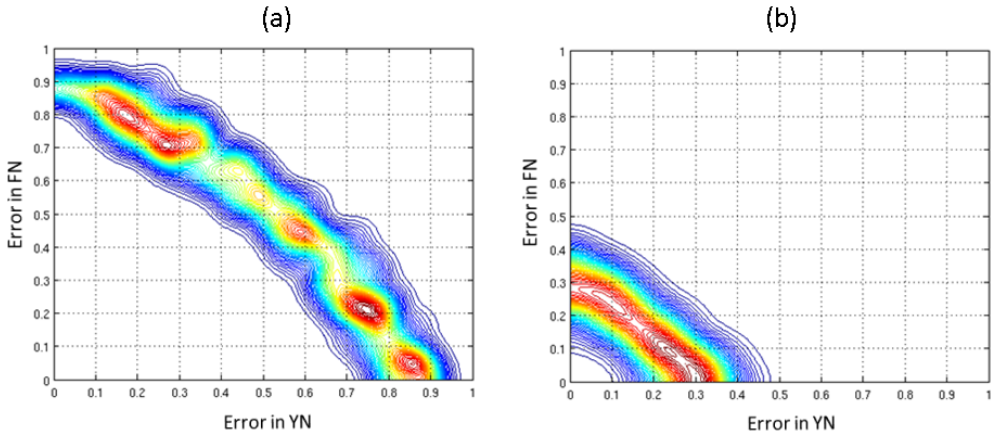


Figure 3.19: Error estimates for the Yeast-Fly pair using the DD model: (a) Without accounting for proteome sampling and (b) accounting for incomplete sampling. Estimates are lower than Human-Yeast. The colour scheme for the density contours is the same as in Figure 3.12.

3.4 Conclusions

The evolutionary mechanisms which have shaped protein interaction networks over millions of years are likely to lead to significant divergence between extant species. However, the results of our study indicate that evolution alone is unlikely to be the sole cause of the low conservation detected between current experimental interaction data sets. We used two independent evolutionary models and gained similar estimates of the expected amount of conservation at the network level between Human and Yeast interactomes. Results from other pairs of species including Human-Fly and Fly-Yeast appear to confirm that the conserved interaction sets are significantly larger than alignment results from experimental data-sets. This indicates that error has an important role to play. In our strategy, we use simple error models to introduce both false positives and false negatives into the simulated networks. The alignments obtained from these simulated networks with error are remarkably similar to real alignments in terms of size as well as density and connectivity allowing us to estimate likely error rates in existing interaction data-sets independent of any further biological information (GO annotation, co-expression etc.). While our initial estimates are high, a more realistic analysis catering for incomplete sampling of proteomes at the node level provide error rates which are in general agreement with previous studies. We find that the effect of false positives on alignment is negligible while

missing links (false negatives) account for most of the deterioration in the results. Thus for alignments at the interactome level, data incompleteness poses a far greater threat than low-quality data within reasonable limits and should perhaps be the main focus. It is speculative to propose a time-frame for bridging this gap based on the recent rate of experimental interaction detection studies. Still, our analysis indicates that we might be a considerable distance away from reliable species-wide interactome comparison studies, which would probably need >90% coverage for each species. Finally, even taking into account the effect of error on alignment, network evolution at currently accepted rates leaves little interaction conservation between species after the several hundred million years. This could potentially pose an unavoidable obstacle for systems-level species comparison studies.

Chapter 4

Conservation of a temporally ordered process

4.1 Introduction

In the previous chapters I have discussed the results from my research regarding protein interaction network alignment and network evolution at the whole interactome level. One of the primary conclusions from this work was the high sensitivity of network alignment to false negatives in the interaction data, which makes any inferences about conserved regions very challenging. It is also true that most network alignment studies (and protein interaction network analysis in general) interpret the network as a static graph, which is a highly simplified view. The interactions inside the cell are dynamic in nature, with proteins being expressed and interacting at different times in addition to different cellular locations.

While the current state of available biological data does not generally allow us to form a temporally dynamic view of the entire interactome, it may be possible for biologically well characterized processes to assign a rough temporal ordering to the proteins involved as well as their interactions. One such process is mitosis. Mitosis is the process by which a eukaryotic cell separates the chromosomes in its cell nucleus into two identical sets in two nuclei. It is generally followed immediately by cytokinesis, which divides the nuclei, cytoplasm, organelles

and cell membrane into two cells containing roughly equal shares of these cellular components. Mitosis and cytokinesis together define the mitotic (M) phase of the cell cycle - the division of the mother cell into two daughter cells, genetically identical to each other and to their parent cell. During mitosis the pairs of chromosomes condense and attach to fibers that pull the sister chromatids to opposite sides of the cell. The cell then divides in cytokinesis, to produce two identical daughter cells.

In this chapter we study the conservation of the mitotic network in Human, Yeast and Fly. We found that despite similar numbers of annotated mitotic proteins across the three species, network alignment using several existing algorithms does not detect any conservation. Moreover, this failure of network alignment seems to be associated with low numbers of sequence-based matches for mitotic proteins across species. Our subsequent investigation into the link between sequence and function at the genome level indicates a highly fuzzy relationship which has implications for network alignment studies based only on protein sequence similarity.

4.2 Methods

4.2.1 Temporal labels for mitotic proteins

The process of mitosis is complex and highly regulated. The sequence of events is divided into phases, corresponding to the completion of one set of activities and the start of the next. These phases include interphase, prophase, prometaphase, metaphase, anaphase and telophase (Figure 4.1).

In collaboration with biologists (Private communication, James Wakefield), we first identified the following set of temporally ordered labels corresponding to the major mitotic phases:

A. Prior to M phase

B. Entry into mitosis

C. Prophase

D. Early Pro-metaphase (Spindle and kinetochore formation)

E. Pro-metaphase (kinetochore-MT interactions)

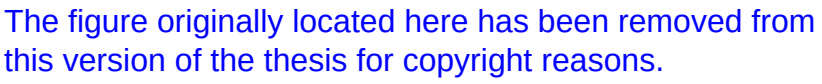


Figure 4.1: Major stages during Mitosis in Human cells (Allensby, 2011). The stages are temporally ordered from top to bottom in the figure.

F. Late Pro-metaphase (metaphase plate formation)

G. Metaphase/Anaphase transition

H. Early Anaphase (Chromosome segregation)

I. Mid Anaphase (central spindle formation)

J. Late anaphase (contractile ring formation)

K. Telophase

L. Cytokinesis

Following this, a detailed list of GO annotation terms related to the above labels was constructed. This was again based on the domain expertise of our biological collaborators who identified GO terms biologically relevant to each label. In general, several GO terms were associated with each of the above labels (see Appendix B for list of GO terms under each label). Finally, we extracted all proteins in each of our species' of interest (Human, Yeast and Fly) which were annotated with any of these GO terms. These sets of proteins with temporal labels (A-L) were used to create the putative mitotic networks (interaction data for Yeast and Fly was downloaded from DIP and for Human from HPRD in May 2010).

4.2.2 Inference of labels using Markov random fields

The mitotic networks built were poorly connected, with many isolated proteins. As current GO functional annotation is sparse, it is quite likely that our initial set of mitotic proteins is incomplete. We therefore introduced additional proteins (not yet annotated as mitotic in GO) into the mitotic network if they displayed the following properties: they interact with two or more mitotic proteins and indirectly link up at least two mitotic proteins which would otherwise be disconnected. The first condition makes it more likely that the additional proteins are also mitotic and the second condition addresses the poor connectivity of the initial networks.

After inserting additional proteins in the networks, we inferred their temporal labels using a Markov random field (MRF) approach. MRF methods provide a framework for probabilistic modelling of dependent random variables. They are widely applied to a variety of problems

with spatial dependencies, such as image analysis (Geman and Geman, 1984), where a picture is considered as a square grid of pixels (i.e an undirected graph) and each pixel corresponds to a variable whose value (i.e color) depends on the values of its neighbourhood pixels. The most probable colouring configurations of the missing pixels can be inferred from the full joint probability distribution. The colours of the missing pixels are predicted simultaneously, allowing prediction in cases where the entire neighbourhoods of pixels have to be predicted. MRF is thus particularly suited for a guilt-by association approach. The framework for protein function prediction based on MRF was originally proposed by Deng et al. (2003). Given a network with N proteins and a set E of pair-wise interactions, each node is coloured depending on whether the corresponding protein performs or does not perform a particular function (e.g. one GO term, or one temporal label in our analysis), where the colouring of unannotated proteins remains unknown. The colouring is encoded in an N-dimensional binary vector $\mathbf{x}$, i.e. $x_i = 1$ if the i^{th} protein has a particular temporal label, $x_i = 0$, if it does not. The aim of the inference is to assign each unannotated protein to one of the two possible states. The MRF model entails that the probability of state x of the network given a vector θ of model parameters is,

$$P(\mathbf{x}|\theta) = \frac{1}{Z(\theta)} exp(U(\mathbf{x}, \theta)), \tag{4.1}$$

where -U is known as the energy function and $Z(\theta)$ is a normalizing constant that depends on θ. In a homogeneous second order MRF, U can be written as (Besag, 1993; Sharan et al., 2007)

$$U(\mathbf{x}, \theta) = \sum_{i=1}^{N} G_1(x_i) + \sum_{i=1}^{N} \sum_{j=i+1}^{N} G_2(x_i, x_j), \tag{4.2}$$

where G_1 and G_2 are problem-dependent functions. G_1 takes one value per state, without considering the interactions of the protein, i.e. $G_1(1) = \alpha$ and $G_1(0) = 0$. The function G_2 is equal to zero if proteins i and j do not interact. For interacting proteins, like Deng et al. (2003), we used three classes of interactions. If both of the interacting proteins have the temporal label of interest then $G_2(1, 1) = \beta^{11}$. If only one of them has the label then $G_2(1, 0) = G_2(0, 1) = \beta^{10}$, and when none of them has the label, $G_2(0, 0) = \beta^{00}$. The number of protein pairs in these

three classes is given by N_{11}, N_{10} and N_{00}, respectively. The energy function of this MRF is,

$$\alpha \sum_{i=1}^{N} x_i + \beta^{11} N_{11} + \beta^{10} N_{10} + \beta^{00} N_{00},$$

which can be written in terms of elements of $\mathbf{x}$ as,

$$U(\mathbf{x}, \theta) = \alpha \sum_{i=1}^{N} x_i + \beta^{11} \sum_{(i,j)\in E} x_{ij} + \beta^{10} \sum_{(i,j)\in E} [x_i(1-x_j) + x_j(1-x_i)] + \beta^{00} \sum_{(i,j)\in E} (1-x_i)(1-x_j)], \tag{4.3}$$

with $\theta = (\alpha, \beta^{11}, \beta^{10}, \beta^{00})$. There are two ways of colouring the network that differ only in the value of the i^{th} protein. By inserting Equation 4.2 in 4.1 and setting $\beta^1 = (\beta^{11} - \beta^{10})$ and $\beta^0 = (\beta^{10} - \beta^{00})$, the log-odds are written as,

$$\begin{aligned} log \frac{P(x_i = 1 | \mathbf{x}_{-i}, \alpha, \beta^1, \beta^0)}{P(x_i = 0 | \mathbf{x}_{-i}, \alpha, \beta^1, \beta^0)} &= \alpha + \beta^1 \sum_{j \in S_i} x_j + \beta^0 \sum_{j \in S_i} (1 - x_j) \\ &= \alpha + \beta^1 M_{i1} + \beta^0 M_{i0}, \end{aligned} \tag{4.4}$$

where $\mathbf{x}_{-i}$ denotes $\mathbf{x}$ without the i^{th} element and S_i the set of proteins that interact with protein i. This equation has two predictors M_{i1} and M_{i0} counting the number of neighbouring proteins of protein i that do and do not have the label, respectively, and three unknown parameters. The conditional probability that unannotated protein i has the label can be calculated when the right side of the logistic equation is known. In this way we can sample the state of each unannotated protein when we know the parameters and the states of its neighbours. The problem that some or all neighbours have an unknown state can be circumvented by repeated sampling of states, starting from an initial configuration, until convergence using Gibbs sampling (Geman and Geman, 1984).

We used a Bayesian strategy developed by Kourmpetis et al. (2010) and draw from the joint posterior density of $\mathbf{x}, \alpha, \beta^0, \beta^1$ using an MCMC algorithm, starting from an initial configuration. This method uses the pseudolikelihood function (PLF) in 4.5, which is the product of the conditional probabilities across nodes, rather than the full likelihood which has an intractable

normalizing constant.

$$PLF = (\mathbf{x}|\alpha, \beta^1, \beta^0) = \prod_{i=1}^{N} P(x_i|\mathbf{x}_{-i}, \alpha, \beta^1, \beta^0). \tag{4.5}$$

A uniform prior is used as the joint prior distribution of the model parameters. This is Gibbs sampling in which, at iteration, t, the elements of $x^{(t)}$ corresponding to unannotated proteins are updated conditionally on the values of the parameters α, β^0, β^1, and the parameters are updated conditionally on $x^{(t)}$. The parameter update uses the adaptive MCMC algorithm called the Differential Evolution Markov Chain (DEMC) (ter Braak and Vrugt, 2008) as follows. A candidate point $\theta^* = (\alpha^*, \beta^{0*}, \beta^{1*})$ is obtained using the equation:

$$\theta^* = \theta + \gamma(Z_{R1} - Z_{R2}) + \mathbf{e} \tag{4.6}$$

where θ denotes the current state of the parameter vector, $\gamma \sim U(\gamma^*/2, \gamma^*)$ is the scaling parameter and $\gamma^* = 2.38/\sqrt{2d}$ is the optimal step size (Braak, 2006), where d is the parameter dimension. In our problem, $d = 3$ and therefore $\gamma^* = 0.97$. Z_{R1}, Z_{R2} are uniformly selected from past samples of the Markov Chain as stored in a matrix $\mathbf{Z}$ and $e \sim MVN(0, 10^{-4})$. θ^* is accepted using a Metropolis step, with probability:

$$r = \min\left(1, \frac{PLF(x^{(t)}|\theta^*)}{PLF(x^{(t)}|\theta)}\right) \tag{4.7}$$

The $\mathbf{Z}$ matrix is initialized in the following way. First, the Maximum Penalized Pseudolikelihood Estimates of θ, $\hat{\mu}$ and $\hat{\Sigma}$ are obtained by logistic regression using the brglm R package (Kosmidis, 2007). Then $m = 10d$ parameter values are sampled from $N(\hat{\mu}, \hat{\Sigma})$ and stored in $\mathbf{Z}$, where d is the dimension of the parameter vector. During the simulation, the state of θ is appended to $\mathbf{Z}$ in every iteration (Braak, 2006). Convergence was tested by performing multiple independent runs from dispersed starting points. We note here that the use of a single chain for DEMC as proposed by Kourmpetis et al. (2010) may not be theoretically justified as the resulting chain does not satisfy the Markov property.

For a given network, the above MRF-based inference method was carried out separately for each temporal label. Unannotated proteins were then annotated with labels having posterior

probability greater than 0.5.

4.2.3 Network alignment methods

To assess the conservation of the mitotic networks in Yeast, Human and Fly, we used three alignment methods: NwBlast, MaWISH and IsoRANK (all three methods were discussed in Chapter 1, Section 1.4.6.1 and 1.4.6.2). Network alignment with all three methods was carried out using BLAST E-values as the node similarity criteria. Details are given in the results section.

4.2.4 Functional similarity measures

In Chapter 2 we developed a simple protein functional similarity measure, which we demonstrated produces better network alignment results than sequence similarity. Here we investigated the link between sequence and function using two independent semantic similarity measures reported in literature, which are applicable to GO annotations. Lin (1998) used an information-content based approach to define semantic similarity between terms of a corpus. The information content of a term depends upon the frequency of its usage in the corpus. The semantic similarity for two terms is calculated through the information content of each term separately as well as the information content of their most informative common ancestor term. Given two terms t_1 and t_2, Lin's measure uses the information content of the shared parents of the two terms, as defined in Equation 4.8, where $S(t_1, t_2)$ is the set of parental concepts shared by both t_1 and t_2. As GO allows multiple parents for each concept, two terms can share parents by multiple paths. Taking the minimum $p(t)$, where there is more than one shared parent and calling this p_{ms} for probability of the minimum subsumer,

$$p_{ms} = min_{t \in S(t_1,t_2)}\{p(t)\} \tag{4.8}$$

Lin's measure uses both the information content of the shared parents, and that of the query terms. In this case, as $p_{ms} \geq p(t_1)$ and $p_{ms} \geq p(t_2)$, this value varies between 1 (for similar

concepts) and 0,

$$Sim(t_1, t_2) = \frac{2 * \ln\, p_{ms}(t_1, t_2)}{\ln\, p(t_1) + \ln\, p(t_2)} \tag{4.9}$$

This term-based semantic similarity can be used to calculate similarity between two genes A and B using, for example the maximum approach; the similarity of the proteins is the maximum semantic similarity between any pair of terms in the set of terms annotating A and the set of terms annotating B.

Given that the depth of the shared parent nodes may not be a suitable criteria for some limited cases in which the terms to be compared are close to the root, measures have been developed that take into account other aspects of the ontology structure. Wang et al. (2007) developed a new definition that considers the local relationships in the sub-graph generated by the terms, rather than their global positions in the directed acyclic graph (see Section 2.2.1 for a discussion of GO structure and annotation domains). This formula determines the semantic similarity of two GO terms based on both the locations of these terms in the GO graph and their semantic relations with their ancestor terms, addressing the drawbacks in previous approaches. Then, they designed an algorithm to measure the functional similarity of two genes based on the semantic similarities among the GO terms annotating these genes. To measure the semantic similarity of GO terms, they first encode the semantics of a GO term into a numeric format. Since the semantics (biological meanings) of a GO term are determined by its location in the entire GO graph and its semantic relations with all of its ancestor terms, they use the DAG (a subgraph of an ontology) starting from the specific GO term and ending at any of the root term (biological process, cellular component or molecular function) to represent this term.

Formally, a GO term A can be represented as $DAG_A = (A, T_A, E_A)$ where T_A is the set of GO terms in DAG_A, including term A and all of its ancestor terms in the GO graph, and E_A is the set of edges (semantic relations) connecting the GO terms in DAG_A. To encode the semantics of a GO term in a measurable format to enable a quantitative comparison of two term's semantics, they define the semantic value of term A as the aggregate contribution of all terms in DAG_A to the semantics of term A. Terms closer to term A in DAG_A contribute more to its semantics, while terms farther from term A in DAG_A contribute less as they are more

general terms. Therefore, they define the contribution of a GO term t to the semantics of GO term A as the S-value of GO term t related to term A. For any term t in $DAG_A = (A, T_A, E_A)$, its S-value related to term A, $S_A(t)$, is defined as:

$$\begin{cases} S_A(A) = 1 \\ S_A(t) = max\{w_e * S_A(t') | t' \in children(t)\} \quad \text{if } t \neq A \end{cases} \tag{4.10}$$

where w_e is the semantic contribution factor for edge $e \in E_A$ linking term t with its child term t'. In DAG_A, GO term A is the most specific term and its contribution to its own semantics is defined as one. Other terms in DAG_A are more general and, hence, contribute less to the semantics of GO term A. Therefore, $0 < w_e < 1$. After obtaining the S-values for all terms in DAG_A, the semantic value of GO term A, $SV(A)$, is calculated as:

$$SV(A) = \sum_{t \in T_A} S_A(t) \tag{4.11}$$

Given $DAG_A = (A, T_A, E_A)$ and $DAG_B = (B, T_B, E_B)$ for GO terms A and B respectively, the semantic similarity between these two terms, $S_{GO}(A, B)$, is defined as

$$S_{GO}(A, B) = \frac{\sum_{t \in T_A \cap T_B} (S_A(t) + S_B(t))}{SV(A) + SV(B)} \tag{4.12}$$

where $S_A(t)$ is the S-value of GO term t related to term A and $S_B(t)$ is the S-value of GO term t related to term B.

To accurately measure the functional similarity between two genes, one must also consider the contributions from the semantically similar terms that annotate these two genes respectively. The semantic similarity between one term go and a GO term set $GO = \{go_1, go_2, , go_k\}$, $Sim(go, GO)$, is defined as the maximum semantic similarity between term go and any of the terms in set GO. That is,

$$Sim(go, GO) = max_{1 \leq i \leq k}(S_{GO}(go, go_i)) \tag{4.13}$$

Finally, given two genes G_1 and G_2 annotated by GO term sets $GO_1 = \{go_{11}, go_{12}, , go_{1m}\}$

and $GO_2 = \{go_{21}, go_{22}, , go_{2n}\}$ respectively, their functional similarity is calculated as,

$$Sim(G_1, G_2) = \frac{\sum_{1 \leq i \leq m} Sim(go_{1i}, GO_2) + \sum_{1 \leq j \leq n} Sim(go_{2j}, GO_1)}{m + n} \tag{4.14}$$

4.3 Results

4.3.1 Initial mitotic networks

The mitotic networks for Yeast, Human and Fly were created by first downloading proteins annotated with GO terms related to any of the temporal labels discussed in Section 4.2.1. The interactions within these proteins were then extracted from the complete DIP Yeast and Fly networks and HPRD Human network (all interaction data downloaded in May 2010). The number of extracted proteins from each temporal category are given in Table 4.1.

Table 4.1: Number of proteins in each temporal category for Yeast, Human and Fly

Label	Yeast	Human	Fly
A	1	5	19
B	28	17	13
C	14	10	19
D	2	17	196
E	7	0	2
F	0	6	12
G	13	16	0
H	7	3	22
I	0	0	1
J	0	0	0
K	1	2	2
L	40	31	61
Total	113	105	347

Figure 4.2 depicts the mitotic network constructed for Yeast. This figure only contains mitotic proteins which have known interaction to at least one other mitotic protein.

Even a cursory look at Figure 4.2 reveals several salient features of the network. The interactions predominantly take place between proteins with the same label. This fits well with the generally held view that proteins with similar functions tend to interact with each other. Table 4.1 indicates that while there are substantial differences between the three model species

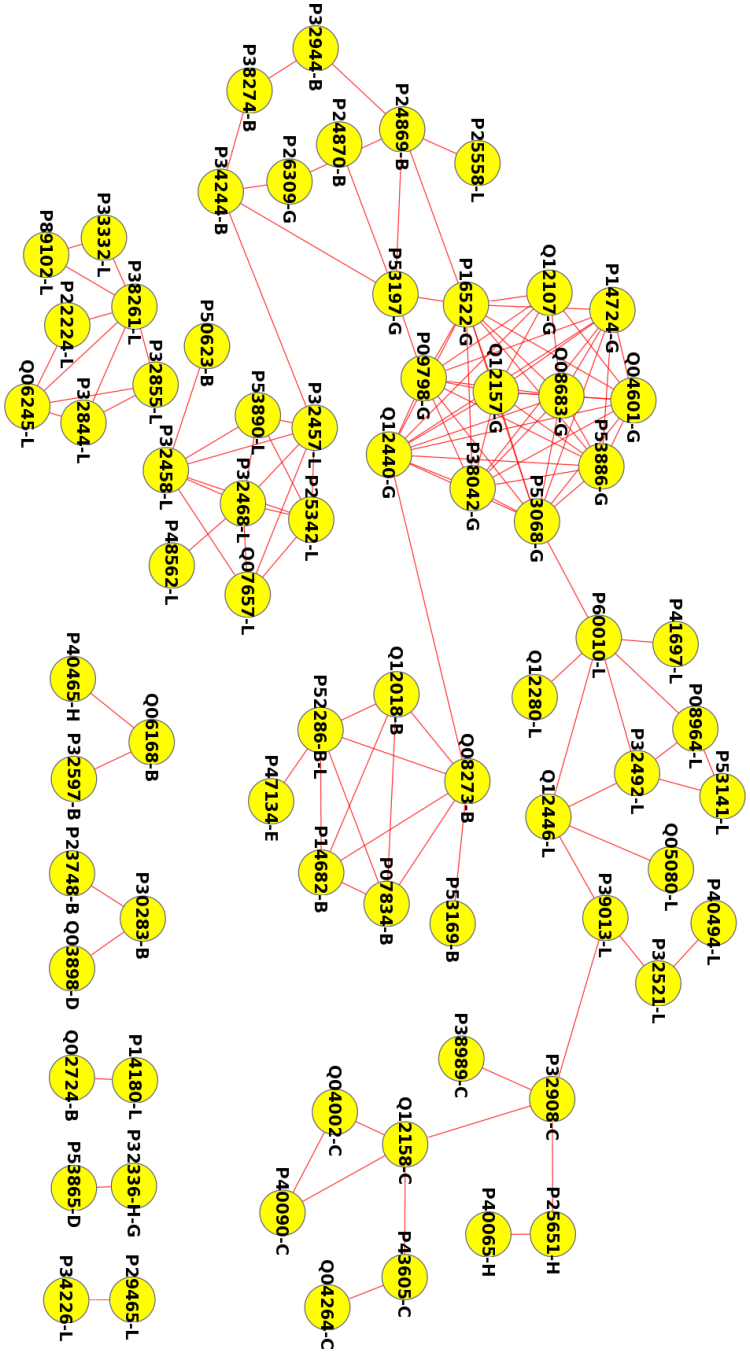


Figure 4.2: Yeast mitotic network. The node labels consist of the Uniprot accession number of the protein and its temporal label (A-L), separated by "-". Most interactions are between proteins with the same temporal label.

in terms of the number of mitotic proteins in each category, there are common patterns as well. This is especially true for Yeast and Human which have comparable numbers of proteins in categories B, C, G, H and L. The tendency of proteins to interact with identically-labelled proteins is also common in the three species (see Figure 4.3), as measured by the following propensity score: For a given label, let x be the number of links within this label, and y be the number of links with other labels, then,

$$Propensity = \frac{x - y}{x + y}, \text{ where } x + y > 0. \tag{4.15}$$

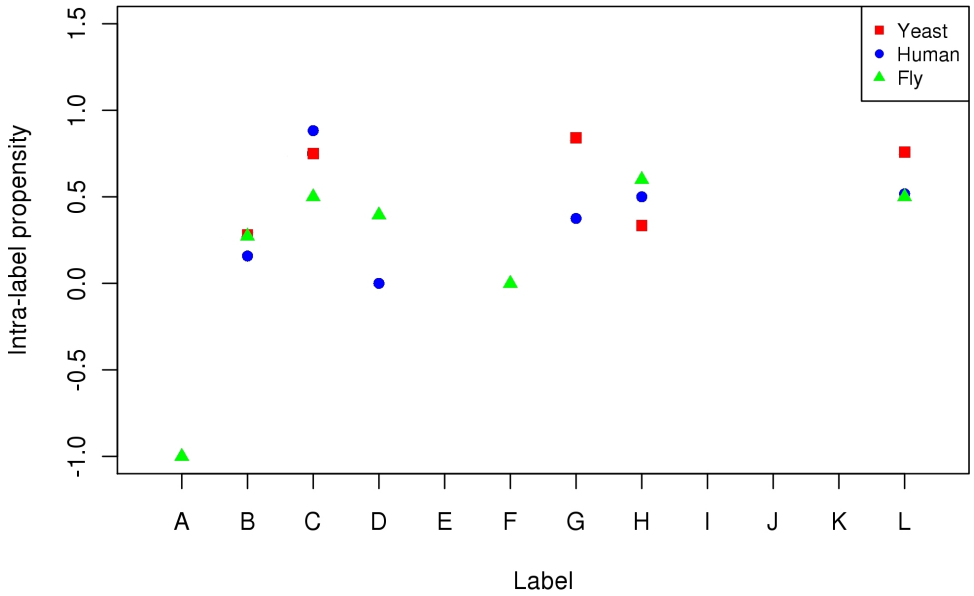


Figure 4.3: Propensity of intra-label interactions in the Yeast, Human and Fly mitotic networks. Propensity is defined as in 4.15. Almost all labels for each species show positive propensities for intra-label links.

Referring back to Figure 4.2, we observe that the network is rich in a few particular labels (e.g L and G) while others are entirely absent which was partly explained by the fact that we failed to retrieve any proteins from GO for that particular label. This could be a potential consequence of variable degrees of specificity of the GO terms under each label: A label that covers only highly specific GO terms, would be less likely to contain many proteins. We tested if some labels contain more specific GO terms by plotting the maximum depth (or GO level)

of all GO terms in each label. The maximum depth for a GO term is the maximum number of steps required to reach it from the root node (Biological Process in this case) in the GO directed acyclic graph and can be used as a rough proxy for how specific the term is. As seen in Figure 4.4, labels L and H contain GO terms that are more general (lower GO level) which could be the reason for a higher number of proteins retrieved for these labels. Still, there is not a clear pattern, as label D, which retrieved a large number of proteins from Fly, contains some of the most specific GO terms among all labels.

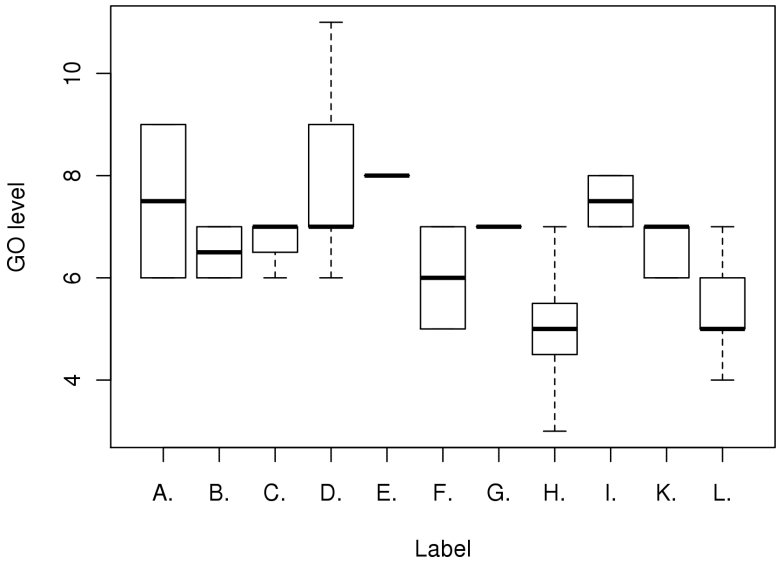


Figure 4.4: Maximum depth of GO terms for each label. For each label, the maximum depth of each GO term covered by it is determined in the GO directed acyclic graph (the root node, Biological Process, is at level 0.).

We note that some labels are also under-represented in the network in Figure 4.2 because there are no known interactions linking proteins annotated to them with the rest of the network. A related observation is the fact that the mitotic networks for all three species are disconnected and do not form a single connected component. We addressed these issues to some extent later on by predicting labels for proteins that link to the mitotic network but do not yet have mitotic annotations.

4.3.2 Temporal label prediction

As mentioned earlier, to alleviate the problem of disconnected mitotic networks, we inserted additional proteins, having no mitotic annotations. These additional proteins do, however have non-mitotic GO annotations. Figure 4.5, gives a histogram of the maximum depth of GO annotation for each protein in this additional set (Yeast).

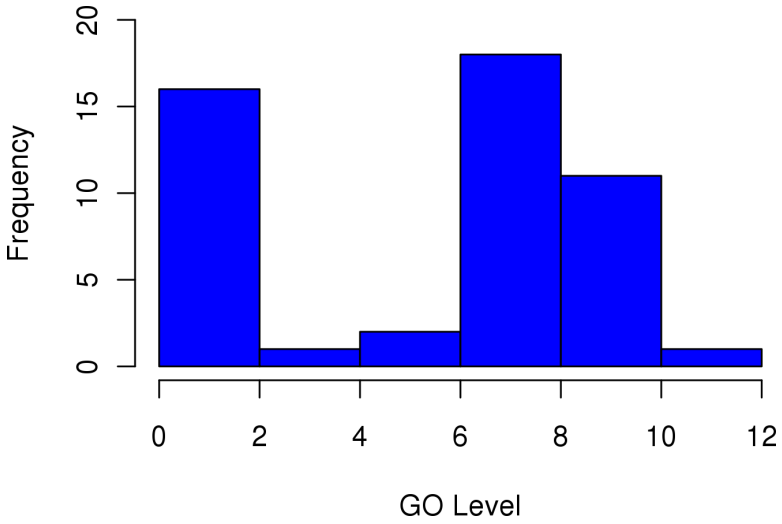


Figure 4.5: Histogram of maximum depth of GO annotation for proteins in the additional set. The additional set consists of proteins that do not yet have mitotic annotations but are responsible for linking up at least two mitotic proteins. In the figure, the y-axis (Frequency) stands for the number of proteins.

While more than half of these proteins are well-annotated (to a depth of > 6), around a third are very generally characterized (depth < 2) making it more likely that they could be assigned mitotic annotations in the future. For this study however, we inserted the entire set into the mitotic network and predicted their temporal labels using the MRF method described earlier as well as the majority vote (MV) method. In the latter case, a protein was assigned the most frequent label among its immediate neighbours. In case of ties, multiple labels were assigned. To assess the relative performance of the two methods, leave-one-out cross validation was employed. This was done by hiding the label of an annotated protein and then predicting it (along with all the unannotated proteins). This process was repeated for all annotated proteins in the network and the accuracy of the method was defined as the proportion of correctly predicted labels. The number of labelled and unlabelled nodes in the networks are given in

Table 4.2 while the inference accuracies are compared in Table 4.3. The random field based method clearly outperforms majority vote, especially when the number of unlabelled nodes is large (Fly).

Table 4.2: Statistics for networks induced by temporally labelled proteins and their indirect neighbours.

	Nodes		
Species	Labelled	Unlabelled	Edges
Yeast	92	51	312
Human	64	86	240
Fly	152	149	571

Table 4.3: Prediction accuracy of temporal labels using the majority vote and MRF methods.

Species	MV	BMRF
Yeast	0.630	0.728
Human	0.343	0.375
Fly	0.276	0.710

The label inference process described above led to well-connected mitotic networks for Yeast, Human and Fly. For each species, the issue of isolated mitotic proteins was resolved to a great extent as the newly added proteins with the inferred labels join up the majority of mitotic proteins into a large connected component. As mentioned earlier, the mitotic networks from the three species share similar number of proteins for several temporal labels and also exhibit a high tendency for intra-label interactions. These observations, along with the fact that mitosis is a fundamental cellular process led us to rigorously test the conservation of the mitotic networks by carrying out network alignment.

4.3.3 Mitotic network alignment

To assess the degree of conservation of the mitotic process across species, we first aligned the complete Human interaction network to the Yeast and Fly networks using local alignment methods (MaWISH, NwBLAST) and a global alignment method (IsoRANK). Note that we do not align just the mitotic networks, but the complete networks for each species (which

subsume the mitotic networks). This is because our aim is to test whether network alignment correctly aligns the relevant regions of two networks (in this case, whether the mitotic part of the interactome of one species is aligned to the mitotic part of the other species). The network alignment results were analyzed in terms of the number of mitotic proteins in one species aligned to mitotic proteins in another species (here, mitotic proteins are the ones labelled by GO as well as those for which we inferred temporal labels using the MRF approach in the previous section). This analysis differs for the local and global network alignment methods as follows:

The local alignment methods output a list of pairs of conserved clusters, one from each species. For a given mitotic protein x in species A, if it is present in one of the conserved clusters detected through alignment, and the corresponding cluster from species B also contains at least one mitotic protein, then we classify x as correctly aligned.

Global alignment methods output a one-to-one mapping of proteins across two species. In this case we classify a mitotic protein x in species A as correctly aligned if the alignment method maps it to a mitotic protein in species B.

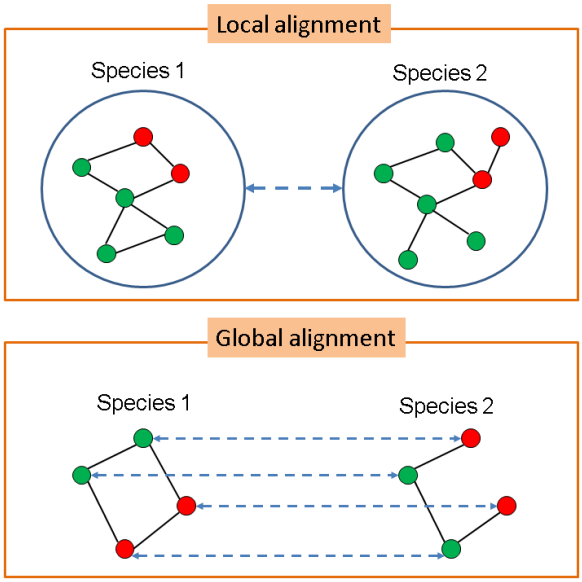


Figure 4.6: Local and global network alignment results. Local alignment outputs pairs of conserved modules. In this case, both modules have two mitotic proteins (in red) each, so two mitotic proteins are classified as correctly aligned. Global alignment explicitly outputs one-to-one mappings of aligned proteins. In this case only one of the two mitotic proteins in species 1 is aligned to a mitotic protein in species 2.

As shown in Table 4.4, a very small number of mitotic proteins from Human are aligned to mitotic proteins in either Yeast or Fly. This is true even when using a lenient sequence similarity threshold (BLAST E-value $< 10^{-5}$) for the alignment and the fact that our mitotic set also contains proteins with inferred mitotic labels. These results are quite surprising given that one would expect the mitotic process to be significantly conserved between the model organisms. Instead, we find that network alignment using all three methods maps most mitotic proteins in Human to non-mitotic proteins in Yeast and Fly and fails to detect any significant level of conservation at the interaction level.

Table 4.4: Number of mitotic proteins in one species aligned to mitotic proteins in the other species as a result of network alignment (Human against Yeast and Human against Fly).

		H-Y		H-F	
Method	E-value	Human	Yeast	Human	Fly
MaWISH	10^{-5}	12	23	5	3
	10^{-10}	10	19	5	3
NwBLAST	10^{-5}	19	33	8	11
	10^{-10}	17	33	6	9
IsoRANK	10^{-5}	13	13	6	6
	10^{-10}	10	10	6	6

Since network alignment in this case was carried out using protein sequence similarity, this implies that mitotic proteins in one species do not share sequence similarity preferentially with mitotic proteins in the other species. We explicitly tested this possibility by plotting the pattern of sequence similarity of mitotic proteins in Human to all proteins in Yeast and Fly (in this case mitotic proteins are only the ones labelled by GO, to avoid any bias due to the addition of proteins with inferred labels). The complete Human proteome was first BLASTed against the complete Yeast and Fly proteomes. For each Human mitotic protein, we investigated whether any mitotic protein in the other species was present in its BLAST results and if so whether it was the top sequence match. The E-value cut-off for the BLAST runs was 10, which is the default used by BLAST and is extremely lenient. As shown in Figure 4.7, for the majority of Human mitotic proteins, the corresponding most sequence similar proteins in either Yeast or Fly are non-mitotic proteins.

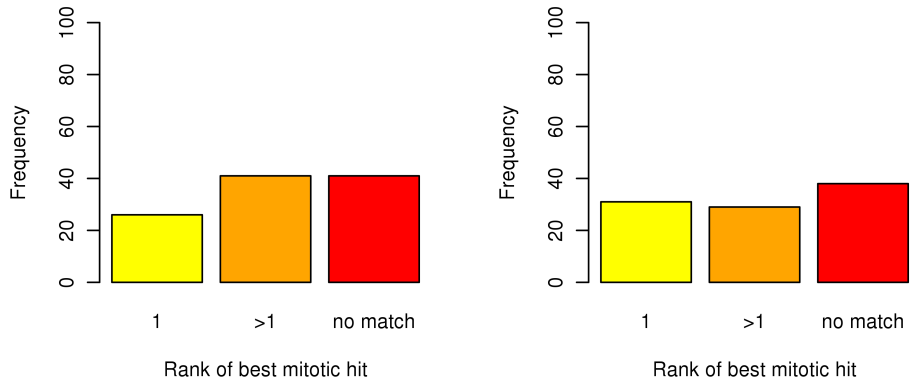


Figure 4.7: Pattern of sequence similarity between Human mitotic proteins and (a) Yeast, (b) Fly proteins. Rank '1' means that the best sequence match for a Human mitotic protein was a Yeast (or Fly) mitotic protein. Rank '>1' means that the best match for a Human mitotic protein was not a Yeast(or Fly) mitotic protein, although a Yeast(or Fly) protein was among the BLAST results. 'No match' means that for a Human mitotic protein there was no Yeast(or Fly) mitotic protein within an E-value cut-off of 10.

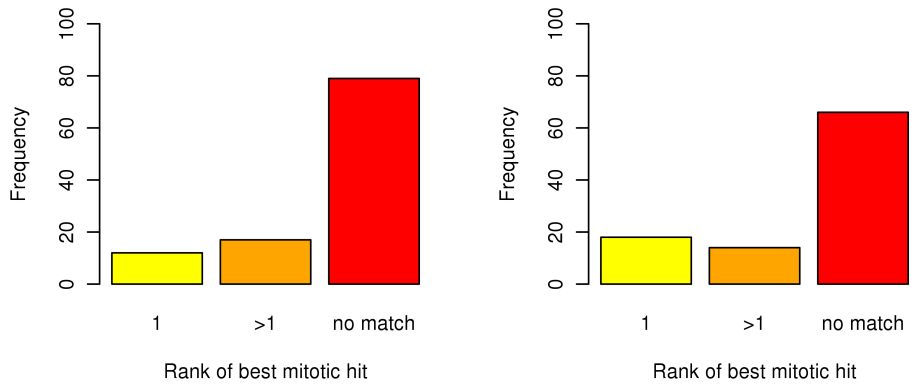


Figure 4.8: Pattern of sequence similarity between Human mitotic proteins and (a) Yeast, (b) Fly mitotic proteins with the same temporal label. The axes here have the same meaning as in Figure 4.7.

The results are even worse if the additional constraint is added that the temporal labels of the mitotic proteins being mapped should be the same. As shown in Figure 4.8, only 11 Human proteins are sequence-similar to some Yeast mitotic protein with the same label. For proteins with such highly similar functions in two species not to exhibit high sequence similarity has

significant implications for network alignment. In the following sections I explore in more detail the nature of the relationship between protein sequence and function.

4.3.4 Sequence versus function

As an initial test, we investigated the extent of agreement between the Lin and Wang functional similarity measures on the Human-Yeast pair. This was carried out by calculating the functional similarity for all possible pairs of proteins in the Human and Yeast pair using both methods. The scatter-plot in Figure 4.9 indicates that the two measures are in fairly good agreement for this genome-level comparison (Spearman's correlation coefficient: 0.89).

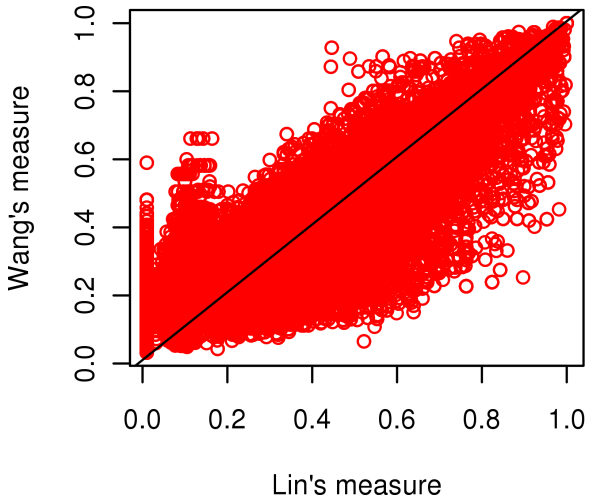


Figure 4.9: Scatterplot of functional similarity values of all pairs of proteins in Human and Yeast using Lin's measure and Wang's measure. The two scores are highly correlated with a Spearman's ρ value of 0.89.

Given the high correlation between the Lin and Wang measure, we subsequently show results only for Wang's measure when comparing functional and sequence similarity. First, all-against-all BLAST searches were carried out using the entire proteomes with the following combinations: Human against Human, Human against Yeast and Yeast against Yeast. The BLAST bit score for each pair of proteins was then chosen as the proxy for sequence similarity. Bit scores are normalized, which means that the bit scores from different alignments can be compared, even if different scoring matrices have been used. The E-value, on the other hand, gives an indication of the statistical significance of a given pairwise alignment and reflects the size of the database

and the scoring system used. So while E-values are preferred for detecting potential orthologs between one pair of species, here we want to compare the relationship between sequence and function in several pairs of species making the bit scores a better choice.

It has been suggested in a previous study that the correlation between sequence and function might be stronger for the molecular function (MF) domain of GO (Lord et al., 2003). We therefore started out by using this domain. Figure 4.10 plots the Wang functional similarity values (calculated using only MF annotations) against the log of the bit scores. For within-species as well as across species comparisons, the correlation between sequence and functional similarity is quite weak, confirmed by the relatively low Spearman's correlation coefficient values in Table 4.5.

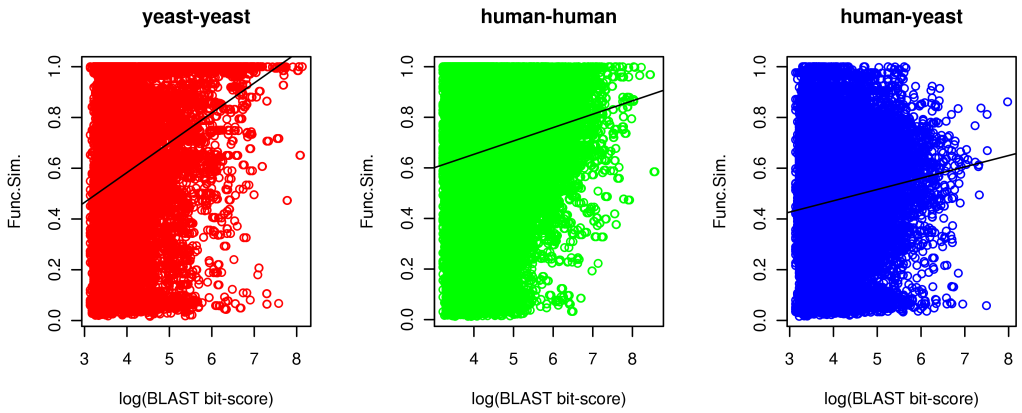


Figure 4.10: BLAST bit scores vs. functional similarity scores using the molecular function domain of GO. At the entire genome level, there is a weak correlation between sequence and functional similarity.

Table 4.5: Correlation between BLAST sequence similarity bit scores and functional similarity scores using the molecular function domain of GO.

Species	Spearman's ρ	p-value
Yeast-Yeast	0.302	<2.2e-16
Human-Human	0.141	<2.2e-16
Human-Yeast	0.115	<2.2e-16

Repeating the test with the biological process (BP) domain gives similar results, although the correlation is slightly lower within Yeast, and slightly higher within Human and across

Yeast and Human (Figure 4.11 and Table 4.6). It is worth noting here that for both domains, the correlation between sequence and function for proteins within Yeast is much higher than either within Human or across Yeast and Human.

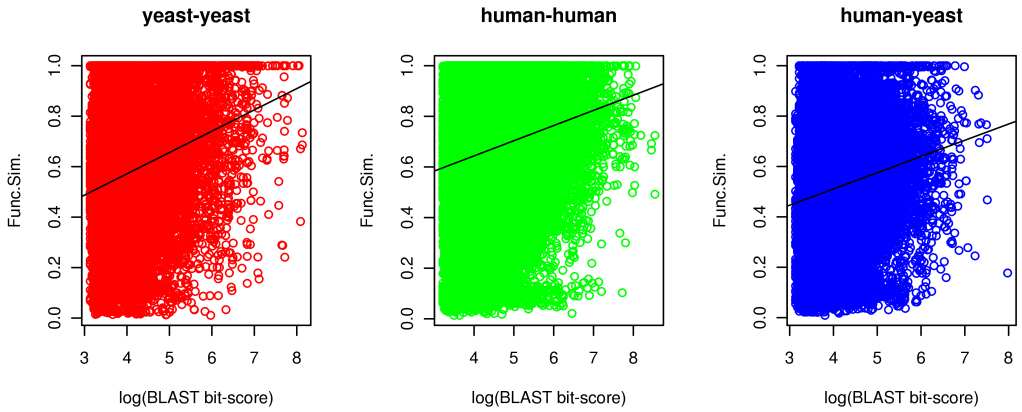


Figure 4.11: BLAST bit scores vs. functional similarity scores using the biological process domain of GO

Table 4.6: Correlation between BLAST sequence similarity bit scores and functional similarity scores using the biological process domain of GO.

Species	Spearman's ρ	p-value
Yeast-Yeast	0.255	<2.2e-16
Human-Human	0.159	<2.2e-16
Human-Yeast	0.184	<2.2e-16

4.3.4.1 Effect of annotation quality

The quality of functional annotation of gene products in GO can vary widely. For every single annotation of a molecule to a GO term, GO provides an evidence code indicating how that annotation is supported. Although evidence codes do reflect the type of work or analysis described in the cited reference which supports the GO term to gene product association, they are not necessarily a classification of types of experiments/analyses. GO evidence codes can be broadly classified into experimental, computational and those supported by author statements or curatorial judgements. Each of these classes have typically several subclasses of codes. While it is a

highly subjective exercise to assign relative confidence scores to annotations based on evidence codes, it is generally believed that annotations with experimental evidence codes are likely to be biologically meaningful. All of the results regarding the relationship between sequence and functional similarity discussed so far in this chapter were generated using all available GO annotations for each protein. Here I describe the results obtained when functional similarity between proteins is calculated using only experimentally supported annotations. Table 4.7 shows the Spearman correlation coefficients between sequence and functional similarity, using the biological process and molecular function domains.

Table 4.7: Correlation between BLAST sequence similarity bit scores and functional similarity scores using only experimentally supported GO annotations.

Species	ρ-BP	ρ-MF	p-value
Yeast-Yeast	0.281	0.472	<2.2e-16
Human-Human	0.194	0.251	<2.2e-16
Human-Yeast	0.220	0.290	<2.2e-16

Comparison with Tables 4.6 and 4.5 indicates that using only the experimentally supported portion of GO annotations leads to an improved correlation between sequence and functional similarity. While the difference is not very striking for the biological process domain, the molecular function domain shows a substantial improvement (especially for the Yeast-Yeast test case). However, this improvement does come at the cost of annotation density. As a relatively small portion of GO annotations have experimental codes (Table 4.8 gives the number of proteins in Human and Yeast with GO annotations), the number of protein pairs for which functional similarity can be calculated deteriorates rapidly.

Table 4.8: Number of proteins in Human and Yeast with GO annotations in the biological process and molecular function domains.

	All evidence codes		Experimental codes	
Species	BP	MF	BP	MF
Human	20381	23671	4789	3462
Yeast	4882	4127	3858	2760

In any case, the correlation is not strong enough to justify the use of sequence information

alone to align networks, as indicated by our comparative analysis of the mitotic network in three species.

4.4 Conclusions

Network alignment studies have generally been carried out previously to detect conservation at the whole interactome level. As discussed in previous chapters, such large-scale studies are highly sensitive to incompleteness in the interaction data and thus could result in a very skewed picture. Here we studied the conservation of the mitotic network in model species, which is one of the most important biological processes in eukaryotic cells. Exploiting the fact that the stages involved in the mitotic process are highly ordered in time, we also studied the pattern of interaction within and across temporally distinct stages and the extent to which this pattern is similar in the Yeast, Human and Fly networks.

We first alleviated the problem of insufficient mitotic labels in downloaded data by predicting labels for unannotated proteins that link up multiple mitotic proteins. This was carried out using a Markov random field method which was found to outperform the majority vote method, especially in the presence of large numbers of unannotated proteins. This label inference step resulted in much larger and better connected mitotic networks for the three species. However, pairwise network alignment (local and global) using several existing methods failed to match up the mitotic networks to any reasonable extent. We discovered that this was a result of insufficient sequence-similarity matches between the mitotic networks of several species.

Further investigation in the nature of the relationship between sequence and function similarity indicated a weak, although positive correlation. The correlation improves when only experimentally supported functional annotations are used. Still, this imperfect link between sequence and function introduces another dimension to the network alignment problem. Alignments based on sequence-similarity of proteins might not be biologically the most relevant, whereas those based on functional similarity have other constraints: Functional characterization of proteins is by nature relatively subjective, and the quality and detail of annotation varies widely across species. Thus in our view systems-level alignment studies need to be aware not only of the biases introduced due to error and incompleteness in interaction data, but also the

crucial choice of the node-mapping criteria.

Chapter 5

Conclusions and future work

Understanding the organization of protein interaction networks and to establish how they contribute to cellular and organism phenotypes is one of the major challenges in the post-genomic era. With ever-expanding interaction datasets, analysis of these networks has moved on from exploratory studies within single species to the alignment of multiple networks. The aim of network alignment is to quantify the extent of conservation across organisms at the systems level and to identify sets of interactions or network modules that may be functionally important. These goals are somewhat analogous to what sequence alignment has achieved for us in terms of biological understanding at the gene and protein level. However, alignment of such large-scale networks is much more challenging task: The graph matching problem underlying network alignment is by itself computationally very hard, whereas noise and incompleteness in interaction data makes any inferences subject to a great deal of uncertainty. A crucial element of the network alignment process is the node-mapping criteria which is used to aid alignment by identifying, for each protein in one network, a set of possible matches in one or more other networks. In general, research has focussed on better and faster graph matching algorithms and not much attention has been given to the node-mapping criteria used. To our knowledge, all network alignment studies have by default relied on sequence similarity (usually BLAST E-values) for this purpose.

In this thesis, I have for the first time, used functional similarity of proteins across pairs of species to carry out network alignment. Protein function is an inherently non-quantitative concept. Still, the availability of ever-expanding functional annotation resources such as the

Gene Ontology in conjunction with semantic similarity measures from the information theory literature, provides an opportunity of studying functional relationships within and across species quantitatively. I devised a simple GO-based functional similarity measure, and showed that the conserved modules detected using this measure outperform the results from sequence similarity based network alignment using several existing methods. I also extended the application of this functional similarity to the more general problem of module detection in networks. I show that it is possible to detect high quality modules by starting out from seeds that are conserved across two or more species and employing additional biological information such as gene expression. This method has the advantage that it can improve module coverage as networks for more species become available. I also proposed simultaneous clustering of multiple networks as an alternative to network alignment in the presence of extremely noisy interaction data.

Research in protein network alignment is tied closely to the question of how protein networks evolve and how initially similar networks diverge over time as a result of these evolutionary mechanisms. While a complete understanding of these mechanisms is no doubt still a long way in the future, the gene duplication-divergence model and its variants provide an intuitive and elegant system with well-motivated biological foundations. Here I devised a method to explain the low conservation normally detected in existing network alignment studies of model organisms. I simulate network evolution followed by comparison of the alignment results between empirical and simulated interaction networks to infer the likely rates of error in real datasets. The novelty of the method lies in the fact that it does not need any additional biological information and brings together several key ideas in the analysis of protein networks. Moreover, the method can be used to compare competing error models in terms of their ability to explain the observed properties of empirical network alignments. Using this method in conjunction with approximate Bayesian computation, I inferred parameters for two uniform random error models and estimated error rates ranging from 20% to 55% for interaction datasets of model organisms (Yeast, Human and Fly). Another conclusion of this analysis was that incompleteness in interaction datasets dramatically affects alignment quality whereas the presence of spurious interactions within reasonable limits may not be a key concern for such studies.

Comparative analyses at the interactome-level have traditionally viewed the network as a static graph, which is a gross over-simplification, although one necessitated by the lack of rel-

evant data. The interaction networks inside the cell are in fact highly dynamic, with sets of interactions separated in time and/or space. Alignment of static graphs would thus at best only provide rough estimates of the true of level of conservation of these dynamical systems in different species. One can begin to study some temporal characteristics of interaction patterns in well studied processes. In this thesis we focussed on cell mitotis, and defined a set of temporally ordered labels for the major stages involved in this fundamental biological process. Our aim was to study the conservation of the dynamics of this process in multiple species by generating a series of time-evolving alignments. However our results indicated that despite some common characteristics in the mitotic networks of several model species, such as the tendency of interaction within same time labels, network alignment fails even at the complete mitotic network level, precluding any possibility of aligning sub-networks for each label separately. We concluded that this failure of network alignment was related to the lack of sufficient node-to-node mappings between mitotic proteins, detectable across species by sequence similarity. This is quite surprising for proteins that share such a well-defined function across different species. We verified these results by exploring the relation between sequence and function at the genome level, within and across species. This analysis detected only a weak correlation between the measures, although the correlation was somewhat improved when using only experimentally supported functional annotations. These results could have implications for the network alignment field as a whole. Our study seems to indicate that the choice of this criteria could be crucial and also opens up the possibility for different node mapping measures, based on the goals of the matching problem.

The work presented in this thesis has significant scope to be expanded upon and here I identify several areas which can be the subject of future research. First, the relationship between sequence and function, and the effect of a choice between these on network alignment results raises the question whether it is reasonable, or indeed, desirable to aim for a single, 'best' alignment between two networks. Proteins are versatile molecules and display multi-functionality. It would thus seem overly restrictive to constrain possible alignments between two networks based purely on a single measure such as sequence similarity. It may be interesting to explore the potential of carrying out alignments between the same sets of species with various node-mapping measures and determine if the conflicting alignments shed light on different

aspects of the system. Another relevant area of research, which has already received significant attention is modelling protein network evolution. While models exist, some of which were used by us in this thesis, improvements are certainly possible and also needed. For instance, the basic duplication divergence model assumes a uniformly random gene duplication at each step followed by some edge wiring. This process assumes complete independence between nodes, whereas in reality, it is reasonable to expect that the duplication of a gene is strongly influenced by its context, requiring the addition of dependence into the model. Finally, as more data becomes available and we get a clearer picture of the interaction network as an ever-changing dynamic system, it would be an exciting, though challenging problem to incorporate this into the alignment framework. To measure conservation in multiple species, one would need to not only compare the entire set of interactions, but whether the dynamics of how these interactions are organized and re-organized in time are also conserved.

Appendix A

GO enrichment of conserved modules

In chapter 2, I analysed the GO enrichment of sets of proteins found in conserved modules detected by network alignment. While that analysis was restricted to GO level 2 for sake of clarity, here I give the list of 50 most significantly enriched GO terms for each set of proteins (at all GO levels).

GO Term	Description	P-value	# Proteins
GO:0006468	protein phosphorylation	1.39E-059	81
GO:0016310	phosphorylation	1.68E-055	81
GO:0006796	phosphate metabolic process	3.33E-053	94
GO:0006793	phosphorus metabolic process	1.21E-048	94
GO:0006464	protein modification process	1.71E-042	136
GO:0043412	macromolecule modification	1.02E-035	137
GO:0050789	regulation of biological process	1.69E-034	206
GO:0065007	biological regulation	3.44E-034	221
GO:0019538	protein metabolic process	7.02E-030	177
GO:0050794	regulation of cellular process	1.92E-029	182
GO:0044267	cellular protein metabolic process	2.26E-026	156
GO:0044260	cellular macromolecule metabolic process	3.16E-023	252
GO:0043170	macromolecule metabolic process	1.28E-022	255
GO:0023052	signaling	6.02E-022	63
GO:0023033	signaling pathway	6.02E-022	63
GO:0009987	cellular process	1.05E-021	334
GO:0023034	intracellular signaling pathway	1.46E-020	57
GO:0007165	signal transduction	1.77E-019	53
GO:0019222	regulation of metabolic process	1.08E-017	141
GO:0051726	regulation of cell cycle	6.46E-017	49
GO:0065009	regulation of molecular function	1.99E-016	43
GO:0007049	cell cycle	3.53E-016	68
GO:0050790	regulation of catalytic activity	6.15E-016	38
GO:0010564	regulation of cell cycle process	1.91E-015	40
GO:0035556	intracellular signal transduction	2.14E-015	38
GO:0016043	cellular component organization	3.10E-014	172
GO:0071842	cellular component organization at cellular level	7.51E-014	152
GO:0051301	cell division	9.01E-014	56
GO:0044238	primary metabolic process	1.49E-013	261
GO:0071840	cellular component organization or biogenesis	2.01E-012	178
GO:0031323	regulation of cellular metabolic process	2.29E-012	120
GO:0045859	regulation of protein kinase activity	6.47E-012	18
GO:0044265	cellular macromolecule catabolic process	9.48E-012	63
GO:0051128	regulation of cellular component organization	9.68E-012	43
GO:0022402	cell cycle process	1.43E-011	69
GO:0071841	cellular component organization or biogenesis at cellular level	1.46E-011	167
GO:0009057	macromolecule catabolic process	1.78E-011	64
GO:0044248	cellular catabolic process	1.87E-011	85
GO:0050896	response to stimulus	3.39E-011	104
GO:0006996	organelle organization	3.42E-011	120
GO:0043549	regulation of kinase activity	3.62E-011	18
GO:0071900	regulation of protein serine/threonine kinase activity	5.43E-011	15
GO:0010695	regulation of spindle pole body separation	5.53E-011	9
GO:0006984	ER-nucleus signaling pathway	6.15E-011	25
GO:0007264	small GTPase mediated signal transduction	8.06E-011	24
GO:0051338	regulation of transferase activity	1.02E-010	18
GO:0042325	regulation of phosphorylation	1.12E-010	19
GO:0060255	regulation of macromolecule metabolic process	2.82E-010	111
GO:0048518	positive regulation of biological process	2.89E-010	50

Figure A.1: GO enrichment of Yeast proteins in conserved modules found by aligning Yeast to Human using sequence similarity.

GO Term	Description	P-value	# Proteins
GO:0009987	cellular process	1.08E-033	605
GO:0050789	regulation of biological process	5.64E-032	315
GO:0050794	regulation of cellular process	1.45E-029	282
GO:0065007	biological regulation	3.15E-029	337
GO:0090304	nucleic acid metabolic process	1.40E-027	285
GO:0080090	regulation of primary metabolic process	4.70E-027	229
GO:0060255	regulation of macromolecule metabolic process	9.92E-027	222
GO:0044260	cellular macromolecule metabolic process	1.79E-026	425
GO:0019222	regulation of metabolic process	1.18E-025	241
GO:0043170	macromolecule metabolic process	7.36E-025	429
GO:0032774	RNA biosynthetic process	5.83E-024	88
GO:0019219	regulation of nucleobase, and nucleic acid metabolic process	9.23E-024	180
GO:0051171	regulation of nitrogen compound metabolic process	1.35E-023	180
GO:0031323	regulation of cellular metabolic process	3.99E-023	220
GO:2000112	regulation of cellular macromolecule biosynthetic process	4.20E-023	196
GO:0010556	regulation of macromolecule biosynthetic process	5.00E-023	196
GO:0006139	nucleobase, nucleoside, nucleotide and nucleic acid metabolic process	1.25E-022	304
GO:0031326	regulation of cellular biosynthetic process	1.27E-022	198
GO:0009889	regulation of biosynthetic process	2.12E-022	198
GO:0006906	vesicle fusion	9.94E-022	33
GO:0010468	regulation of gene expression	3.33E-021	189
GO:0071842	cellular component organization at cellular level	1.20E-020	265
GO:0045449	regulation of transcription	1.95E-019	156
GO:0016071	mRNA metabolic process	1.72E-018	86
GO:0048284	organelle fusion	2.04E-018	40
GO:0006350	transcription	2.43E-018	133
GO:0006996	organelle organization	4.14E-018	213
GO:0006366	transcription from RNA polymerase II promoter	7.94E-018	44
GO:0016070	RNA metabolic process	1.39E-017	177
GO:0006351	transcription, DNA-dependent	5.44E-017	54
GO:0016043	cellular component organization	9.02E-016	285
GO:0016050	vesicle organization	1.33E-015	37
GO:0006379	mRNA cleavage	2.58E-015	21
GO:0006397	mRNA processing	5.06E-015	67
GO:0043144	snoRNA processing	1.60E-014	23
GO:0006944	cellular membrane fusion	3.28E-014	36
GO:0061025	membrane fusion	3.28E-014	36
GO:0044238	primary metabolic process	3.46E-014	448
GO:0034641	cellular nitrogen compound metabolic process	4.52E-014	307
GO:0034645	cellular macromolecule biosynthetic process	9.74E-014	215
GO:0009059	macromolecule biosynthetic process	2.11E-013	215
GO:0006807	nitrogen compound metabolic process	3.11E-013	309
GO:0043632	modification-dependent macromolecule catabolic process	1.76E-012	59
GO:0071840	cellular component organization or biogenesis	2.39E-012	293
GO:0016074	snoRNA metabolic process	2.60E-012	24
GO:0051252	regulation of RNA metabolic process	2.65E-012	109
GO:0071841	cellular component organization or biogenesis at cellular level	3.58E-012	277
GO:0031123	RNA 3'-end processing	5.45E-012	33
GO:0006355	regulation of transcription, DNA-dependent	9.06E-012	106

Figure A.2: GO enrichment of Yeast proteins in conserved modules found by aligning Yeast to Human using functional similarity.

GO Term	Description	P-value	# Proteins
GO:0009987	cellular process	1.56E-016	279
GO:0008152	metabolic process	9.38E-009	230
GO:0044237	cellular metabolic process	5.33E-011	229
GO:0044238	primary metabolic process	1.92E-011	219
GO:0043170	macromolecule metabolic process	1.03E-014	205
GO:0044260	cellular macromolecule metabolic process	2.07E-015	203
GO:0065007	biological regulation	2.01E-015	160
GO:0019538	protein metabolic process	2.25E-026	151
GO:0050789	regulation of biological process	2.89E-016	149
GO:0071840	cellular component organization or biogenesis	1.45E-007	140
GO:0016043	cellular component organization	6.94E-008	132
GO:0050794	regulation of cellular process	3.79E-013	129
GO:0071841	cellular component organization or biogenesis at cellular level	3.17E-006	128
GO:0044267	cellular protein metabolic process	6.06E-019	125
GO:0043412	macromolecule modification	1.02E-027	112
GO:0006464	protein modification process	2.56E-029	106
GO:0071842	cellular component organization at cellular level	2.06E-004	103
GO:0019222	regulation of metabolic process	1.44E-007	99
GO:0060255	regulation of macromolecule metabolic process	7.09E-006	84
GO:0080090	regulation of primary metabolic process	3.66E-005	84
GO:0050896	response to stimulus	2.34E-006	78
GO:0006793	phosphorus metabolic process	2.66E-033	72
GO:0006796	phosphate metabolic process	1.74E-035	71
GO:0009056	catabolic process	9.65E-007	69
GO:0006468	protein phosphorylation	2.67E-048	68
GO:0016310	phosphorylation	3.60E-045	68
GO:0044248	cellular catabolic process	1.24E-007	66
GO:0051716	cellular response to stimulus	1.56E-007	58
GO:0006950	response to stress	1.14E-004	58
GO:0009057	macromolecule catabolic process	1.30E-007	49
GO:0033554	cellular response to stress	5.98E-006	49
GO:0022607	cellular component assembly	2.32E-005	49
GO:0023052	signaling	3.49E-014	47
GO:0023033	signaling pathway	3.49E-014	47
GO:0022402	cell cycle process	2.01E-005	47
GO:0044265	cellular macromolecule catabolic process	7.07E-007	46
GO:0048519	negative regulation of biological process	2.08E-006	45
GO:0006508	proteolysis	2.42E-010	43
GO:0071844	cellular component assembly at cellular level	2.29E-004	43
GO:0007049	cell cycle	3.24E-006	42
GO:0023034	intracellular signaling pathway	2.31E-012	41
GO:0065008	regulation of biological quality	9.04E-005	40
GO:0051726	regulation of cell cycle	1.13E-012	39
GO:0051246	regulation of protein metabolic process	5.90E-008	37
GO:0048523	negative regulation of cellular process	1.66E-004	37
GO:0051603	proteolysis involved in cellular protein catabolic process	8.87E-012	36
GO:0065003	macromolecular complex assembly	6.31E-004	36
GO:0051301	cell division	1.14E-005	35
GO:0042221	response to chemical stimulus	9.74E-005	35

Figure A.3: GO enrichment of Yeast proteins in conserved modules found by aligning Yeast to Fly using sequence similarity.

GO Term	Description	P-value	# Proteins
GO:0006412	translation	1.33E-043	99
GO:0044260	cellular macromolecule metabolic process	1.11E-036	335
GO:0043170	macromolecule metabolic process	2.40E-035	338
GO:0006366	transcription from RNA polymerase II promoter	6.01E-031	50
GO:0034645	cellular macromolecule biosynthetic process	4.13E-030	200
GO:0009059	macromolecule biosynthetic process	1.13E-029	200
GO:0008380	RNA splicing	2.36E-025	55
GO:0000398	nuclear mRNA splicing, via spliceosome	6.14E-025	42
GO:0000377	RNA splicing, via transesterification reactions with bulged adenosine	1.28E-024	42
GO:0000375	RNA splicing, via transesterification reactions	3.28E-024	43
GO:0006351	transcription, DNA-dependent	1.66E-023	53
GO:0044238	primary metabolic process	7.63E-023	347
GO:0019538	protein metabolic process	3.52E-022	195
GO:0016070	RNA metabolic process	4.44E-022	147
GO:0044237	cellular metabolic process	3.22E-020	358
GO:0009987	cellular process	2.08E-019	420
GO:0032543	mitochondrial translation	2.85E-019	37
GO:0016071	mRNA metabolic process	4.07E-017	68
GO:0006397	mRNA processing	4.28E-017	58
GO:0032774	RNA biosynthetic process	5.44E-017	63
GO:0008152	metabolic process	5.91E-017	361
GO:0032568	general transcription from RNA polymerase II promoter	6.53E-017	20
GO:0034622	cellular macromolecular complex assembly	7.63E-017	75
GO:0060255	regulation of macromolecule metabolic process	2.32E-016	151
GO:0044267	cellular protein metabolic process	2.52E-016	163
GO:0065003	macromolecular complex assembly	3.01E-016	80
GO:0044249	cellular biosynthetic process	3.72E-016	217
GO:0010499	proteasomal ubiquitin-independent protein catabolic process	3.45E-015	14
GO:0009058	biosynthetic process	4.03E-015	217
GO:0010468	regulation of gene expression	8.04E-015	133
GO:2000112	regulation of cellular macromolecule biosynthetic process	1.68E-014	134
GO:0010556	regulation of macromolecule biosynthetic process	1.88E-014	134
GO:0080090	regulation of primary metabolic process	2.77E-014	150
GO:0031326	regulation of cellular biosynthetic process	4.73E-014	135
GO:0051123	RNA polymerase II transcriptional preinitiation complex assembly	4.74E-014	14
GO:0071842	cellular component organization at cellular level	4.91E-014	185
GO:0009889	regulation of biosynthetic process	6.49E-014	135
GO:0007005	mitochondrion organization	1.14E-013	50
GO:0019222	regulation of metabolic process	2.00E-013	158
GO:0006413	translational initiation	7.64E-013	20
GO:0070897	DNA-dependent transcriptional preinitiation complex assembly	8.55E-013	15
GO:0006906	vesicle fusion	1.33E-012	22
GO:0051246	regulation of protein metabolic process	3.98E-012	57
GO:0022607	cellular component assembly	5.48E-012	86
GO:0071844	cellular component assembly at cellular level	6.76E-012	81
GO:0090304	nucleic acid metabolic process	9.04E-012	180
GO:0034621	cellular macromolecular complex subunit organization	1.16E-011	81
GO:0006417	regulation of translation	1.61E-011	43
GO:0043933	macromolecular complex subunit organization	2.42E-011	85

Figure A.4: GO enrichment of Yeast proteins in conserved modules found by aligning Yeast to Fly using functional similarity.

Appendix B

GO terms in mitotic temporal labels

The complete list of all GO terms under each mitotic temporal label used in Chapter 4.

A. Prior to M phase

1. *centriole replication*
2. *spindle pole body duplication*

B. Entry into mitosis

1. *G2/M transition of mitotic cell cycle*
2. *meiotic G2/MI transition*

C. Prophase

1. *spindle pole body separation*
2. *lamin depolymerization*
3. *meiotic nuclear envelope disassembly*
4. *mitotic chromosome condensation*
5. *mitotic nuclear envelope disassembly*

D. Early Pro-metaphase (Spindle and kinetochore formation)

1. *establishment of meiotic spindle localization*
2. *establishment of mitotic spindle localization*
3. *meiotic spindle organization*
4. *mitotic spindle organization*
5. *meiotic spindle stabilization*
6. *mitotic spindle stabilization*
7. *kinetochore organization*

E. Pro-metaphase (kinetochore-MT interactions)

1. *attachment of spindle microtubules to kinetochore during meiotic chromosome segregation*
2. *attachment of spindle microtubules to kinetochore during mitosis*
3. *attachment of spindle microtubules to meiotic chromosome*
4. *attachment of spindle microtubules to mitotic chromosome*
5. *sister chromatid biorientation*

F. Late Pro-metaphase (metaphase plate formation)

1. *meiotic metaphase plate congression*
2. *mitotic metaphase plate congression*

G. Metaphase/Anaphase transition

1. *exit from mitosis*
2. *mitotic metaphase/anaphase transition*

H. Early Anaphase (Chromosome segregation)

1. *centromere separation*
2. *chromosome separation*

3. *male meiosis chromosome segregation*
4. *meiotic chromosome movement towards spindle pole*
5. *mitotic chromosome movement towards spindle pole*
6. *sister chromatid segregation*
7. *chromosome segregation*
8. *meiotic chromosome segregation*

I. Mid Anaphase (central spindle formation)

1. *spindle midzone assembly*
2. *spindle elongation*

J. Late anaphase (contractile ring formation)

1. *contractile ring maintenance involved in cytokinesis during cell cycle*
2. *contractile ring localization involved in cytokinesis during cell cycle*

K. Telophase

1. *meiotic chromosome decondensation*
2. *meiotic nuclear envelope reassembly*
3. *mitotic chromosome decondensation*
4. *mitotic nuclear envelope reassembly*
5. *mitotic nuclear pore complex reassembly*
6. *nuclear pore organization*

L. Cytokinesis

1. *cytokinesis during cell cycle*
2. *cytokinesis*
3. *cytokinetic process*
4. *barrier septum formation*

5. *cell plate formation*
6. *cell separation during cytokinesis*
7. *cell septum edging catabolic process*
8. *cellular bud neck septin ring organization*
9. *contractile ring contraction involved in cytokinesis during cell cycle*
10. *contractile ring localization involved in cytokinesis during cell cycle*
11. *contractile ring maintenance involved in cytokinesis during cell cycle*
12. *cytokinesis, completion of separation*
13. *cytokinesis, initiation of separation*
14. *cytokinesis, site selection*
15. *formation of actomyosin apparatus involved in cytokinesis*
16. *medial membrane band formation*
17. *membrane addition at site of cytokinesis*
18. *primary cell septum biogenesis*
19. *primary cell septum disassembly*
20. *regulation of contractile ring contraction involved in cytokinesis during cell cycle*
21. *selection of site for barrier septum formation*

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. 29

Allensby B 2011 Diagrammatical example of mitosis. `http://www.ba-education.com/for/science/dnabiology.html`. 99

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. 15

Aloy P and Russell RB 2002 Potential artefacts in protein-interaction networks. *FEBS Letters* **530**(1-3), 253 – 254. 9

Aloy P and Russell RB 2006 Structural systems biology: modelling protein interactions. *Nature Reviews Molecular Cell Biology* **7**(3), 188–197. 9

Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W and Lipman DJ 1997 Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.. *Nucleic acids research* **25**(17), 3389–3402. 42

Aranda B, Achuthan P, Alam-Faruque Y, Armean I, Bridge A, Derow C, Feuermann M, Ghanbarian AT, Kerrien S, Khadake J, Kerssemakers J, Leroy C, Menden M, Michaut M, Montecchi-Palazzi L, Neuhauser SN, Orchard S, Perreau V, Roechert B, van Eijk K and Hermjakob H 2009 The IntAct molecular interaction database in 2010. *Nucl. Acids Res.* p. gkp878. 69

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. 40

Bader G and Hogue C 2003 An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinformatics* **4**(1), 2. 22

Bader GD and Hogue CW 2002 Analyzing yeast protein-protein interaction data obtained from different sources.. *Nature biotechnology* **20**(10), 991–997. 4

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. 9

Barabasi AL and Albert R 1999 Emergence of Scaling in Random Networks. *Science* **286**(5439), 509–512. 17

Barabasi AL and Oltvai ZN 2004 Network biology: understanding the cell's functional organization. *Nature Reviews Genetics* **5**(2), 101–113. 13, 17

Barabsi AL and Albert R 1999 Emergence of Scaling in Random Networks. *Science* **286**(5439), 509–512. 17

Batada NN, Hurst LD and Tyers M 2006a Evolutionary and Physiological Importance of Hub Proteins. *PLoS Computational Biology* **2**(7), e88+. 64

Batada NN, Reguly T, Breitkreutz A, Boucher L, Breitkreutz BJ, Hurst LD and Tyers M 2006b Stratus not altocumulus: A new view of the yeast protein interaction network. *PLoS Biol* **4**(10), e317. 29

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. 29

Beaumont M, Zhang W and Balding DJ 2002 Approximate Bayesian Computation in Population Genetics. *Genetics* **162**(4), 2025–2035. 76

Berg JM, Tymoczko JL and Stryer L 2006 *Biochemistry (Biochemistry (Berg))* sixth edition edn. W. H. Freeman. 2

Besag J 1993 Statistical analysis of dirty pictures*. *Journal of Applied Statistics* **20**(5), 63–87. 101

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. 7

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. 8

Braak CJ 2006 A markov chain monte carlo version of the genetic algorithm differential evolution: easy bayesian computing for real parameter spaces. *Statistics and Computing* **16**, 239–249. 103

Brändén C and Tooze J 1991 *Introduction to protein structure*. Garland Publishing, New York. 2

Breitkreutz BJ, Stark C and Tyers M 2003 The grid: The general repository for interaction datasets. *Genome Biology* **4**(3), R23. 9

Brohee S and van Helden J 2006 Evaluation of clustering algorithms for protein-protein interaction networks. *BMC Bioinformatics* **7**(1), 488. 27

Brown KR and Jurisica I 2005 Online Predicted Human Interaction Database. *Bioinformatics* **21**(9), 2076–2082. 29

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. 10, 70

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. 25, 83

Chen PY, Deane CM and Reinert G 2008 Predicting and validating protein interactions using network structure. *PLoS Comput Biol* **4**(7), e1000118. 28

Chiang T, Scholtens D, Sarkar D, Gentleman R and Huber W 2007 Coverage and error models of protein-protein interaction data by directed graph analysis. *Genome Biology* **8**(9), R186. 9

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. 4

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. 30

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. 8

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. 13

Deane CM, Salwiski , Xenarios I and Eisenberg D 2002 Protein Interactions. *Molecular & Cellular Proteomics* **1**(5), 349–356. 10

Deng M, Chen T and Sun F 2004 An integrated probabilistic model for functional prediction of proteins.. *J Comput Biol* **11**(2-3), 463–475. 26

Deng M, Mehta S, Sun F and Chen T 2002 Inferring DomainDomain Interactions From ProteinProtein Interactions. *Genome Research* **12**(10), 1540–1548. 27, 70

Deng M, Zhang K, Mehta S, Chen T and Sun F 2003 Prediction of protein function using proteinprotein interaction data. *Journal of Computational Biology* **10**(6), 947–960. 26, 101

Dezs Z, Oltvai ZN and Barabsi AL 2003 Bioinformatics Analysis of Experimentally Determined Protein Complexes in the Yeast Saccharomyces cerevisiae. *Genome Research* **13**(11), 2450–2454. 20

Dongen S 2000 A cluster algorithm for graphs. Technical report, Amsterdam, The Netherlands, The Netherlands. 21

Dutkowski J and Tiuryn J 2007 Identification of functional modules from conserved ancestral protein protein interactions. *Bioinformatics* **23**(13), i149–158. 34

Eden E, Navon R, Steinfeld I, Lipson D and Yakhini Z 2009 Gorilla: a tool for discovery and visualization of enriched go terms in ranked gene lists. *BMC Bioinformatics* **10**(1), 48. 60

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. 8

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. 8

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

Evlampiev K and Isambert H 2007 Modeling protein network evolution under genome duplication and domain shuffling. *BMC Systems Biology* **1**(1), 49. 18

Felsenstein J 1989 PHYLIP - Phylogeny Inference Package (Version 3.2). *Cladistics* **5**, 164–166. 34

Finn RD, Mistry J, Tate J, Coggill 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. 27, 70

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. 33

Flannick J, Novak AF, Do CB, Srinivasan BS and Batzoglou S 2008 Automatic parameter learning for multiple network alignment *RECOMB*, pp. 214–231. 35

Frey BJ and Dueck D 2007 Clustering by Passing Messages Between Data Points. *Science* **315**(5814), 972–976. 22

Geman S and Geman D 1984 Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **PAMI-6**(6), 721–741. 101, 102

Girvan M and Newman MEJ 2002 Community structure in social and biological networks. *Proceedings of the National Academy of Sciences of the United States of America* **99**(12), 7821–7826. 45

Goll J and Uetz P 2006 The elusive yeast interactome. *Genome Biology* **7**(6), 223. 10

Güldener U, Münsterkötter M, Kastenmüller G, Strack N, van Helden J, Lemer C, Richelles J, Wodak SJ, García-Martínez J, Pérez-Ortín JE, Michael H, Kaps A, Talla E, Dujon B, André B, Souciet JL, De Montigny J, Bon E, Gaillardin C and Mewes HW 2005 CYGD: the Comprehensive Yeast Genome Database.. *Nucleic Acids Res.* 49

Guo X and Hartemink AJ 2009 Domain-oriented edge-based alignment of protein interaction networks. *Bioinformatics* **25**(12), i240–1246. 34

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. 28, 29

Hart GT, Ramani A and Marcotte E 2006 How complete are current yeast and human protein-interaction networks?. *Genome Biology* **7**(11), 120. 4, 6, 10, 38

Hartwell LH, Hopfield JJ, Leibler S and Murray AW 1999 From molecular to modular cell biology.. *Nature* **402**(6761 Suppl), C47–C52. 15, 20

Higham DJJ, Rasajski M and Przulj N 2008 Fitting a Geometric Graph to a Protein-Protein Interaction Network.. *Bioinformatics.* 73

Hishigaki H, Nakai K, Ono T, Tanigami A and Takagi T 2001 Assessment of prediction accuracy of protein function from proteinprotein interaction data. *Yeast* **18**(6), 523–531. 25, 26

Hu P, Janga SC, Babu M, Daz-Meja JJ, Butland G, Yang W, Pogoutse O, Guo X, Phanse S, Wong P, Chandran S, Christopoulos C, Nazarians-Armavil A, Nasseri NK, Musso G, Ali M, Nazemof N, Eroukova V, Golshani A, Paccanaro A, Greenblatt JF, Moreno-Hagelsieb G and Emili a 2009 Global Functional Atlas of Escherichia coli Encompassing Previously Uncharacterized Proteins. *PLoS Biol* **7**(4), e1000096. 83

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. 10

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. 3

Huynen M, Snel B, Lathe W and Bork P 2000 Predicting protein function by genomic context: quantitative evaluation and qualitative inferences.. *Genome Res* **10**(8), 1204–1210. 24

Ispolatov I, Krapivsky PL and Yuryev A 2005a Duplication-divergence model of protein interaction network. *Phys. Rev. E* **71**(6), 061911. 17

Ispolatov I, Krapivsky PL, Mazo I and Yuryev A 2005b Cliques and duplication–divergence network growth. *New Journal of Physics* **7**, 145. 18, 71

Ito T, Chiba T, Ozawa R, Yoshida M, Hattori M and Sakaki Y 2001 A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proceedings of the National Academy of Sciences of the United States of America* **98**(8), 4569–4574. 70

Jeong H, Mason SP, Barabási AL and Oltvai ZN 2001 Lethality and centrality in protein networks.. *Nature* **411**(6833), 41–42. 13, 64

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. 28

Kanehisa M and Goto S 2000 KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research* **28**(1), 27–30. 33

Kasahara M 2007 The 2r hypothesis: an update. *Current Opinion in Immunology* **19**(5), 547 – 552. Hematopoietic cell death/Immunogenetics/Transplantation. 17

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. 30

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. 30

Keshava Prasad TS, Goel R, Kandasamy K and Keerthikumar S 2009a Human Protein Reference Database 2009 update. *Nucl. Acids Res.* **37**, D767–772. 49

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, Ramabadran S, Chaerkady R and Pandey A 2009b Human Protein Reference Database2009 update. *Nucleic Acids Research* **37**(suppl 1), D767–D772. 9

Kim PM, Korbel JO and Gerstein MB 2007 Positive selection at the protein network periphery: Evaluation in terms of structural constraints and cellular context. *Proceedings of the National Academy of Sciences* **104**(51), 20274–20279. 12

Kosmidis I 2007 brglm: Bias reduction in binary-response glms `http://go.warwick.ac.uk/kosmidis/software`. 103

Kourmpetis YAI, van Dijk ADJ, Bink MCAM, van Ham RCHJ and ter Braak CJF 2010 Bayesian markov random field analysis for protein function prediction based on network data. *PLoS ONE* **5**(2), e9293. 102, 103

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. 33

Lee H, Tu Z, Deng M, Sun F and Chen T 2006 Diffusion kernel-based logistic regression models for protein function prediction.. *OMICS* **10**(1), 40–55. 26

Lee PH, Huang CH, Fang JF, Tsai J and Ng KL 2005 Study of the protein-protein interaction networks via random graph approach. *Cognitive Informatics, IEEE International Conference on* **0**, 110–119. 17

Lhost G 2011 Protein-protein, protein-drug interactions and ms `http://www.specmetcrime.com/noncovalent_complexes_in_mass_s.htm`. 5

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. 35, 36

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. 13

Lin D 1998 An Information-Theoretic Definition of Similarity *Proceedings of the 15th International Conference on Machine Learning*, pp. 296–304. 104

Liu Y, Liu N and Zhao H 2005 Inferring proteinprotein interactions through high-throughput interaction data from diverse organisms. *Bioinformatics* **21**(15), 3279–3285. 28

Lord PW, Stevens RD, Brass A and Goble CA 2003 Investigating semantic similarity measures across the gene ontology: the relationship between sequence and annotation. *Bioinformatics* **19**(10), 1275–1283. 117

Luciano, Rodrigues FA, Travieso G and Boas VPR 2007 Characterization of complex networks: A survey of measurements. *Advances in Physics* **56**(1), 167–242. 13

Maldonado E, I H, P C, L W and D R 1990 Factors involved in specific transcription by mammalian rna polymerase ii: role of transcription factors iia, iid, and iib during formation of a transcription-competent complex.. *Mol Cell Biol* **10**(12), 6335–6347. 54

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. 8

Marcotte EM, Xenarios I, van der Bliek 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. 8

Maslov S and Sneppen K 2002 Specificity and stability in topology of protein networks. *Science* **296**(5569), 910–913. 16

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. 30

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. 9

Mintseris J and Weng Z 2005 Structure, function, and evolution of transient and obligate proteinprotein interactions. *Proceedings of the National Academy of Sciences of the United States of America* **102**(31), 10930–10935. 3

N. Przulj, O. Kuchaiev AS and Hayes W 2010 Geometric Evolutionary Dynamics of Protein Interaction Networks *Proceedings of the 2010 Pacific Symposium on Biocomputing, Big Island, Hawaii, January 4-8.* 72

Narayanan M and Karp RM 2007 Comparing protein interaction networks via a graph match-and-split algorithm.. *Journal of computational biology : a journal of computational molecular cell biology* **14**(7), 892–907. 42, 43

Newman MEJ and Girvan M 2004 Finding and evaluating community structure in networks. *Phys. Rev. E* **69**(2), 026113. 21

Nowicki K and Snijders TAB 2001 Estimation and prediction for stochastic blockstructures. *Journal of the American Statistical Association* **96**(455), pp. 1077–1087. 17

Obayashi T, Hayashi S, Shibaoka M, Saeki M, Ohta H and Kinoshita K 2008 COXPRESdb: a database of coexpressed gene networks in mammals. *Nucleic Acids Research* **36**(suppl 1), D77–D82. 46

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. 30

Overbeek R, Fonstein M, DSouza 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. 7

Parrish J, Yu J, Liu G, Hines J, Chan J, Mangiola B, Zhang H, Pacifico S, Fotouhi F, DiRita V, Ideker T, Andrews P and Finley R 2007 A proteome-wide protein interaction map for campylobacter jejuni. *Genome Biology* **8**(7), R130. 83

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. 8, 30

Penrose M 2003 *Random Geometric Graphs.* Oxford University Press. 17

Phizicky E and Fields S 1995 Protein-protein interactions: methods for detection and analysis. *Microbiol. Rev.* **59**(1), 94–123. 6

Przeworski M 2003 Estimating the Time Since the Fixation of a Beneficial Allele. *Genetics* **164**(4), 1667–1676. 76

Puig O, Caspary F, Rigaut G, Rutz B, Bouveret E, Bragado-Nilsson E, Wilm M and Sraphin B 2001 The tandem affinity purification (tap) method: A general procedure of protein complex purification. *Methods* **24**(3), 218 – 229. 5, 6

R Development Core Team 2011 *R: A Language and Environment for Statistical Computing* R Foundation for Statistical Computing Vienna, Austria. ISBN 3-900051-07-0. 63

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. 19, 76

Reichardt J and Bornholdt S 2006 Statistical mechanics of community detection. *Phys. Rev. E* **74**(1), 016110. 21

Resnik P 1995 Using Information Content to Evaluate Semantic Similarity in a Taxonomy *IJCAI*, pp. 448–453. 41

Rives AW and Galitski T 2003 Modular organization of cellular networks. *Proceedings of the National Academy of Sciences of the United States of America* **100**(3), 1128–1133. 15

Ruepp A, Brauner B, Dunger-Kaltenbach I, Frishman G, Montrone C, Stransky M, Waegele B, Schmidt T, Doudieu ON, Stmpflen V and Mewes HW 2008 CORUM: the comprehensive resource of mammalian protein complexes. *Nucleic Acids Research* **36**(suppl 1), D646–D650. 49

Saeed R and Deane C 2008 An assessment of the uses of homologous interactions. *Bioinformatics* **24**(5), 689–695. 38

Saito R, Suzuki H and Hayashizaki Y 2002 Interaction generality, a measurement to assess the reliability of a proteinprotein interaction. *Nucleic Acids Research* **30**(5), 1163–1168. 10, 69

Saito R, Suzuki H and Hayashizaki Y 2003 Construction of reliable proteinprotein interaction networks with a new interaction generality measure. *Bioinformatics* **19**(6), 756–763. 10

Scannell DR, Butler G and Wolfe KH 2007 Yeast genome evolutionthe origin of the species. *Yeast* **24**, 929–942. 18

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. 41

Schwikowski B, Uetz P and Fields S 2000 A network of protein-protein interactions in yeast.. *Nature biotechnology* **18**(12), 1257–1261. 24, 26

Segal E, Wang H and Koller D 2003 Discovering molecular pathways from protein interaction and gene expression data. *Bioinformatics* **19**(suppl 1), i264–i272. 24

Sharan R and Ideker T 2006 Modeling cellular machinery through biological network comparison. *Nature Biotechnology* **24**, 427–433. 32

Sharan R, Suthram S, Kelley RM, Kuhn T, McCuine S, Uetz P, Sittler T, Karp RM and Ideker T 2005 Conserved patterns of protein interaction in multiple species. *Proceedings of the National Academy of Sciences of the United States of America* **102**(6), 1974–1979. 33, 50

Sharan R, Ulitsky I and Shamir R 2007 Network-based prediction of protein function. *Mol Syst Biol*. 101

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. 35

Skrabanek L, Saini H, Bader G and Enright A 2008 Computational prediction of proteinprotein interactions. *Molecular Biotechnology* **38**, 1–17. 10.1007/s12033-007-0069-2. 8

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. 20, 66

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. 90, 93

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. 7

Tanaka R, Yi TM and Doyle J 2005 Some protein interaction data do not exhibit power law statistics. *FEBS Letters* **579**(23), 5140 – 5144. 13

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. 24

Tatusov RL, Galperin MY, Natale DA and Koonin EV 2000 The cog database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Research* **28**(1), 33–36. 76

ter Braak C and Vrugt J 2008 Differential evolution markov chain with snooker updater andfewer chains. *Statistics and Computing* **18**, 435–446. 10.1007/s11222-008-9104-9. 103

ThermoScientific 2011 Schematic summary of a standard co-immunoprecipitation assay. `http://www.piercenet.com/Proteomics/browse.cfm?fldID=9C471132-0F72-4F39-8DF0-455FB515718F`. 7

Uetz P, Giot L, Cagney G, Mansfield TA, Judson RS, Knight JR, Lockshon D, Narayan V, Srinivasan M, Pochart P, Qureshi-Emili A, Li Y, Godwin B, Conover D, Kalbfleisch T, Vijayadamodar G, Yang M, Johnston M, Fields S and Rothberg JM 2000 A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. *Nature* **403**(6770), 623–627. 70

Uetz P, Titz B and Cagney G 2008 Experimental methods for protein interaction identification and characterization In *Protein-protein Interactions and Networks* (ed. Dress A, Vingron M, Myers G, Giegerich R, Fitch W, Pevzner PA, Panchenko A and Przytycka T) vol. 9 of *Computational Biology* Springer London pp. 1–32. 10.1007/978-1-84800-125-1_1.4

Venkatesan K, Rual J, Vazquez A, Stelzl U, Lemmens I, HirozaneandKishikawa T, Hao T, Zenkner M, Xin X, Goh K, Yildirim MA, Simonis N, Heinzmann K, Gebreab F, Sahalie JM, Cevik S, Simon C, de Smet A, Dann E, Smolyar A, Vinayagam A, Yu H, Szeto D, Borick H, Dricot A, Klitgord N, Murray RR, Lin C, Lalowski M, Timm J, R K, Boone C, Brn P, Cusick ME, Roth FP, Hill DE, Tavernier J, Wanker EE, Barabasi A and Vidal M 2009 An empirical framework for binary interactome mapping. *Nat Meth* **6**, D767–772. 93

von Mering C, Jensen LJ, Snel B, Hooper SD, Krupp M, Foglierini M, Jouffre N, Huynen MA

and Bork P 2005 STRING: known and predicted protein-protein associations, integrated and transferred across organisms.. *Nucleic acids research.* 81

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. 4, 11

Wang JZ, Du Z, Payattakool R, Yu PS and Chen CF 2007 A new method to measure the semantic similarity of GO terms. *Bioinformatics* **23**(10), 1274–1281. 42, 105

Wilkins MR and Kummerfeld SK 2008 Sticking together? falling apart? exploring the dynamics of the interactome. *Trends in Biochemical Sciences* **33**(5), 195 – 200. 29

Winzeler EA, Shoemaker DD, Astromoff A, Liang H, Anderson K, Andre B, Bangham R, Benito R, Boeke JD, Bussey H, Chu AM, Connelly C, Davis K, Dietrich F, Dow SW, El Bakkoury M, Foury F, Friend SH, Gentalen E, Giaever G, Hegemann JH, Jones T, Laub M, Liao H, Liebundguth N, Lockhart DJ, Lucau-Danila A, Lussier M, M'Rabet N, Menard P, Mittmann M, Pai C, Rebischung C, Revuelta JL, Riles L, Roberts CJ, Ross-MacDonald P, Scherens B, Snyder M, Sookhai-Mahadeo S, Storms RK, Véronneau S, Voet M, Volckaert G, Ward TR, Wysocki R, Yen GS, Yu K, Zimmermann K, Philippsen P, Johnston M and Davis RW 1999 Functional characterization of the S. cerevisiae genome by gene deletion and parallel analysis.. *Science (New York, N.Y.)* **285**(5429), 901–906. 63

Xenarios I, Salwinski L, Duan XJ, Higney P, Kim SM and Eisenberg D 2002a DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions. *Nucl. Acids Res.* **30**(1), 303–305. 49

Xenarios I, Salwnski L, Duan XJ, Higney P, Kim SM and Eisenberg D 2002b DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions. *Nucleic Acids Research* **30**(1), 303–305. 9

Yeang CH, Ideker T and Jaakkola T 2004 Physical network models. *Journal of Computational Biology* **11**(2-3), 243–262. 29

Yu H, Kim PM, Sprecher E, Trifinov V and Gerstein M 2007 The Importance of Bottlenecks in

Protein Networks: Correlation with Gene Essentiality and Expression Dynamics. *PLoS Computational Biology* **preprint**(2007), e59.eor+. 64

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. 9

Zaslavskiy M, Bach F and Vert JP 2009 Global alignment of protein-protein interaction networks by graph matching methods. *Bioinformatics* **25**(12), i259–1267. 35

Zhang J June 2003 Evolution by gene duplication: an update. *Trends in Ecology and Evolution* **18**, 292–298(7). 17

Zotenko E, Mestre J, O'Leary DP and Przytycka TM 2008 Why Do Hubs in the Yeast Protein Interaction Network Tend To Be Essential: Reexamining the Connection between the Network Topology and Essentiality. *PLoS Comput Biol* **4**(8), e1000140+. 64