Communities and Homology in Protein-protein Interactions

Anna Lewis

Balliol College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Michaelmas 2011

This thesis is dedicated to the inspirational Jen Weterings.

Wishing her well in her ongoing recovery.

Acknowledgements

With thanks to Charlotte, Nick and Mason, for being the best supervisors a girl could ask for.

With thanks to all those who have helped me with the work in this thesis. In particular Sumeet Agarwal, Rebecca Hamer, Gesine Reinert, James Wakefield, Simon Myers, Steve Kelly, Nicholas Lewis, members of the Oxford Protein Informatics Group and members of the Oxford/Imperial Systems and Signalling group.

With thanks to all those who have made my DPhil years a wonderful experience, in particular my family, my Balliol family, and my DTC family.

Abstract

Knowledge of protein sequences has exploded, but knowledge of protein *function* is needed to make use of sequence information, and this lags behind. A protein's function must be understood in context and part of this is the network of interactions between proteins. What are the relationships between protein function and the structure of the interaction network? In the first part of my thesis, I investigate the functional relevance of clusters, or *communities*, of proteins in the yeast protein interaction network. Communities are candidates for biological modules. The work I present is the first to systematically investigate this structure at multiple scales in such networks. I develop novel tests to assess whether communities are functionally homogeneous, and demonstrate that almost every protein is found in a functionally homogeneous community at some scale.

The evolution of protein sequences is well-studied, but comparatively little is known about the evolution of protein function. Such knowledge is needed to understand when it is appropriate to annotate newly sequenced proteins by transferring functional information from homologs–i.e. evolutionarily related proteins. In the second part of my thesis, I assess the success of transferring protein-protein interactions across species and use this to estimate the rate at which interactions are lost in evolution. At levels of sequence similarity associated with functional annotation transfer, I demonstrate that protein-protein interaction transfer is unreliable.

The relevance of community structure for understanding protein function and the low conservation of individual interactions, suggests a possible role for communities in the evolution of cellular function. I discuss this possibility in my conclusions.

Contents

List of Figures

Chapter 1

Introduction

1.1 Overview

1.1.1 Systems biology and protein-protein interactions

Systems biologists seek to supplement the successes of molecular biology with a more systems level approach [166]. Molecular biology has tended to focus on in depth investigations of individual molecules, whereas the systems biology approach is to study how individual components interact together in order to bring about system-level properties. One example is how cardiac cells interact to produce the beat of the heart [237]. Another example is how macro-molecules in the cell, such as proteins, interact to bring about cellular function [20].

Systems biology aims to be a post-genomic science: it is, in part, a response to the realisation that the gap between sequencing DNA and identifying genes, and understanding what function the products of those genes have, is a very large one. In particular, the vast amount of sequence data collected across the tree of life has allowed a fairly good understanding of the many ways in which sequence can change, but this does not straightforwardly translate to insight into how the functions of gene products can change and develop [66, 343]. The interactions of gene products – in particular of proteins – are at an interesting position in the hierarchy of biological organisation from genotype to phenotype, in that they are both of functional relevance and have a close link to sequence: genetic sequence determines protein sequence, which determines protein structure, which is thought to largely determine protein-protein interactions [40, 348].

Studies of protein-protein interactions can give insight both into how complex cellular behaviour arises from the behaviour of individual macro-molecules and into how this complex cellular behaviour evolved. It is therefore hardly surprising that over the last few years considerable effort has been made to determine experimentally protein-protein interactions, and that large data sets incorporating tens of thousands of interactions are now available. In this chapter, I review the study of protein-protein interactions. In Chapter 2, I introduce some of the mathematical tools relevant to subsequent chapters.

1.1.2 Communities

The set of protein-protein interactions within a species can be considered as a network (a *Protein-protein Interaction Network* or PIN), with proteins as nodes and known interactions between them as edges. The patterns of interaction between proteins are anticipated to be highly structured, possibly at many different scales: one can investigate not only the properties of an individual component (a protein and its interacting partners) and the properties of the whole (the entire PIN), but also mid-level structure. Consider the analogy of the friendship network between pupils in a school: one might expect class groups and, at a larger scale, year groups to appear as dense regions in the network. Densely connected regions in a network are often called *communities* [91] and, as the school example illustrates, they potentially exist at multiple scales. Are there communities in PINs that have functional significance, and if so at what scales? What, if anything, can a study of their structure tell us about the functional organisation of the cell? One hypothesis is that a community of proteins is a *module*: it carries out a particular cellular task, in comparative isolation from the rest of the system. These questions are the starting point for Chapter 3.

1.1.3 Homology

Homology, which means similarity through common descent, occurs on many scales, from genetic sequence to anatomy. The high degree of observed protein sequence similarity between proteins in different species gives a strong expectation that discoveries about protein function made in one species will provide understanding in another [66]. Are protein-protein interactions conserved through evolution? There now exist considerable data for a few species, which enables comparative studies. If protein-protein interactions are well-conserved through evolution, then we can be more confident that knowledge of protein function gained in one species can be 'transferred' to other species. If they are not, then this goes some way to explaining the large amount of phenotypic divergence that exists despite the very high degree of sequence conservation. This is the topic of Chapter 4.

1.1.4 Communities and homology of protein-protein interactions

It is often assumed that functional modules are also evolutionary modules [125] – that evolution tinkers with the connections between modules rather than with modules themselves. But is it the case that nature instead tinkers with interactions within modules, as the effects of these are fairly isolated and hence less likely to result in disruption? In this thesis, I do not examine the role of communities in evolution, but the work presented does suggest some lines of enquiry that would explore these questions. I discuss this in Chapter 5.

1.2 Protein-protein interaction data

1.2.1 Proteins

Proteins are large macro-molecules composed of at least one polypeptide (i.e. a chain of amino acids held together by peptide bonds). The word ‘protein’ comes from the Greek meaning ‘primary’, a fitting name as proteins are of primary importance to cellular life, carrying out most of the cellular tasks. For example, major classes of proteins include enzymes, ion channels, antibodies, transcription factors, hormones, chaperones, and cytoskeletal constituents.

In the simplest case, a sequence of DNA is transcribed into RNA, which is translated via the genetic code into a polypeptide. We now know that the connection between DNA sequence and polypeptide chain can be significantly more complicated than this. The human genome is estimated to have about $20,000 - 25,000$ protein coding genes [308], though the number of proteins is much larger because of multiple splice variants.

Polypeptides encoding proteins typically fold into a unique three dimensional structure, which is thought to be determined primarily by the amino-acid sequence [29]. There are two common structural motifs within proteins, alpha-helices and beta-sheets, which are referred to as secondary structure.

Proteins come in many different shapes and sizes, and this diversity is matched by the functions in which proteins participate. Proteins can be classified based on their amino-acid sequence, their three-dimensional structure, the molecular function(s) they perform, the biological process(es) in which they partake, etc. Such classifications are obviously not independent of one another.

Proteins do not carry out their functions in isolation but rather act in concert with other cellular constituents, including other proteins.

1.2.2 What is a protein-protein interaction?

Protein-protein interactions are of diverse types. For example, some proteins are never found except when interacting with each other [e.g. 294] (such interactions are termed *obligate*), while other interactions are *transient* [e.g. 294]. Examples of obligate interactions include multi-subunit enzymes; examples of transient interactions include hormone-receptor and enzyme-inhibitor interactions. The distinction between obligate and transient is not black and white, as many intermediate strengths and durations of interaction are possible [220]. Protein-protein interactions likely depend on conditions or life-cycle-stage. In the case of multi-cellular organisms they are also likely dependent on cell-type. Some protein-protein interactions may only occur when the proteins have particular post-translational modifications [178].

What, then, counts as a protein-protein interaction? The literature has rather side-stepped the issue of delineating what a protein-protein interaction is and is not (e.g. does it have to be specific? If so, how specific?). Instead, the outputs from various experimental protocols, suitably filtered for obvious false-positives (i.e. interactions reported but not truly there), are taken as definitions of protein-protein interactions. In this way, a protein-protein interaction is largely synonymous with 'a true positive from one of the main methodologies'.

In the next subsection, I outline the main experimental protocols and discuss the currently available data sets. To supplement experimentally derived data, tools for computationally predicting protein-protein interactions have been developed (see Appendix A).

1.2.3 Experimental protocols

1.2.3.1 Yeast-two-hybrid screens

The yeast-two-hybrid (y2h) technique was first proposed by Fields and Song [85], and studies using this technique and its variants are amongst the largest contributors to currently available data sets. A good review of the technique is found in Brückner et al. [43]. Prior to Fields and Song [85], almost all reports of protein-protein interactions relied on biochemical techniques. Yeast-two-hybrid allowed the report of interactions in living cells. The method takes advantage of the fact that some transcription factors (i.e. proteins that are involved in the control of transcription of DNA) are composed of two separate domains, a binding domain and an activating domain, which need to be in close proximity before transcription can occur. In y2h, one protein (the 'bait') is expressed as a fusion with the binding domain, and another protein (the 'prey') is expressed as a fusion with the activating domain. It is only if the two proteins physically interact that the binding domain and the activating domain can activate transcription of some reporter gene(s) [85]. The expression of the reporter gene, which can for example allow growth on a specific medium or express a fluorescent protein [43], is then a signal that the bait and prey interact.

High-throughput studies have been performed in *Saccharomyces cerevisiae* [145, 326, 360], *D. melanogaster* [90, 107], *C. elegans* [185], and *Homo sapiens* [279, 327]. These studies have not tested for the existence of all possible interactions, and several sources of false-positive and false-negative errors are known.

False positives in y2h screens include the fact that the proteins are over-expressed in a non-native cellular localisation: such detected interactions might be real in the sense that the proteins bind specifically to each other, but never happen under normal cellular conditions [334]. More recent two-hybrid approaches allow the screen to be performed in mammalian cells [317], where this should be less of an issue. As the method relies on fusion proteins, misfolding of the bait or prey is possible, leading to increased affinity for certain targets [196].

False negatives can arise if the signal from the reporting genes is insufficiently strong, perhaps due to low protein abundance. The fusion proteins might misfold, blocking relevant interaction sites. Non-yeast proteins might not receive the correct post-translational modifications in yeast [43]. Self-interacting proteins can be hard to identify, as the baits interact with each other and the preys interact with each other, resulting in reduced concentrations of bait/prey interactions [104]. Classical y2h requires the expression of proteins in the nucleus, which does not happen for e.g. membrane proteins, so interactions involving membrane proteins are missed. Variants such as the split ubiquitin y2h [43] have been applied on a large scale to *S. cerevisiae* membrane proteins [217]. Some bait proteins can themselves be activators, meaning that an interaction with the prey-activating domain fusion protein is not needed for reporter gene transcription [85]. Such proteins are called ‘auto-activators’, and in classical y2h they must be excluded from screening for interactions. Various variants of the classical screen have been suggested that allow auto-activators to be included [43] and are incorporated in more recent studies [e.g. 360].

1.2.3.2 Purification and identification of complexes

Associations between proteins are also identified using biochemical purification of complexes followed by some method for identifying the members of the complex (for example, by mass spectrometry). The dominant low-throughput assay for purifying proteins in a complex is co-immunoprecipitation [29]. A high throughput purification technique known as tandem affinity purification (TAP) was introduced in 2001 by Puig et al. [260].

In co-immunoprecipitation assays an antibody that targets a known protein that is believed to be part of a larger complex is identified, and it is used to precipitate out that protein and any others that may associate with it. Often more than one antibody is required, as the part of the protein to which the antibody binds, the epitope, may be covered under cellular conditions (for example by the proteins that associate with the protein of interest). Co-immunoprecipitation assays can be based on endogenous proteins. In such cases, it is perceived as a 'gold standard' of association studies, though not suitable for high-throughput screens as some prior knowledge of the bait protein is needed.

TAP permits the high-throughput purification of complexes under cellular conditions. A TAP tag is fused to a bait protein, and the fusion is introduced into cells on a plasmid. Prey proteins then associate with the bait. The TAP tag allows the fusion protein and associated preys to be selected and purified using an affinity column. Those prey proteins that bind tightly enough to the bait to survive the purification can then be identified, typically by mass spectroscopy [260]. The TAP method is reviewed in e.g. Collins and Choudhary [60]. It has been applied on a large scale to *S. cerevisiae* [100, 101, 131, 170] and to *E. Coli* [9, 99]. A study focused on disease associated genes in *H. sapiens* cell lines has also been published [82].

In contrast to y2h screens that report binary physical interactions, the purification of complexes reveals only biochemical association. In converting a set of bait-prey associations into binary interactions, a choice has to be made about whether to report all bait-prey pairs only (the 'spoke model'), to additionally report all prey-prey pairs (the 'matrix' model), or to report some intermediate number [14, 73].

False positives can be generated from contaminants, which can bind to the column matrices used for affinity purification or form non-specific interactions with the bait protein [260]. Non-specific interactions are made more likely if the fusion of the tag to the bait alters its fold, exposing hydrophobic surfaces that are prone to interact non-specifically [196].

False negatives arise if the epitope tag does not end up on the surface of the bait protein or if it causes the protein to misfold [260]. In general the proteins must be abundant enough and the complexes must be sufficiently stable to withstand the purification procedure [260], though some have suggested modified protocols that are able to detect less stable interactions [e.g. 225, 341]. Self-interactions are often either removed [82] or not reported due to the lack of untagged baits [104]. In addition to the bias towards abundant proteins, there is a bias against membrane proteins, which are hard to purify [334]. An additional source of false-negatives is from failure to identify proteins by mass spectrometry.

1.2.3.3 Other methods

There are several other low-throughput technologies. Examples include crystallised complexes found in the protein data bank [294], fluorescence resonance energy transfer microscopy [158], and a technique called protein-fragment complementation assay (PCA) which relies on two proteins of interest being fused to complementary fragments of a reporter fragment [314].

1.2.3.4 Literature curation of small-scale experiments

The results of many small-scale experiments have been assembled into datasets by the manual curation of reports of interactions in the literature. For example in the 'LC' data set of Reguly et al. [269], over $9,000$ papers were curated, and $11,334$ *S. cerevisiae* protein-protein interactions were reported. The Human Protein Reference Database (HPRD, [161]), aims to manually curate protein-protein interactions in *H. sapiens*, including from small-scale experiments [161]. Such data sets clearly have biases towards proteins that have been extensively studied before, for example disease related proteins in *H. sapiens* and essential proteins in *S. cerevisiae* [119]. In Sambourg and Thierry-Mieg [287] a very strong correlation is demonstrated between how well-studied a protein is and how many interacting partners it has. This is true not only for the literature curated binary physical interactions from the BioGRID database [307], but also for the union of three high-throughput y2h studies [145, 326, 360].

1.2.3.5 Databases

Many protein-protein interaction databases have been set up to collate published data. The main extant databases are IntAct [7], MINT (the Molecular Interactions database [48]) and BioGRID [307]. The *H. sapiens*-specific HPRD [161] also collates many interactions. There are several other databases [15, 147, 194, 285].

The amount of protein-protein interaction data has grown steadily over the past decade. Figure 1.1 illustrates the growth in the protein-protein interaction data curated in the IntAct database (and is reproduced from Supplementary Figure 2 from the latest paper published by the IntAct group [7]). It is clear from this figure that the data remains concentrated in a few model species.

The different databases have different criteria for collation, and these can change with time. For example in BioGRID the matrix model is generally used for co-complex type data (see Section 1.2.3.2) [307], whereas in IntAct and MINT, the spoke model is now used [7, 48] with spoke-expanded co-complex interactions available for separate download. To give an example of the difference curation protocols can make, at the time of writing (October 25th 2011), the IntAct database contains $287,648$ interactions, of which $183,355$ are binary interactions and an additional $104,293$ are 'spoke expanded co-complexes' – i.e. the difference between the matrix model and spoke model for TAP type data. A comparison of these numbers to Figure 1.1 demonstrates that at least some spoke-expanded co-complex data was present in IntAct, as the total number of interactions reported in 2009 is larger than the total number of non-spoke expanded co-complexes reported on the IntAct website at the time of writing. Although this is not clear from their website nor published papers, it appears that recently a clear separation of spoke and spoke-expanded data has been made, whereas before (including the data I used in Chapter 3, downloaded in January 2010) a mixture of the two types was present dependent on the study being curated. For example, there are currently $104,990$ interactions reported for *S. cerevisiae*, of which $49,682$ are spoke-expanded interactions. The figure of roughly $65,000$ visible in Figure 1.1 must include some but not all of the spoke-expanded data.

Structured ontologies have now been developed for depositing interactions in databases [160], though these have not been adopted uniformly [307].

1.2.4 Choice of data set

Researchers have numerous choices for constructing protein-protein interaction data sets to be used in a study. One very important choice is whether to include only interactions detected more than once to try and filter out false-positive interactions. The significance of this choice was made clear by Hakes et al. [119], where it is shown that keeping multiply observed interactions may lead to a more *reliable* data set but not a *representative* data set. As discussed below in Section 1.3.2, they show that results can depend on whether or not this filtering step is applied.

A second choice is whether to use data sets that combine physical association data (largely reported via y2h screens) and association data (largely reported via TAP type data). Owing to the very different protocols, one might expect these data sets to be complementary to each other. Yu et al. [360] demonstrated that the two data types are indeed complementary: interactions detected by y2h data are more likely to be transient signalling and inter-complex interactions.

Something a researcher has little control over, but which can have profound effects on results, is the data-handling protocols used by the authors of high-throughput studies [e.g. 282]. In the case of the two large *S. cerevisiae* TAP studies [101, 170], it has been shown that the striking differences in the results can be attributed largely to

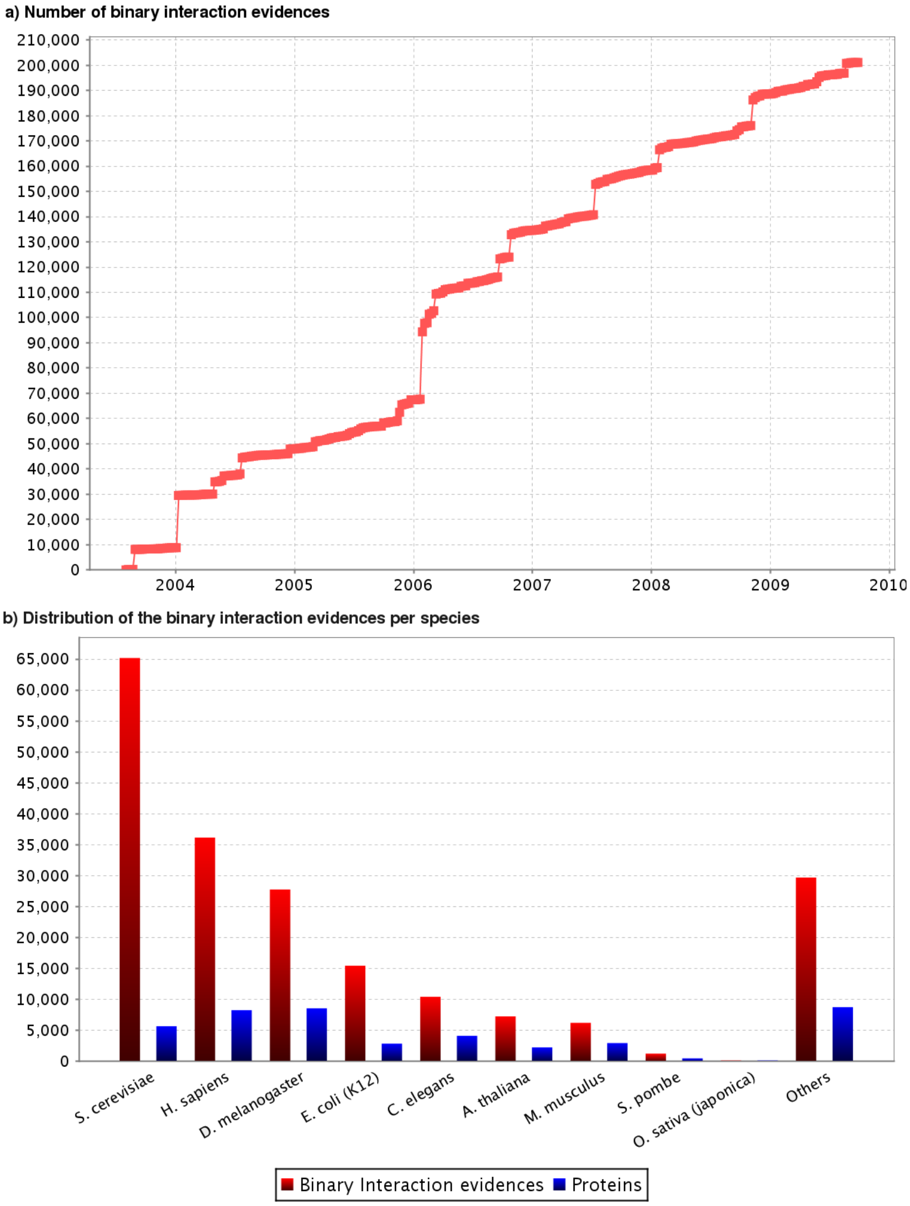


Figure 1.1: **The size of the IntAct database, as reproduced from Supplementary Figure 2 of Aranda et al. [7]. (This figure is reproduced under the terms of the Creative Commons Attribution Non-Commercial License).**

different data handling protocols [351]. This issue is compounded by the unavailability of the raw data for high-throughput studies in many cases [119].

1.2.5 Estimates of error rates

In Section 1.2.3 I outlined some of the potential sources of error and bias in protein-protein interaction detection methods. Here I review ways of estimating the magnitude of these errors in individual and combined data sets.

Initial error rates for high-throughput data sets relied heavily on gold standard sets of interactions. These were used either to define what should be expected of truly interacting proteins in terms of e.g. co-expression of the proteins [16, 70, 72] or topological patterns in the PIN [51], or used straightforwardly in intersection assays [e.g. 73, 78, 334], reviewed in Hart et al. [124]. These initial rates of false-positives were about 50% or even higher [73, 305]. An issue with gold-standard data sets is their reliability and representativeness. The most common choice of gold-standard data set for error analysis was the MIPS complex data [213]. Given that y2h screens were designed to detect transient interactions, interpreting their low overlap with co-complex data as evidence of very high false-positive rates may be unfair [136, 360]. In von Mering et al. [334], the authors found for the MIPS complex data that interacting proteins had a high tendency to be of the same functional type, and they used this to conclude that interactions between proteins of different functional types in high-throughput studies consist mostly of false-positives. This neglects the possibility that the MIPS data is itself biased – including the possibility that functional annotations may be based on interaction data, reinforcing this bias.

An alternative method to using gold-standard sets of interactions assumes that certain sets of interactions are more likely to be true positives. Examples of properties include functional similarity of the two proteins [51], the same cellular components annotated to both proteins [305], or the existence of homologous proteins from other species interacting [70, 283]. Such studies, based on assumptions as to the properties interacting proteins might be expected to have, risk those assumptions being wrong (for example, it does not appear that proteins that have similar expression profiles are more likely to interact [32]).

More recently, the estimates of false-positive rates in high-throughput studies – which were originally very high [73, 305] – have been revised down by methodology that pays closer attention to the data generating process. In Huang et al. [137], the analysis of 'interaction sequence tag' counts suggested false-positive rates of about 25% in *S. cerevisiae* and between 40% and 45% in *D. melanogaster* [137]. However, this methodology relies on raw data that is not always available. An 'empirical framework' for assessing error [332, 360] estimated the false-positive rates of y2h screens to be between $0\% - 26\%$, with 26% the value for the Ito-full data set [145], which was previously estimated to have the highest false-positive rate [73, 305]. This empirical framework relies on performing additional experiments and having access to many experimental protocol details, so is not generalisable.

In contrast to the case of high-throughput y2h studies, low-throughput literature-curated interactions are considered to have had negligible false-positive rates [63] – indeed, they were often used as gold-standard data sets. Cusick et al. [63] argued, based on a 're-curation' exercise, that the error rate in curation is as high as 45%. In this exercise, they counted it against an interaction that it was reported only once. However, a paper in response argued that the re-curation error rate was actually $2 - 9\%$ [286]. Cusick et al. [63] also demonstrated that there is a small overlap of interactions in literature-curated data sets due to a lack of overlap of publications examined. This suggests that false-negative rates are the largest concern, and that combining different databases will be necessary if one is aiming for as comprehensive a data set as possible.

A simple method to estimate the total false-negative rates, and hence the total

Table 1.1: **Estimates of the total number of protein-protein interactions in several species, in thousands of interactions**. When a range of numbers is given, this corresponds to ranges given in the original papers. It does not have a uniform interpretation across the different studies. The study of Huang et al. [137] first estimates the number of interactions per protein, and then estimates the total number of interactions from both the mean and median of this estimate.

	S. cerevisiae	*C. elegans*	*D. melanogaster*	*H. sapiens*
Stumpf et al. [313]	14 – 38	220 – 266	72 – 78	589 – 723
Sambourg and Thierry-Mieg [287]	32 – 43	-	-	-
Huang et al. [137], from mean	30	610	325	-
Huang et al. [137], from median	137	1250	613	
Venkatesan et al. [332]	-	-	-	74 – 200

number of interactions, extrapolates existing data sets to unstudied proteins [313]. The most common method relies on having two data sets assumed to be sampled independently of each other. In conjunction with estimated false-positive rates of the two data sets, the total number of interactions can be estimated via the hypergeometric distribution [73, 113, 124, 287]. Sambourg and Thierry-Mieg [287] illustrated that literature-curated and high-throughput studies are not in fact independent samplings from the true set of *S. cerevisiae* interactions, and they correct for this by restricting the literature-curated data set to very well studied proteins. They estimated the total number of binary *S. cerevisiae* interactions (i.e. excluding association type data) to be $37,600$, which is larger than earlier estimates (e.g. Sprinzak et al. [305] estimated $10,000 - 20,000$ interactions). In Table 1.1 I give a summary of estimates for the total number of interactions given in some of the papers mentioned in this section.

1.2.6 Considered as a network

We can represent a PIN by an adjacency matrix $\mathbf{A}$, where $A_{ij} = 1$ if proteins i and j interact, and $A_{ij} = 0$ otherwise. Considering sets of interactions as a network is useful for two main reasons. First, visual representation of a network can enable patterns to be spotted (this is mostly applicable to subsets of the PIN). Second, structural patterns in the PIN can be explored using the tools developed in graph theory and network science.

A few points are worth noting concerning the representation of protein-protein interactions as networks in practice:

- Although both y2h and TAP detect directed relationships (i.e. the bait-prey relationship is asymmetrical), the resulting PINs are almost always treated as undirected, i.e. symmetrical.

- Most analyses of PINs treat interactions as either present or absent. There are of course various ways one could consider weighting the edges – for example, by the probability that the interaction is indeed truly there [335].

- Despite the very different nature of the binary physical relationships reported by y2h screens and the co-association evidence that TAP experiments provide (see Section 1.2.3), these data types are frequently combined.

- Existing interaction data is not complete, and coverage of it is biased (see Section 1.2.3).

As tools for understanding the cell, PINs are clearly limited. Some of their most obvious shortcomings are that they do not incorporate any temporal or spatial information and many of the critical molecules of life are not proteins. In addition, current publicly available data is not available for different cell types or different cell states.

Protein interaction networks have been extensively studied since the availability of large-scale data sets. To give an idea of the scale of this endeavour, several studies of PINs have received over a thousand citations (e.g. [149, 202]). The structure of a PIN, often in conjunction with other types of biological data, has been used to make grand claims about cellular function and about the evolution of cellular complexity (epitomised and summarised in [20]).

In this thesis I use the term PIN when I am explicitly interested in the network structure of protein-protein interactions. I use the term *interactome* to mean simply the total set of protein-protein interactions under consideration.

1.3 Protein-protein interactions and function

There have been numerous attempts to connect PIN structure with biological function. Here I discuss the most prominent examples.

1.3.1 Predicting protein function

We expect the study of biological networks to reveal aspects of functional organisation, through seeing how proteins interact with each other to bring about particular cellular tasks. At the lowest level of structural organisation, we therefore anticipate that proteins that interact are involved in similar processes. One practical benefit of this is that for poorly-characterised proteins, we can predict their function based on the function of their interacting partners (reviewed in Sharan et al. [293] and Wang and Marcotte [340]).

There are many aspects of 'biological function'. Principally, these include the cellular tasks in which the protein is involved (e.g. molecular transport, transcription, metabolism) and which phenotypes result on disruption of the protein (e.g. disease, reviewed in Ideker and Sharan [142]). Function is itself not a clearly defined concept, and there is certainly no one way to correctly quantify 'similarity of function'. The most comprehensive attempt at cataloguing the function of gene products is the Gene Ontology, GO [12], which maintains a set of terms and the relationships between them (see Section 2.4), which third parties (for example, model organism databases) then annotate to gene products. The simplest use of protein-protein interactions to predict function is to assign to a protein the functional term(s) of its interacting partners [309].

This procedure, termed ‘guilt-by-association’, has many variants and has proven very hard to beat [293]. An alternative or extension of guilt-by-association methods is to incorporate indirect connections [e.g. 54] or execute some sort of label propagation method [e.g. 224], though it has recently been argued that these methods simply recapitulate information that was lost when the raw data sets were thresholded prior to analysis [106].

1.3.2 Degree distribution

The *degree* of a node is defined as the number of interacting partners it has (i.e. the number of edges connected to it). Many real-world networks were found to have degree distributions with heavy tails – i.e. the majority of nodes in a given network have few connections, but a small number of nodes (called ‘hubs’) have a very large number of interactions [235]. Under a simple model of a random network, known as the Erdős-Rényi graph, where each node has some probability p of being connected to every other, the distribution of node connectivities (the ‘degree distribution’) is binomial. This is very different to the heavy-tailed degree distributions often observed in real networks, and this was construed as an exciting finding that revealed deep truths about complex systems [19]. A particular form of heavy tailed distribution is a power-law distribution (often called ‘scale-free’ in the literature [19]), whose characteristic feature is that it is a straight line when plotted on a log-log plot. Power laws appear in physics in diverse places, particularly in the study of critical phenomena in statistical physics [357]. Scholars started to report power laws in numerous real-world networks [57]. The demonstration that a simple model – that of network growth with *preferential attachment* of new nodes to nodes which already had many connections – produced a power-law degree distribution added to the excitement [19].

It has since been shown that power laws are not the best statistical fit for many of the degree distributions [57], including those of PINs, where log-normal or stretched exponential distributions (depending on species) give a better statistical fit [310]. It has also been demonstrated that sub-networks of networks with power-law degree distributions need not have power-law degree distributions [311], particularly pertinent for the case of PINs where data is known to be far from complete. It has also been shown that many network models (random, exponential, power law, truncated normal), when sampled by first randomly selecting nodes and then selecting interactors of those nodes (a not unrealistic model of sampling for protein-protein interactions), give power-law degree distributions [121]. A thorough critique of the ubiquity of networks with power-law degree distributions is found in Fox Keller [94].

What is the biological relevance of heavy-tailed degree distributions? They have been used as a basis for strong claims about the evolution of cellular complexity, and I review these in Section 1.4.4. Here I focus on the connections between degree distributions and biological function.

It was found that hubs in the *S. cerevisiae* PIN were more likely to be essential proteins. (In this study, an 'essential protein' is defined as one that produces a lethal phenotype when the gene encoding it is engineered to not be transcribed [149]). This finding is referred to as the 'centrality-lethality' rule, and it has been reproduced in other data sets [e.g. 116]. A subsequent analysis (using a different definition of how important a protein is: the extent to which it is evolutionarily constrained) found that this effect is very small, though statistically significant, and is restricted to genes involved in the cell cycle and transcription [117]. In *H. sapiens*, the equivalent centrality-lethality rule might be that disease genes have a larger number of interactions. Proteins involved in cancer have been reported to be more likely to be hubs [150], whereas proteins involved in disease (including cancer!) have been found not to show this trend [110].

Maslov and Sneppen [202] reported that hub-hub interactions are under-represented (they conjectured that this could have arisen because such structure minimises unfavourable cross-talk between densely interconnected regions of the network centered on the hubs), but it has been show that the presence of this effect depends on how the PIN dataset is constructed [23, 119].

Proteins are expressed at different times within the cell, and this information can be overlaid onto a PIN in an attempt to investigate dynamic organisation in cells. This was performed using mRNA co-expression as a proxy for protein-coexpression for *S. cerevisiae* in Han et al. [120], where it was claimed that two distinct types of hub were identifiable: 'party' hubs, which tend to be co-expressed with their interacting partners, and 'date' hubs, which are not. Party hubs were proposed to function locally, coordinating particular biological processes. Date hubs were proposed to operate more globally, by connecting disparate biological processes. There have been several subsequent papers that claim to refute or attempt to restate the findings, with the balance of evidence suggesting that the initial findings, which were reported on a small data set, were both dubious and particular to that data set [1, 23, 24, 31, 347].

1.3.3 Motifs

Network *motifs*, first introduced in Milo et al. [218], are small patterns of connected nodes. Some motifs were found to be over-represented (compared to random) in particular networks, and these were hypothesised to have some functional significance [218], though this putative significance has been questioned [e.g. 143]. An analysis of the biological significance of motifs in the *S. cerevisiae* PIN appeared in [355], where it was found that constituents of motifs, particularly highly interconnected motifs, were more likely to be evolutionary conserved. Turanalp and Can [325] investigated recurring functional interaction patterns in *S. cerevisiae* PIN motifs. Motifs are particularly susceptible to noise in interaction data [68], and there is also controversy about judging whether motifs are over-represented [10, 219].

1.3.4 Communities

Considerable recent attention has been given to the modularity of the cell's functional organisation [120, 125, 268]. A *module* is often construed as a group of components that carry out a functional task fairly independently from the rest of the system. It is thought that such modules yield robust and adaptable systems [3]. There is also much suggestive evidence that modules within the cell are themselves the building blocks of a higher level of structural organisation [e.g. 13, 276, 358].

Within the networks literature, a great many algorithms have been proposed that locate dense regions in a network, often called *communities* (see Section 2.1, and reviewed in e.g. Porter et al. [259] and Fortunato [91]). A community is loosely defined as a group of nodes that are more closely linked with themselves than with the rest of the network.

Communities are potentially good candidates for functional modules, and many studies report running one of the myriad algorithms for detecting community structure on PINs [e.g. 44, 50, 75, 184, 191, 199, 210, 253]. If communities are good candidates for biological modules, their functional relevance is potentially twofold: gathering evidence for the modular organisation of the cell, and helping predict the function of proteins about which little is known.

Whether or not communities are good candidates for biological modules is typically assessed through ascertaining whether proteins in communities are in some way functionally homogeneous. This can be done by searching for terms in a structured vocabulary – usually the Gene Ontology (GO, [12]) or Munich Information Centre for Protein Sequences categories (MIPS, [211]) – that are significantly over-represented within communities. Almost all studies assess this over-representation using a hypergeometric test (an exception is [253], which defines a measure of redundancy of annotation). If such terms exist, the identified communities are said to be 'enriched' for biological function. In all the studies of which I am aware, the great majority of communities detected are found to be enriched.

However, as argued in Chapter 3, the literature-standard test of functional homogeneity is not sufficiently strict, as it does accommodate the fact that communities consist of many pairs of interacting proteins, and hence does not ascertain whether communities 'add value' beyond these pairwise relationships (Lewis et al. [181]). In Chapter 3 I undertake a closer investigation of the functional relevance of community structure in PINs.

The relevance of communities for protein function prediction is still an open question: many studies refer to the relevance of community structure in predicting protein function, but very few actually use communities in this way. The most direct way of doing this, annotating uncharacterised proteins with terms that are over-represented in communities, has been shown to be less effective than guilt by association [303] (where a protein is annotated with the terms of its interacting partners, discussed in Section 1.3.1). Our research into the distribution of particular protein functional classes in communities hints at some ways in which community membership could best be used in a protein function prediction algorithm (Section 3.10).

1.4 Protein-protein interactions and evolution

For a review of protein evolution in more general contexts, see Pál et al. [244]; for a review of the evolution of protein-protein interactions see Levy and Pereira-Leal [179]. The majority of this section, with the exception of 1.4.5.1, is very similar to content I wrote for a paper published jointly with Ramazan Saeed and Charlotte Deane [182].

1.4.1 Evolutionary claims in the literature

Some bold claims have been made about the origins of cellular complexity based on observations of biological network structure. For example, on the implications of a 'scale-free' degree distribution in metabolic networks: 'Therefore, the evolutionary selection of a robust and error-tolerant architecture may characterize all cellular networks, for which scale-free topology with a conserved network diameter appears to provide an optimal structural organization.' [148]. In Barabási and Albert [19], the authors state that the success of their proposed preferential-attachment model 'indicates that the development of large networks is governed by robust self-organising phenomena that go beyond the particulars of the individual systems.'

It is, however, not at all clear that such grandiose claims are appropriate. Words such as 'optimality' are connected with the selection of particular attributes by natural selection. Claims for direct selection have to be backed by a demonstration that the attribute in question could not have evolved without direct selection, via genetic drift, mutation and recombination. For example, Lynch [192] argued that biologically realistic null hypotheses need to be considered to make any claims about selected complexity. Indeed, in the case of network structural motifs, whether or not the network growth process would itself lead to observed motif counts needs to be addressed [reviewed in 302]. It is also possible that complex protein-interaction architectures could be a by-product of selection for something else: Fernandez and Lynch [84] argued that this is indeed the case, and that the accumulation of mildly deleterious mutations produces – as a side effect – selection for protein-protein interactions.

1.4.2 Interactions effecting rate of evolution

Many models of evolution assume that sequences will change more slowly if they are under more constraints [87, 241]. Specifically in the context of protein evolution, Zuckerkandl [365] proposed the notion of 'fitness density': the rate of evolution should be inversely proportional to the fraction of amino-acid residues engaged in specific functions. One might therefore expect that (a) proteins with more interactions would evolve more slowly and (b) residues at interaction interfaces should evolve more slowly than other surface residues.

Initial work appeared to confirm (a) [95], however, since then, this claim has been challenged because it is sensitive to biases in interaction data [35] and confounding independent variables (notably expression rate [2]); is not particular to number of interactions but to other network features [116]; and does not stand up to the data [25, 117, 151, 282]. A review of employed methodologies concluded that the number of translation events (which is well indicated by expression level, protein abundance, and codon adaptation index), rather than the number or patterns of protein-protein interactions, was the key determinant of evolutionary rate [74].

Work based on structures of proteins crystallised together has been more conclusive with regard to (b): interface residues are more evolutionarily conserved than other surface residues [46, 328], although this effect is moderate. In an individual interface, some residues, known as *hot spots*, are thought to dominate the interaction binding energy [55, 221], and these are found to be even more conserved than other interface residues [162]. Proteins involved in obligate interactions are more conserved evolutionarily than those involved in transient interactions, which are in turn more evolutionarily conserved than those not known to be involved in any interactions [318].

1.4.3 Co-evolution of interacting proteins

In a recent review of co-evolution Pazos and Valencia [251] stressed the necessity of distinguishing between *co-evolution*, the existence of mutual selective pressure inferred from similarity of evolutionary histories, and *co-adaptation*, the molecular mechanisms that would explain co-evolutionary changes. Evidence of co-adaptation would be needed to infer direct physical interactions. Not all cases of co-evolution will be from co-adaptation, due to confounding factors such as similar expression patterns or common function.

The genomes of different organisms can be compared to give information about likely functional association between proteins. If the same genes tend to occur as neighbours in multiple organisms, then one can infer functional association between them [e.g. 65]: if two proteins cannot perform their cellular function without each other, then when one is lost, there will be no evolutionary advantage to keeping the other, so they will be lost from the genome as a pair. Patterns of presence and absence of genes in different organisms, termed *phylogenetic profiles*, are the simplest clue of protein co-evolution. Similarly, profiles encoding the presence and absence of protein domains can be used to detect functional associations [243]. Additional patterns in phylogenetic profiles, such as anti-correlation [222] and correlations between triplets of proteins [37], can also give information about functional associations.

Interacting proteins are often transcribed as a single unit (operon) in bacteria. In Huynen et al. [141] it is shown that the products of $63 - 75\%$ of co-regulated genes tend to interact physically, which suggests that the regulation of proteins co-evolves with the proteins and their interactions.

It is also possible to compare the *phylogenetic trees* of proteins, i.e. the estimated evolutionary relationships of families of proteins. The motivation for such an approach is that the phylogenetic trees of, for example, ligands and their receptors are more similar than would be expected under standard models for the rate of sequence change, which indicates some degree of co-evolution [250]. This relationship is even more striking when the phylogenetic trees are built from the sequence of the interacting interfaces rather than the whole protein [220]. The same study shows that residues at the interface of obligate complexes tend to evolve slowly, allowing co-evolution of the partner interface, whereas transient interfaces tend to have an increased rate of residue substitution, leaving little evidence of correlated mutations across the interface [220]. It is also possible to concentrate on the domains within proteins, and this can be used to infer which domains are responsible for a given interaction [152]. As these approaches rely on generating reliable phylogenetic trees, they are well placed to take advantage of the growing amount of available sequence information. A generalisation of this approach comes from the acknowledgement that, because proteins can interact with many different partners, considering only pairwise interactions will never give a complete picture of protein co-evolution. Rather, co-evolution depends on all of the different interactions in which a protein may be engaged, so comparing proteins not only pairwise, but against all other proteins can give a better idea of the co-evolution of a given pair [153].

Other generalisations come from investigating the similarities of protein structures, rather than just sequences. Williams and Lovell [348] offered an integrated view of sequence and structural divergence, claiming that both co-evolution following sequence changes and structural accommodation of non-compensated substitutions can be accommodated in the same framework. The majority of the methods discussed above assess similarity based on nucleotide or amino-acid substitutions, but there is some evidence that the role of insertions and deletions of short stretches of nucleotide (indels) is important. Indels are particularly common on the surfaces of proteins [see 28] and are thus suspected to play a large role in 'rewiring' PINs. The importance of indels has been highlighted by a study finding that proteins from families which are thought to possess recent indel mutations tend to score higher on a range of measures designed to assess how central a protein is within the PIN [133].

The circumstances under which one can infer direct physical interactions from the observation of protein co-evolution are not clear [251]. Hakes et al. [118] argued that there is no evidence that co-evolution is due to co-adaptation, arguing for the importance of common evolutionary forces (notably expression levels) as responsible for co-evolution. However, Kann et al. [155] argued that there is some evidence for co-adaptation alongside more general evolutionary forces. A better understanding of the molecular mechanisms underpinning compensatory changes will help distinguish these correlations and perhaps the conditions under which co-adaptation can be inferred. A better understanding of how other protein features can influence functional association will also be vital.

1.4.4 Models for protein-protein interaction network evolution

Many models have been proposed for the growth of protein interaction networks (reviewed in Stumpf et al. [312]). The literature follows a clear pattern: a model for network growth is proposed; it is shown to match some aspect of the data; this is used to make a claim about how cellular complexity arose. This is followed by new studies that point out some lack of fit between a pre-existing model and the data, and a new or modified model is proposed. In this section, I review some of the most influential models, though as discussed in Section 1.4.1, developing a model that fits the data is not the same as discovering how cellular complexity arose. In addition to possibly providing explanations of the ways in which networks evolve, such generative models have a role to play in the generation of ensembles of networks used for assessment of the statistical significance of observed patterns.

The field of modelling network growth was re-vitalised in 1999 with the proposal of the preferential attachment model of Barabási and Albert [19], which picked up on ideas dating back to the 1950s [69, 296]. It was observed that, if new nodes attached themselves to old nodes with a probability proportional to the number of interactions of the old nodes, a power law distribution of node degree would result in the limit as network size goes to infinity. Such distributions were, at the time, being reported in many different types of network [231], including PINs [79]. This would fit the observation that older proteins have more interactions [354]. The problem with the model of growth by linear preferential attachment as a model of PIN growth is that

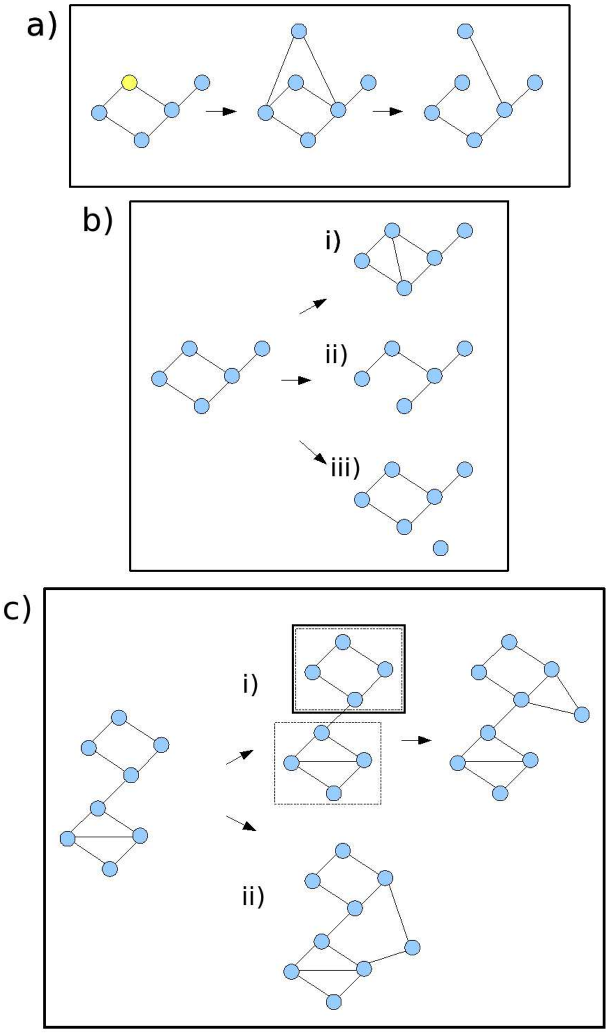


Figure 1.2: **PIN evolution models.** a) *Duplication-Divergence.* A node is chosen to be copied. The new node is given edges to the same set of nodes as the chosen node (duplication). Some fraction of edges are then lost (divergence). b) *Asymmetric gain and loss of interactions.* Three move types are possible. i) Addition of a link. A link is made between one node chosen at random and another chosen proportional to node degree. ii) Removal of link. A protein is chosen uniformly at random, and one of its edges is chosen uniformly at random to be removed. iii) A new node is added with zero edges. The probabilities of these three moves are chosen such that the mean node degree stays the same and the network grows at some empirically inferred rate. c) *Crystal Growth Model.* After an initial seeding phase, either i) Modules are computed, one is chosen, and a new node is added to this chosen module or ii) A new node is put into its own module, and connects to other modules which have few edges (anti-preferential attachment rule).

it does not correspond to any clear biological mechanism, with the possible exception of horizontal gene transfer in prokaryotes.

What are the likely factors underlying PIN evolution? Errors in replication can result in a change in copy number of proteins – from individual genes being duplicated or lost [reviewed in e.g. 363] to the whole genome being duplicated [reviewed in e.g. 157, 289]. 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) [363]. Gene duplication was hypothesised to be disadvantageous in complexes in particular, and evidence for fewer single-gene duplication events in gene families encoding complexes has been found in support of this [248].

Many PIN evolution models have been based on this idea of duplication followed by divergence [144, 331] (see Figure 1.2a). There are many different variants, but all share in common that a node is selected to be copied (*duplication*), some fraction of the nodes edges are replicated in the duplicated node, and some more added (*divergence*). Models that allow for whole genome duplication events have also been proposed [81]. Both preferential-attachment and duplication-divergence models produce the heavy-tailed degree distributions found in PINs, and also both match the data that suggest that proteins of high degree tend to connect to proteins of low degree (node disassortativity) [165]. Duplication-divergence models generate some level of hierarchical modularity (whereby small densely connection groups of proteins, termed modules, are nested into larger modules, which are in turn nested into larger modules, etc), though not as much as suggested in the data [165]. The duplication-divergence model has been argued to not account for hubs that have multiple interaction interfaces [164] and not produce enough triangles in the network unless heritable interaction sites are modelled [105].

Despite the successes of duplication-divergence models in capturing many aspects of the empirical PIN, it has been claimed that gene duplication and divergence may in fact have played only a limited role in the evolution of PINs, as the dynamics of the gain and loss of individual interactions is thought to happen at a much shorter time-scale [25, 336]. Berg et al. [30] proposed a model based on the addition and loss of individual edges. They found that the rate of addition and loss of edges depends asymmetrically on the number of connections of both interacting partners, which is to be expected because when a new link is formed, typically only one node undergoes a mutation with the other remaining unchanged (see Figure 1.2 b). By building a stochastic model based on these observations, they matched the degree distribution and node disassortativity found in data. Beltrao and Serrano [25] investigated factors that determine interaction turnover and show that the less specific an interaction is (judged on the diversity of structures of the interacting partners) the faster the interactions change. They suggest that power-law degree distributions can be explained partially by the cell's need for a diversity of specificity of interaction types (i.e. some proteins are highly non-specific binders and hence have very large numbers of interacting partners).

In a separate critique of the popular models discussed above, Kim and Marcotte [165] investigated the age-dependent evolution of proteins and claimed this cannot be accounted for by duplication-divergence or preferential attachment models. Instead, they propose a *crystal growth model* based on a) interaction probability increasing with availability of unoccupied interaction surface, b) tightly connected groups of proteins developing as the network grows, c) once a protein is committed to such a group, further connections tend to be made with other members of that group (see Figure 1.2c). In this model, proteins are more likely to link to proteins of a similar age, as observed in real PINs. The model uses modules in networks as a key idea in PIN evolution, an idea that is not that well explored elsewhere (though see Li and

Maini [183] for an abstract model of network growth based on modules).

It is not clear how best to assess different models of network evolution and growth, though these have matured somewhat beyond matching degree distributions [165, 215, 266, 350].

1.4.5 Between-species comparisons

Protein-protein interactions can be compared between species in order to investigate how they have evolved and to help understand differences and similarities between species. Insights gleaned can then be used to predict interactions in species for which there exists little data [e.g. 338].

Between-species comparisons are made possible by the identification of the 'same' protein in multiple species: protein sequences are compared, and attempts at identifying orthologs can be made. *Orthologs* are proteins that share a common ancestor protein but were split by a speciation event. The high degree of observed protein sequence homology gives a strong expectation that discoveries about protein function made in one species will provide understanding in another [316]. The extent of homology of protein function is of both practical and theoretical importance, as it underlies the reliance on a few model organisms and provides insight into the maintenance and diversification of protein function through evolution [66, 343].

1.4.5.1 Interologs

To what extent are protein-protein interactions conserved through evolution? A high degree of conservation makes it viable to transfer interactions across species well separated on the tree of life. This is particularly pertinent given the cost of gathering experimental data and the concentration of that data in few species. If, however, there is a low degree of conservation of protein-protein interactions, then – given the very high degree of conservation of protein sequences – this would suggest that interactions can be lost and gained rapidly with little sequence change. This in turn could help explain how small changes in protein sequence, on occasion, bring about large phenotypic changes.

The homology of protein-protein interactions can be investigated by seeking evidence of *interologs*. Interologs are pairs of interacting proteins: A interacting with B in one species and A' interacting with B' in another, where A' is an ortholog of A and B' is an ortholog of B (see Figure 1.3). As illustrated in Figure 1.3, one can seek evidence for interologs by inferring interactions *across* species (when the source and target species are different) or *within* species.

Across-species interologs were first introduced in Walhout et al. [338]. Since then, many studies have used inferred interactions on the basis of homology to make interaction predictions [e.g. 41, 42, 77, 97, 102, 138, 139, 150, 177, 185, 255, 346, 356].

Despite the prevalent use of such inferences, relatively little published work has investigated the reliability of transferring interactions across species. Published success rates for transferring interactions vary from less than 5% [98] to 100% [262], and many values in between have also been reported [169, 204, 216, 267, 359]. As I discuss below, these differences can be explained in part by methodological choices.

Matthews et al [204] used *S. cerevisiae* as a source species for inferring interactions in *Caenorhabditis elegans*. Testing their predictions (and re-checking the *S. cerevisiae* interactions), they found that between 16% and 31% of the inferences were correct. They also found no detectable correlations between the extent of sequence-similarity and the likelihood of an interaction being conserved. Using only one-to-one ortholog matching (i.e. allowing each protein in one species to be judged a homolog of at most one protein in the other) and taking into account errors in their data, a conservation rate of between 34% and 64% was reported between *H. sapiens* and mouse transcription factor-transcription factor interactions [267] (the study did not investigate the extent of sequence homology). A recent study comparing two yeasts, *S. cerevisiae*

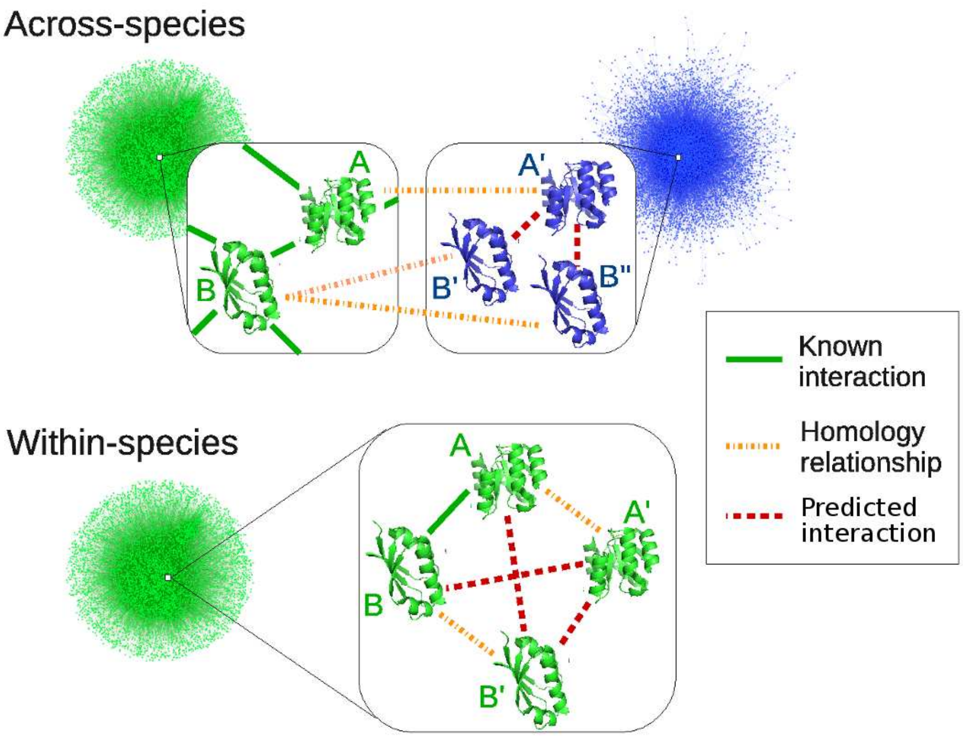


Figure 1.3: **Interactions can be predicted (red dashes) from known interactions (green) and homology relationships (orange dashes).** Interactions can be inferred both *across* species and *within* species. **Across species**: The interaction between proteins A and B in the source species (the interaction network is in green) is used to infer interactions between any homologs of A and any homologs of B in the target species – in this case, interactions between $A' - B'$ and $A' - B''$. These predicted interactions can then be compared to the interactions in the target species (the blue network). The homologous interactions $A - B$ and $A' - B'$ are called *interologs*. **Within species**: The interaction $A - B$ is used to infer interactions between the homologs of A and of B – in this case, $A - B'$, $A' - B$, and $A' - B'$. We call the inferences from $A - B'$ and $A' - B$ 'one-same' inferences and the inference from $A' - B'$ a 'both-different' inference.

and *Kluyveromyces waltii*, excluded duplicated genes, and found that 43 of 43 tested interactions were conserved [262].

These experimental investigations should be compared to larger-scale investigations performed on pre-existing data. Yu et al. [359] performed two relevant experiments. In the first, they transferred binary interactions from *C. elegans* to co-complex data in *S. cerevisiae* and found less than a 10% conservation rate when the joint E-value, J_E, was between 10^{-100} and 10^{-50}, where J_E is the geometric mean of the two E-values, $J_E = \sqrt{(}E_{\text{val}}(A, A')E_{\text{val}}(B, B'))$, and the E-value ($E_{\text{val}}$), or *Expect-value*, gives a measure of how often one would expect to see a query-hit pair by chance. The conservation was higher at more stringent E-values and lower at less stringent values. In the second, they transferred interactions from *C. elegans*, *Drosophila melanogaster*, and *H. Pylori* binary data – along with *S. cerevisiae* co-complex data – to *S. cerevisiae* co-complex data and found that the conservation rate was just over 50% when $J_E < 10^{-70}$. It is unclear how large a contribution was made by using *S. cerevisiae* as both the target and source species. A direct comparison of binary data to binary data was performed in [98], where an overlap of 63 of a possible 1405 interactions (a conservation rate of 4.5%) was found between large-scale *H. sapiens* and *D. melanogaster* data sets. A study of the binary data available in 2006 by Mika and Rost [216] found a low level of conservation of interactions across species: accuracy never exceeded 20%, even for the most sequence-similar homologs. This study also reported that within-species interactions were more conserved than across-species interactions. Similar results were also reported in [169]. These results were surprising in light of the long-standing belief that proteins arising from gene-duplication events must diverge in function in order to be conserved, whereas proteins that arise from a speciation event have evolutionary pressure to maintain the function of the ancestral protein [193].

Errors in the interaction data can have a substantial impact on results. Most obviously, false negatives in the target species' interaction data set will cause some transferred interactions to be judged as non-conserved when the data in the target species is simply missing. However, except for Ref. [267], which examines one type of protein (transcription factors) in one pair of species (mouse and human), none of these studies investigated the role of errors in the data when assessing conservation. In Chapter 4, we consider this problem, and we thereby arrive at more reliable estimates of protein-protein interaction conservation rates [180].

1.4.5.2 PIN alignment

In addition to comparing pairs of interacting proteins, there is a large body of literature that attempts to align PINs – i.e. to find sets of interactions that are conserved in one or more species. This can be done in either a local fashion, in which small groups of proteins in one species are aligned with small groups in another one or more species and a given protein can appear in more than one locally aligned region [76, 88, 186, 205, 277], or globally, in which each protein in one species is matched to some number (zero, one, or more) of proteins in another species [88, 187, 297, 362]. The idea behind local network alignment is that those sets of interactions that are co-conserved likely have some biological significance. Indeed, the methods tend to be tested as to how well locally aligned sets of proteins correspond to known protein complexes [e.g. 76]. A review is found in Sharan and Ideker [292].

Some of the proposed algorithms for alignment model the evolution of interactions directly by using a model for PIN growth and/or a phylogenetic reconstruction of the protein family relationships to infer the PIN of the ancestral species [76, 130, 205].

1.5 Protein-protein interactions, function, and evolution

It is hoped that the study of protein-protein interactions can give insight both into how proteins co-ordinate to bring about biological function and into how and why these patterns of co-ordination evolved (though, as discussed in Section 1.4.1, great care must be taken in making claims concerning the direct selection of particular traits). As an off-shoot of this investigation, methods that predict protein function and protein-protein interactions have been developed. In the above two sections, I have tried to separate these functional and evolutionary aspects, but they are of course intertwined. Here I discuss a few cross-over points.

Much of the evolutionary study of protein-protein interactions relies on first identifying evolutionary relationships between proteins. However, there is no simple way to establish which proteins are indeed orthologs. Indeed, one could say that it is important to incorporate functional information into models of orthology. For example, Bandyopadhyay et al. [18] used protein-protein interactions to help identify orthologs.

Modules are expected to be evolutionarily cohesive (where that term has no precise meaning in the literature), as the functional task associated with a given module presumably needs many of the same proteins in different organisms. However, this has been questioned by Snel and Huynen [299], who found that there is not that much evidence for the evolutionary cohesiveness of some candidates for modules: complexes, metabolic pathways, and known operons in *E. Coli.* Data incompleteness and noise is one possible explanation for this. Another is that the observed evolutionary flexibility actually reflects the functional flexibility (where, for example, there are many shared components between modules [see 100]). Whether evolution tinkers with the interactions between or within modules is also up for discussion: one recent study that runs models of network evolution backwards found that complexes have been significantly rewired over time and that new edges tend to form within existing complexes [228]; another recent study found that interactions within modules are more conserved than interactions between modules [364].

In Chapter 3 of this thesis, I explore the function of communities in PINs, motivated by the idea that communities are potentially good candidates for biological modules. Evolutionary considerations, which are tied to the idea of a module, lurk in the background. In Chapter 4, I explore the evolutionary conservation of protein-protein interactions without incorporating network structure. Further directions for this research (discussed in Chapter 5) include the evolutionary conservation of communities at the levels of proteins, protein-protein interactions, and protein function.

This thesis also contains four appendices: Appendix A contains a summary of methods proposed in the literature to predict protein-protein interactions computationally; Appendices B and C contain examples of communities discussed in Chapter 3; Appendix D contains figures supplemental to those provided in Chapter 4.

Chapter 2

Tools and techniques

In this chapter, I discuss the choice of tools and techniques employed in later chapters and explain the ideas on which they are based. This necessitates a brief review of various techniques, discussion of known problems and limitations, and an assessment of when particular choices are appropriate.

2.1 Community detection

2.1.1 Introduction

One often refers to groups such as families, villages, cities, nations, cultural groups, friendship groups, and business organisations as social communities. This intuitive notion of community is of a group of individuals with strong connections to other members of the group and sparser connections to the rest of the social network. Within the social sciences, there has been a long tradition of the mathematical study of *community structure*, arguably dating from the 1920s [274, 345].

The field of community detection was invigorated in 2002 by Girvan and Newman [108], a paper that bought the community detection problem to the attention of the statistical physics community. Since then, several thousand papers have been published on the topic, and it is now one of the most active sub-fields in network science [259]. The notion of a community is not precisely defined, and different approaches thus rely on some arbitrary or common-sense notion of a community. Reviews of the field appear in e.g. Porter et al. [259] and Fortunato [91].

Biological systems possess structure at many different scales. In investigating community structure of PINs it is hence natural to seek out a method that allows communities to be identified at multiple scales or resolutions. The most popular method for detecting communities, modularity maximisation [91], was found to impose a limit on the minimum size of communities that could be found [92]. A generalisation of the method allows communities to be detected at multiple resolutions [270]. In the following sections I motivate our choice of this method, introduce some of the mathematical details, and briefly mention some alternative approaches. As our focus in Chapter 3 is with assessing the functional significance of detected communities, I end this review by discussing the work that has been done in assessing the outputs of community detection algorithms.

2.1.2 Choosing a method

Hundreds of different community detection algorithms have been proposed [91]. How does one make a choice between them?

There are three main features that make an algorithm attractive for a particular use. The first is that it is capable of dealing with networks of the size in hand in a reasonable amount of time. The second is that it performs well in practical contexts. The third is that the theoretical underpinnings of the method have been studied sufficiently that it is well-behaved and its properties are understood.

The performance of community detection algorithms is typically assessed against synthetic benchmark networks that are designed with community structure ‘planted’ within them [91] as well as using real-world networks with ‘known’ community structure. The three elements of such a test are the output of an algorithm, a set of benchmark networks with known community structure, and a comparison measure to assess the similarity of the two. (I discuss comparison measures in Section 2.2.)

The most comprehensive benchmarking of algorithms appears in Lancichinetti and Fortunato [173], though this is for graphs without hierarchical community structure. Three algorithms were found to have comparably high performance. One of these is a method for optimising modularity referred to as the Louvain method, proposed by Blondel et al. [34] and outlined in Section 2.1.3.3. This algorithm is a locally greedy algorithm – it trades off accuracy for speed, so it is surprising that it outperforms the much slower and supposedly more accurate methods for optimising modularity such as annealing.

Because this method also has a tunable resolution parameter, and is very fast (running in a second or less on the networks considered here), and through being a member of the family of algorithms that maximise modularity has theoretical properties that are at least partially understood, it is our choice of method.

It is of interest to note that the review of Fortunato [91] concludes that most modern algorithms yield similar results in practical applications (though others contest this [111]), and stresses that the real challenge is to interpret the communities once they are detected.

2.1.3 Modularity maximisation

The quantity known as ‘modularity’ was proposed in Newman and Girvan [236] to measure the quality of a partition of the nodes of a network into communities. It is based on the idea that a random network is not expected to have any community structure, so the possible existence of communities is interrogated via a comparison of the network of interest (represented by adjacency matrix $\mathbf{A}$), to a suitably chosen *null model*, $\mathbf{P}$, where the null model is a network that shares some of the same properties as the original network. The modularity Q is then given by

$$Q = \frac{1}{2m}\sum_{ij}(A_{ij} - P_{ij})\delta(c_i, c_j), \tag{2.1}$$

where the sum runs over every pair of nodes, m is the total number of edges of $\mathbf{A}$ and c_i is the community assignment of node i. The δ-function yields 1 if nodes i and j are in the same community and 0 otherwise, so partitions that have a large modularity will have many edges in which $A_{ij} > P_{ij}$ inside communities and few edges in which $A_{ij} < P_{ij}$ inside communities.

The conventional choice of null model, known as the *Newman-Girvan* null model [236], ensures that the expected degree sequence of $\mathbf{P}$ matches that of $\mathbf{A}$, in recognition of the importance of heavy-tailed degree distributions in many real-world networks:

$$P_{ij} = \frac{k_i k_j}{2m}, \tag{2.2}$$

where $k_i = \sum_j A_{ij}$ is the degree of node i. Numerous modifications and generalisations of modularity have been proposed [91].

Although originally introduced to judge the quality of a partition returned by an independent algorithm, directly maximising modularity is now by far the most widely used class of methods to detect communities in graphs [91]. It has been proved that maximising modularity is an NP-hard problem [39]. Algorithms must hence employ some computational heuristic to find a 'good' enough solution (i.e. a large local maximum of Q) in a reasonable amount of time.

An issue with techniques based on modularity maximisation is the possibility of an exponentially large number of distinct partitions, all of which have values close to the maximum value of modularity, but which can be structurally quite different from each other [111]. This is sometimes a serious issue with real-world applications.

2.1.3.1 Resolution limit

Fortunato and Barthelemy [92] showed that community detection algorithms based on modularity maximisation contain a *resolution limit*. Modularity optimisation often does not detect communities that are smaller than some characteristic size that depends on the number of nodes in the network. Roughly speaking, communities with a total number of interactions $\sqrt{m}$ or smaller will tend to be merged with other communities (recall that m is the total number of edges in the network).

2.1.3.2 A multi-resolution generalisation

Modularity maximisation can be adapted to incorporate a *resolution parameter*, allowing one to explore the network at different resolutions in order to find communities of different sizes.

Reichardt and Bornholdt [270] showed that optimising modularity is equivalent to minimising the energy of an infinite-range q-state Potts model. In statistical mechanics, the Potts model is a model of interacting 'spin states', where q is the number of spin-states (not all of which need be occupied). If the Potts spin variables are c_i, the energy H of the system is given by

$$H = -\frac{1}{2m} \sum_{ij} J_{ij} \delta(c_i, c_j), \tag{2.3}$$

where $\mathbf{J}$ is the matrix of couplings between the spins. The components of $\mathbf{J}$ are given by

$$J_{ij} = A_{ij} - \lambda P_{ij}, \tag{2.4}$$

where λ is the resolution parameter and A_{ij} and P_{ij} are, as above, the adjacency matrix of the network and a null model, respectively. Other choices of $\mathbf{J}$ are possible; see e.g. Traag and Bruggeman [320] and Ronhovde and Nussinov [278]. Minimising the energy is equivalent to maximising the modularity if P_{ij} is chosen to be the standard Newman-Girvan null model and λ is set to be 1. By tuning the resolution parameter λ, the typical size of communities changes: at $\lambda = 0$, all nodes are placed in the same community; at $\lambda = \infty$, all of the nodes are placed in different communities; at intermediate values one finds communities at different scales or resolutions.

2.1.3.3 Algorithms for maximising modularity

A large variety of algorithms have been applied to find good approximations of the modularity maximum. Many of these techniques are reviewed in Fortunato [91], and they include greedy algorithms, simulated annealing, extremal optimisation, spectral optimisation and more.

Here I outline the 'Louvain' method of Blondel et al. [34], because this is the method we employ in Chapter 3 (see Section 2.1.2 for why we chose this method). The method is based on the iteration of two phases. Initially all nodes are placed in their own community. The first phase considers for each node i whether adding it to the community of any of its neighbours increases the modularity and makes the move with the largest gain in modularity if this is so. This process is applied repeatedly and sequentially until no more moves result in increased modularity (the order of the nodes can therefore make a difference). In the second phase, a new network is built whose nodes are the communities found in the first phase, with edge weights between them equalling the summed weight of edges between the nodes in the two communities considered. These two phases are then repeated until no moves lead to an increase in modularity, at which point the algorithm terminates. It is freely available at `www.lambiotte.be/codes.html`. The algorithm is exceptionally fast, running in a second or less on networks of the size considered in this thesis (i.e. several thousand nodes).

2.1.4 Other multi-resolution approaches

2.1.4.1 Traditional clustering techniques

One clan of traditional hierarchical clustering algorithm starts with a definition of a similarity measure definable between nodes and clusters [126]. In *agglomerative clustering*, each node is initially in its own cluster, and the two nodes with the highest similarity are merged. The similarities between clusters are computed again, and the process is repeated. *Divisive algorithms* are also possible, where the nodes are all initially placed in the same cluster, and clusters are split by removing edges or other structures with low similarity. These algorithms output a tree like structure known as a *dendrogram*. They were argued in Newman [232] not to perform well as methods for community detection on many large real-world networks, based on an analysis of the performance of one such method on a common bench mark data set.

Partitional clustering techniques start with a pre-assigned number of clusters. Each node is embedded in a metric space, such that distance between nodes is a measure of their dissimilarity. The nodes are then assigned to clusters to optimise a given cost function. The most popular technique in this family is *k-means clustering* [197].

2.1.4.2 k-clique percolation

Palla et al. [245] proposed a method based on the notion of a k-clique. A k-clique is a complete sub-graph of k nodes that are fully connected to each other. Two k-cliques are called ‘adjacent’ if they share all but one node. The algorithm finds *k-clique communities*, where such a community is a k-clique and all of the k-cliques which are connected to it through being adjacent to each other. By varying k, one can find nested communities. k-cliques with values of k between three and six are usually used [91]. Despite the popularity of this method, it has numerous problems associated with it, summarised in Fortunato [91]: it has an overly stringent notion of a community, which overlooks other dense regions that aren't quite as well connected as cliques; in networks with few cliques, few communities will be returned, and in networks with many cliques, trivial structure will be returned; it has a fundamental issue that, rather than looking for dense sub-graphs, the method looks for sub-graphs that contain many cliques, which is often a different type of structure to community structure.

2.1.4.3 MCL

Introduced in van Dongen [330], the Markov Clustering (MCL) algorithm is based on a type of flow diffusion on the network. The *transfer matrix*, $\mathbf{T}$, whose entries T_{ij} give the probability a random walker starting at node j moves to node i, is taken as the starting point. In an 'expansion step', this matrix is raised to some power (normally two). In an 'inflation step', each entry of the resulting matrix is raised to some power α, and the resulting matrix is normalised such that the elements of each column correspond to probability values. These two steps are then repeated until the algorithm produces a separation of the network into disjoint segments. These segments are interpreted as communities.

2.1.4.4 Hierarchical random graphs

The hierarchical structure of networks is explored in Clauset et al. [56] via the introduction of a class of *hierarchical random graphs*. These are defined by dendrograms that have a probability p attached to each node in the dendrogram. Nodes i and j have a probability of being linked equal to the probability p associated to their lowest common split in the dendrogram. The likelihood for a given model (with a certain dendrogram structure and set of probabilities p) is sampled using a Markov chain Monte Carlo method, such that an ensemble of model configurations is returned.

A class of possible organisations of the network, with well defined-probabilities, are hence returned. It is not clear what information one can extract from averaging over the ensemble of hierarchical random graphs or whether this approach suggests that we should alter the ways we think about relevant graph partitions [91].

2.1.4.5 Dynamical communities

'*Stability*' emerges as a natural quality function if one takes a dynamical view of community detection: a random walk on a network is expected to be trapped for long periods of time in good communities before being able to escape [171, 172]. The stability of a partition depends on the length of the random walk, such that the characteristic size of the communities grows with time.

It has been shown that the Potts model formalism with the Newman-Girvan null model is a linear approximation of the stability function, with the time parameter inversely proportional to the resolution parameter λ [171, 172]. This dynamical approach to modularity thus helps motivate the choice of the Newman-Girvan null model, as it is shown to arise naturally within this framework.

2.1.5 What next after detecting communities?

In contrast to the vast amount of effort put into the development of new community detection methods, there has been comparatively little work on assessing the significance of the outputs of community detection algorithms. Mark Newman, whose co-authored paper in 2002 sparked recent interest in the field [108], has written that, 'The development of methods for finding communities within networks is a thriving sub-area of the field (of network science), with an enormous number of different techniques under development. Methods for understanding what the communities mean after you find them are, by contrast, still quite primitive, and much needs to be done if we are to gain real knowledge from the output of our computer programs' [234].

There are two broad approaches one might take to investigate the significance of community structure. The first is an internal measure, where one uses the network structure itself, perhaps in combination with randomised versions of this structure, to investigate the extent to which a network does in fact have community structure. The second is to enquire whether the detected community structure has any relevance, which is done by investigating its relationship with external properties.

There are very few theoretical results concerning when a community or a network partition is to count as significant. Borrowing results from spin glass theory, Reichardt and Bornholdt [271, 272] have calculated the expectation of the modularity for an Erdös-Rényi random network. Initial attempts at investigating the probability distribution of maximum modularity for the Newman-Girvan null model have also been made [264].

In the absence of many theoretical results, one possible approach is to assess the robustness of a given partition to the addition of random error or noise [reviewed in 91]. Various measures have been proposed that assess precisely this [see e.g. 103, 156]. These measures have no absolute significance, so they need to be compared to the values of these measures achieved with suitably randomised versions of the network. Rather than introducing noise, one can also introduce quality functions and compare directly to suitably randomised graphs [see e.g. 33]. Because many of the algorithms are non-deterministic, another approach is to assess how sensitive an algorithm is to initial conditions [see e.g. 171, 203]. One question of considerable importance is whether the output of the algorithm as a whole is assessed or whether each community is assessed one at a time; see e.g. Lancichinetti et al. [175] for an approach that incorporates this second option. Such community-specific approaches are attractive, as it may not be the case that the whole network is modular.

Multi-resolution methods can present an additional set of challenges for assessing the significance of partitions. This is particularly the case if there are no clearly ‘correct’ partitions, such as one would expect from a network designed to have clear hierarchical structure. In general, a measure of the partition is proposed (for example, the number of communities) and then plotted against resolution parameter, such that plateaus in the measure can be used to identify ‘stable’ partitions [8, 278]. It is also possible to define a measure for particular communities; for example, Pons and Latapy [258] propose to investigate the range of values of the resolution parameter over which a community ‘lives’. Because the outputs of community detection methods can be noisy and can depend on initial conditions, investigating correlations between partitions rather than relying on measures of the partitions themselves can be helpful; see Ronhovde and Nussinov [278]. This is particularly the case for large graphs, where minimal shifts of nodes between communities can introduce a large amount of noise.

What of external measures of the significance of communities? As mentioned above, the focus of the community detection literature has not been the assessment of the functional relevance of communities. The most frequent analyses compare returned partitions with some properties (labels) of the nodes – for example, which research division an individual is a member of [108], the college year group to which they belong [323], or what functional class is annotated to a protein [e.g. 44, 75, 184, 191, 210, 253]. Such a comparison can be done by comparing the partition of nodes by a community detection algorithm to the partition of the nodes by property (e.g. research division), for which one needs a way to compare partitions (see Section 2.2). Alternatively, one can assess each value of the property (e.g. each functional type of a protein), and assess whether that particular node type is over-represented in any communities. The latter approach is the norm for assessing communities in protein-protein interaction networks, and this will be discussed in Chapter 3.

There are very few studies that also study the functional significance of detected communities. Bassett et al. [21] demonstrated dynamically changing modular structure associated with a learning task in networks derived from functional MRI data.

Guimera and Amaral [114] detected communities in metabolic networks and then defined particular ‘node roles’. A node is characterised as having a *participation coefficient* and *within-community degree*, this two-dimensional parameter space is carved up into various roles, and the differing behaviour of these roles with different functional characteristics is assessed. Additional quantities that capture a node’s relationship to partitions can be defined. In the context of protein-protein interaction networks, Agarwal et al. [1] suggested focusing on interaction roles rather than node roles.

2.2 Comparing partitions

One might need to compare partitions of a set of objects into groups to compare:

1. the output of a community detection algorithm to a ‘ground truth’ of community membership;

2. the output of a community detection algorithm to some independent grouping of the nodes to ascertain whether or not the groupings are statistically related;

3. multiple runs of the same algorithm to discern the robustness of community structure to changes in the initial conditions.

There are a suite of measures used to compare partitions, see Ref. [207]. I closely follow this reference in this brief summary.

Suppose that one has two partitions of n objects, $\mathcal{X} = X_1, X_2, ...X_{n_X}$ and $\mathcal{Y} = Y_1, Y_2, ...Y_{n_Y}$, where n_X and n_Y are the number of clusters in each partition. The number of objects in clusters X_i and Y_j are n_i^X and n_j^Y, respectively.

There are three main types of comparison measures: those based on counting pairs, those based on matching clusters, and those based on information theory.

Measures based on counting pairs are based on four variables: the number of pairs that are in the same/different clusters in both $\mathcal{X}$ and $\mathcal{Y}$, N_{11}/N_{00}; the number of pairs of objects that are in the same cluster in $\mathcal{X}/\mathcal{Y}$ but different clusters in $\mathcal{Y}/\mathcal{X}$, N_{01}/N_{10}. The sum of these variables is always equal to $n(n-1)/2$. There have been various measures proposed that are different functions of these variables, such as the Rand measure [265], Wallace's coefficient [339], the Jaccard measure [27], and many more. These measures do not take into account that the two partitions could be generated by chance alone, and various measures have been proposed that attempt to account for this – for example the Fowlkes-Mallows index [93] and the Adjusted Rand coefficient [140]. These measures have 'baselines' which are the expected values of the measure under a suitable null hypothesis. The two main issues with such measures are that a) the appropriate choice of null model is unclear (for example, whether this should depend on how the data are generated), and b) the baselines can vary sharply, and it is unclear whether any linearity in the measures can be assumed above the baseline [207].

Set-matching measures find the clusters that best match each other in each partition and add up the contributions of matches found [e.g. 206]. The generic problem with such measures is that they are not particularly discriminative, as they ignore what happens to the unmatched part of each cluster [207].

The third set of measures is based on the idea that if two partitions are similar, then one needs little information to infer one given the other. To define these measures, one needs the concepts of entropy and conditional entropy [195]. The uncertainty about which cluster a randomly chosen object is in is given by the entropy:

$$H(\mathcal{X}) = -\sum_{i=1}^{n_X} P(i) \log P(i), \tag{2.5}$$

where $P(i)$ is the probability that a randomly chosen object is in cluster X_i (it is simply $P(i) = n_i^x/n$). If $P(i) = 0$, the value of the corresponding element of the sum is taken to be zero. The conditional entropy of $\mathcal{X}$ given $\mathcal{Y}$ is

$$H(\mathcal{X}|\mathcal{Y}) = -\sum_{i=1}^{n_X}\sum_{j=1}^{n_Y} P(i,j)\log(P(i|j)), \tag{2.6}$$

where $P(i,j)$ is the probability that an object belongs to clusters X_i and Y_j, and $P(i|j)$ is the conditional probability that an object is in X_i given that we know it is in Y_j.

The *mutual information* of two partitions $\mathcal{X}$ and $\mathcal{Y}$ is defined as

$$MI(\mathcal{X},\mathcal{Y}) = H(\mathcal{X}) - H(\mathcal{X}|\mathcal{Y}). \tag{2.7}$$

It is a measure of the mutual dependence of two random variables [195]. Dividing by the mean of $H(\mathcal{X})$ and $H(\mathcal{Y})$ has been proposed as a way of normalising the mutual information [96], and the resulting measure has been used in the community detection literature [64].

In the context of comparing partitions, Meilă [207] presented the measure *variation of information* VI. It is defined as

$$VI(\mathcal{X},\mathcal{Y}) = H(\mathcal{Y}|\mathcal{X}) + H(\mathcal{X}|\mathcal{Y}). \tag{2.8}$$

The main advantage that the variation of information has over the mutual information is that it is a mathematical metric, meaning that it can be thought of as a distance between elements: it is always non-negative, it is symmetric, it is zero if and only if the partitions being compared are identical, and it obeys the triangle inequality. This enables treatments past pairwise comparisons between clusterings into the aggregate set of relationships between clusterings, as its values can be manipulated via addition, multiplication, averaging, etc, which is necessary for the third application enumerated above. VI grows with the maximum number of clusters in either partition, which reflects the idea that partitions can get more diverse as the number of clusters increases. VI can be normalised, $nVI = VI/\log(n)$, to produce a distance that varies between zero and one. The VI is very useful as a comparative measure – e.g. comparing the similarity of $\mathcal{X}$ and $\mathcal{Y}$ to the similarity of $\mathcal{X}$ and $\mathcal{Z}$ – particularly given that all those measures based on pair counting (such as the Rand index) must be suitable corrected for chance in making such comparisons, and there are problems associated with such corrections.

How should one interpret the absolute values of any of these measures? For example, is a value of 0.3 large? For the second application enumerated above (where detected communities are compared to some independent grouping of the nodes to ascertain whether or not the groupings are independent), making suitable comparisons to randomised distributions may be appropriate, à la the tests performed in Traud et al. [323].

A related problem is how to assess the similarity of a particular cluster in $\mathcal{X}$ to a particular cluster in $\mathcal{Y}$, which is necessary if one wants to 'track' a community through many different partitions – for example with changing resolution parameter. We need to do exactly this in Chapter 3 for visualisation purposes. Various options are possible, and the literature on set matching measures for the aggregate similarity of partitions is relevant. We use a method based on the overlap of shared nodes [246]. (A convention based on edges rather than nodes gives nearly identical results.) For each pair of communities $\{X_i, Y_j\}$, define

$$W_{ij} = \frac{|X_i \cap Y_j|}{|X_i \cup Y_j|}, \tag{2.9}$$

where $|B|$ denotes the cardinality (number of elements) of the set B. Begin with $\mathcal{X}$ as the partition at the highest resolution and $\mathcal{Y}$ at the next highest resolution. Starting with the largest value of W_{ij}, we relabel community i as community j. Relabelling proceeds with the next largest W_{ij}, as long as community i is not yet relabelled, until all communities have been relabelled. If $|n_Y| > |n_X|$ (which will be unlikely in our case, as $\mathcal{Y}$ is at a lower resolution), we introduce a new label. The old $\mathcal{Y}$ becomes the new $\mathcal{X}$, and the new $\mathcal{Y}$ is the partition at the next highest resolution.

2.3 Assessing predictive ability

In Chapter 3, we would like to assess whether any topological properties of a community are predictive of that community being functionally homogeneous. In Chapter 4, we would like to see whether some properties of proteins are predictive of whether those proteins will interact. If we treat the general case of a binary classification task, one has the (binary) output of a diagnostic test or model and would like to compare it to the (binary) true outcome of known cases. The possible outcomes of a binary classification task can be represented in a contingency table (see Table 2.1). Various summary statistics can be defined on the basis of the values in this table:

- The fraction of all of the actual positives that are predicted as positive, the *true positive rate*, TPR (also known as the *sensitivity* or *recall*): $\frac{TP}{TP+FN}$.
- The fraction of all of the actual negatives that are predicted as positive, the *false positive rate*, FPR: $\frac{FP}{FP+TN}$
- The fraction of all of the actual negatives that are predicted as negative, the *true negative rate*, TNR (also known as the *specificity*): $\frac{TN}{FP+TN}$. Note that $FPR = 1 - TNR$.
- The fraction of all of the predicted positives that are actually positive, the

positive predictive value, PPV, also known as the *precision*: $\frac{TP}{TP+FP}$.

- The fraction of all of the predicted negatives that are actually negative, the *negative predictive value*, NPV: $\frac{TN}{TN+FN}$.

Table 2.1: Possible outcomes of a binary classification task

		actual value	
		1	0
prediction outcome	1	TP, ‘hit’	FP, ‘false alarm’
	0	FN, ‘miss’	TN, ‘correct rejection’

In many applications, including those considered in this thesis, the output of a predictive model will be a continuous variable. A threshold τ is applied to this output such that a positive is predicted if the output value is above the threshold and a negative is predicted otherwise. All of the quantities defined by the contingency table and the summary statistics based on these can hence take on different values for different values of τ. In order to overcome subjectivity in choice of τ, various ways of assessing the performance of a classifier with changing τ have been proposed.

A Receiver Operating Characteristic (ROC) curve shows all of the possible pairs of FPR (one minus specificity) and TPR (sensitivity) as τ varies [see e.g. 83]. The area under the ROC curve, called the Area Under Curve (AUC), is a measure of the quality of a classifier. A perfect classifier would achieve an AUC of 1, and a random classifier would be expected to achieve 0.5. Values of the AUC below 0.5 indicate that one could use the classifier in the opposite way to the one tested to achieve a performance better than random. The AUC is very popular because it is an objective, non-parametric and easy-to-calculate measure that enables straightforward comparisons of classifiers. Several caveats are, however, necessary. The measure assumes that researchers are interested in all conditions in which a model could operate: from very high false positive rates to very high false negative rates. Partial ROC curves have been proposed to focus attention on, for example, the regime of low false positive rate [e.g. 319]. A recent criticism in Hand [123] is that the measure does not allow the user to specify the relative costs of false positives and false negatives, but rather selects this relative cost to be a function of the output of the model. The measure can be quite noisy when applied to small data sets [122]. Caution must be employed in using ROC curves and the AUC when the size of the positive true class is much smaller than that of the negative true class (as is often the case). This is because large changes in the number of FPs can lead to small changes in FPR when the number of TNs is very large. If this is the case, then one can either sample sets of negative instances of the same size as that of positive instances set and plot ROC curves for many such sampled sets, or adopt an alternative measure, such as the Precision-Recall (PR) curve.

The PR curve plots precision (PPV) against recall (TPR) [45]. The precision must be compared with the precision that one would expect of a random classifier, which is simply the fraction of the whole set that is actually positive. A disadvantage of PR curves is that they cannot be summarised into a single number in a straightforward manner. Various relationships between the ROC curve and the PR curve can be demonstrated; see Davis and Goadrich [67], where it is shown that if a classifier produces a ROC curve that is 'above' that of another classifier, then it will also produce a PR curve that is 'above' the other.

2.4 Similarity measures based on structured ontologies

It is impossible to uniquely define – let alone uniquely quantify – similarity in biological function, but for a variety of applications it is useful to have some proxy for this. In order to systematise knowledge about gene products, structured vocabularies have become prominent. The most comprehensive is the Gene Ontology (GO) [12].

GO terms are related to each other through a directed acyclic graph (DAG), as illustrated in Figure 2.1. There are various relationships defined in GO, by far the most prominent of which is the ‘is a’ relationship. Proteins are annotated with the most specific terms that are known about them. It is then possible to add to this set their parent terms by following the structure of the DAG up to the root node. The GO has three separate sub-ontologies:

- The *Cellular Component* sub-ontology describes locations at the levels of sub-cellular structures and macromolecular complexes.
- The *Molecular Function* sub-ontology describes the jobs a gene product does or the ‘abilities’ that it has.
- The *Biological Process* sub-ontology describes ‘processes’, by which is meant a recognized series of events or molecular functions.

Anyone can add to the set of maintained annotations of the terms in these sub-ontologies to gene products. Annotations must include a reference to the source of the evidence for that annotation as well as an indication of what type of evidence it is. There are various evidence codes, such as ‘inferred from experiment’, ‘inferred from sequence orthology’, and ‘inferred from electronic annotation’.

The Munich Information Center for Protein Sequences (MIPS) [212] provides another set of functional annotations. There are twenty eight primary functional categories, and each of these has further sub-divisions, such that the resultant structure of the aggregate scheme is a tree.

A starting point for defining a semantic similarity measure between two sets of ontology terms is to think about the semantic similarity of two terms. Measures based on the distance of the two terms within the ontology suffer from the fact that it is by no means clear that it is appropriate to consider each edge as a uniform distance [273]. To get around this, Resnik [273] proposed an information content measure, based on

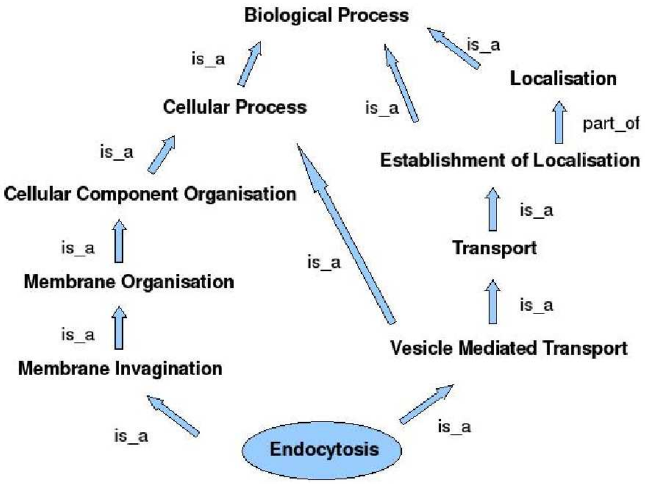


Figure 2.1: **Structure of the Gene Ontology.** Terms are related to each other through a directed acyclic graph.

the idea that the extent to which terms share information in common captures an intuitive idea of similarity. In an 'is a' ontology, high shared information in common is indicated by a highly specific common ancestor concept. The information content IC of two terms t_i and t_j with common ancestral terms $A(c_i, c_j)$ is

$$IC(t_i, t_j) = \max_{c \in A(c_i, c_j)}(-\log(p(c)), \tag{2.10}$$

where $p(c)$ is the probability that term c is annotated to any object in the ontology. Note that the self-similarity of a concept does not have to be 1. One feature of this measure is that it disregards how far away the two terms are from their most specific common ancestor. Thus, for example, if chocolate cake is a cake is a snack, and a digestive is a biscuit is a snack, then $IC(\text{chocolate cake}, \text{digestive}) = IC(\text{cake}, \text{biscuit})$. However, one might consider that chocolate cake and digestives are more dissimilar than cakes and biscuits. Lin [188] proposed a way of taking this into account by normalising by the self-similarities of the two terms:

$$IC_N(t_i, t_j) = \frac{2IC(t_i, t_j)}{IC(t_i, t_i) + IC(t_j, t_j)}. \tag{2.11}$$

Note that under IC_N the self-similarity of two terms is $IC_N(t_i, t_i) = 1$.

To compare the similarity of proteins, one needs to generalise to measures that compare two *sets* of terms, $\{t_i\}$ and $\{t_j\}$. Most proposed measures consider all pairwise similarity measures between the two sets of terms (usually Resnik's IC), and define some function of them, for example the mean [190], the maximum [291], or the mean of maxima [290]. Pandey et al. [247] took an axiomatic approach to the problem and demonstrated that these measures all fail to satisfy some properties that one would desire of a similarity measure. These desirable properties of the similarity of two sets of terms are (a) that the relationship is symmetrical, (b) that adding a common annotation should not decrease the similarity between the proteins, (c) if new annotations are added to a protein, the similarity of this protein to any other should not decrease (i.e. the similarity measure should rely on positive evidence only), (d) a set of annotations should be at least as similar to itself as to any other set.

Pandey et al. [247] suggested using a straightforward generalisation of Resnik's information content measure defined on whole sets rather than on a composite of pairwise similarity measures – they show this measure obeys the properties listed above. One considers the probability that an object is annotated with the intersect of the two sets of terms, $p(\{t_i\} \cap \{t_j\})$, to define the information content between these sets as

$$IC^s(\{t_i\}, \{t_j\}) = -\log(p(\{t_i\} \cap \{t_j\})). \tag{2.12}$$

One can normalise this measure in exactly the same way as above:

$$IC^s_N(\{t_i\}, \{t_j\}) = \frac{2IC^s(\{t_i\}, \{t_j\})}{IC^s(\{t_i\}, \{t_i\}) + IC^s(\{t_j\}, \{t_j\})}. \tag{2.13}$$

The two measures IC^s and IC^s_N have contrasting strengths and weaknesses. When considering functional annotations to proteins, one must take into account that there are a vast number of missing annotations – both in terms of proteins about which nothing is known and for proteins whose functions are only partially known (which one might suspect to be the overwhelming majority of all proteins). In the cake and biscuit example above, the normalised measure judges 'cake' and 'biscuit' as more similar than 'chocolate cake' and 'digestive', even though it might well be the case that one simply doesn't know what kind of cake and what kind of biscuit is involved in the first comparison because of lack of annotation. The normalised version thus tends to judge things with non-specific annotations as similar, despite the fact that they may simply be lacking annotations. This also goes for the self-similarity of proteins. With the normalised measure, every protein has a self-similarity of 1 even if it has only a very generic term annotated to it such as 'cellular process'. One could consider this a strength of the normalised measure, as with the un-normalised measure, two proteins that are both members of a common functional type will be judged less similar than two proteins that share more specialised functions. When using the un-normalised measure, large sets of proteins that all have nearly identical annotations hence have a low similarity score to each other, which is likely to be the case for proteins in large complexes where in practice the annotations of all of the subunits are nearly identical. In such cases, the un-normalised measure is likely to be a more conservative measure of functional similarity, relative to what one might intuit. Conversely, the normalised measure risks being overly generous in the case of poorly annotated proteins. We opt to use both measures in our investigation. We divide the measures by $\log(n)$, where n is the total number of proteins in the data set, to get a quantity whose upper bound is 1.

Chapter 3

The Function of Communities in Protein-Protein Interaction Networks at Multiple Scales

Most of the results presented in this chapter were published in BMC Systems Biology [181]. This chapter also includes new material, in particular the inclusion of an additional measure of functional similarity; a look at the robustness of the partitions found (Section 3.4.1); results from the literature-standard test of functional homogeneity (Section 3.6); a new test of functional homogeneity that considers chains of interacting proteins (Section 3.7.1); and an investigation into the distribution of protein functional types in communities (Section 3.10).

3.1 Introduction

The idea that network communities have some relationship to functional modules is entrenched in the literature (see Section 1.3.4). Indeed, it is such a strong hypothesis that finding communities enriched with functional terms has been used to assess how good a given community detection algorithm is [50]. Nonetheless, this remains an under-tested hypothesis, and one that deserves greater attention given its centrality to how it is thought biological function emerges from the interactions of parts to make a whole.

In Section 2.1, we summarised the current state of the community detection literature: there has been a lot of work on proposing new algorithms, but less work done to assess the significance of communities found [259, 323]. This is especially true of communities found in protein-protein interaction networks. Myriad studies have been published that employ a particular community detection algorithm, and then assess the functional homogeneity of communities found by searching for terms in a structured vocabulary – usually the Gene Ontology (GO, [12]) or Munich Information Centre for Protein Sequences categories (MIPS, [211]) – that are significantly over-represented within communities (e.g. [44, 75, 184, 191, 210, 253]). If such terms exist, the identified communities are said to be 'enriched' for biological function (see Section 1.3.4). This reliance on one yardstick (that of functional homogeneity as assessed by enrichment of functional terms) for investigating the biological relevance of network community structure is unsatisfactory for two main reasons. First, it is not a stringent enough test for what it was designed to show, namely that community structure adds insight compared to considering simpler topological features. Second, there are additional aspects of the link between community structure and biological function that are worth investigating: in particular, the patterns of the distribution of different functional classes of proteins in communities.

In Section 2.1, I introduced the idea of multi-scale community detection. As there are many scales of potential functional relevance within the PIN – one might expect to find smaller communities embedded inside progressively larger ones [259] – it seems natural to apply a method which allows one to uncover structure at many different scales. Prior to our paper [181], there had been no study published to our knowledge that investigated the community structure of PINs at multiple scales.

In this chapter, we probe the functional relevance of communities at multiple resolutions (scales) in two *S. Cerevisiae* (yeast) PINs. There are three main issues we consider. First we investigate the functional homogeneity of communities found, and observe how this changes with the scale at which we probe the network. Second, we consider the relationship of multi-scale community structure to a particular protein: it is possible to see which other proteins co-occur in communities with a protein of interest at different resolutions. Perhaps it co-occurs robustly with a small group of proteins at high resolution but also with a larger set of proteins at a lower resolution. Both groups are of potential interest in understanding what role the protein plays. This is particularly pertinent for poorly annotated proteins, as their potential functions can be revealed through clustering into communities [303]. Third, we explore the distribution of proteins of different functional types in communities.

3.2 Data sets and data processing

3.2.1 Protein-protein interaction data sets

Protein-protein interactions are of two fundamentally different types (see Section 1.2.2). The Molecular Interactions ontology [128] recognises two distinct types of interactions: *physical associations* (henceforth denoted P) and *associations* (henceforth denoted A). The main experimental type for the former is yeast-two-hybrid screens (y2h, see Section 1.2.3.1). The main type of experiment to fall under the latter is based on tandem affinity purification (TAP, see Section 1.2.3.2). These interaction types are known to have very different properties [294, 334]. Additionally, the networks constructed using these two types of interactions have different global properties (see Table 3.2). We thus investigate the two networks, based on A type and P type interactions, independently.

Here we use the BioGRID (`www.thebiogrid.org`, downloaded January 2010,

Table 3.1: Numbers of yeast interactions of A and P type in the three databases used.

	A	P
Intact	23632	26611
MINT	13347	10407
BioGRID	35716	13142
Union	48348	33342

[306]), IntAct (`www.ebi.ac.uk/intact`, downloaded January 2010, [159]), and MINT databases (`mint.bio.uniroma2.it/mint`, downloaded January 2010, [361]) to assemble our protein interaction networks. These databases overlap in the interactions deposited in them: we consider the union. We use only interactions between proteins that have an SGD identification (Saccharomyces Genome Database, `www.yeastgenome.org`, [53]).

We divide interactions on the basis of their type (A or P) and hence assemble two networks. Numbers of interactions are given in Table 3.1. The IntAct database [159] gives interaction types from the Molecular Interaction ontology [128] directly. The MINT database [361] uses the Molecular Interaction detection type ontology, the broad categories of which are biophysical, biochemical, and protein complementation assay. The biochemical techniques give evidence of association (A type interactions), and the biophysical and protein complementation assays give evidence of physical interactions (P type). The BioGRID database [306] uses its own evidence types. Those giving evidence of P type interactions are reconstituted complex, PCA, Co-crystal structure and yeast-two-hybrid. Those giving evidence of A type interactions are affinity capture, biochemical activity, co-fractionation, co-purification and Far Western. (Details of these experimental types can be found on the BioGRID website, `www.thebiogrid.org`.)

Of the potential 6607 proteins in the yeast proteome (`www.yeastgenome.org`), there are 5002 proteins connected by A type interactions, and 5692 connected by P type interactions. There are only 5947 interactions in the intersection of the A and P

Table 3.2: **Network statistics of the A and P networks**

Network	A	P
Number of nodes	4980	5669
Number of edges (of which self-edges)	48,330 (868)	33,321 (941)
Mean degree	19.1	11.5
Density of interactions	0.0039	0.0021
Mean local clustering coefficient	0.22	0.10

interactions sets. Here we only study the largest connected component of these two networks. Some summary statistics for the networks are shown in Table 3.2. The A network is denser (density is the number of interactions divided by the number of possible interactions), and has a higher mean local clustering coefficient. A node has a high clustering coefficient, c, if its neighbours are also neighbours of each other [231, 344]. It is defined for each node as

$$c = \frac{N_{\text{triangle}}}{N_{\text{triple}}}, \tag{3.1}$$

where N_{triangle} is the number of triangles of which the node is a member, and N_{triple} is the number of connected triples in which the node is the central node. (A connected triple is a single node with edges running to an unordered pair of other nodes.)

3.2.2 GO

Th Gene Ontology, GO, is a structured vocabulary, whose terms are annotated to proteins by researchers (see Section 2.4). Here, we use the Biological Process subontology annotations to yeast, which are maintained by the SGD consortium [53]. Terms are related to each other through a directed acyclic graph (DAG, a directed graph with no directed cycles). Proteins are annotated with the most specific terms that are known about them. It is then possible to add their parent terms to this set by following the structure of the DAG up to the root node. Of the 6346 yeast proteins in the GO annotation set, 5347 have biological process annotations (excluding the root node). The mean number of annotations per protein is 17.2. The majority of proteins in the PINs we consider have GO biological process annotations: 4394 proteins in the P network and 4610 in the A network. We carried out the same analysis using the molecular function and cellular component sub-ontologies and obtained similar results.

We calculate the functional similarity between pairs of proteins using these GO annotations and the measures defined in Equations 2.12 and 2.13 in Section 2.4, and refer to these un-normalised and normalised measures as G and N, respectively. The two measures differ in their consideration of lack of specific annotations, with the un-normalised measure being in general more conservative (see Section 2.4).

3.2.3 MIPS

We compared our GO results to those gained from the use of MIPS terms (`www.helmholtz-muenchen.de/en/ibis`, [211], an alternative set of functional annotations to GO). Here we only use the top level of the MIPS hierarchy, which has 28 terms. MIPS terms are annotated to 4431 of the proteins in the A and 4231 of the proteins in the P network. We apply the un-normalised functional similarity measure given by Equation 2.12 to the MIPS data. We refer to this similarity measure as M.

3.2.4 Chemoinformatics screen of growth rates data

Hillenmeyer et. al. published the growth rates of gene-knockout strains under 418 different conditions [129]. The data is given in the form of a vector L_i with components

$$L_i^t = \log(\mu_i^c/\mu_i^t), \tag{3.2}$$

Table 3.3: **Pairwise similarities of proteins in the A and P networks under the four different similarity measures, G, N, C, and M**

	A		P	
Measure	All pairs	Interacting pairs	All pairs	Interacting pairs
G (un-normalised GO)	0.13	0.31	0.12	0.25
N (normalised GO)	0.16	0.39	0.15	0.30
C (correlated phenotypes)	0.036	0.12	0.036	0.077
M (MIPS)	0.084	0.18	0.083	0.16

where the parameter μ_i^c is the mean growth rate of strain i under different control conditions, and μ_i^t is the growth rate under one of the 418 treatment conditions. We use the results from the homozygous strains. As many gene deletions are lethal, data are only available for 3625 proteins, of which 3184 are in the A network and 3422 are in the P network.

This data is used to give a measure of the functional similarity of two proteins i and j, C_{ij}, by calculating the Pearson correlation coefficient of L_i with L_j.

3.3 Pairwise properties of proteins

Community structure, if of any biological relevance, should uncover patterns that are more than the sum of effects from pairs of interacting proteins. In Table 3.3, we show the pairwise similarity of proteins in each network under our four different measures of functional similarity (two based on GO, one on MIPS, and one on correlated growth rates; see Section 3.2). For each of the four measures, the similarity of pairs known to interact with either A or P type interactions is much higher than a randomly chosen pair of proteins. This not only helps motivate the investigation of the connection between functional similarity of proteins and the topology of the network, but also demonstrates the necessity of taking into account pairwise properties when assessing any additional information that one can gain by studying communities.

3.4 Communities

We apply the Potts community detection algorithm (discussed in Section 2.1.3.2, using the Louvain algorithm [34]) to the A and P networks separately. The algorithm partitions the network into disjoint communities, where the size of the communities is influenced by a resolution parameter λ. When λ is small, large communities are found. As it increases, higher resolution structure in the form of smaller communities becomes visible.

We investigate partitions of the network in the range $0.1 \leq \lambda \leq 1000$, and sample at intervals of 0.01 on a logarithmic scale (we hence report results for $-1 \leq \log(\lambda) \leq 3$). At $\lambda = 0$, all nodes in our set will be assigned to the same community. As we increase λ, communities split and become smaller. If we allow λ to increase eventually each node will be assigned to its own community.

Figure 3.1 shows the communities that we find in the A and P networks as the resolution parameter λ is varied. As λ increases, more and smaller communities are found (see Table 3.4). At $\lambda = 1$ (i.e. $\log(\lambda) = 0$), which corresponds to standard Newman-Girvan modularity [233], most communities contain a few hundred proteins. By $\log(\lambda) = 3$ however, almost all proteins are in communities of size three or smaller. As shown in Figure 3.1, some sets of nodes are classified in the same community through large changes in the resolution parameter and hence represent particularly inter-connected parts of the network.

Figure 3.1 can be contrasted with Figure 3.2, which illustrates similar calculations on an Erdős-Rényi random network and a network designed to possess strong communities. In the former, not much structure is present; in the latter, there are very distinct blocks. To produce these images, we adopt a convention for ordering the proteins (explained in Section 2.2).

The two networks, A and P, contain very different types of interactions, and they can therefore be used to identify different aspects of the cell's functional organisation.

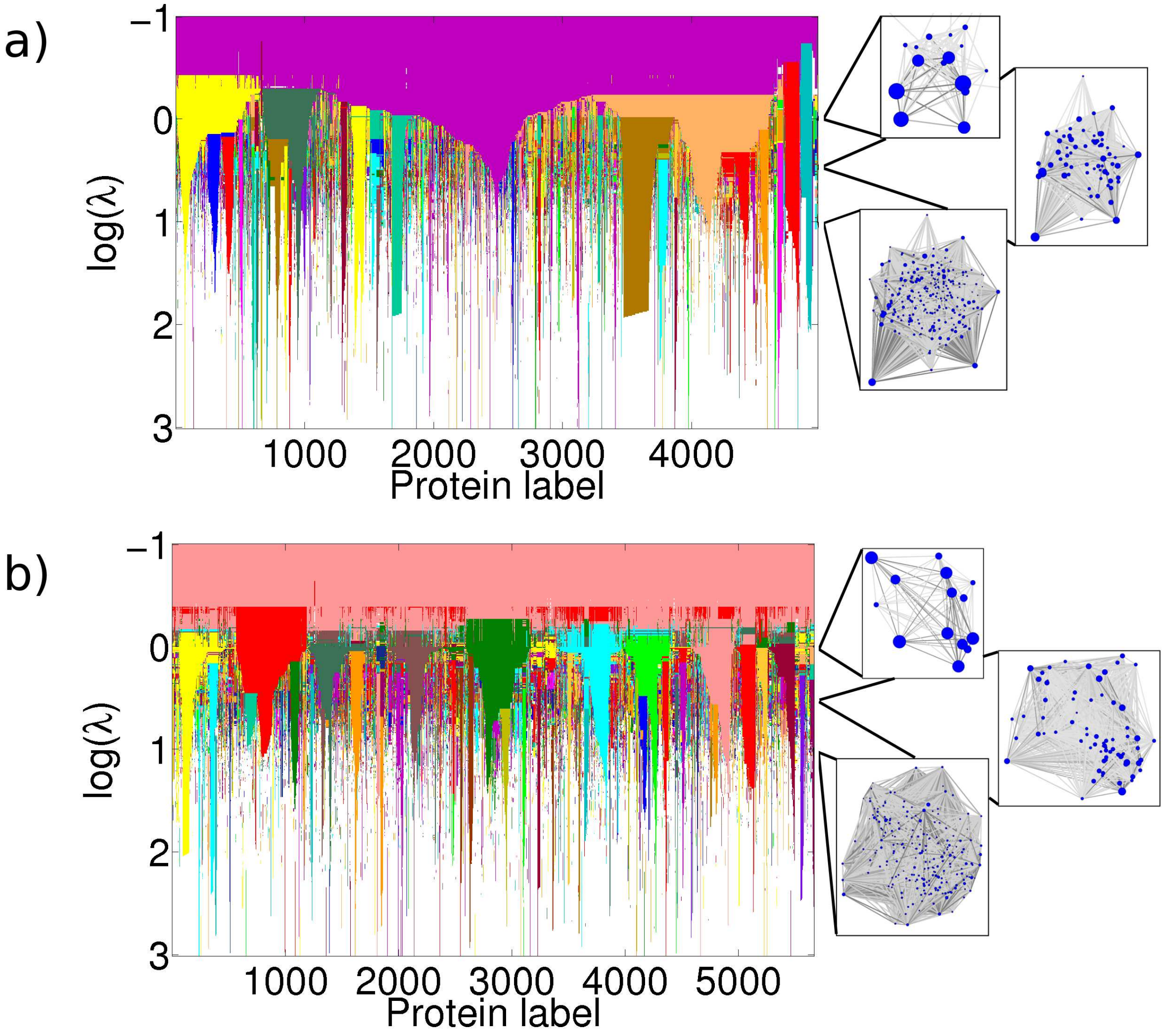


Figure 3.1: **Communities identified in the A and P Networks.** Communities identified in a) the A network and b) the P network. When the resolution parameter λ is very small, all nodes are assigned to the same community (which is analogous to viewing the network at a great distance). As λ is increased (viewing the network at progressively closer distances), more structure is revealed. The figures on the right hand side show visualisations of the networks' partition into communities at three different values of λ. Each circle represents a community, with size proportional to the number of proteins in that community, positioned at the mean position of its constituent nodes. (We acknowledge the authors of Ref. [322] for use of code to generate these plots. The node positions were determined via a standard force directed network layout algorithm [154].) The shade of the connecting lines is proportional to the number of edges between two communities. The main figure shows the communities that we find as we vary the resolution parameter. We identify communities as the same through changing resolution parameter, and hence colour them the same, according to a convention described in Section 2.2 (only communities of size 50 or more are shown). Note that the ordering of proteins is not the same in the two figures.

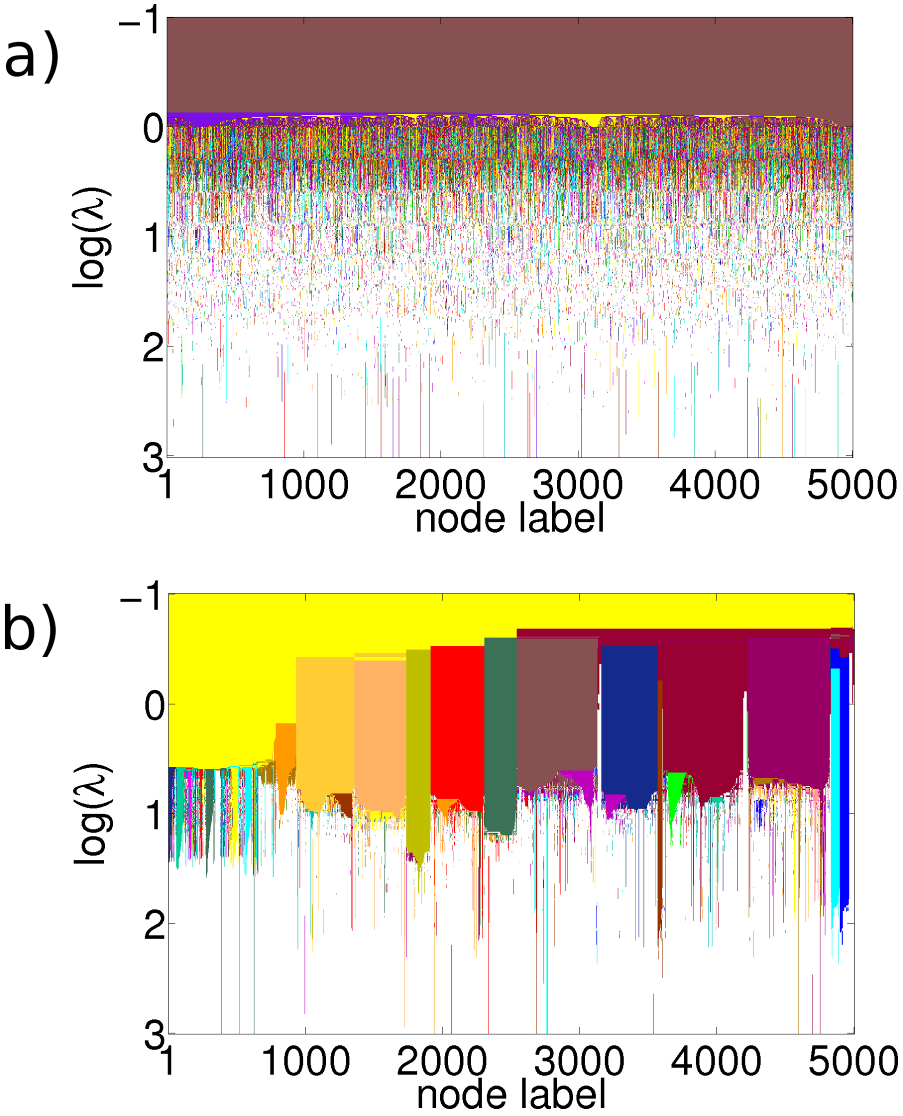


Figure 3.2: **As for Figure 3.1, but for a) An Erdős-Rényi random network and b) a network with strong community structure.** Both networks were designed to be of approximately the same size as the A and P networks (5000 nodes). The probability that two nodes are connected in the random network is the same as for the A network. We generated the network with community structure from code available at http://sites.google.com/site/santofortunato/inthepress2 (reported in [174]). The parameters that we chose matched the statistics of the A network (average degree of 19, maximum degree of 1182), with additional parameters chosen as suggested default values (the exponent for the degree distribution is 2, the exponent for the community size distribution is 1, and the mixing parameter is 0.2).

Table 3.4: **Mean size of communities in the A and P networks**. Communities of size three or fewer proteins are excluded from these calculations.

$\log(\lambda)$	**mean size of communities**	
	A	P
-0.5	621	2834
0	293	405
0.5	73	79
1	22	26
1.5	11	10
2	6.4	6.5
2.5	5.2	4.9
3	4.4	4.4

The A network is also much denser than the P network. Clustering into communities would be dominated by A type interactions if the two interaction types were considered together, thereby making it very hard to pick out any structures given by P type interactions (as occurs in [256]).

3.4.1 Robustness of partitions found

As discussed in Section 2.1.3, a paper published recently pointed out a general problem with maximisation of modularity and similar quality functions: there are potentially many partitions all sharing nearly identical values of the modularity [111]. As finding the exact maximum of modularity is known to be an NP-hard problem [39, 127], the heuristics that have to be used in practice return varying partitions, and the assignments of nodes to specific partitions can be potentially very different from one another.

As mentioned in Section 2.1.3.3, the Louvain algorithm we use is sensitive to the order in which the nodes appear in the input list. This enables an investigation of the potential different partitions all with similar local maxima of modularity. We hence run the algorithm for 100 different node orderings at each value of λ, to test

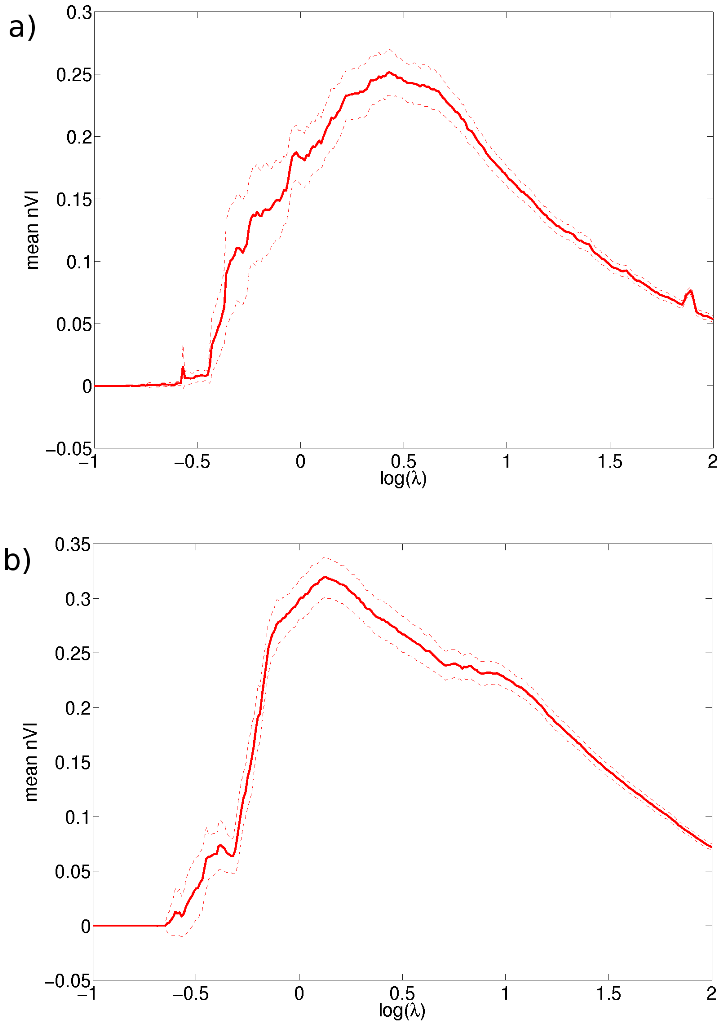


Figure 3.3: **The mean normalised variation of information (nVI) based on one hundred different runs of the community detection algorithm at each value of the resolution parameter λ, for a) the A network and b) the P network.** An nVI of 0 would indicate that all the partitions were identical. The dashed curves show plus and minus one standard deviation.

the extent to which partitions can differ in the networks we consider.

There are many ways to compare two partitions, see Section 2.2. For the reasons outlined in that section, we employ the normalised Variation of Information, nVI, to make the $(100\times99)/2$ comparisons at each value of λ. The results are shown in Figure 3.3. Values of nVI are normalised to lie between 0 and 1, with 0 indicating identical partitions. At intermediate values of λ the nVI has a maximum. For comparison to some other measures, in the A network at $\log(\lambda) = 0.5$ where the mean nVI is 0.25 (the resolution for which there is the most variability as judged by the mean nVI), the mean normalised Mutual Information is about 0.75 and the mean Adjusted Rand about 0.6 (see Section 2.2 for an introduction to these measures).

How large are these nVI values? To gain some intuition, we randomly rewired the A network 100 times, keeping the number of interactions of each protein constant, such that the rewired networks differed in 10% of their interactions. We ran the community detection algorithm (keeping node ordering constant) on these networks, and found the nVI values between the partitions to be slightly higher (about 0.05 larger over the full range of resolution-parameter values) than those shown in Figure 3.3.

3.5 Examples of communities found at multiple resolutions

To motivate a systematic probing of the function of communities in protein-protein interactions at multiple scales, we give two examples of communities in this section. We 'eyeball' their significance not by any statistical test but by looking at the short protein descriptions found on the SGD website (`www.yeastgenome.org` [53]).

Consider the community at $\log(\lambda) = 0$ that is marked as the blue block in Figure 3.1 for the A network (over node labels approximately 0 to 500). This contains 528

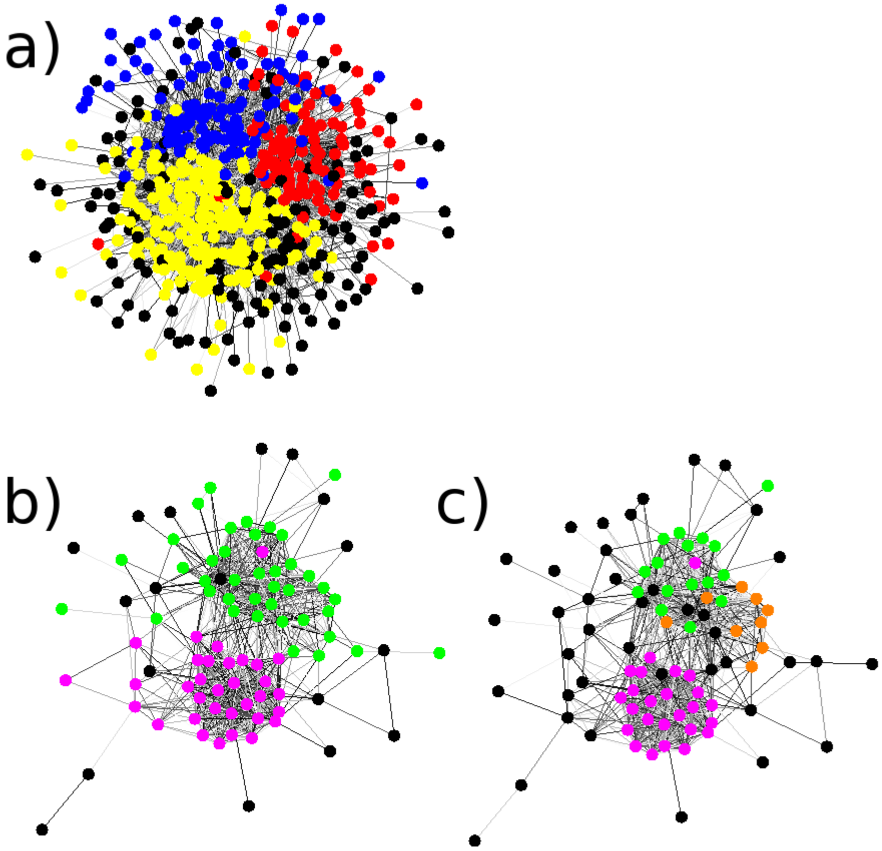


Figure 3.4: **Examples of communities found** a) A representation of a community in the A network at resolution parameter value $\log(\lambda) = 0$, with nodes (proteins) coloured according to the partition of this community at $\log(\lambda) = 0.5$. The colours are the same as for Figure 3.1 a), where this group of proteins has labels roughly in the range $0 - 500$. Almost all of the nodes have some relationship to the ribosome. The proteins in the yellow community are mostly ribosomal subunits, those in the red community are mostly pre-cursors to and processors of the small ribosomal subunit, and those in the blue community have similar roles to those in the red community but for the large subunit. The shading of the edges has no significance; its purpose is to ease visualisation. Black nodes are not located in one of the three largest communities discussed in the text. b) A representation of a community at $\log(\lambda) = 0.5$, with nodes (proteins) coloured according to the partition of this community at $\log(\lambda) = 0.75$. The proteins identified at the lower resolution parameter value almost all play some role in transcription initiation. At the higher resolution parameter value, more structure is revealed: the pink community consists mostly of proteins from the RNA polymerase II mediator complex and the green community mostly consists of proteins from the TFIID and SAGA complexes. c) Partition at a higher resolution parameter value ($\log(\lambda) = 1.6$). The green community from b) has split into the SAGA complex (green) and the TFIID complex (orange). The names and descriptions of the proteins in the communities in these examples are given in Appendix B. The node positions for visualisation were computed in the same way as for Figure 3.1.

proteins, the majority of which are known to have some relationship to the ribosome, see Tables B.1 – B.4. Figure 3.4 a) shows this community, where we have coloured nodes according to the community partition at the higher resolution $\log(\lambda) = 0.5$. The colours – red, yellow, and blue – are the same as in Figure 3.1, where most of the community present at $\log(\lambda) = 0$ has split into three communities at $\log(\lambda) = 0.5$. The blue community consists of 107 proteins, which are largely precursors to and processors of the large ribosomal unit. The red community consists of 95 proteins, which have a similar function but for the small ribosomal subunit. The yellow community has 190 proteins, 93 of which are constituents of the ribosome and the remainder of which are either of unknown function or associate to the ribosome.

An illustration of the biological relevance of community structure at three partitions is given in Figures 3.4 b) and c). We show a community of 90 proteins at $\log(\lambda) = 0.5$, and display its partition into communities at b) $\log(\lambda) = 0.75$ and c) $\log(\lambda) = 1.6$. Almost all of the proteins in the community at $\log(\lambda) = 0.5$ play some role in transcription initiation. At $\log(\lambda) = 0.75$, this community has split into two main smaller communities: the pink community contains constituent proteins of the RNA polymerase II mediator complex and the green community contains components of the closely related SAGA and TFIID complexes [353]. At $\log(\lambda) = 1.6$, this second community has split into the SAGA and TFIID complexes. We give short descriptions of the proteins in these communities in Appendix B.

These examples illustrate that the results of multi-scale community detection can have intuitive biological relevance. We proceed in the remainder of this chapter to elucidate this connection between network structure and biological function.

3.6 The standard assessment of biological relevance: functional enrichment

There is a literature-standard way of assessing whether communities in protein-protein interaction networks are functionally homogeneous and hence candidates for biological modules (Section 1.3.4). A community is judged to be functionally homogeneous if at least one term is enriched in the community, where enrichment is determined in a comparison between the subset of proteins in a community compared to all proteins in the data set using a cumulative hypergeometric distribution [38]. Consider a population of size M, in which K elements have a particular feature. If N draws are made without replacement from this population, the probability that up to x of the drawn samples have the feature is given by the cumulative hypergeometric distribution:

$$F(x|M,K,N) = \sum_{i=0}^{x} \frac{\binom{K}{i}\binom{M-K}{N-i}}{\binom{M}{N}}. \tag{3.3}$$

In our case M is the total number of proteins in the network, K is the number of proteins annotated with the term in question, N is the number of proteins in the community, and x is the number of proteins in the community with the particular term. The probability, p, that a term would appear x or more times in the community of interest by chance is then

$$p = 1 - F(x-1|M,K,N). \tag{3.4}$$

This test is performed for multiple terms, so to control for multiple testing we must apply a correction. Here we apply the Bonferroni correction, which simply means multiplying the resultant p values by the number of tests performed, and is a standard approach [301].

We assess the communities we find using this test. In common with the literature, we judge a community functionally enriched if at least one term has a value of p (after being corrected for multiple testing) below 0.05. Figure 3.5 shows the number of proteins found in communities of size four or more (black), and the number of proteins in such communities which are found to be functionally enriched under this literature standard test (red). This test tends to find the very large communities found at low resolution to be functionally enriched. for example, the community of 4873 proteins present at $\log(\lambda) = -0.74$ in the A network is functionally enriched using this test.

This test, although useful for identifying which terms are enriched within which communities, is not a satisfactory test of whether communities are functionally homogeneous. This is because it fails to take into account that pairs of interacting proteins are more likely to be annotated with the same functional terms than non-interacting pairs (see Section 3.3). One must control for the presence of interacting protein pairs, to see whether communities do not end up enriched for function just because they contain many such pairs.

3.7 Functional homogeneity of communities

A community necessarily contains many more interacting pairs than a randomly chosen set of proteins. We thus compare the pairwise functional similarities of all interacting pairs of proteins in a community to the same measure for all interacting pairs in the network, thereby controlling for the number of interacting pairs.

To capture the pairwise similarity between two proteins that interact $\{ij\}$, we use z-scores:

$$z_{\{ij\}} = \frac{S_{\{ij\}} - \mu}{\sigma}, \tag{3.5}$$

where S stands for one of our four similarity measures – based on GO (G and N),

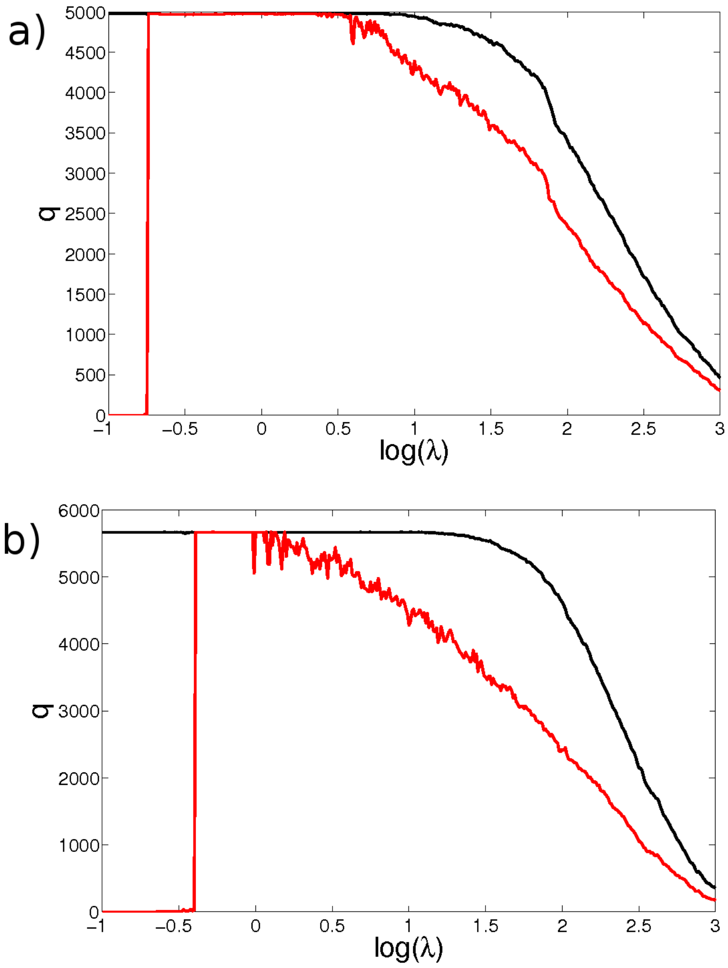


Figure 3.5: **Number of proteins in communities of size four or more that are enriched for at least one GO biological process term at the** 0.05 **significance level (red-curve) and in total (black curve) for a) the** A **network, and b) the** P **network.** At the standard resolution parameter value of $\log(\lambda) = 0$, almost every protein is in a community that is functionally enriched in both networks. Many published studies have applied this test to the output of community detection algorithms and found that most communities are enriched for at least one functional term. This is taken as evidence for the modular organisation of the cell.

MIPS (M), or correlated growth rates, (C) – μ is the mean, and σ the standard deviation of all the values of S for which proteins i and j interact in the network of interest (A or P).

A desirable quality for our test of functional homogeneity is the ability to compare communities found at different resolutions in an even-handed manner. It is inherent in the nature of a statistical test that the significance of the test statistic under consideration (for example, the difference between the sample mean and the population mean) depends on the sample size: if one has a larger sample size, one can judge smaller differences to be 'significant'. To determine the aggregate z-score z_{agg} for the mean of a set of individual z-scores z_{ind} one calculates $z_{\text{agg}} = \sqrt{N}\mu(z_{\text{ind}})$, where N is the number of interacting pairs in the community and $\mu(z_{\text{ind}})$ is the mean of their z_{ind} [208]. Hence, given a $\mu(z_{\text{ind}})$, a larger and hence more significant z_{agg} is achieved for a larger sample size (i.e., larger N). In order to separate the effects of the number of interactors in the community from functional homogeneity, we thus choose to base assessment of functional homogeneity on the $\mu(z_{\text{ind}})$, in our case $\mu(z_{\{ij\}})$ (where $z_{\{ij\}}$ is defined in Equation 3.5). We judge as 'significant' all those communities that have $\mu(z_{\{ij\}})$ above 0.3, and call such communities "functionally homogeneous". We stress that this is not strictly an assessment of statistical significance, as we are choosing to ignore sample size. The value of 0.3 is somewhat arbitrary: communities would be judged to be significant at the 0.05 significance level if they contained 30 or more interacting pairs.

We now assess how many communities are judged functionally homogeneous, looking in particular at how our results vary with resolution parameter.

The black curves in Figure 3.6 illustrate for a) the A network and b) the P network, i) the number of communities of size four or more as the resolution parameter changes, and ii) how many proteins are in those communities. The coloured curves represent, for each of our four measures of functional similarity, the number of com-

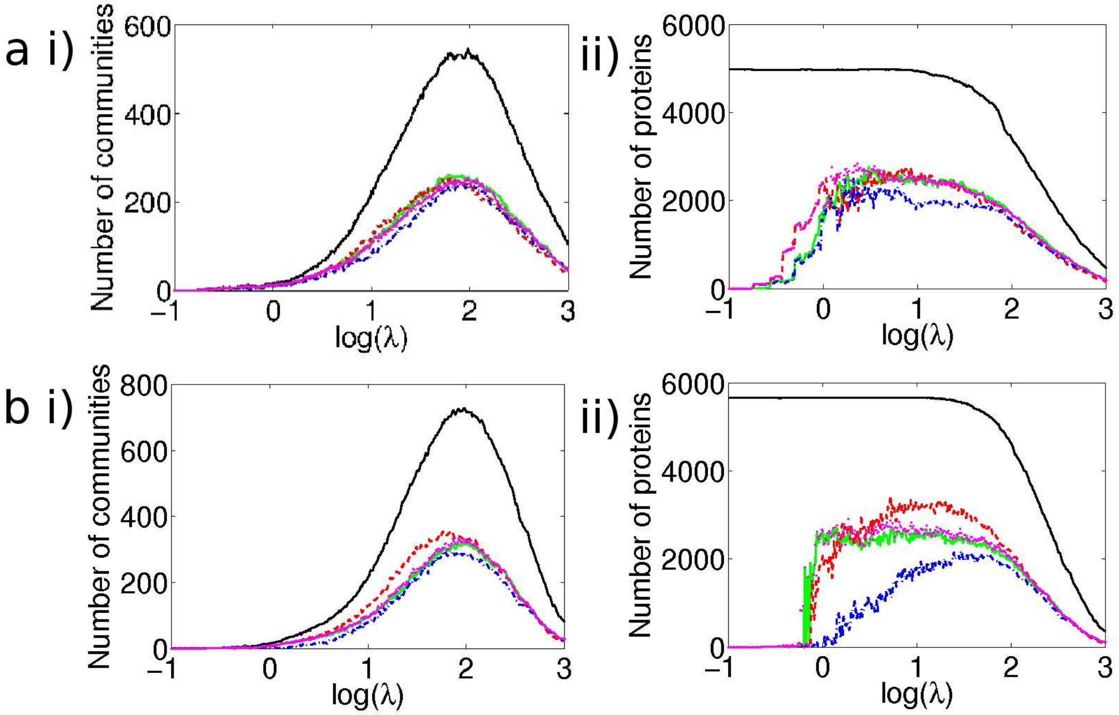


Figure 3.6: **For a) the A network b) the P network, i) the number of communities of size four or more and ii) the number of proteins in such communities and the fraction of these that are judged functionally homogeneous.** i) The number of communities of size four or more with changing resolution parameter (solid black curve) ii) The number of proteins p in communities of size four or more (solid black curve). Also shown are the numbers of communities/proteins in such communities judged to be functionally homogeneous according to the GO G similarity measure (green curves), the GO N similarity measure (dotted pink curves), the MIPS measure (dot-dashed blue curves), and the correlated growth similarity measure (dashed red curves). At values of $\log(\lambda) \leq 0.5$, relatively few proteins are in communities judged to be functionally homogeneous. The curves are similar for both networks, and they show a similar proportion of proteins in functionally homogeneous communities. One difference is that there are more proteins in functionally homogeneous communities at a lower value of $\log(\lambda)$ for the P network.

munities judged to be functionally homogeneous, and the number of proteins in the communities that we judged to be functionally homogeneous. We find that the large communities present at small values of the resolution parameter λ are not judged to be functionally homogeneous. As λ is increased, larger numbers of proteins occur in functionally homogeneous communities, peaking in the range $0.5 \leq \log(\lambda) \leq 1$ for the A network. At $\log(\lambda) = 0.5$, the mean community size is 73 proteins, and 2541 of 4980 proteins are in functionally homogeneous communities as judged by our GO similarity measure G.

We find that the functional homogeneity is very similar under the G and N measures, and show their correlation for communities detected in the A network at $\log(\lambda) = 0.5$ in Figure 3.8. The three communities which can be seen to have a substantially lower functional homogeneity under the G measure than the N measure are exactly as anticipated – they are for large sets of proteins that all share near identical annotations, which we would nonetheless consider to be fairly complete annotations: one of these communities is dominated by ribosomal proteins, one by mitochondrial ribosomal proteins, and one by proteasomal proteins. Due to the very close correlation between the G and N measures, we henceforth refer only to the G measure.

The shapes of the curves in both Figure 3.6 a) and b) are in the most part similar for the measures considered. Indeed, we find that the overlap between the communities judged to be functionally homogeneous between any two of the G, C, and M measures is high; for example, it is 70% between the GO and correlated growth rates measure over almost the entire range of the resolution parameter in both the A and P networks (see Figure 3.7). Given that the correlated growth similarity measure represents a very different data type to the GO and MIPS annotations, this agreement gives us confidence in the similarity measure we use for GO and MIPS. As we use only the top level of the MIPS functional annotations, we capture less information

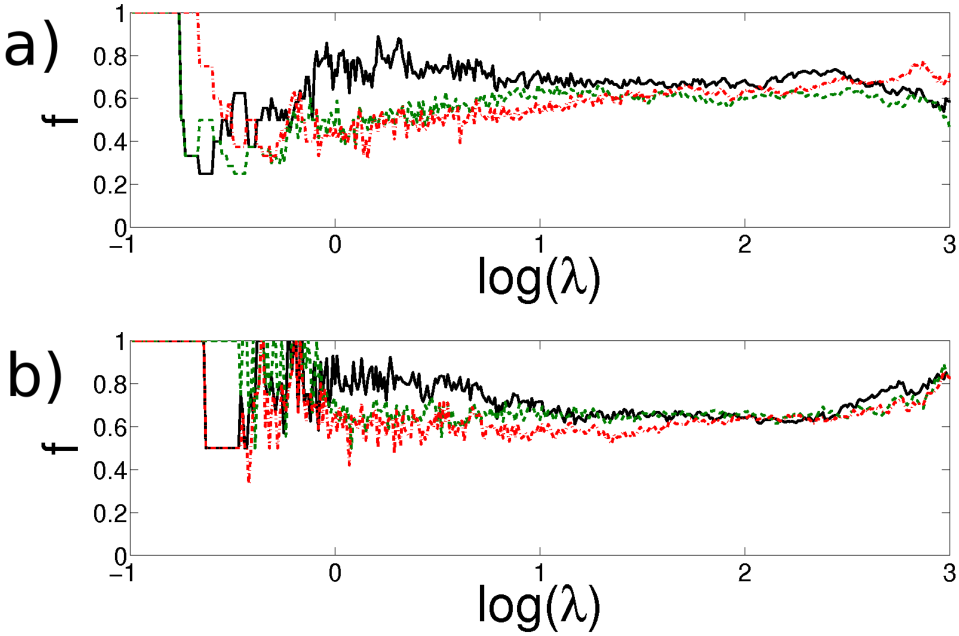


Figure 3.7: **Agreement in assessment of functional homogeneity between pairs of similarity measures.** For a) the A network and b) the P network, the fraction, f, of communities that are either both judged as functionally homogeneous or both judged as not functionally homogeneous under the G and C measures (black curve), the G and M measures (dark green dashed curve), and the M and C measures (red dot-dashed curve). The large overlap between the measures derived from ontologies (G and M) with the measure derived from a single large-scale experiment (C) gives confidence in our ontology-derived measures.

than the GO measure, so it is unsurprising that fewer communities are found to be functionally homogeneous using this measure.

The P network shows a similar pattern to the A network. One difference is that communities start to be judged as functionally similar at a slightly lower resolution. That there are comparably many functionally homogeneous communities in the P network as the A network is of interest, as communities found in P networks have previously been found to be poor choices for predicting function on the basis of enrichment of terms [303].

For almost all proteins, there is some value of the resolution parameter that assigns them to a functionally homogeneous community. In fact, 4652 out of 4980 A

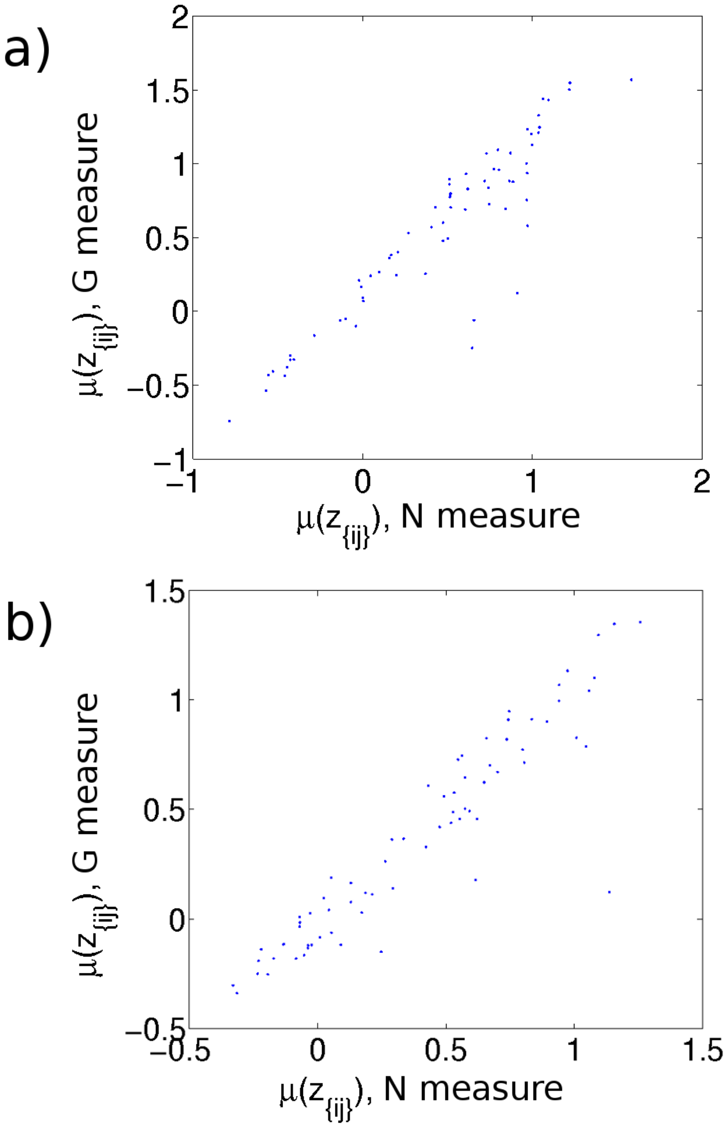


Figure 3.8: **The functional homogeneity of communities is very similar using the G and N measures of similarity**. Here we show for a) the A network and b) the P network the mean of the individual z-scores, $\mu(z_{\{ij\}})$, for the un-normalised measure based on GO (G) and the normalised measure based on GO (N) for communities at $\log(\lambda) = 0.5$. The three communities with substantially lower values under the G measure are as anticipated (see text).

proteins and 5647 out of 5669 P proteins are in such communities at some value of the resolution parameter that we considered.

3.7.1 Beyond pairwise measures

Tests of the functional homogeneity of communities need a comparison class, i.e. a set of objects relative to which one wishes to assess functional homogeneity. Studies published previous to ours had considered as a comparison class the whole set of proteins (see explanation of the literature standard test in Section 3.6). This effectively treats communities as 'bags of proteins', and compares each such bag to the whole. We have argued that this is not a strict enough test as a community contains many pairs of interacting proteins, and interacting proteins are known to be more functionally similar than non-interacting pairs (see Section 3.3). The literature-standard test thus leaves open the possibility that communities are no more functionally homogeneous than a bag of interacting pairs of proteins. In the previous section, we introduced our tests designed to check whether communities are more than a bag of interacting pairs.

One could consider that these tests are still hampered by being purely pairwise. The interactions within a community are not independent, but in some sense 'close' to each other. Here, we introduce a larger control class: that of sets of proteins joined together along a path. In Section 2.1.3.2 we discussed how it has been shown that modularity is a linear approximation of a quality function known as 'stability'. The optimisation of the stability quality function returns as communities sets of nodes that a random walker on the network visits within a certain time span. This time span is inversely proportional to the resolution parameter λ. As a comparison class for a community, which is very closely related to the set of nodes a random walker visits, we consider the set of nodes a random walker *that performs minimal back-tracking steps* visits. Such a walk is given by the depth-first search algorithm [315]. A depth-first search starts at a node chosen uniformly at random, then follows unexplored edges as far as possible before it needs to back-track, at which point it returns to the last node it visited that had unexplored edges. We use MATLAB's `graphtraverse` function to generate such sets of nodes, which we henceforth refer to as *chains*. For a community of size n nodes, we generate 100 such chains (each starting from a randomly chosen node) also of length n.

To compare the functional homogeneity of a community to the comparison class of a set of chains, we find the mean of the functional homogeneity of all pairs of proteins in the community, m_c, and the same value for each i of the hundred chains, m_i. We then define the z-score z_c for the chains-based functional homogeneity of community c to be

$$z_c = \frac{m_c - \mu_i}{\sigma_i}, \tag{3.6}$$

where μ_i and σ_i are the mean and standard deviation of the set m_i. We calculate these scores for each of the four functional homogeneity measures and for both the A and P networks, and show our results in Figure 3.9.

In the P network, the very large community present around $\log(\lambda) = -0.5$ (which contains the whole network except for ten proteins) is judged functionally homogeneous under the M and C measures. This problem arises when creating samples for comparison that are almost the size of the whole network: the samples are all more or less identical and hence have a vanishing standard deviation, which leads to a correspondingly high z-score if the actual community has an even slightly higher value. This problem only arises when communities are almost, but not quite, the size of the whole network. The same communities tend to be judged functionally homogeneous compared to chains under all four measures (see Figure 3.10. Results for the N measure are extremely similar to those for the G measure and are hence not shown).

A comparison to Figure 3.6 illustrates that, with the exception of the issue just outlined, a similar proportion of communities and proteins within those communities are judged as functionally homogeneous under this chain measure as with our interactors measure introduced in the previous Section. It is the case that communities judged functionally homogeneous under our interactors-based test also tend to be judged homogeneous under our chains-based test (see Figure 3.11).

Both the chains-based test and our interactors-based test take into account different aspects of a community that could make it inappropriate to compare the functional homogeneity of a community to a 'bag of proteins'. The interactors-based test controls for the presence of more interactions between the proteins than one would expect if the proteins were chosen at random; the chains-based test controls for the fact that interactions within a community are not independent of each other, but are in some sense 'close'. Using these tests, we show that communities give additional insights compared both to sets of interacting pairs, and to groups of nodes joined together in a chain.

In the following sections, when we refer to 'functionally homogeneous communities', we refer to the results of our interactors-based test.

3.8 Use of topological properties to select functionally homogeneous communities

Almost all proteins are in functionally homogeneous communities at some value of the resolution parameter, so we would like to devise a method to identify these resolutions. We investigate whether any easily-calculated topological properties of the communities can act as indicators of functional homogeneity. Given a protein of interest, we can then use such measures to identify 'good' resolutions without the need to assess functional homogeneity.

We tested 26 topological properties – see the list in Table 3.5 – for their ability

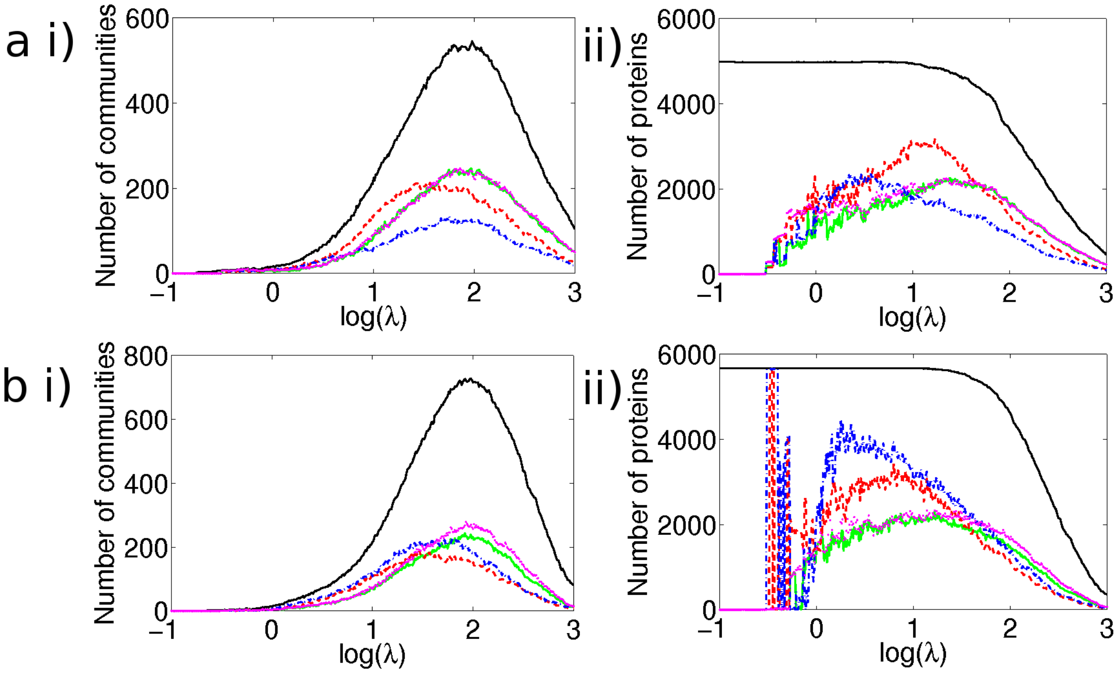


Figure 3.9: **For a) the A network and b) the P network, i) the number of communities of size four or more and ii) the number of proteins in such communities and the fraction of these that are judged functionally homogeneous using our chains-based test.** i) The number of communities of size four or more with changing resolution parameter (solid black curve) ii) The number of proteins p in communities of size four or more (solid black curve). Also shown are the numbers of communities and the numbers of proteins in such communities judged to be functionally homogeneous according to the GO G similarity measure (green curves), the GO N similarity measure (dotted pink curves), the MIPS measure (dot-dashed blue curves), and the correlated growth similarity measure (dashed red curves). A comparison to Figure 3.6 shows that there are similar proportions of proteins in functionally homogeneous communities judged using our chains-based test as using our interactors-based test. This reinforces the idea that communities yield additional insights that one cannot obtain by just using simple topological features.

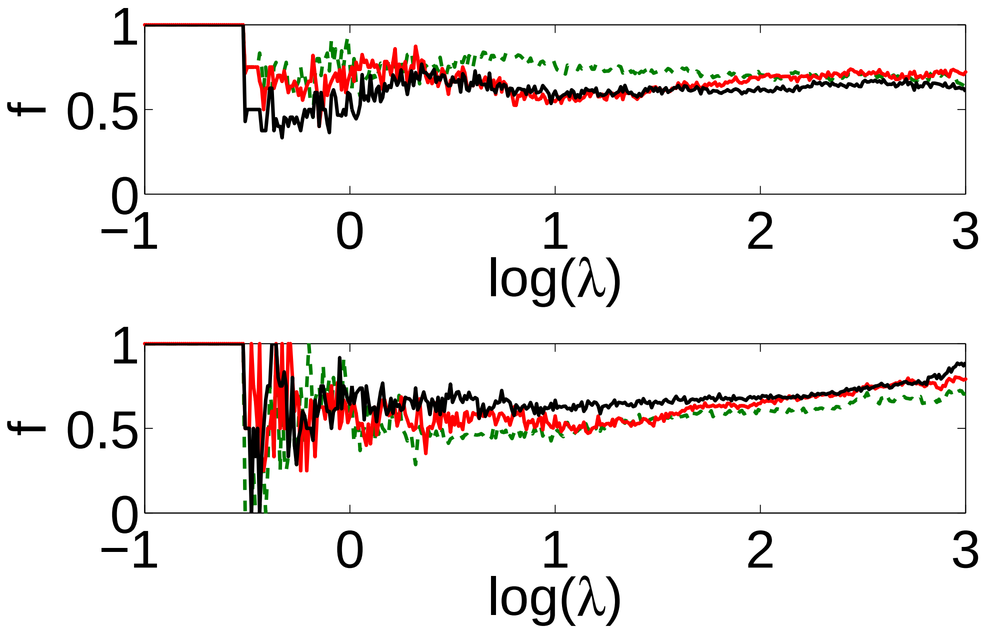


Figure 3.10: **Agreement in assessment of functional homogeneity, where this is as compared to chains, between pairs of similarity measures.** For a) the A network and b) the P network, the fraction f of communities of size four or more that are either both judged as functionally homogeneous or both judged as not functionally homogeneous compared to chains, under the G and C measures (black curve), the G and M measures (dark green dashed curve), and the M and C measures (red dot-dashed curve).

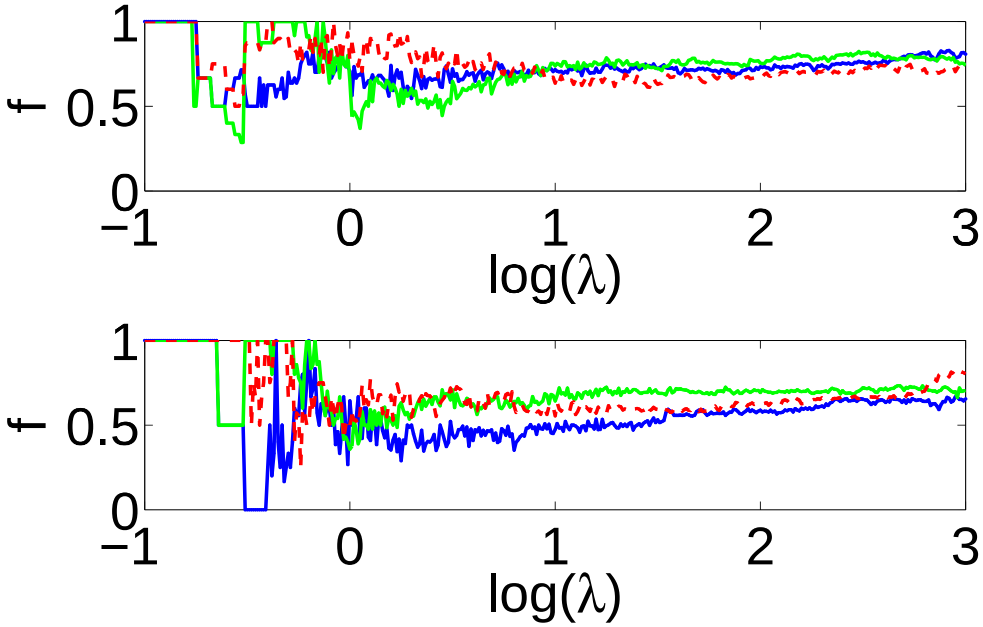


Figure 3.11: **Agreement in assessment of functional homogeneity for our interactors-based test and our chains-based test under the different similarity measures.** For a) the A network and b) the P network, the fraction f of communities of size four or more that are either both judged as functionally homogeneous or both judged as not functionally homogeneous for both the interactors-based test and the chains-based test, under the G measure (green curve), the M measure (dot-dashed blue curve), and the C measure (dashed red curve). The tests tend to pick out similar sets of communities as functionally homogeneous under all the similarity measures.

to predict functional homogeneity using the area under the ROC curve (AUC, see Section 2.3)). Examples of diagnostics tested include the mean local clustering coefficient, betweenness measures, and network diameter. Any topological properties that correlate well with functional homogeneity can then be used to predict functionally homogeneous communities. We use each topological property as a classifier by predicting communities as functionally homogeneous when the value of that property is above a threshold, which we vary allowing us to construct a Receiver Operating Characteristic (ROC) curve (see Section 2.3). A ROC curve plots the number of communities correctly predicted as functionally homogeneous versus the number falsely predicted [83]. We calculate the AUC for each diagnostic at each value of λ, and report the mean of this quantity over resolutions between $0 \leq \log(\lambda) \leq 3$ (we exclude $-1 \leq \log(\lambda) < 0$, as the results are very noisy due to the small number of communities present). An AUC of 0.5 would be expected from a random classifier. AUCs of greater than 0.5 imply that higher values of the diagnostic are predictive of functional homogeneity. AUCs of less than 0.5 imply the diagnostic would be predictive if instances with values of that diagnostic *below* the threshold were used (i.e. that the property and functional homogeneity are negatively correlated).

In general, the AUCs for the P network are lower than those for the A network (Table 3.5), perhaps because there is more potentially usable information in the A network as it is significantly denser (see Table 3.2).

We find that the mean local clustering coefficient is the most useful of the topological properties tested in the prediction of functional homogeneity for all three similarity measures in the P network and for both the G and C measure in the A network. (Recall from Section 3.2.1 that a node has a high clustering coefficient if its neighbours are also neighbours of each other.) Figure 3.12 shows for a) the A network and b) the P network the ROC curves for using the mean clustering coefficient of nodes in a community as a predictor of functional homogeneity for each of the three

Table 3.5: **Topological network diagnostics tested and AUCs.** The network topology measures tested and their associated AUCs. We report the results for using each of these as a predictor for functional homogeneity as judged under the three measures of functional similarity – GO (G), correlated growth rates (C), and MIPS (M) – for both the A and P networks. The AUCs are given as the mean performance over the range $0 \leq \log(\lambda) \leq 3$. The clustering coefficient (definition given in the text, equation 3.1) is the best predictor in all cases save for the MIPs measure in the A network. (The topological properties were computed from code developed by Gabriel Villar.)

	A			P		
Network topology measure	G	C	M	G	C	M
Mean degree	0.6476	0.6142	0.513	0.6062	0.5373	0.5387
Degree assortativity coefficient [230]	0.6913	0.6277	0.4799	0.6541	0.5517	0.5181
Clustering coefficient [62]	**0.7186**	**0.6613**	0.5521	**0.665**	**0.5829**	**0.5725**
Global mean Soffer clustering coefficient [300]	0.4857	0.4819	0.3915	0.5395	0.4735	0.4461
Local mean Soffer clustering coefficient [300]	0.4784	0.4662	**0.3892**	0.5312	0.4654	0.454
Mean geodesic node betweenness centrality [342]	0.46	0.4973	0.5045	0.4954	0.5094	0.4959
Mean closeness centrality [342]	0.5275	0.5524	0.4877	0.5053	0.4919	0.4815
Mean eigenvector centrality [342]	0.5601	0.5722	0.5312	0.5658	0.5551	0.5246
Mean information centrality [342]	0.5191	0.5429	0.5253	0.5432	0.5456	0.517
Mean geodesic distance [62]	0.3839	0.3717	0.4274	0.4823	0.4945	0.5066
Diameter [342]	0.4457	0.4042	0.4366	0.5074	0.5004	0.5079
Mean harmonic geodesic distance [62]	0.4088	0.4042	0.5024	0.4709	0.4834	0.4995
Energy [62]	0.5237	0.4982	0.4568	0.5265	0.4976	0.5114
Entropy [62]	0.5655	0.5327	0.5077	0.5428	0.5127	0.528
Off-diagonal complexity [163]	0.5941	0.5457	0.5081	0.5827	0.5054	0.5237
Cyclomatic number [163]	0.6331	0.5733	0.5173	0.6146	0.53	0.5425
Connectivity [163]	0.6437	0.5766	0.5245	0.6324	0.5334	0.5468
Number of spanning trees [163]	0.4525	0.4531	0.4451	0.4584	0.4516	0.4491
Medium articulation [163]	0.5659	0.4463	0.5295	0.5754	0.507	0.5592
Efficiency complexity [163]	0.5316	0.5343	0.4911	0.5078	0.4945	0.4982
Graph index complexity [163]	0.6564	0.6492	0.5211	0.599	0.5469	0.525
Density	0.6541	0.6553	0.5277	0.6227	0.5676	0.5235
Efficiency [176]	0.579	0.5896	0.4964	0.5336	0.5071	0.4865
Fraction of articulation vertices [324]	0.5065	0.5028	0.5216	0.5064	0.5062	0.5091
Largest eigenvalue	0.6054	0.5663	0.4941	0.5619	0.5041	0.5185
Rich club coefficient [59]	0.5428	0.5896	0.4988	0.5486	0.5209	0.4868

similarity measures.

In the A interactions set some experimental results are described using the matrix model and some using the spoke model (Section 1.2.3.2). This choice could cause artefactual topological features, so the extent to which we find particular topological features correlating with functional homogeneity could be sensitive to annotation choice. We are therefore encouraged that the same trends in predictive ability are evident in the P network, for which there is no such element of discretion.

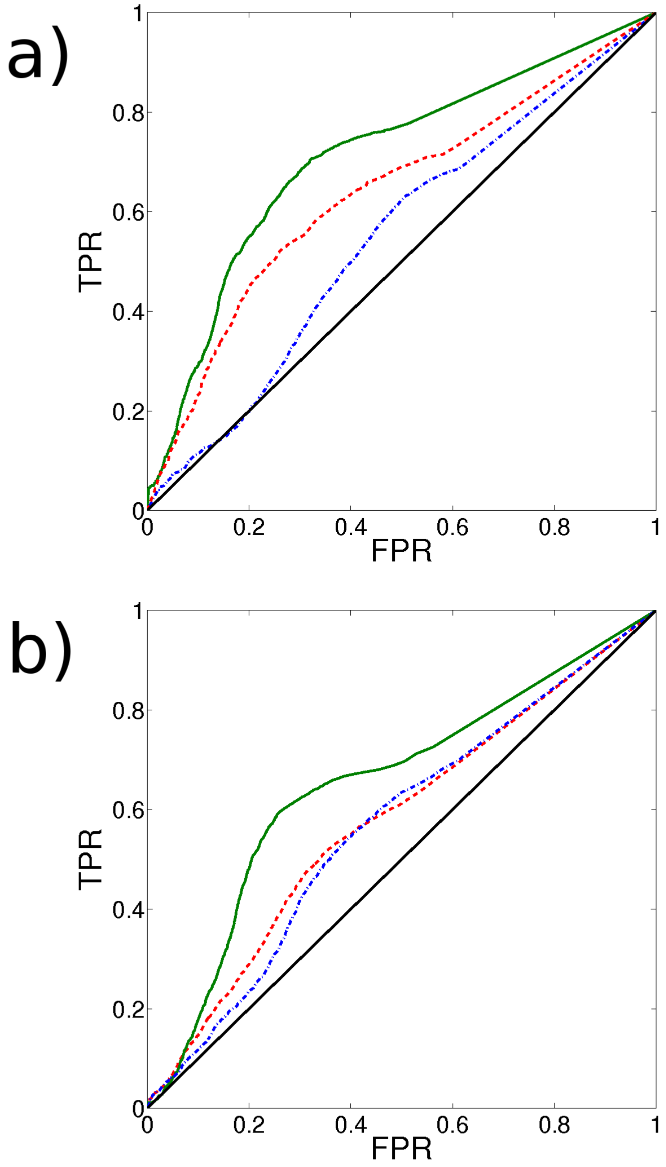


Figure 3.12: **ROC curves for using mean clustering coefficient to pick out functionally homogeneous communities in a) the A network and b) the P network.** The Receiver Operating Characteristic (ROC) curve for using mean clustering coefficient as a predictor of functional homogeneity under the GO measure (solid green curve), MIPS measure (dot-dashed blue curve), and correlated growth measure (dashed red curve). We plot the false positive rate (FPR) versus the true positive rate (TPR). As for Table 3.5, we use the mean FPR and TPR rates for resolutions between $0 \leq \log(\lambda) \leq 3$. A random classifier would give the solid black line. For the A network under the GO measure, a true positive rate of 70% is achievable with a false positive rate of 30%. For both networks, the best predictive ability is achieved for the GO measure, and the worst for the MIPS measure (see Table 3.5 for AUCs). The AUCs for the P network are in general lower than those for the A network (see Table 3.5).

3.9 Tracing the community membership of a particular protein

Multi-resolution community detection and characterisation is relevant from the global viewpoint, where one can investigate the aggregate functional organisation of the proteome (as we have done in previous Sections). It is also relevant from a local perspective, where the community membership of particular proteins can be traced through changing the resolution parameter. As mentioned in Section 3.1, this could be particularly useful for poorly characterised proteins, as the proteins with which it co-occurs in communities at different resolutions can be indicative of its function.

We thus now investigate a protein-centred view of multi-resolution community detection. We consider, for an example protein, the properties of the communities to which it is assigned through changing resolution parameter (see Figure 3.13). The size of the communities, their mean similarity under the G and C measures, and the mean clustering coefficient are shown. The protein is a member of the ESCRT-I complex. (Figure 3.14 gives a further four examples.) Note the robust properties of the communities in the A network over resolution-parameter values of approximately $1 \leq \log(\lambda) \leq 2.5$, despite the tendency for them to be further partitioned as λ increases. At these resolutions, the protein is in the same community as other members of the complex, as well as a few other very closely associated proteins. Beyond $\log(\lambda) = 2.5$, the complex is broken up, as reflected in the drop in mean similarity values. The community present over $0.7 \leq \log(\lambda) \leq 1.4$ in the P network contains many proteins associated to the complex (in addition to the complex itself). Above the step observable at $\log(\lambda) = 1.4$, only members of the complex are present. In Appendix C, we give the names and brief functional descriptions of proteins that occur in some of the same communities for this example (as well as the four other examples given in Figure 3.14). These five examples all show the following behaviour.

- In general, as is expected, the size of the community to which a protein is assigned decreases with increasing resolution. There is often a large range of resolutions over which the community has constant size. Such communities are particularly resilient to being partitioned at increasing resolutions, despite the tendency for them to be further partitioned.

- The community similarity under the G and C measures often shows a close correlation.

- At higher resolutions, there tends to be a higher community similarity, as might be expected. This is, however, not always the case: community similarity can decrease at higher resolutions. In these instances, a group of proteins has been partitioned beyond the point at which function is shared – possibly through the exclusion of proteins involved in the same processes that do not necessarily directly interact with each other.

- Although there is often a large overlap between the community membership in the A and P networks, this need not be the case. For example, in Figure 3.14 c), the depicted protein occurs with other proteins in the same complex in the A network, whereas in the P network it occurs with members outside the complex that are nonetheless involved in the same process (see Appendix C). The functional homogeneity of communities can also be different: sometimes the protein occurs in many functionally homogeneous communities in the A network and not the P network, and sometimes vice versa. This is unsurprising given the very different nature of A and P interactions. By treating them separately, we are able to pick out both patterns.

- These figures suggest, as should be no surprise given the results in Section 3.8, that clustering is a good proxy for functional homogeneity when looking at

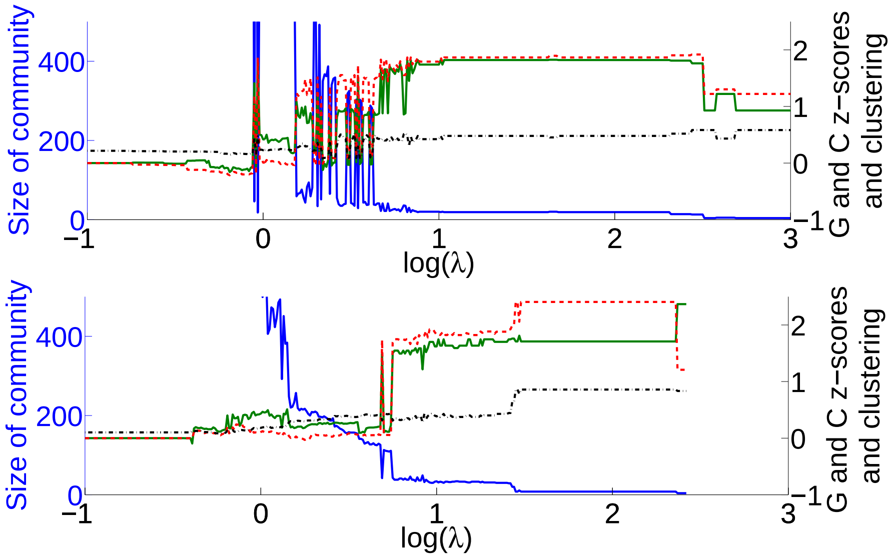


Figure 3.13: **Tracing the community membership of a particular protein through changing resolution.** For the example protein YCL008C, we show the size (solid blue curve), mean local clustering coefficient (dot-dashed black curve), mean z-score under the GO measure (solid green curve), and correlated growth measure (dashed red curve) with changing resolution for the A network (top) and P network (bottom). Long plateaus in these properties represent robust communities.

individual proteins, and in the absence of much functional information could guide which resolution(s) should be targeted for investigation.

3.10 Investigating particular protein functions

In the preceding sections we have focused on the aggregate functional homogeneity of communities. In this section we investigate the distribution of particular protein functional types in communities. We do not in general anticipate that one module carries out a task well-captured by a particular functional annotation: functional annotation schemes were not necessarily designed to reflect the sort of tasks a biological module might be anticipated to carry out. Consider the example of proteins involved in transport. Almost every module's task would be anticipated to need proteins to move cellular components around. Likewise, a module entirely composed of proteins

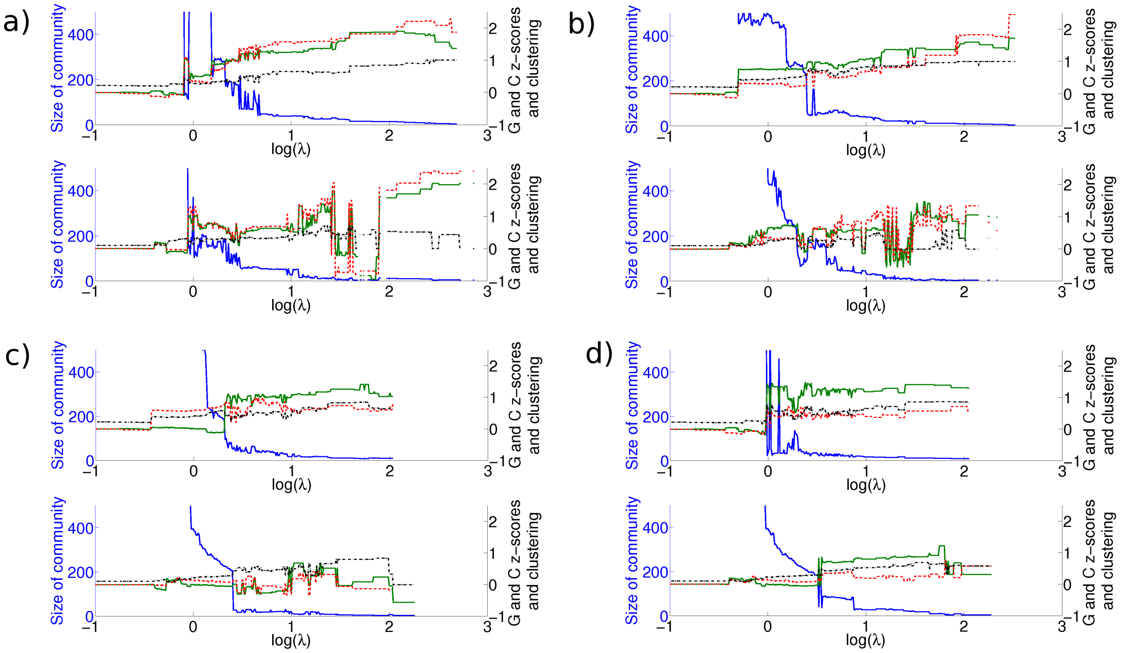


Figure 3.14: **Further examples as per Figure 3.13.** These figures display the same information as Figure 3.13, but for the proteins a) YAL002W, b) YAL011W, c) YAL016W, and d) YAL021C. We show the size (solid blue curve), mean local clustering coefficient (dot-dashed black curve), mean z-score under the GO measure (solid green curve), and correlated growth measure (dashed red curve) with changing resolution for the A network (top) and P network (bottom). Gaps appear whenever the protein is assigned to a community of size three proteins or less. We give the names of proteins in several example communities, chosen as motivated by these figures, in Appendix C.

involved in transport is not anticipated (the transport of what for what purpose?). There may well be functional annotations that do indeed better correspond to the tasks that biological modules carry out. In this section we seek to give a more fine grained view of the connection between community structure and biological function by investigating the different behaviour of proteins in different functional classes. There is a large class of questions of potential interest: Are proteins of the same type typically concentrated in one or a few communities? How does this change with changing resolution parameter? Do pairs of functional types co-occur preferentially within communities? Are there any 'natural' resolutions for proteins of a particular functional class?

Here we focus on a small but broad set of protein types, which are the GO biological process terms within the yeast GO slim [132] that are annotated to at least 100 yeast proteins, see Table 3.6. A GO slim is a slimmed-down version of the whole GO ontology to a small set of terms designed to give a representative and broad overview of the complete set of annotations. It is not the case that they are the top level of the hierarchy; indeed, it is possible for one GO slim term to be the parent of another.

3.10.1 Interactions and functional classes

In Section 3.3, we showed that a pair of proteins that interact have a higher functional similarity (under the four measures considered) than a randomly selected pair of proteins. This suggests that proteins interact with proteins that carry out similar functions, but is this the case for all functional types?

Figure 3.15 shows the tendency for proteins of two functional types, s and t, to be connected by protein-protein interactions. This tendency is calculated by counting the number of times term s is annotated to a protein and term t is annotated to a protein with which this protein interacts, and dividing by the total number of proteins term s is annotated to multiplied by the total number of proteins term t is annotated to. For

Table 3.6: GO slim terms used and the numbers of proteins in each network annotated to these terms.

	GO slim term	**Number of proteins**	
		A **network**	*P* **network**
1	generation of precursor metabolites and energy	127	130
2	DNA metabolic process	337	344
3	transcription	108	108
4	protein modification process	441	45
5	cellular amino acid and derivative metabolic process	173	177
6	transport	770	830
7	response to stress	431	444
8	mitochondrion organization	120	131
9	cytoskeleton organization	141	143
10	cell wall organization	126	133
11	signal transduction	143	148
12	membrane organization	197	206
13	RNA metabolic process	616	614
14	vesicle-mediated transport	272	277
15	cellular homeostasis	103	112
16	response to chemical stimulus	271	281
17	cellular lipid metabolic process	176	200
18	cellular protein catabolic process	120	124
19	cellular carbohydrate metabolic process	184	203
20	heterocycle metabolic process	120	133
21	cofactor metabolic process	105	121
22	chromosome organization.	332	333

both A and P networks, this tendency is high when $s = t$ (higher numbers down the diagonal), indicating that to a large extent, proteins tend to interact with proteins involved in similar biological processes. However, this tendency, called *homophily* in the context of social networks [342], is by no means uniform for all functional types and is indeed hardly present for certain functional classes in both the A and the P network – for example proteins annotated with 'cellular lipid metabolic process' (term 17) and 'cytoskeleton organization' (term 10) in the A network. There are also some pairs of functional classes with a high tendency to be connected by interactions, in particular 'transcription' (term 3) and 'RNA metabolic process' (term 13) in the A network.

There are some striking similarities for the A and P network, but also some notable differences. For example, proteins annotated with 'RNA metabolic process' (term 13) are much more likely to have A type interactions between themselves than is the case for P type interactions.

These patterns of interaction between different functional classes should be borne in mind when considering which functional types of protein are assigned to the same communities.

3.10.2 Distribution of functional classes in communities

We now ask, are proteins of a given functional type concentrated within one or a few communities, or are they spread evenly throughout the network?

The distribution of proteins annotated with a particular function across communities can be summarised by using the entropy (as introduced in Section 2.2). The entropy of a term t, E_t, is

$$E_t = \sum_i p_t(i) \log[p_t(i)], \tag{3.7}$$

where $p_t(i)$ is the fraction of proteins annotated with term t that are in community

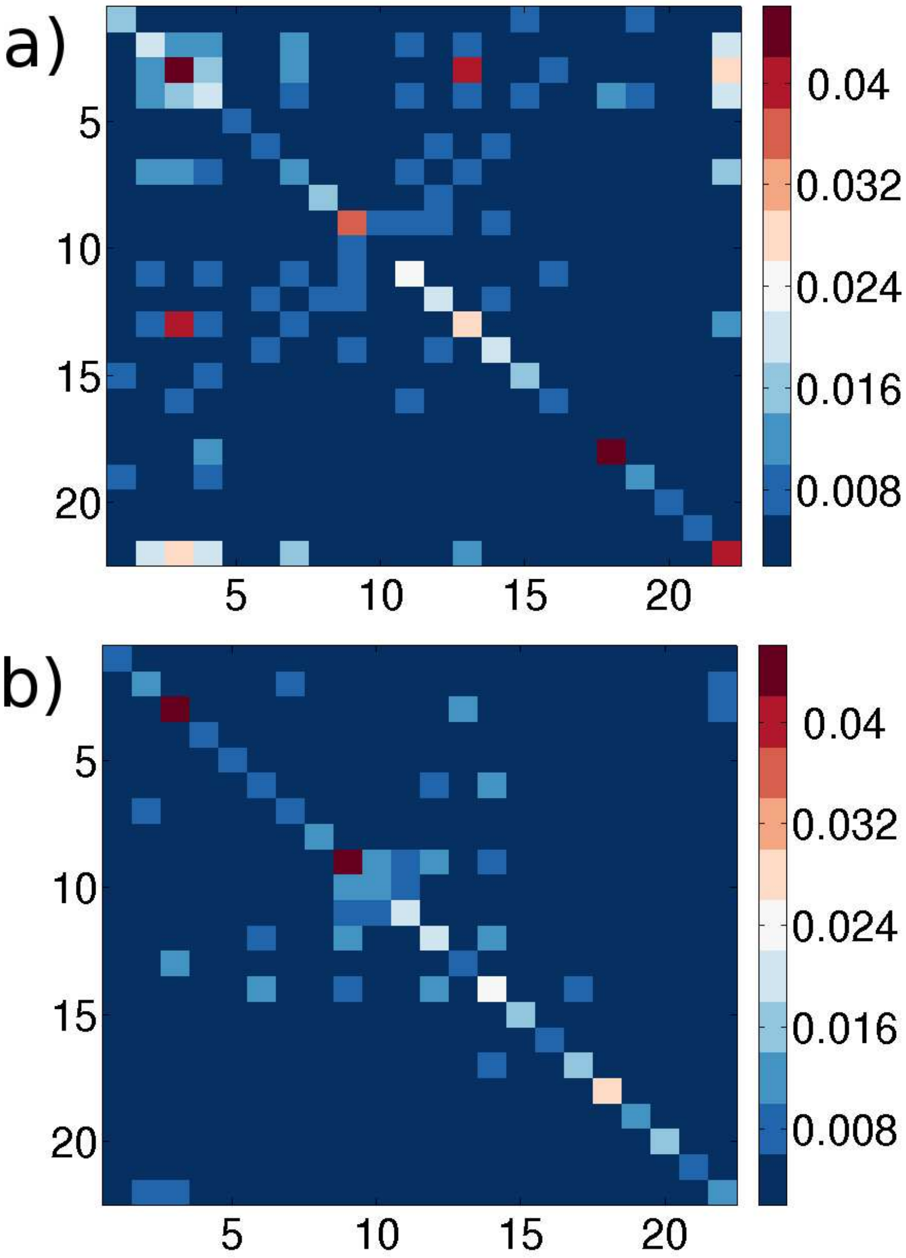


Figure 3.15: **Tendency for pairs of GO slim terms to be connected by a)** A **and b)** P **interactions.** These values are normalised by the number of possible connections between proteins of different functional classes. Proteins of some functional types show a high tendency to interact with proteins of the same type.

i. If a term is spread homogeneously through communities, it has a high entropy. If a term is concentrated in one or a few communities, it has a low entropy.

We calculate the entropy of the 22 GO slim functional classes with changing resolution parameter. As entropy is expected to change as the number of terms in the sum (in this case, the number of communities) changes, we normalise the entropies at each value of the resolution parameter using their z-score:

$$z_{E_t}(\lambda) = \frac{E_t(\lambda) - \mu_E(\lambda)}{\sigma_E(\lambda)}, \tag{3.8}$$

where $\mu_E(\lambda)$ and $\sigma_E(\lambda)$ are the mean and standard deviation of the entropies of all terms at resolution parameter λ. The z_{E_t}-scores allow us to compare the concentration of particular functional annotations within communities, and how this changes with resolution parameter λ.

Different terms display different behaviours. We pick out a few terms that show differing behaviours for the P network in Figure 3.16. Proteins annotated with the term 'transcription' are highly concentrated within particular communities at $\log(\lambda) = 0$, but are more evenly spread through communities at increasing values of λ; those annotated with 'cellular protein catabolic process' are most concentrated at an intermediate value of $\log(\lambda) = 1$; those annotated with 'response to stress' are consistently evenly spread; proteins involved in 'cell wall organisation' are most concentrated relative to other terms at the highest resolution parameter value that we investigate. The tendency for some functional types to be most concentrated within communities at higher values of resolution perhaps indicates that these processes are 'multi-faceted', in the sense that the cell may require this process independently for different tasks.

There is a very high similarity in the $z_{E_t}(\lambda)$-scores between the A and P networks (not shown). This similarity, notwithstanding the large differences between A and P

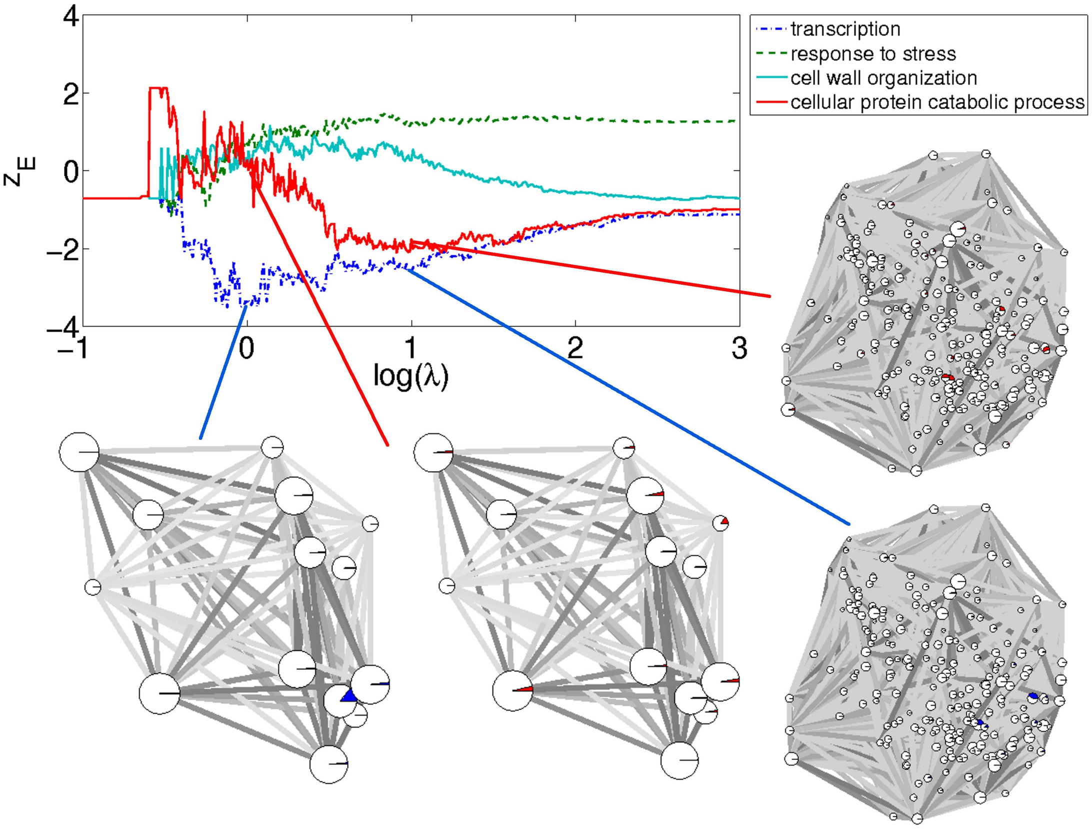


Figure 3.16: **The distribution of four GO slim terms throughout communities in the P network, as measured by the z-scores of the entropy of the distribution of these terms throughout communities.** The 'bubble' plots are as for Figure 3.1; the pie-charts illustrate the fraction of proteins in that community annotated with the term in question. The blue bubble plots are for the term 'transcription', the red 'cellular protein catabolic process'. Low z_{E_t}-scores indicate that proteins annotated with term t tend to be concentrated in a few communities. Concentrating on the two left-most bubble plots, in the plot on the right many of the communities have red content, whereas in the plot on the left very few of the communities have blue content: transcription proteins are highly concentrated within communities (i.e. have a low entropy) at this resolution. Terms show different tendencies to be concentrated in communities at different resolution parameters.

interaction types (in terms of their overall network structure, see Table 3.2, and the functional classes that the interactions tend to join, see Figure 3.15), gives weight to any identified differences between GO terms.

3.10.3 Co-occurrence of functional classes in communities

In the previous section, we investigated the distribution of individual terms across communities. Here we investigate the distribution of pairs of terms across communities. Calculating the co-occurrence of terms based on the distribution of proteins in communities does not take into account the known tendencies of proteins of particular functional types to interact with each other (as seen in Section 3.10.1). We hence find the ratio of the fraction of interactions that connect proteins annotated with terms s and t that are within communities, $F_{\{st\}}(\lambda)$, to the fraction of all interactions that are within communities, $F_{\{all\}}$:

$$W_{st}(\lambda) = \frac{F_{\{st\}}(\lambda)}{F_{\{all\}}(\lambda)}. \tag{3.9}$$

Figure 3.17 shows the values of W_{st} for both networks at $\log(\lambda) = 0.5$ and $\log(\lambda) = 1.5$. It is evident from these figures that, even after taking into account the pairwise tendencies of different classes of protein to interact, the co-occurrence of terms within communities is not homogeneous, particularly for the A network. Comparing Figures 3.17 and 3.15, one sees that the protein types that tend to interact also tend to have these interactions inside communities. Some particularly strong connections are evident at $\log(\lambda) = 1.5$: for example, 'mitochondrion organization' (term 8) and 'membrane organization' (term 12) in the A network; 'cell wall organization' (term 10) and 'heterocycle metabolic process' (term 20) in the P network; and 'heterocycle metabolic process' (term 20) and 'cofactor metabolic process' (term 21) in the P network. The first and third of these are not surprising, though the second one is not obvious. This serves to highlight the relevance of community structure: it recapitulates much of what we might expect, but can also bring to our attention patterns that would not otherwise be obvious.

3.10.4 Functionally homogeneous communities and functional classes

One might expect proteins involved in particular processes to show different propensities to lie in functionally homogeneous communities. We investigate what fraction of each type of protein lie in communities judged functionally homogeneous under the GO measure G through changing resolution parameter. We pick out a few examples for the A network in Figure 3.18. As these examples show, the propensity for proteins of a particular functional type to be in communities judged functionally homogeneous differs substantially (contrast 'chromosome organisation' and 'cellular lipid metabolic process'). Additionally, there are some indications that the resolutions of most interest can depend on the type of protein under investigation. Proteins annotated with 'RNA metabolic process' (term 13) are more likely to be found in functionally homogeneous communities at $\log(\lambda) = 0.8$, where the mean size of communities is 30. In contrast, proteins involved in vesicle-mediated transport (term 14) are found in greater numbers in functionally homogeneous communities at $\log(\lambda) = 1.7$, where the mean size of communities is 10.

3.11 Conclusions

If protein interaction networks are to aid understanding of how biological function emerges from the concerted action of many proteins, then it is crucial to explore connections between network structure and biological function.

We find that community structure does indeed help identify sets of proteins that

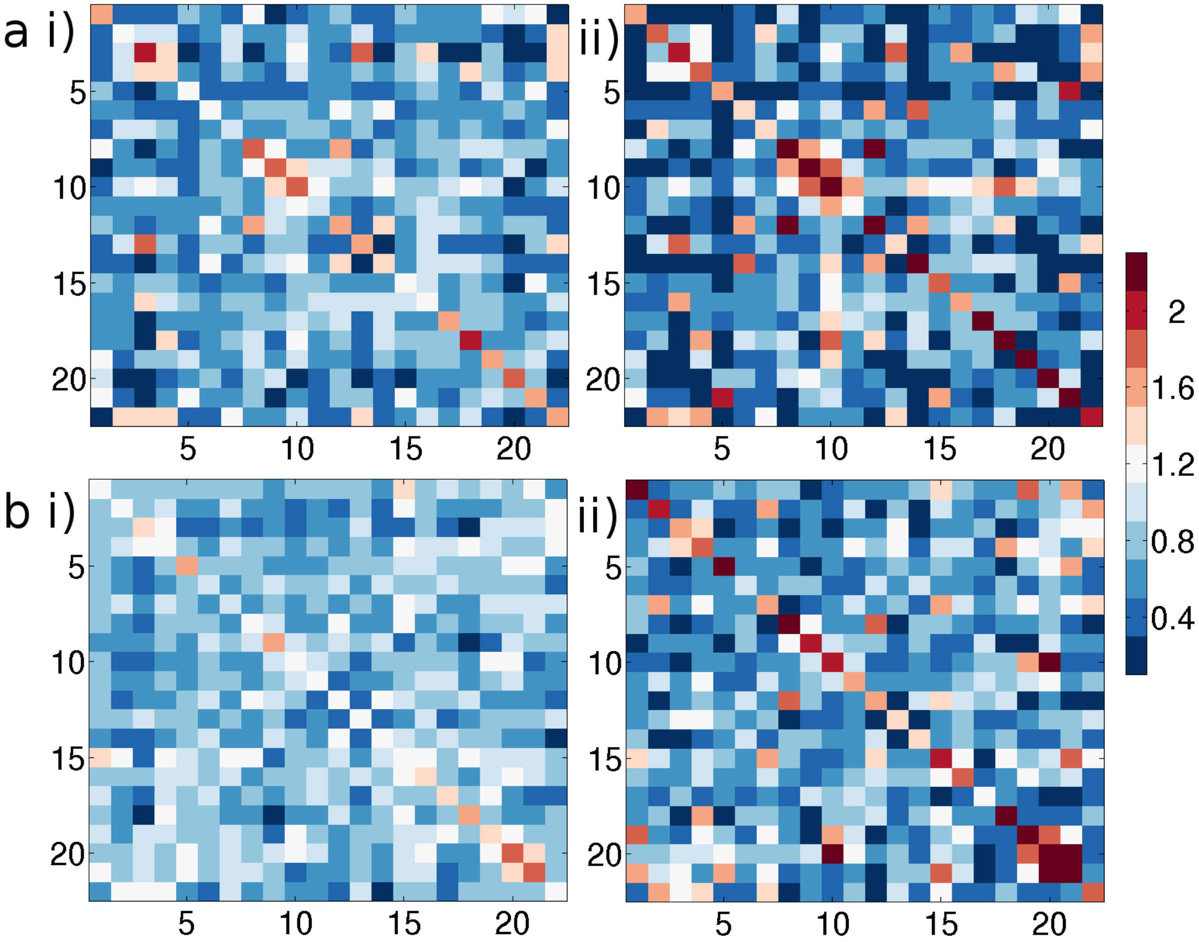


Figure 3.17: **The fraction of interactions connecting proteins of different types found within communities, W_{st}, in a) the A network and b) the P network, at i) $\log(\lambda) = 0.5$ and ii) $\log(\lambda) = 1.5$.** Interactions between proteins of the same functional type are the most likely to be found within communities, though there are exceptions (large off diagonal values). More heterogeneity of W is present at $\log(\lambda) = 1.5$ than at $\log(\lambda) = 0.5$. There are several notable differences between the A network and the P network. For example, at $\log(\lambda) = 1.5$ interactions between proteins annotated to 'heterocycle metabolic process' (term 20) and 'cofactor metabolic process' (term 21) are very likely to be in the same community in the P network, but not in the A network.

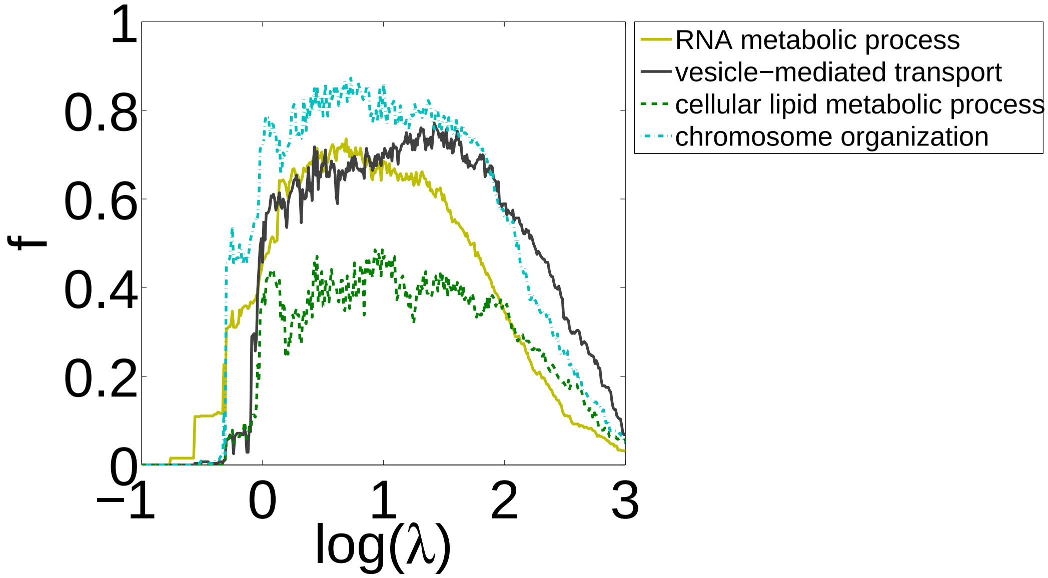


Figure 3.18: **Fraction f of proteins of particular types in functionally homogeneous communities in the A network** With changing resolution parameter proteins of particular types have consistent differences as to how often they are found in functionally homogeneous communities. For example, proteins involved in 'chromosome organisation' are far more likely to be in functionally homogeneous communities than proteins annotated with 'cellular lipid metabolic process'. There are also some features that suggest resolutions that may be of particular interest for a given process. For example, $\log(\lambda) = 1.7$ (for which the mean size of communities is 10) for proteins involved in 'vesicular mediated transport' and $\log(\lambda) = 0.8$ (where the mean size of communities is 30) for proteins annotated with 'RNA metabolic processes'.

act together, and that this connection between network structure and biological function depends on what network scales are probed. We do not expect there to be any single scale of interest in this middle-scale structure of the protein interaction network; although previous studies had applied community detection algorithms to protein interaction networks, no study had investigated this structure at multiple resolutions. We find that 4652 of 4980 proteins in the A network and 5647 of 5669 proteins in the P network are in functionally homogeneous communities at some value of the resolution parameter, as judged under a similarity measure based on GO annotations. The number of proteins in functionally homogeneous communities peaks at about $\lambda = 3$ for the A network (where the mean size of communities is 73, compared to the standard 'modularity' resolution of $\lambda = 1$, at which the communities have a mean size of 293). For the P network the peak is less pronounced, with the actual maximum occurring at $\lambda = 7$ (i.e. $\log(\lambda) = 0.86$).

Having a good measure of functional homogeneity is central for our analysis. We approach this issue by using four different characterisations of functional similarity: two based on the GO and one on the MIPS structured vocabularies, and one based on the growth rates of gene knock-out strains under different chemical conditions [129] (an independent and objective characterization of biological function). The prevalent method in the literature for assessing functional homogeneity of a group of proteins is inappropriate for communities, as the number of interacting pairs in a group must be taken into consideration. By defining similarity at the pairwise level, we have developed a fair test of functional homogeneity through a comparison of interacting pairs. We also capture the aggregate functional similarity of two proteins, overcoming the need to assess functional homogeneity on a term by term basis (although this is, of course, also possible once communities of particular interest have been identified). Our tests of functional homogeneity – which are not statistical tests in the conventional sense because of our desire to exclude the effects of sample size – using the four measures of similarity show a high level of agreement with each other, giving us confidence in our chosen measures of functional similarity.

As the functional knowledge of proteins is far from complete (even for well-characterised organisms such as yeast), we also search for topological properties of communities that are correlated with functional homogeneity. Through a characterization of the communities using 26 topological properties, we identify the mean clustering coefficient of a community as a good predictor of functional homogeneity, with a true positive rate of 70% achievable with a false positive rate of only 30% for the A network using the GO similarity measure G.

We have illustrated the utility of our framework for biologists who are interested in a particular protein. In a chosen interaction network, one can determine the community membership of the protein of interest at multiple resolutions. Even if there is a dearth of functional information, the easily-calculated mean local clustering coefficient can suggest resolutions of particular interest.

An investigation of the distribution of proteins of different functional types through communities uncovered some striking differences. As communities are broken up at increasing resolution-parameter values, some functional types of proteins become more concentrated within communities, whereas the distribution for others becomes more homogeneous with increasing resolution-parameter values. This is the first hint of the importance of different resolution parameters for proteins of different functional types. We also found that some functional types are never concentrated within a few communities but are instead always spread fairly evenly throughout communities. We find that interactions connecting proteins of the same functional type (for some but not all functions) are very likely to be found within communities, and that this tendency becomes more pronounced at higher resolutions. We also investigated whether proteins of some functional types are more likely to be found in functionally homogeneous communities than others, and found this was indeed the case.

Throughout this study, we have investigated two separate yeast protein interaction networks: that based on associations (the A network; mostly TAP-type data), and that based on physical associations (the P network; mostly yeast-two-hybrid data). We find that the two networks have similar properties with respect to their community structure, despite their very different global topological properties. Rather than regarding the yeast-two-hybrid data as of an inferior quality [303], we start from the basis that it is of a fundamentally different type and should thus be treated separately. We find similar percentages of functionally homogeneous communities in both networks.

In conclusion, we have linked the community structure of two yeast protein interaction networks with biological function by probing different scales of network structure. The identified communities are candidates for biological modules within the cell. We have illustrated how this connection can be used to select groups of proteins that likely participate in similar biological functions. Through tracing the community membership of some example proteins and investigating protein functional classes, we have highlighted that there are different scales of interest in the community structure of PINs, and that the scale (or scales) of primary interest depend on which proteins or processes one is interested in.

Chapter 4

What evidence is there for the homology of protein-protein interactions?

4.1 Introduction

In Section 1.4.5.1 interologs were introduced. They are pairs of interacting proteins: A interacting with B in one species and A' interacting with B' in another, where A' is an ortholog of A and B' is an ortholog of B (see Figure 1.3). In this chapter we ask, how much evidence is there for interologs? We do so by investigating the evidence for the homology of binary protein-protein interactions using data from six species: baker's yeast *S. cerevisiae* (SC), nematode worm *C. elegans* (CE), fruitfly *D. melanogaster* (DM), human *H. sapiens* (HS), fission yeast *S. pombe* (SP) and mouse *M. musculus* (MM). The first four species we investigate because there exists considerable data for them, the last two because these species are evolutionarily close to *S. cerevisiae* and *H. sapiens* respectively, and thus represent an interesting point of comparison.

We first calculate observed conservation rates for interactions across species and discuss the effects of potential bias in the interaction data.

We then attempt to address the sources of error that could cause the observed conversation rates to be underestimates. We decouple the effects of interaction completeness from the conservation of interactions through evolution and thereby arrive at estimates for both. Using the assumptions of our model and definitions of homology frequently employed for transferring functional annotations, we show that the fraction of interactions that are conserved is low even when interactome errors are taken into account. If strict definitions of homology are employed, the number of conserved interactions across species is low. We emphasise that our estimates of the fraction of conserved interactions do not consider the biases in the interaction data and are hence probably overestimates.

Using our estimates for the fraction of interactions conserved across species we produce estimates for the rate at which interactions are lost through evolution – the first, to our knowledge, based on large-scale data sets and comparing species that are well separated on the tree of life – finding rates of about 0.001 per million years between the most sequence-similar proteins.

As we find that it is only at less stringent sequence similarities that significant numbers of interactions can be inferred correctly, but that at these similarities the fraction of correct inferences is low, we investigate several protein properties to see if they can select those inferences that are likely to be correct. We investigate properties that should be available in the absence of all information about the proteins save their sequences. Although we do not find any properties that suggest reasons for the low observed conservation rates, we do demonstrate that some of the properties can help select conserved interactions, particularly if used in combination with one another.

Finally, we revisit the finding reported by Mika and Rost [216] that within-species interactions are more conserved than across-species interactions. This finding was inconsistent with the established belief that orthologs tend to maintain the same function whereas paralogs tend to evolve new functions (see e.g. [316] and [229] for a recent discussion of this). By carefully controlling for factors that advantage within-species interaction prediction, we argue that within-species interactions are less reliable than across-species ones.

Functional annotations are often transferred using definitions that are not particularly strict (e.g. [89, 112, 288]). We argue that the low success of interaction transfer at comparable levels of sequence similarity cannot be explained solely by interactome errors. Unless a very stringent definition of ortholog is employed, the rate of evolutionary change of interactions is too high to allow transfer across species that are well separated on the tree of life. At such stringent definitions, the number of conserved interactions is low. The common practice of transferring interactions on the basis of homology between such distant species [41, 42, 77, 102, 137, 138, 150, 177, 185, 255, 346, 356] must be treated with caution.

4.2 Data sets and data processing

4.2.1 Protein-protein interaction data

There are two primary types of protein-protein interactions; see Section 1.2.3: (1) direct protein-protein interaction data (2) evidence that proteins participate in the same complex. These different types of interaction have a different nature; for example, they are predisposed to be identified between different protein functional classes [334]. Because the ratios of direct protein-protein interactions to within-complex interactions differ substantially by species (within-complex data is concentrated within *S. cerevisiae* [100, 101, 131, 170]), we investigate only direct protein-protein interactions.

Several publicly available databases gather interaction data from multiple sources

Table 4.1: **Protein-protein interaction data for the six species investigated**. Low-throughput interactions are those interactions that have supporting evidence in publications that report fewer than one hundred interactions. The *S. cerevisiae* network is more complete than those of the other species: a much higher fraction of *S. cerevisiae* proteins have protein-protein interaction data, and each protein is involved in more interactions. The approximate number of proteins only considers one protein isoform per gene (we report the number of unique STRING identifiers [147]).

	SC	CE	DM	HS	MM	SP
# interactions	44266	7275	20334	45695	2911	1155
Fraction of low-throughput interactions	0.15	0.15	0.08	0.61	0.75	0.90
# proteins in interactome	5782	3988	6514	9597	2101	793
Mean # of interactions for proteins in interactome	7.6558	1.8242	3.1216	4.7614	1.3855	1.4565
# proteins (approximate)	6490	19522	13520	20763	21427	4806
Mean # of interactions for all proteins	6.8206	0.3727	1.504	2.2008	0.1359	0.2403

[15, 49, 147, 159, 161, 194, 285, 306]. We assembled our interaction lists from four of the largest databases: BioGRID (`www.thebiogrid.org` [306]; downloaded in June 2010), IntAct (`www.ebi.ac.uk/intact` [159]; downloaded in June 2010), MINT (`mint.bio.uniroma2.it/mint` [49]; downloaded in June 2010), and HPRD (`hprd.org` [161]; downloaded in July 2010). We use a locus-based approach; in other words, we consider only one protein isoform per gene and achieve this by mapping all protein identifiers to the identifiers used in STRING [147].

From these databases we select only direct protein-protein interaction data, thereby excluding all indirect association data, such as from tandem affinity purification experiments. We used interactions with ‘physical association’ evidence type from the IntAct database; ‘biophysical’ or ‘protein complementation’ assay type from the MINT database; ‘reconstituted complex’, ‘PCA’, ‘Co-crystal structure’ or ‘yeast-two-hybrid’ from the BioGRID database; and all interactions from the HPRD, as it only contains binary interaction data.

We amalgamate the interaction data from these sources. Table 4.1 gives the data set sizes for the species that we investigate. This data combines results from low-throughput and high-throughput studies. We give an indication of the relative contributions of low- and high-throughput studies by calculating the fraction of interactions that are reported by a study that observed fewer than one hundred interactions. As indicated in Table 4.1, there are many more interactions per protein reported for *S. cerevisiae* than for any other species, and the interaction data for *S. pombe* and *M. musculus* are particularly sparse. Comparing the sizes of the interactomes of these data sets to the estimates of the total sizes of the interactomes surveyed in Section 1.2.5 (see Table 1.1), it is clear that the *S. cerevisiae* interactome might not be far from complete, whereas the coverage of the other interactomes is low.

4.2.2 Homology data

Detecting homologs is an unsolved problem [321], so one must adopt some operational definition. Sequence similarity lies at the heart of judging whether sequences are homologous [36], though more advanced techniques incorporate additional information such as phylogenetic-tree analysis and gene-tree/species-tree reconciliation [280, 321, 333]. A conservative operational definition has the advantage that false-positive homologs will be minimised, but the disadvantage that many true homologs will be missed. In the context of inferring functional annotations from a source species to a target species, a conservative definition of homology will lead to low numbers of predictions. We consider three different operational definitions of homology: `blastp` [5] reciprocal hits; `blastp` reciprocal best hits; and EnsemblCompara GeneTrees [333].

The most common tool used to identify potentially homologous protein sequences on large scales is `blastp` [5]. Use of this method enables one to connect the success of interolog prediction with the blast E-value, which is the most common diagnostic used to measure sequence similarity. The E-value (E_{val}) gives a measure of how often one would expect to observe a particular hit by chance when a query sequence is compared to a database of potential hit sequences.

We downloaded amino acid sequences for the proteins of the species considered from the NCBI (`ftp://ftp.ncbi.nih.gov/refseq/release`). We ran `blastp` using default parameters (except for setting the maximum number of hits retrieved to be

1000000 and the E-value cut-off to be 10^{-6}). Note that the default settings include a filter for low-complexity regions. This is important for large-scale analysis, as otherwise the hits found for proteins with large amounts of low-complexity sequence will not be comparable to those without (in smaller-scale analysis the outputs of `blastp` can be checked individually for the effects of low-complexity sequence). For each query, we selected the hit with the lowest E-value and only kept pairs that were found as 'query-hit' and as 'hit-query' ('reciprocal hits'). These homology relationships are thus many-to-many. In Table 4.2 we give the numbers of reciprocal hits found at two different similarity cut-offs, $E_{\text{val}} \leq 10^{-10}$ and the more stringent $E_{\text{val}} \leq 10^{-70}$. In Figure 4.1 we illustrate for the reciprocal-hit homologs for the example species pair *D. melanogaster* and *H. sapiens* some of the relationships between percentage sequence identity (the percentage of residues in the aligned region of the query-hit pair that are identical), alignment coverage (the minimum of the fraction of the query covered by the alignment and the fraction of the hit covered by the alignment), the product of the lengths of the proteins, and the E-value. The E-value is designed to control for the length of proteins (as, in general, it is easier for longer proteins to match other proteins). However, a small residual correlation may remain between $E_{\text{val}}(A, A')$ and the product of the lengths of A and A'. In our data this does not appear to be an issue (see Figure 4.1 F). We tested for a linear relationship for the *D. melanogaster* – *H. sapiens* matches, and found a correlation of -0.03 (Pearson's coefficient). We did not test for a non-linear relationship.

Rather than choosing a particular sequence-similarity cut-off, we investigate the success of interolog inferences at different E-value thresholds. At each E-value threshold we consider all sequence similar pairs with an E-value at or below that threshold as homologs. We also consider using minimum percentage sequence identity values as an operational definition of homology.

An alternative approach to defining homologs and then making inferences is to de-

Table 4.2: **Reciprocal-hits homology relationships at two different E-value thresholds.**

Number of homology relationships, $E_{\text{val}} \leq 10^{-10}$							
target species		SC	CE	DM	HS	MM	SP
source species	SC	9752	15427	20373	34988	31443	16327
	CE	15427	103265	47023	78067	70543	17919
	DM	20373	47023	51434	149693	134237	22749
	HS	34988	78067	149693	217629	557652	41976
	MM	31443	70543	134237	557652	495248	40304
	SP	16327	17919	22749	41976	40304	6577
Number of proteins involved in homology relationships, $E_{\text{val}} \leq 10^{-10}$							
source species	SC	2446	2062	2428	2547	2435	3561
	CE	2516	9233	5374	5441	5309	3055
	DM	3362	5658	6366	7138	6914	4007
	HS	4428	7728	9435	10229	12671	5811
	MM	4211	7320	8913	13264	10756	5616
	SP	3260	2191	2619	2837	2815	1903
Number of homology relationships, $E_{\text{val}} \leq 10^{-70}$							
source species	SC	3349	1085	1448	1961	1795	2515
	CE	1085	8669	3294	4320	3924	1252
	DM	1448	3294	3702	7546	6714	1721
	HS	1961	4320	7546	32581	77188	2553
	MM	1795	3924	6714	77188	62405	2434
	SP	2515	1252	1721	2553	2434	791
Number of proteins involved in homology relationships, $E_{\text{val}} \leq 10^{-70}$							
source species	SC	1202	473	687	741	683	1479
	CE	525	4284	1527	1585	1484	669
	DM	757	1610	2350	2904	2720	977
	HS	988	2116	3903	5882	10143	1376
	MM	912	1893	3467	10536	6196	1321
	SP	1359	586	820	922	911	763

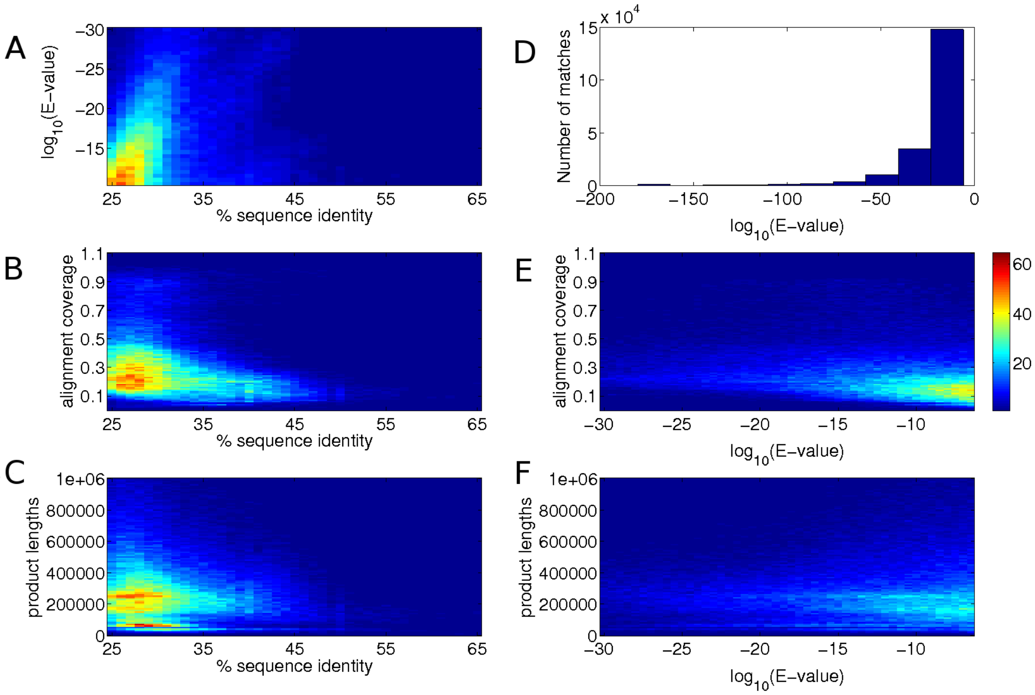


Figure 4.1: **For the reciprocal-hits matches found between *D. melanogaster* and *H. sapiens*, relationships between some sequence-similarity properties.** All subplots except (D) give two dimensional histograms for the number of matches (see colourbar) found at different values of two properties: (A) E-value and percentage sequence identity (pid), (B) alignment coverage and percentage sequence identity, (C) product of the lengths of the query and hit proteins and percentage sequence identity, (E) alignment coverage and E-value, and (F) product of query and hit lengths and E-value. Subplot (D) is a histogram of the number of matches at different E-values.

Table 4.3: **Number of reciprocal-best-hits homology relationships.** As this is a one-to-one orthology definition, the number of homology relationships and the number of proteins involved in homology relationships are the same.

target species		SC	CE	DM	HS	MM	SP
source species	SC	-	1230	1626	1749	1746	2666
	CE	1230	-	2761	2886	2875	1477
	DM	1626	2761	-	4347	4332	2037
	HS	1749	2886	4347	-	12579	2225
	MM	1746	2875	4332	12579	-	2205
	SP	2666	1477	2037	2225	2205	-

fine a joint sequence-similarity measure. This has been investigated in the literature: the results are very similar to those achieved when first defining homologs [216, 359]. For completeness, we also report results for using the product of E-values as a joint sequence-similarity measure.

In addition to studying reciprocal-hit matches, we also consider only reciprocal best hits. Two sequences are considered each others' reciprocal best hits if the first is the best hit when the second is queried against the database and the second is the best hit when the first is queried against the database. The reciprocal-best-hit criterion gives one-to-one query-hit matches. We also require that both hit-query and query-hit E-values must be 10^{-10} or lower. We give the numbers of reciprocal-best-hit matches in Table 4.3, and give a histogram of the E-values for these homologs in Figure 4.2. The reciprocal-best-hits method suffers from being dependent on the precise database used for the queries. There is also no guarantee that the closest-sequence homolog is the closest functional homolog.

We additionally consider homologs as defined by EnsemblCompara GeneTrees [333]. This method is based on the inference of multiple potential gene tree topologies; it penalises those topologies which are inconsistent with known species relationships. We report the numbers of orthologs defined by EnsemblCompara GeneTrees [333] in Table 4.4. (EnsemblCompara GeneTrees does not include *S. pombe*). We also use the manually-curated orthologs between *S. pombe* and *S. cerevisiae* that are reported in

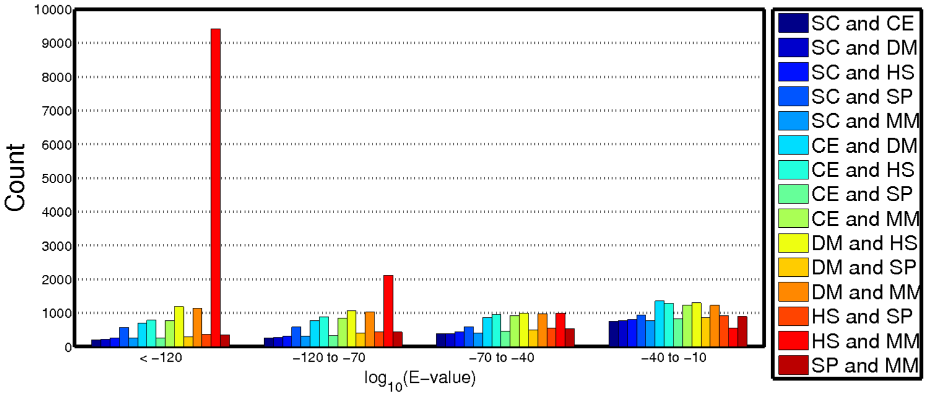


Figure 4.2: **Histogram of the E-values for reciprocal-best-hit homologs.** There are an anomalously large number of very sequence similar reciprocal-best-hit homologs between *H. sapiens* and *M. musculus*. This is because of all the species pairs considered these are by the far the most closely related (they diverged about 90 million years ago, compared to 760 million years ago for the next most recently diverged pair, *S. cerevisiae* and *S. pombe*). They also have large genomes compared to the other species (see Table 4.1).

Ref. [352]. There are 4966 homology relationships reported between 3875 *S. cerevisiae* proteins and 3657 *S. pombe* proteins.

4.3 Interactions conserved across species

4.3.1 The evidence

From an interaction $A - B$ in the source species, we infer all interactions $A' - B'$ in the target species, where A' is a sequence homolog of A and B' is a sequence homolog of B (see Figure 1.3). We consider all six species as source species but exclude *M. musculus* and *S. pombe* as target species because of the sparsity of data in these organisms. (We do, however, consider them as target species for *H. sapiens* and *S. cerevisiae*, respectively.) For the reciprocal-hits data, we investigate the effect of the E-value as an operational definition of homology (meaning that both $E_{\text{val}}(A, A')$ and $E_{\text{val}}(B, B')$ must be below a similarity threshold). Each interaction in the target

Table 4.4: **Homology relationships as defined by EnsemblCompara GeneTrees [333].**

Number of homology relationships						
target species		SC	CE	DM	HS	MM
source species	SC	-	5276	5109	6170	4943
	CE	5276	-	13334	12656	10210
	DM	5109	13334	-	12465	10196
	HS	6170	12656	12465	-	16227
	MM	4943	10210	10196	16227	-
Number of proteins involved in homology relationships						
source species	SC	-	2315	2346	2456	2363
	CE	3589	-	5623	5645	5501
	DM	3514	5945	-	6189	5903
	HS	4119	7577	7822	-	12461
	MM	3812	7006	7252	12807	-

species can conceivably be predicted more than once, but we consider only one inference to it. Hence, when we report the number of transferred interactions that are correct, we always give the number of unique interactions that are predicted correctly.

We compute the number of inferred interactions that are correct by counting how many of them are found in the interaction set of the target species (see Figure 4.3 A). The fraction of correct inferences observed, denoted $O_{s,t}$, is the number of correct inferences divided by the total number of inferences (see Figure 4.3 B). As can be seen in Figure 4.3 A, large numbers of correct inferences are only made at relatively lax E-values (to the right side of the figure). However, as would be expected and is shown in Figure 4.3 B, only a small fraction of the inferences are correct at these lax E-value cut-offs. Appendix D contains variants of Figure 4.3. Figure D.1 shows the same figure as Figure 4.3 with the axes scaled differently for each target species. Figure D.2 shows the results for using a threshold of percentage sequence identity. Figure D.3 shows the results for using the geometric mean of E-values as a joint similarity measure (see Section 1.4.5.1). It is similar to Figure 4.3, as is to be expected from previous results in this literature [216, 359].

It is important to compare the success of inferring interactions using homology

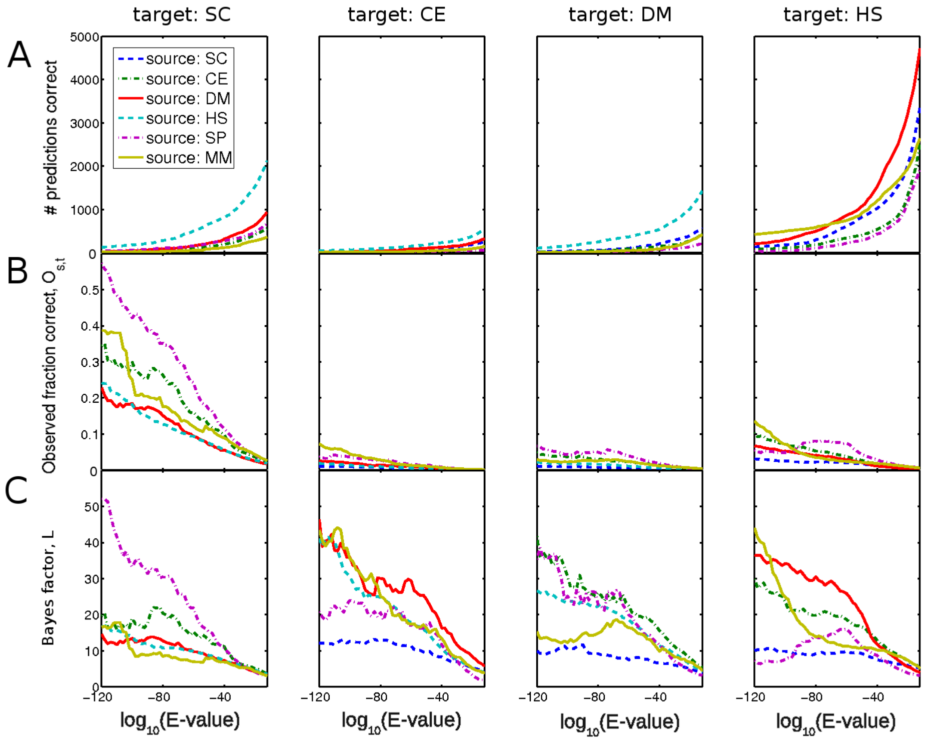


Figure 4.3: **Large numbers of correct inferences are only observed when the fraction of correct inferences is very low.** We show the results of inferring interactions from *S. cerevisiae* (SC), *C. elegans*, *D. melanogaster* (DM), *H. sapiens* (HS), *S, Pombe* (SP), and *M. musculus* (MM) to the first four of those species. (A) Number of correct interolog inferences across species, for different `blastp` E-value cut-offs. (B) Fraction of all inferences that are observed in the interactions of the target species, $O_{s,t}$. (C) The Bayes Factor L. This indicates how much better it is to use the inferences than to select random pairs of proteins in the target species that have homologs in the source species interactome. (A) and (B) together indicate that it is only at lax E-values that one makes significant numbers of correct inferences, but this is a very small fraction of the total number of inferences made at these E-values. The *S. cerevisiae* data-set coverage is significantly higher than that of other species, so one obtains larger values for inferences to *S. cerevisiae*.

relative to that achieved with random guesses – i.e. how often randomly chosen pairs of proteins will actually interact. One must define what class of inferences are ‘random’: we first consider a random inference as one between any two proteins in the target species, given that they both have homologs in the source species interactome.

Following the work of Jansen et al [146] and Yu et al [359], we consider the Bayes Factor L for an interolog inference (from interacting proteins A and B to an interaction between their homologs A' and B') to be a true prediction. Note that in Jansen et al. [146] and Yu et al. [359] the term ‘likelihood ratio’ is used instead of Bayes Factor. The Bayes Factor, which is a function of the source species and target species interaction data (int_s and int_t), relates the odds of finding a conserved interaction (a *positive*) before and after knowing the interaction data:

$$L(\mathrm{int}_s, \mathrm{int}_t) = \frac{D_{\mathrm{posterior}}}{D_{\mathrm{prior}}} \,,$$

where $D_{\mathrm{posterior}}$, which denotes the odds of finding a positive (i.e. the ratio of the probability of finding a positive to that of finding a negative) *after* we have inferred interactions, is given by

$$D_{\mathrm{posterior}} = \frac{P(\mathrm{pos}|\mathrm{int}_s, \mathrm{int}_t)}{P(\mathrm{neg}|\mathrm{int}_s, \mathrm{int}_t)} = \frac{P(\mathrm{pos}|\mathrm{int}_s, \mathrm{int}_t)}{1 - P(\mathrm{pos}|\mathrm{int}_s, \mathrm{int}_t)} \,.$$

The quantity $P(\mathrm{pos}|\mathrm{int}_s, \mathrm{int}_t)$ is the probability of finding a positive after we have considered the interaction data int_s and int_t. This quantity is estimated by the observed fraction of correct inferences $O_{s,t}$. The quantity D_{prior}, the prior odds of finding a positive in the target species given that there exist homologs of both proteins in the source species interactome, is given by

$$D_{\mathrm{prior}} = \frac{P(\mathrm{pos})}{P(\mathrm{neg})} = \frac{P(\mathrm{pos})}{1 - P(\mathrm{pos})} \,,$$

where $P(\text{pos})$ estimates the number of correct inferences among all possible inferences *before* we consider the interaction data (but assuming that we know which proteins are in the source species interactome). The number of possible inferences is equal to every pair of proteins in the target species, each of which have a homolog in the source species interactome. If there are n proteins in the target species with homologs in the source species interactome, then this is $(n^2 + n)/2$ (including self-interactions).

Predictions are more likely to be true for higher values of the Bayes Factor L. A Bayes Factor of $L = 1$ designates that prediction is no better than guessing that there is an interaction between any pair of proteins in the target species, provided both of them have homologs in the source species interactome.

Figure 4.3 C gives the Bayes Factor L, a measure of the performance of transferred interactions to be correct compared to random. The Bayes Factor indiciates a performance of only a few times better than random at lax E-values, and it is not much larger even at very strict E-values (and very few correct predictions are made at such strict E-values). The Bayes Factor is generally higher for inferences across species that diverged more recently. For example, inferences between *S. cerevisiae* and other species have a low Bayes Factor; inferences from *S. pombe* to *S. cerevisiae* have a higher Bayes Factor.

An alternative comparison to random inference is possible by rewiring the interactions in the source species while fixing the number of interactions for each protein. By keeping constant the number of times each protein appears in the interaction list, we ensure that differences we identify are due to the interactions themselves rather than to the properties of the proteins. We perform this rewiring of the source species interactions ten times for each species pair. This comparison controls for biases in protein appearance in the source species interaction list. (Such biases could either result from the data-gathering process or reflect the underlying biology.) We give the ratio of the number of correct inferences from the actual source species interactions

Table 4.5: **Across species: How does inferring interactions from the source species interactome compare to inferring interactions from randomised versions of the source species interactome?** We give the ratio of the fraction of correct inferences $O_{s,t}$ from the real interaction data compared to randomly rewired data for the reciprocal-hits homologs. (The number of interactions in which each protein participates is preserved in the randomization.) The numbers in parentheses give the standard deviations over 10 rewirings.

$E_{\text{val}} \leq 10^{-10}$					
target species		SC	CE	DM	HS
source species	SC	-	2.3 (0.17)	2.1 (0.091)	1.9 (0.092)
	CE	2.3 (0.18)	-	2.2 (0.16)	1.6 (0.13)
	DM	2.3 (0.10)	2.1 (0.076)	-	1.9 (0.047)
	HS	2.4 (0.068)	2.1 (0.047)	2.0 (0.072)	-
	MM	2.3 (0.25)	1.8 (0.18)	1.7 (0.44)	2.0 (0.37)
	SP	2.5 (0.21)	1.7 (0.22)	1.7 (0.19)	1.5 (0.092)
$E_{\text{val}} \leq 10^{-70}$					
source species	SC	-	8.9 (1.4)	4.3 (1.6)	5.8 (0.90)
	CE	9.1 (3.4)	-	18 (19)	13 (9.5)
	DM	9.9 (4.6)	16 (11)	-	9.4 (3.0)
	HS	5.0 (0.73)	7.3 (2.0)	6.2 (1.1)	-
	MM	6.4 (5.8)	11 (6.4)	11 (3.4)	6.5 (2.2)
	SP	26 (32)	12 (5.9)	15 (12)	8.0 (1.7)

to the mean of several random sets of interactions for each species pair in Table 4.5. A comparison between Figure 4.3 C and Table 4.5 illustrates that considering the different propensities for proteins to appear in the source species accounts for some of the success of transferring interactions on the basis of homology.

Although there are no standard E-value thresholds that are used to define homology, we draw attention to two thresholds that often appear in the literature. A threshold of 10^{-10} is considered a fairly strict criterion for sequence similarity (it is used by the functional annotation tool `Blast2GO` for their ‘strict’ annotation style [112]) and has been used in this literature [204, 214]. At this threshold, although hundreds or thousands of interolog inferences are correct, the fraction of correct inferences is three percent or less (see Figure 4.3 A and B). This small fraction is a result of the very large total numbers of predictions (between tens of thousands and two million, depending on species pair). An E-value threshold of 10^{-70} is considered strict, and has also been used in the literature [214, 359]. At this E-value cut-off, there are a few hundred correct inferences at most (depending on species pair) and at most 30% correct inferences.

Interactions in the target species can be inferred more than once. We give histograms of the number of inferences to each interaction at E-value thresholds of 10^{-10} and 10^{-70} in Figure 4.4.

We show the results for the EnsemblCompara GeneTrees homologs in Table 4.6 and those for reciprocal-best-hit homologs in Table 4.7. The number of correct predictions from *S. cerevisiae* to *S. pombe* using the manually curated set of orthologs is 373, the fraction correct is 0.0091 and the Bayes Factor is 70.7. The corresponding numbers for *S. pombe* as source and *S. cerevisiae* as target species are 387, 0.3446, and 49.6. The results for all these homology definitions are similar to those for the reciprocal-hits data at stringent E-value thresholds. The reciprocal-best-hits definition of homology is the most strict of the definitions considered, which is reflected in the smaller number of correct inferences and the larger fraction of correct inferences.

The fraction of correct inferences clearly depends on the coverage of the target species interactome – note the much higher fraction of correct inferences to *S. cerevisiae* in Figure 4.3 B and in Tables 4.6 and 4.7. This is expected, and below we investigate how the fraction of correct inferences is altered when we take the coverage of the target species interaction data set into account.

Inferences with *M. musculus* and *S. pombe* as source species achieve higher fractions of correct inferences than the inferences from other species. We hypothesise that this is due to biases in the interactomes that are particularly evident for these species. A very large proportion of interactions in the *S. pombe* and *M. musculus* data sets come from low-throughput studies (see Table 4.1). There is a high correlation between the number of publications in which a protein is mentioned and the number

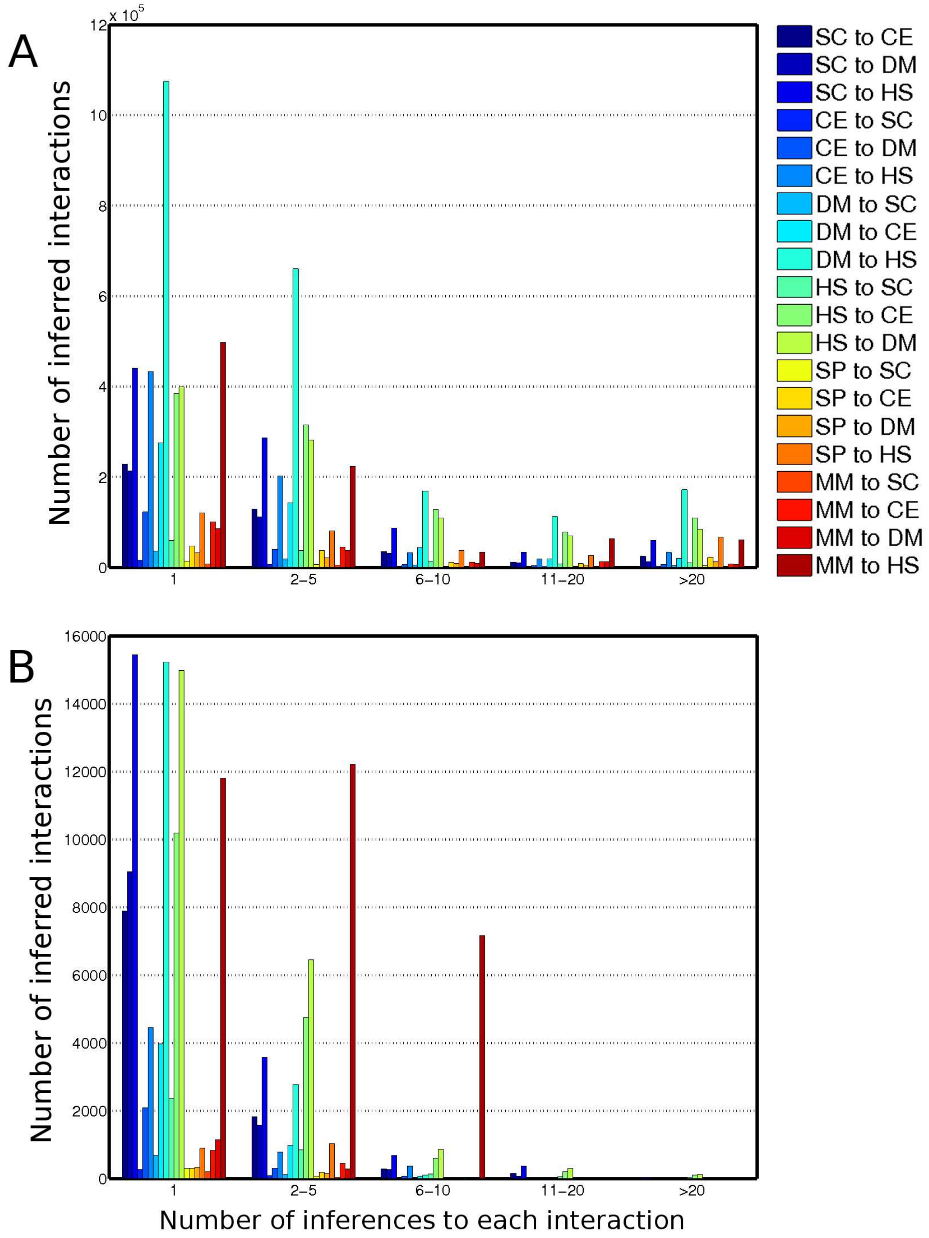


Figure 4.4: **The number of inferences made to each inferred target species interaction at (A) $E_{\textbf{val}} \leq 10^{-10}$ and (B) $E_{\textbf{val}} \leq 10^{-70}$.**

Table 4.6: **Across species inferences using the EnsemblCompara GeneTrees data.** These results show the same quantities as for Figure 4.3 and Table 4.5.

Number of correct inferences					
target species		SC	CE	DM	HS
source species	SC	-	197	349	1601
	CE	203	-	146	421
	DM	338	137	-	841
	HS	1197	265	528	-
	MM	112	55	89	532
Fraction of correct inferences $O_{s,t}$					
source species	SC	-	0.004	0.008	0.025
	CE	0.166	-	0.031	0.047
	DM	0.101	0.013	-	0.042
	HS	0.153	0.013	0.025	-
	MM	0.280	0.042	0.060	0.283
Bayes Factor L					
source species	SC	-	28.9	18.8	31.6
	CE	25.8	-	43.8	42.3
	DM	19.5	66.8	-	53.2
	HS	30.7	55.3	50.7	-
	MM	33.0	62.9	48.2	114
Comparison to rewired source species interactions					
source species	SC	-	19 (11)	13 (2.3)	15 (1.5)
	CE	12 (2.4)	-	27 (10)	24 (16)
	DM	11 (2.4)	32 (18)	-	18 (1.6)
	HS	12 (1.4)	14 (3)	13 (0.93)	-
	MM	18 (7.9)	31 (21)	20 (11)	36 (11)

Table 4.7: **Across species inferences using the reciprocal-best-hits data.** These results show the same quantities as for Figure 4.3 and Table 4.5.

Number of correct inferences					
target species		SC	CE	DM	HS
source species	SC	-	106	166	668
	CE	106	-	106	166
	DM	166	106	-	290
	HS	668	166	290	-
	MM	59	26	37	606
	SP	222	20	26	133
Fraction of correct inferences $O_{s,t}$					
source species	SC	-	0.014	0.020	0.073
	CE	0.275	-	0.076	0.103
	DM	0.214	0.033	-	0.066
	HS	0.335	0.031	0.044	-
	MM	0.488	0.080	0.091	0.281
	SP	0.440	0.062	0.071	0.299
Bayes Factor L					
source species	SC	-	51.3	39.8	69.0
	CE	41.6	-	78.8	66.5
	DM	42.4	113	-	69.6
	HS	76.1	107.6	77.5	-
	MM	55.9	72.7	67.2	107
	SP	62.0	50.8	45.3	73.5
Comparison to rewired source species interactions					
source species	SC	-	38 (29)	23 (6.0)	26 (8.8)
	CE	18 (6.4)	-	26 (10)	25 (11)
	DM	26 (7.8)	55 (37)	-	28 (9.5)
	HS	20 (3.4)	24 (8.4)	23 (6.7)	-
	MM	16 (6.8)	21 (0.36)	30 (0.62)	34 (16)
	SP	21 (4.7)	11 (6.6)	18 (6.9)	41 (33)

of interactions reported for that protein in literature-curated data ($R^2 \approx 0.59$) [287]. This reflects the fact that low-throughput experiments are hypothesis-driven – i.e. particular interactions are tested for if they are of interest to researchers. If hypotheses are formulated in part on what is known about homologous proteins, then one should expect a bias in which homologous interactions are more likely to be reported. This would lead to conservation rates appearing inflated compared to data sampled independently in different species.

In Figure 4.5 and Tables 4.8 and 4.9, we demonstrate that, in the target species, homologs of the source species are considerably more likely to interact than a randomly chosen pair of proteins. This is particularly true for *S. pombe* and *M. musculus*. This suggests that – especially for these two species – interactions are more likely to be reported if there is a homologous interaction in another species. Evidence for the homology of protein-protein interactions will be inflated because of this effect: observed conservation rates depend both on the evolutionary conservation of interactions and on the tendency for researchers to be more likely to look for homologous interactions. Assessing the relative contributions of these two effects is hard, as they manifest in the same way (i.e. in higher observed conservation rates of interactions). Note that the Bayes Factors for inferences from *S. pombe* and *M. musculus* (Figure 4.3 C) are not large compared to the other species, as is the case with the observed fraction of correct inferences (Figure 4.3 B). This is because the Bayes Factor controls for some of this bias by comparing transferred interactions to random guesses between proteins that have homologs in the source species interactome.

4.3.2 Errors in the interactome data

The bias in data-gathering discussed above leads to an overestimate in the fraction of interactions conserved, though errors in the interactome data could lead to the observed rates being underestimates. In particular, one expects the coverage of the

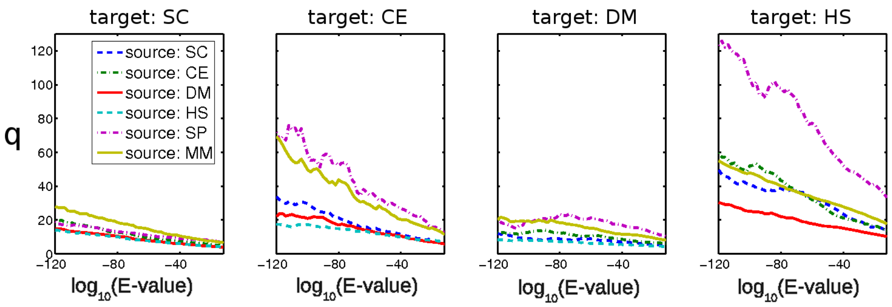


Figure 4.5: **Proteins in the target species that have homologs in the source species interactome are q times more likely to interact than a pair chosen uniformly at random from the target species interactome.** This ratio is the quantity $P(\text{pos})$ divided by the density of interactions in the target species interactome. This indicates a bias such that proteins that have been investigated for protein-protein interactions in one species are not independent of those that have been investigated in another. This is particularly true for *S. pombe* (SP) and *M. musculus* (MM).

Table 4.8: **As for Figure 4.5, but for the EnsemblCompara GeneTrees data.** The density of interactions between proteins in the target species that have homologs in the source species divided by the density of interactions in the target species interactome.

target species		SC	CE	DM	HS
source species	SC	-	3.53	1.60	6.13
	CE	7.06	-	3.02	8.78
	DM	5.31	5.17	-	6.21
	HS	5.39	6.26	2.04	-
	MM	10.8	17.9	5.30	26.2

Table 4.9: **As for Figure 4.5, but for the reciprocal-best-hits data.** The density of interactions between proteins in the target species that have homologs in the source species divided by the density of interactions in the target species interactome.

target species		SC	CE	DM	HS
source species	SC	-	6.86	2.09	8.65
	CE	8.34	-	4.18	13.2
	DM	5.90	7.75	-	7.70
	HS	6.07	7.65	2.40	-
	MM	15.5	30.4	6.01	27.7
	SP	11.5	33.1	6.82	43.8

target species interactome to influence strongly the observed fraction of correct inferences. Previous studies left such effects of interactome incompleteness as possible explanations for the poor performance of interaction transfer on the basis of homology [169, 204, 216, 359]. Here we investigate the magnitude of such effects by considering several possible sources of error.

4.3.2.1 False positives

The effect of false positives in the *source* species leads to an *under*estimation of the fraction of interactions that are conserved, as predictions from false-positive interactions are less likely to be correct. As a simple check of the magnitude of this effect, we simulated for the three species with the largest interactomes false-positive rates in the source species in excess of 50% and found that the observed fractions of correct inferences are not affected greatly by estimated false-positive rates (see Figure 4.6).

The effect of false positives in the *target* species is the opposite of that in the source species: the fraction of interactions conserved will be *over*estimated, as some predictions will be judged to be correct by matching to a false-positive interaction in the target species.

We now show under reasonable assumptions that this overestimation is larger than the underestimation (produced as discussed above by false positives in the source species), provided that $\mathrm{FPR}_s < \mathrm{FPR}_t/(1 - \mathrm{FPR}_t)$, where FPR_s and FPR_t are the false-positive rates in the source and target species, respectively.

One can estimate the magnitude of underestimation from false positives in the source species by assuming that false positives and true positives contribute in a linear fashion to the aggregate fraction of correct inferences:

$$O_{s,t}(\mathrm{data}) = \mathrm{FPR}_s \times O_{s,t}(\mathrm{FP}_s) + (1 - \mathrm{FPR}_s) \times O_{s,t}(\mathrm{TP}_s)\,,$$

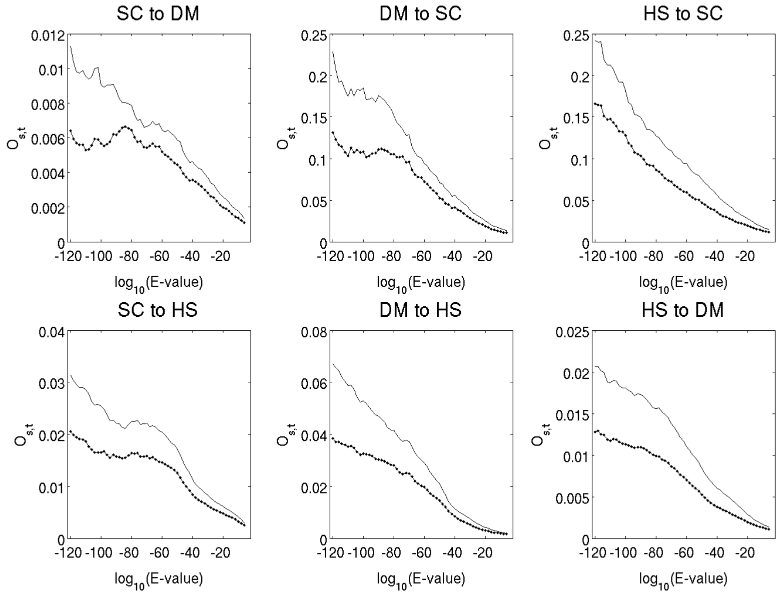


Figure 4.6: **Even rewiring half of the source species interactions does not have a large influence on the observed fraction of correct inferences $O_{s,t}$.** To simulate the effect of false positives in the source species interactions, we randomly rewire half of them (see Materials and Methods). We show results for the actual data (solid curve) and the mean of 10 sets of rewired data (joined-up-dotted curve). The rewiring process simulates a false-positive rate of $(50 + h/2)\%$, where h is the false-positive rate in the data. One can compare the observed fraction of correct inferences for the actual and rewired data to obtain a rough indication of how much the fraction deemed to be correct would differ if the false-positive rate were 0%. We found across the full range of E_{val} thresholds that rewiring half of the data had little impact on the fraction of inferences that were correct. Note that, as discussed in the main text, although false positives in the source species lead to an underestimation of the fraction of correct inferences, false positives in the target species lead to *over*estimation of the fraction of correct inferences.

where FPR_s is the false-positive rate in the source species; and $O_{s,t}(\mathrm{data})$, $O_{s,t}(\mathrm{FP}_s)$, and $O_{s,t}(\mathrm{TP}_s)$ are, respectively, the fraction of correct inferences observed for the data, the fraction that would be observed with 100% false-positive source species interactions, and the fraction that would be observed with 100% true-positive source species interactions. The largest possible underestimation arises with $O_{s,t}(\mathrm{FP}_s) = 0$. The largest underestimation is thus

$$\frac{|O_{s,t}(\mathrm{TP}_s) - O_{s,t}(\mathrm{data})|}{O_{s,t}(\mathrm{TP}_s)} = 1 - (1 - \mathrm{FPR}_s) = \mathrm{FPR}_s\,.$$

Assuming that whether or not an interaction is a false positive and whether or not it is predicted as an inferred interaction are independent assumptions, it follows that the fraction of inferences that are falsely considered to be correct is simply the false-positive rate of the target species interactions:

$$O_{s,t}(\mathrm{TP}_t) = \mathrm{TPR}_t \times O_{s,t}(\mathrm{data}) = (1 - \mathrm{FPR}_t) \times O_{s,t}(\mathrm{data})\,,$$

where $O_{s,t}(\mathrm{TP}_t)$ is the fraction of correct inferences that would be observed if all of the target species data were true positives, and TPR_t and FPR_t are the true- and false-positive rates in the target species. The overestimation caused by false positives in the target species is thus

$$\frac{|O_{s,t}(\mathrm{TP}_t) - O_{s,t}(\mathrm{data})|}{O_{s,t}(\mathrm{TP}_t)} = \frac{|(1 - \mathrm{FPR}_t) - 1|}{1 - \mathrm{FPR}_t} = \frac{\mathrm{FPR}_t}{1 - \mathrm{FPR}_t}\,.$$

Under these assumptions, and provided that $\mathrm{FPR}_s < \mathrm{FPR}_t/(1 - \mathrm{FPR}_t)$, the underestimation caused by false positives in the source species is always less than the overestimation caused by the target species. False-positive rates in the different species interaction sets are unlikely to be so different that this inequality fails to hold, so we do not further consider the possibility that false positives can lead to an underestimation of the conservation of interactions.

4.3.2.2 Coverage of the source species interactions

We hypothesize that the fraction of inferred interactions observed to be correct $O_{s,t}$ is independent of the *coverage* (which is defined as one minus the fraction of false negatives) of the source species interactions. The reason is as follows: although more correct inferences are observed with more interactions in the source species, more incorrect inferences are also made. We tested whether such independence held by sampling the source species interactions. We sub-sample from the interaction lists by randomly selecting 25%, 50%, and 75% of the interactions. At each of these values, we make ten random samplings. The results support our hypothesis; see Figure 4.7.

4.3.2.3 Coverage of the target species interactions

We hypothesised that the fraction of inferred interactions observed to be correct $O_{s,t}$ is directly (i.e. linearly) proportional to the coverage of the target species interactions c_t. For example, if the interaction list of the target species is halved in size, then the fraction of correct inferences should also halve. We tested this hypothesis by sampling from the interaction list of the target species (in the same way as for the source-species interactome) and report the mean coefficients of correlation R^2 between $O_{s,t}$ and c_t: it is 0.98 for the reciprocal-hits definition, 0.99 for the EnsemblCompara GeneTrees homologs, and 0.98 for the reciprocal-best-hits homologs. We give the full set of R^2 values in Tables 4.10, 4.11 and 4.12. All associated p-values are less than 0.05.

The independence of the observed fraction of correct inferences on the source-species interaction coverage and the linear dependence on the target-species interaction coverage help motivate the following simple model for the estimated true rate of conserved interactions:

$$O_{s,t} = E_{s,t} c_t \,, \tag{4.1}$$

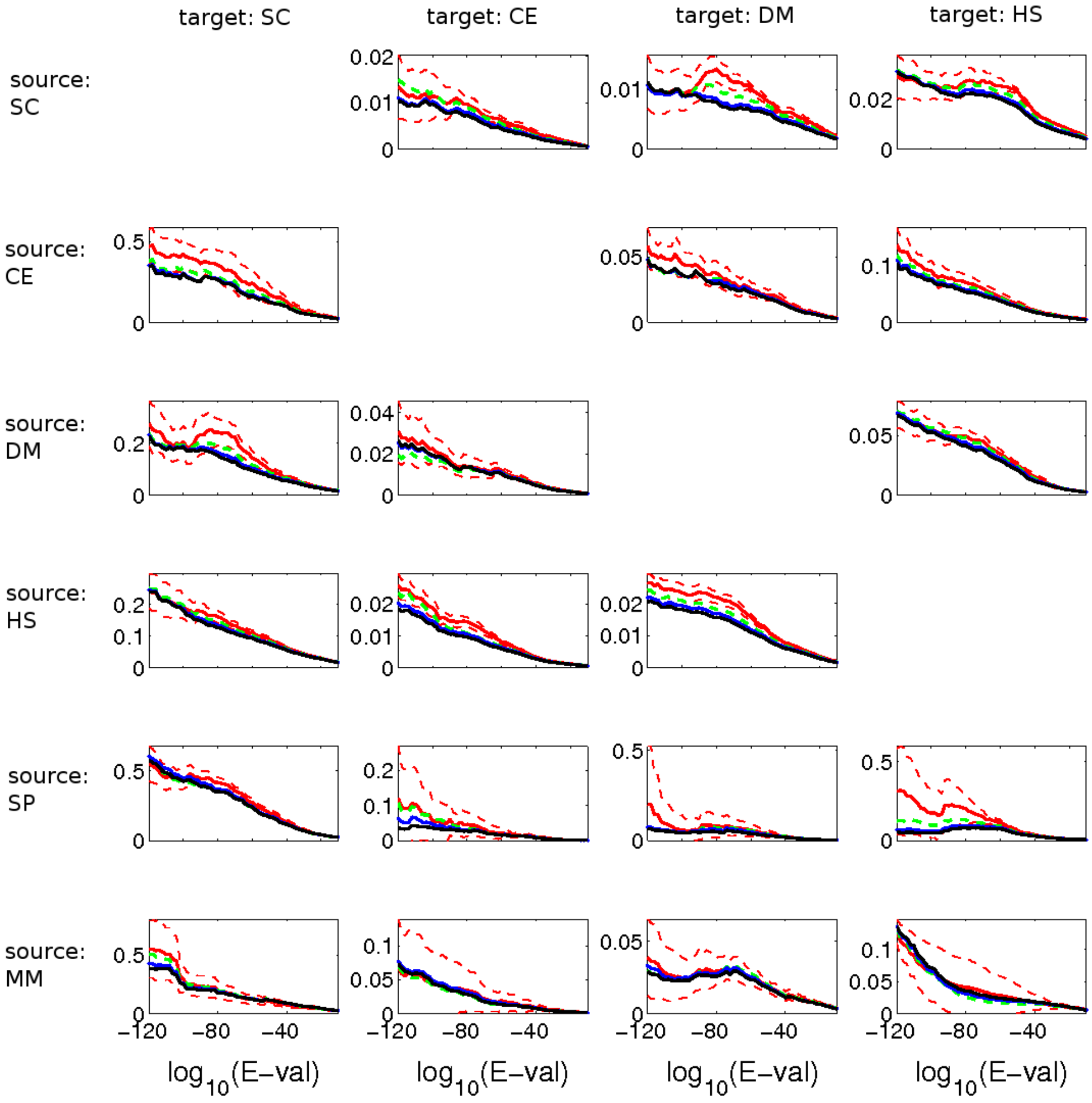


Figure 4.7: **The observed fractions of correct interologs $O_{s,t}$ are largely independent of interaction coverage in the source species.** We sub-sample from the source-species interactomes, and show mean values of $O_{s,t}$ for the actual data (black curve) and when using only 75% (blue dash-dotted curve), 50% (green dashed curve), and 25% (red curve) of the source species interactions. Also shown are the mean ± one standard deviation for the 25% case (dashed red curves). In fact, the values of $O_{s,t}$ actually seem, if anything, to be lower when more interactions are used. Hence, low coverage of the interactions in the source species does not lead to an underestimation of the fraction of correct interologs.

Table 4.10: **Results of tests carried out to examine the hypothesis that the observed fraction of correct inferences $O_{s,t}$ is directly proportional to the coverage of the target species interactions c_t for the reciprocal-hits data.** As we did for the source-species interactome, we sub-sampled from the target-species interactome 10 times by selecting a fraction f of the target species interactions. We investigated $f = 0.25$, $f = 0.5$, and $f = 0.75$. For each of the 10 experiments, we calculated the coefficient of correlation R^2 between $O_{s,t}$ and c_t at these three values of f and also for $f = 1$ (i.e. the complete data set). Here we report the means and standard deviations of the results of the 10 experiments. All the results have an associated p-value of less than 0.05 across all E-value thresholds tested. We show the results at two different E-value thresholds: 10^{-10} and 10^{-70}.

$E_{\text{val}} \le 10^{-10}$					
target species		SC	CE	DM	HS
source species	SC	-	0.997 (0.0020)	0.9980 (0.0020)	0.9998 (0.0002)
	CE	0.9976 (0.0030)	-	0.9980 (0.0029)	0.9996 (0.0003)
	DM	0.9989 (0.0012)	0.997 (0.0032)	-	0.9998 (0.0001)
	HS	0.9996 (0.0006)	0.9981 (0.0024)	0.9993 (0.0009)	-
	MM	0.9982 (0.0016)	0.9898 (0.0104)	0.9971 (0.0021)	0.9995 (0.0004)
	SP	0.9987 (0.0009)	0.9814 (0.0149)	0.9959 (0.0034)	0.9993 (0.0007)
$E_{\text{val}} \le 10^{-70}$					
source species	SC	-	0.9865 (0.0098)	0.9757 (0.0227)	0.9963 (0.0039)
	CE	0.9864 (0.0190)	-	0.9845 (0.0143)	0.9966 (0.0042)
	DM	0.9879 (0.0155)	0.9753 (0.0268)	-	0.9986 (0.0015)
	HS	0.9971 (0.0025)	0.9929 (0.0071)	0.9953 (0.0032)	-
	MM	0.9819 (0.0146)	0.9397 (0.0848)	0.9687 (0.0339)	0.9982 (0.0012)
	SP	0.9900 (0.0083)	0.9265 (0.0716)	0.9627 (0.0283)	0.9930 (0.0066)

Table 4.11: **As for Table 4.10, but for the EnsemblCompara GeneTrees data.** The means and standard deviations of the coefficient of correlation R^2 between $O_{s,t}$ and c_t. All the results have an associated p-value of less than 0.05.

target species		SC	CE	DM	HS
source species	SC	-	0.9928 (0.0067)	0.9973 (0.0019)	0.9993 (0.0006)
	CE	0.9939 (0.0048)	-	0.9918 (0.0088)	0.9973 (0.0022)
	DM	0.9966 (0.0034)	0.9896 (0.0115)	-	0.9985 (0.0012)
	HS	0.9990 (0.0014)	0.9951 (0.0037)	0.9978 (0.0026)	-
	MM	0.9867 (0.0099)	0.9804 (0.0221)	0.9831 (0.0173)	0.9985 (0.0018)

Table 4.12: **As for Table 4.10, but for the reciprocal-best-hits data.** The means and standard deviations of the coefficient of correlation R^2 between $O_{s,t}$ and c_t. All the results have an associated p-value of less than 0.05.

target species		SC	CE	DM	HS
source species	SC	-	0.9919 (0.0078)	0.9926 (0.0108)	0.9983 (0.0012)
	CE	0.9871 (0.0097)	-	0.9887 (0.0131)	0.9938 (0.0056)
	DM	0.9942 (0.0063)	0.9901 (0.0080)	-	0.9968 (0.0026)
	HS	0.9977 (0.0015)	0.9934 (0.0035)	0.9959 (0.0033)	-
	MM	0.9838 (0.0157)	0.9339 (0.0662)	0.9611 (0.0338)	0.9982 (0.0014)
	SP	0.9941 (0.0085)	0.9401 (0.0613)	0.9592 (0.491)	0.9916 (0.0095)

where $O_{s,t}$ is the fraction of inferred interactions observed to be correct, $E_{s,t}$ is the fraction of inferred interactions estimated to be correct (taking into account incomplete interactome coverage), and c_t is the coverage of the target species interactome. We emphasise that this simple model does not take into account the bias in data-gathering processes discussed above. It thus gives estimates expected with biased data; as discussed above, these will be overestimates compared to estimates on data gathered at random. Due to the particularly strong bias associated with the two smallest interactomes (*S. pombe* and *M. musculus*), we estimate $E_{s,t}$ values for these species only with their most closely related species (see below). Focusing just on the four species for which there is the most interaction data, there are twelve equations (one for each pair of species, where order matters) of the form (4.1) for each definition of homology. As there are more unknowns than equations – only the $O_{s,t}$ are known – one cannot solve (4.1) without either making some assumptions or incorporating independent estimates for values of $E_{s,t}$ or c_t. We pursue the former strategy and discuss the latter one.

We make two assumptions to calculate values of c_t, which we then use to solve for values of $E_{s,t}$. First, we assume that the *S. cerevisiae* interactome is complete (which is consistent with the literature; see Table 1.1). Altering this assumption changes all our results by a constant multiple. Second, we assume that the fraction of conserved interactions between a source species x and *S. cerevisiae* is the same as from *S. cerevisiae* to species x; i.e. $E_{SC,x} = E_{x,SC}$. This implies that $c_x = O_{SC,x}/O_{x,SC}$. Making these assumptions allows one to decouple the $E_{s,t}$ values from the c_t values and hence to obtain estimates for both.

We give the estimated values of c_t and the implied total interactome sizes in Table 4.13. These values lie within previous estimates (see Table 1.1). Our estimates of total interactome size, like all others, make a series of assumptions and should therefore be taken as complementary to existing estimates. We estimate the size of the *C.*

Table 4.13: **Estimated interactome coverages and interactome sizes.** We report the means and standard deviations for the reciprocal-hits data over all the E-value thresholds that we investigate. These results assume that the *S. cerevisiae* interactome is complete at 44266 interactions.

	reciprocal hits		EnsemblCompara GeneTrees		reciprocal best hits	
	coverage	interactome size	coverage	interactome size	coverage	interactome size
CE	0.0293 (0.0027)	256000 (24000)	0.024	310531	0.050	150742
DM	0.0707 (0.0214)	349000 (96000)	0.074	308787	0.095	240160
HS	0.1874 (0.0372)	158000 (35000)	0.162	174858	0.217	130204

elegans and *D. melanogaster* interactomes to be larger than that of *H. sapiens.* This is surprising, as the numbers of proteins in the former two organisms are smaller (see Table 4.1). Homologs of *S. cerevisiae* proteins are considerably more likely than random to interact in *H. sapiens* (see Figure 4.5), which is probably due to the high proportion of interactions in *H. sapiens* that come from low-throughput studies (see Table 4.1). This would cause $O_{SC,HS}$ estimates to be higher than expected, and hence, via the equation $c_{HS} = O_{SC,HS}/O_{HS,SC}$, this would cause the c_{HS} estimates to be higher than one might expect. The same effect occurs for *C. elegans*, though to a lesser extent (see Figure 4.5).

We show estimated fractions of interactions conserved in Figure 4.8 and Tables 4.14 and 4.15. As one should expect the estimated fraction of correct inferences is lower between *S. cerevisiae* and the other three species. The estimates are highest for the most stringent definition of homology (reciprocal best hits; see Table 4.15). The extent to which strictness in definition of orthology is important for the transferability of interactions is evident from Figure 4.8: using reciprocal hits at E-values of 10^{-10} and below gives success rates of a few percent, even when interactome incompleteness is taken into account.

One could also solve the set of equations (4.1) by using independent estimates of the coverage of the interactomes c_t. Larger estimates of c_t than ours would give smaller estimates of $E_{s,t}$. The estimated fraction of conserved interactions remains low unless one assumes very small coverages of the target species interactome; this

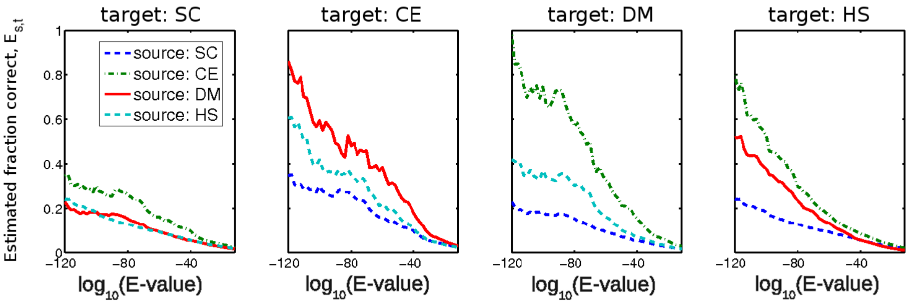


Figure 4.8: **Fraction of interactions estimated to be conserved through evolution $E_{s,t}$, which we calculate by taking interactome coverage into account.** One should expect the lower conservation rates between *S. cerevisiae* (SC) and the other species, given the known evolutionary relationships between these species. We estimate the conservation rates at E-values often associated with the transfer of functional annotations (E-values of about 10^{-10} and below) to be a few percent.

Table 4.14: **Estimated fractions of correct inferences $E_{s,t}$ using the EnsemblCompara GeneTrees data.**

Estimated fraction of correct inferences					
target species		SC	CE	DM	HS
source species	SC	-	0.166	0.101	0.153
	CE	0.166	-	0.433	0.288
	DM	0.101	0.555	-	0.257
	HS	0.153	0.556	0.341	-

Table 4.15: **Estimated fractions of correct inferences $E_{s,t}$ using the reciprocal-best-hits data.**

Estimated fraction of correct inferences					
target species		SC	CE	DM	HS
source species	SC	-	0.275	0.214	0.335
	CE	0.274	-	0.800	0.475
	DM	0.214	0.670	-	0.303
	HS	0.335	0.631	0.467	-

would imply very large total interactome sizes. For example, a 50% success rate for transferring interactions between *S. cerevisiae* and *H. sapiens* at an E-value cut-off of 10^{-70} would imply an interactome size of over half a million interactions for *S. cerevisiae* and nearly three million interactions for *H. sapiens*.

We now consider the extent of conservation between *S. cerevisiae* and *S. pombe*. Making the same assumptions as above, $E_{SC,SP} = E_{SP,SC} = O_{SP,SC}$, the curve shown in dashed-dotted pink in the left-most panel of Figure 4.3 B. We estimate $E_{SC,SP}$ and $E_{SP,SC}$ to be 0.4396 using the reciprocal-best-hits homology definition and 0.3446 for the manually-annotated ortholog data set. The estimated fractions of interactions conserved across *S. pombe* and *S. cerevisiae*, whose last common ancestor existed about 760 million years ago [223], are similar to those between *D. melanogaster*, *H. sapiens*, and *C. elegans*. *D. melanogaster* and *H. sapiens* shared a common ancestor about 830 million years ago [223], and *C. elegans* shared a common ancestor with these two about 960 million years ago [223].

Of all the species pairs one would expect the estimated fraction of correct inferences to be highest between *H. sapiens* and *M. musculus*, as these species shared a common ancestor about 90 million years ago [223]. We report estimates for $E_{HS,MM}$ and $E_{MM,HS}$ in Figure 4.9. At an E-value threshold of 10^{-10}, we estimate $E_{HS,MM}$ to be 3.5% and $E_{MM,HS}$ to be 2.1%. The estimated fraction correct rises above 1 at the most stringent reciprocal hits E-values (this should clearly not be possible!), and is well above 1 for the reciprocal-best-hits data ($E_{HS,MM} \approx 1.45$ and $E_{MM,HS} \approx 1.29$) and the EnsemblCompara GeneTrees data ($E_{HS,MM} \approx 1.75$ and $E_{MM,HS} \approx 2.70$). This could be because our estimates of the coverage of the two species interactomes are too low (which is equivalent to our estimates of the interactome sizes being too high). However, it is far more likely that the estimates of $E_{HS,MM}$ and $E_{MM,HS}$ are too high because of the aforementioned biases in the data-gathering processes. Our model assumes that interactions are sampled independently in different species; how-

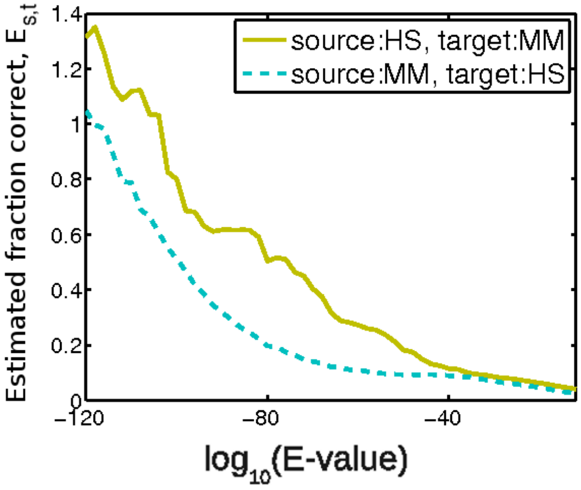


Figure 4.9: **The estimated fraction of correct inferences between *M. musculus* (MM) and *H. sapiens* (HS).** This fraction should clearly not be higher than 1. This feature is likely due to the biases in the data-gathering processes, see text.

ever, if an interaction is known in one species, then researchers might be prompted to search for it in another. This is likely to be particularly true between *H. sapiens* and *M. musculus*.

Our estimates can be compared to the results of studies that experimentally tested for the presence of interologs, which we review from Section 1.4.5.1. Matthews et al [204] tested predictions of inferring from *S. cerevisiae* to *C. elegans* using an orthology definition that was many-to-one (each *S. cerevisiae* protein was considered an ortholog of at most one *C. elegans* protein, but *C. elegans* proteins could have more than one *S. cerevisiae* ortholog). They found that between 16% and 31% of the inferences were correct (c.f. our estimates for the same species pair: 28% using reciprocal-best-hits data and 17% using the EnsemblCompara GeneTrees data). Using one-to-one ortholog matching, a conservation rate of between 34% and 64% was reported between *H. sapiens* and *M. musculus* transcription factor-transcription factor interactions [267]. A recent study comparing two yeasts, *S. cerevisiae* and *Kluyveromyces waltii*, which diverged about 150 million years ago, used one-to-one orthology relationships and found that 43 of 43 tested interactions were conserved [262].

4.3.3 Probability per million years that a duplicated interaction is lost

The results described above can be used to estimate the rate of loss of protein-protein interactions using a simple model. Assume that an interaction that existed in the last common ancestor of the source and target species has a probability p per unit time of being lost in either of the two species. For low p, the probability that we observe an interaction between A' and B' in the target species, *given that we have observed an interaction between A and B in the source species*, is approximately $(1-p)^T$, where T is the number of units of time since the species diverged. There are many ways to estimate T, and we use the mean time and range of times given in Ref. [223].

We seek to show how p varies with the extent of sequence homology. We report results for the EnsemblCompara GeneTrees data, the reciprocal-best-hits data, and the reciprocal-hits data in windows of similarity as judged by E-value. (i.e. $a < E_{\text{val}} \leq b$ for different a and b). We solve the equation $E_{s,t} = (1-p)^T$ to obtain p.

Our calculations suggest that when the divergence time of species is taken into account, the probability per million years of an interaction being lost appears to be fairly independent of species pair (see Figure 4.10; the indicated errors represent ranges in the estimates of T). At the strictest definition of homology that we consider, we find that the rate of change of protein interactions through evolution is about 10×10^{-10} interactions lost per year. One can compare this estimate to the only other estimate we could find in the literature, which gives an estimated rate of $(2.6 \pm 1.6) \times 10^{-10}$ [262]. That study, which is based on a small number of experimentally tested interactions and does not investigate the role of sequence similarity, explicitly excludes the impact of gene duplication, so one would expect a lower rate of protein

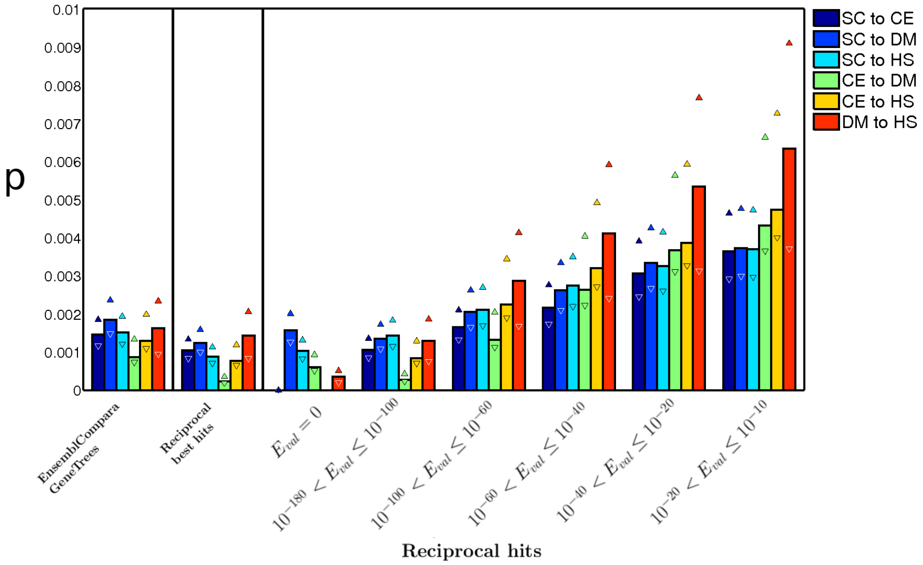


Figure 4.10: **Estimates of the probability p that a duplicated interaction is lost per million years.** If the proteins in two species remain highly similar in sequence, then the probability that both species retain the interaction is higher – i.e. one finds lower values of p at smaller E-values and using the reciprocal-best-hits and EnsemblCompara GeneTrees homology relationships. The divergence time between species is needed to calculate p; we use the estimate and range (shown in triangles) of times given in Ref. [223].

interaction change.

The step from considering the success of inferring interactions across species to inferring the rate at which interactions are lost through evolution is a large one that entails numerous assumptions and abstractions, in addition to those used to estimate values of $E_{s,t}$. First, we suppose that the abstraction to a typical duplicated interaction is a sensible one – i.e. that it makes sense to estimate the rate at which any given duplicated interaction is lost. There are various heterogeneities in protein-protein interactions that might make this questionable. For example, genes that are duplicated might lose interactions faster than genes that are not duplicated. One response is to restrict the enquiry and seek the probability that interactions between non-duplicated genes are lost [262]. Second, we have modelled the loss of interactions as independent of each other, though whether a given interaction is lost will presumably depend on its location in the protein-protein interaction network. Indeed, we present evidence in the next section that some structural network properties can be relevant to the success of interolog inference (also see [135]). Third, we have not taken into account the role of interaction gain through evolution. Fourth, we assume that the homologs we use are in fact true paralogs or orthologs. Our estimates should be considered in light of these caveats. However, given the simplicity of our model, it is encouraging that our estimates for the rate at which interactions are lost is in broad agreement with that of Qian et al [262].

In contrast to the rate of protein sequence evolution, the rate of protein function evolution remains almost unknown [262]. Protein-protein interactions provide a window through which to view this question. Although the rate at which protein-protein interactions are lost within species has been studied [25, 336], the loss rate across species has not received much attention. Consequently, our estimates should be taken as initial ones, and we believe that they are the first ones that are based on large data sets.

4.3.4 Can one select the conserved interactions?

Given the low number of interactions transferable at stringent definitions of homology and the low success rate of transfer of interactions at less stringent definitions, we were motivated to investigate whether there are any properties that can select which inferences are likely to be correct among those made at less stringent definitions of homology (i.e. the reciprocal-hits data). Studies that use transferred interactions in building predicted sets of interactions sometimes also incorporate additional protein properties [137, 150, 255, 356]. Our intention is to investigate the extent to which certain biological properties can explain the lack of interaction conservation at less stringent definitions of homology, rather than to seek an algorithm that accurately predicts protein interactions across species. For this investigation, we focus on the three species for which there exists the most data – *S. cerevisiae*, *D. melanogaster*, and *H. sapiens* – in the hope that the results for these data sets will be influenced less by noise than the smaller data sets.

We are considering the transfer of interactions between interacting proteins A and B in a source species to proteins A' and B' in a target species, where A and A' are homologs and B and B' are homologs. For any given inferred interaction in the target species, there can be multiple possible interactions in the source species from which it could have been inferred. In order to consider properties of the proteins in the source species, it is necessary to state which of these multiple possible interactions is considered to underlie a given inferred interaction $A' - B'$ in the target species. We select, as the 'closest' inference, the one that would be made using the strictest definition of homology (i.e. the one with the minimum value of $\max\{E_{\text{val}}(A, A'), E_{\text{val}}(B, B')\}$).

The first property that we investigated was the size of the family to which a protein belongs. If only one or a few interactions between proteins from one family and proteins from another family is needed for the maintenance of biological function, then one might expect that an inference from or to proteins with many homologs in the other species would be less conserved. We tested how inferences to and from proteins in large protein families affected our results by discarding all predictions in which any of proteins A, B, A', and B' had more than 10 homologs in the other species. This definition of size of family is clearly dependent on the E-value threshold, as a protein's family size becomes smaller at stricter E-values. Our intention was to get an idea of the magnitude of the effect of large families, so we chose one definition of a large protein family (i.e. those of size at least 10). We find that at lax E-values the fraction correct is improved although the number of correct inferences is vastly reduced (see Figure 4.11).

We also investigate the effects of several other properties, which roughly can be divided into three classes: properties of the four proteins A, A', B and B' (e.g. the age

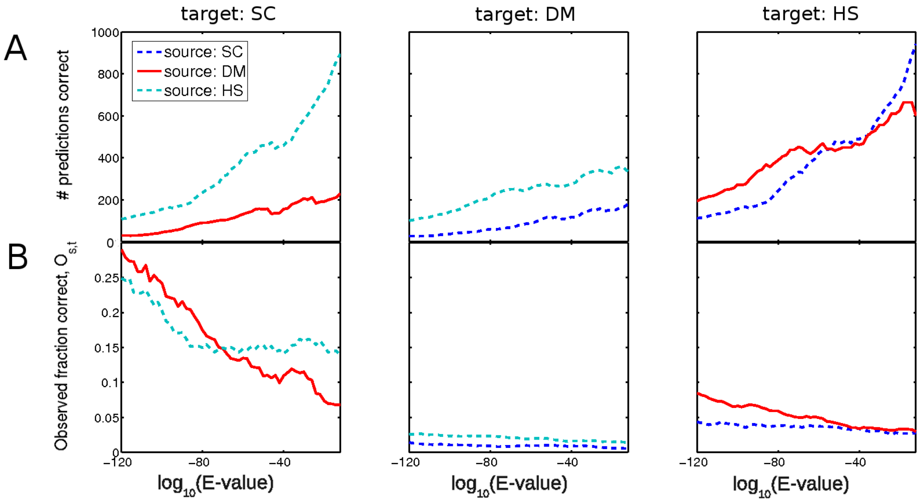


Figure 4.11: **Effects of disallowing inferences from and to large protein families.** This figure is the same as for Figure 4.3 A and B (also see Figure D.1 for the same figure with different scale axes), except that we only make inferences if each of the four proteins A, B, A', and B' has ten or fewer homologs in the other species. One could argue that the low fraction of correct inferences reported in Figure 4.3 B was due in part to allowing inferences from and to large protein families. However, comparing panel B of this figure to Figure 4.3 B illustrates that although the fraction deemed to be correct is somewhat higher at lax E-value cut-offs, this comes only at the great expense of a significant decrease in the number of correct predictions (compare panel A of this figure to Figure 4.3 A). At more strict E-values, the results are unchanged. In other words, imposing a limit on the sizes of the families has a similar effect to imposing a stricter homology cut-off.

of the proteins and the number of domains that make up the proteins); properties of how the interaction $A-B$ is embedded in the source species interaction network (e.g. how many interactions the proteins have); properties of the homology relationships between A and A' and between B and B' (e.g. the similarity of the lengths of proteins A and A').

In inferring $A'-B'$ from $A-B$, we assess the relevance of the following properties (this list is by no means exhaustive):

- The product of the number of homologs of A in the target species and the number of homologs of B in the target species (where homologs are defined as above).

- The product of the number of homologs of A' in the source species and the number of homologs of B' in the source species.

- The total number of inferences to the interaction $A' - B'$.

- The difference in the ages of A and B. We use two proxies for protein age: the lineage specificity of superfamilies (age) [349] (data kindly provided by Sanne Abeln), which is based on the estimated age of the structure of the protein; 'Excess retention' (ER) [282], which counts the number of species in which a protein has orthologs. (We use the Inparanoid database to define orthologs [238]). We were prompted to investigate the difference in the ages of interacting proteins by Refs. [165, 263].

- The difference in the ages of A' and B'.

- The sum of the ages of A and B.

- The sum of the ages of A' and B'.

- The product of the number of domains of A and the number of domains of B. We defined domains via SCOP (Structural Classification of Proteins [227]).

- The product of the number of domains of A' and the number of domains of B'.
- The geodesic edge betweenness centrality of the interaction between A and B [235]. Roughly, this centrality is given by the number of shortest paths between pairs of proteins that pass through the interaction in question.
- The number of triangles in which $A-B$ participates as a fraction of the triangles in which it could participate. This quantity, called the 'matching index' in Ref. [11], gives a measure of local clustering.
- The product of the number of interacting partners of A with the number of interacting partners of B divided by the total number of interactions.
- $\min\{E_{\text{val}}(A, A')E_{\text{val}}(B, B')\}$
- $E_{\text{val}}(A, A')E_{\text{val}}(B, B')$
- $\text{pid}(A, A') + \text{pid}(B, B')$
- $\text{pid}(A, A')\text{pid}(B, B')$
- $g(A, A') + g(B, B')$
- $\text{ac}(A, A') + \text{ac}(B, B')$
- $\text{ls}(A, A') + \text{ls}(B, B')$,

where pid is the percentage sequence-similarity, g is the number of gaps in the sequence alignment, the alignment coverage ac is the minimum of the fraction of the query covered by the alignment and the fraction of the hit covered by the alignment, and the length similarity ls is the length of the shorter sequence divided by that of the longer sequence.

To assess the utility of a property we first build a logistic regression model [134] and then assess the performance of that model using the AUC measure (see Section

2.3). The logistic regression model is used to predict the probability p of an occurrence of an event (in our case, the conservation of an interaction):

$$p = \frac{1}{1 + e^{-(\beta_0+\beta_1 x_1+\cdots+\beta_k x_k)}}, \tag{4.2}$$

where the set of β values are the *regression coefficients*. We used five-fold cross-validation and MATLAB's `glmfit` function to estimate the regression coefficients. When using the AUC measure, if the number of positive instances and negative instances is very dissimilar, samples from the larger set equal to the size of the smaller set should be used (see discussion in Section 2.3). For each property we tested we selected one hundred random sets of non-conserved interactions equal in size to the size of the set of conserved interactions. We report the means and standard deviations over these randomised samples in Figure 4.12. None of the properties achieves a high AUC.

We also assess the performance of a model built using multiple properties. For this, we selected the case of inferences from *H. sapiens* to *S. cerevisiae* as we expect the actual conserved interactions to be well approximated by the observed conserved interactions with *S. cerevisiae* as the target species (see Section 4.3.2.3) and taking *H. sapiens* as source species gives the largest number of inferred interactions. We started with a model that used only the property that achieves the highest AUC, and then added in the property that, when used in conjunction with this property, achieved the highest AUC out of all of the properties tested. We repeatedly added properties in this way until the improvement in the AUC was less than 0.001. The results are shown in Table 4.16. We also indicates whether the regression coefficient for the property is negative or positive. A negative regression coefficient indicates that selecting instances with lower values of the property leads to a higher proportion of conserved interactions. The opposite is true for positive coefficients.

Table 4.16: **Order in which properties are included in a model for predicting interactions from *H. sapiens* to *S. cerevisiae*.**

	Property	**AUC**	**Sign of regression coefficient**
1	Edge betweeness centrality	0.68	negative
2	Maximum E-value	0.7046	negative
3	Matching index	0.7278	positive
4	Sum of ages target species	0.7444	negative
5	ER difference source species	0.7542	negative
6	Age difference target species	0.7618	negative
7	Sum of alignment length	0.7687	positive
8	Product of sequence identities	0.7745	positive
9	Product of size of families target species	0.7788	negative
10	Product of number of domains target species	0.7812	positive
11	Sum of ER source species	0.7833	negative
12	ER difference target species	0.7847	negative
13	Sum of ER target species	0.7875	negative
14	Number of inferences to target interaction	0.7879	positive
15	Number of inferences from source interaction	0.7895	negative

Note that building a model using the maximum E-value and assessing its performance via the AUC measure is equivalent to varying the E-value threshold used to define homologs. This property is the second to be included in the model. The first and third properties are both measures of the local network environment of the interaction in the source species. This indicates the potential importance of network structure in the conservation of interactions, an issue we discuss in Section 5.3. Note that, after the inclusion of the maximum E-value, no further sequence-similarity properties are included. We show ROC curves for models built using the top three properties and the top ten properties in Figure 4.13 (for comparison we also show the ROC curve for using the maximum E-value).

Logistic regression models the log odds of an occurrence of an event as a linear function of the predictor variables. Ideally a linear relationship should exist between variables considered and the log odds, and transformations of the predictor variables can be considered to this end. This would be a way to improve the method described above. Additionally, combinations of variables could be considered as additional pre-

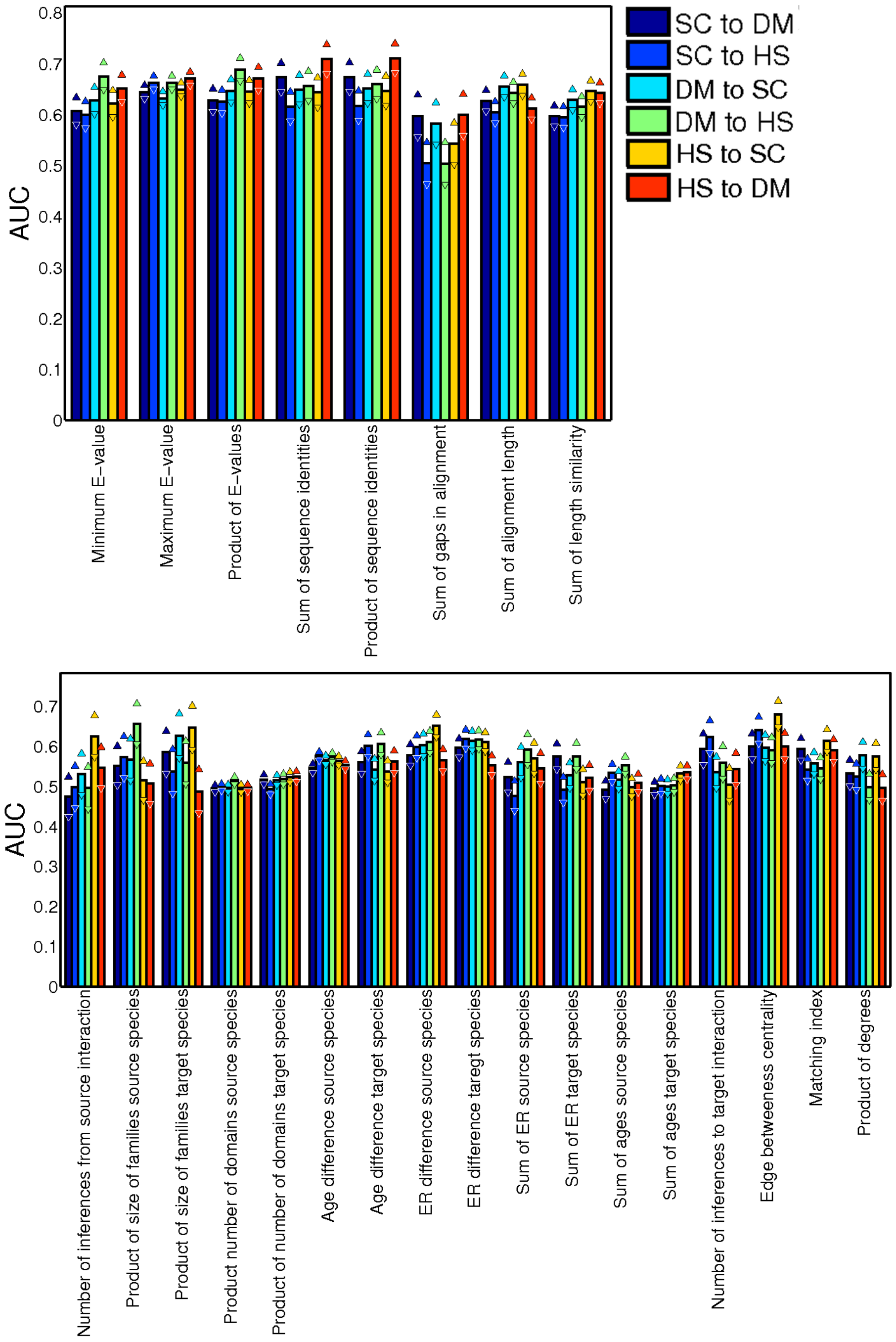


Figure 4.12: **Informativeness of properties for finding conserved interactions, as measured via the AUC.** We investigate the helpfulness of certain properties for selecting the correct inferences. In general, the sequence-similarity properties (shown in the top plot) are more helpful than the others (shown in the bottom plot).

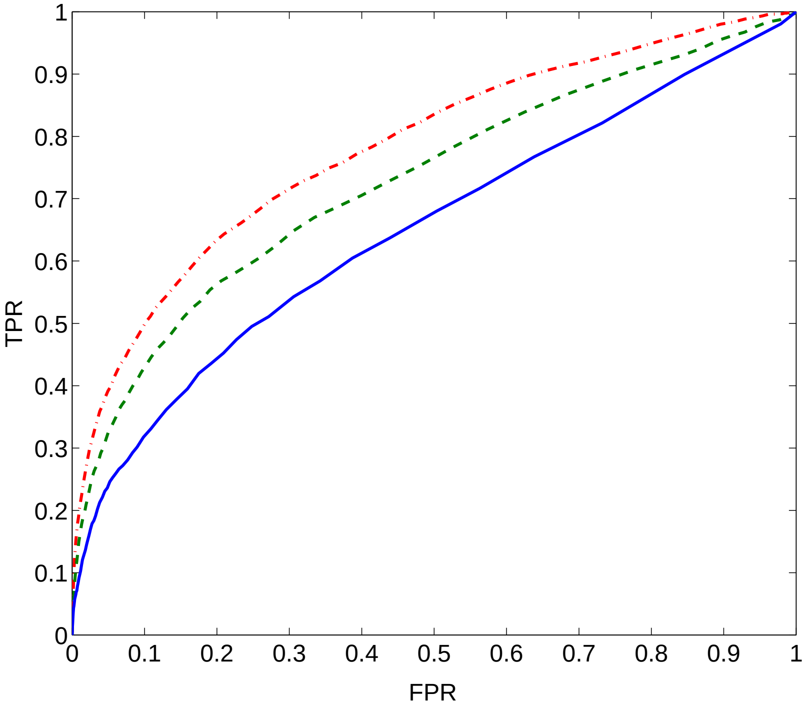


Figure 4.13: **ROC curves for predicting conserved interactions using three models of increasing complexity.** The false positive rate (FPR) versus the true positive rate (TPR) for models built using (a) the predictor maximum E-value (blue curve), (b) the first three properties given in Table 4.16 (dashed green curve), and (c) the first ten properties given in Table 4.16 (red dot-dashed curve).

dictor variables, though one must be careful not to overfit the model to the data when including additional variables. The standard form of logistic regression employed here assumes that the observations are independent of each other. The utility of some of the PIN properties in our model suggests that in this case the observations are not independent of each other. More sophisticated modelling frameworks that included some of the dependencies in both the source-species PIN and the predicted target-species PIN could offer further improvements.

4.4 Interactions conserved within species

We now examine the evidence for the homology of protein-protein interactions within a species. Our principal aim is to compare this evidence to that for across-species inferences.

Two possibilities exist when investigating the homology of interactions within a species. Interactions $A - B$ and $A' - B'$ are homologous; we refer to these as *both-different* conserved interactions. Additionally, interactions $A - B$ and $A - B'$ are homologous; we refer to such interactions as *one-same* conserved interactions. See Figure 1.3.

Evidence of conserved interactions across species comes from interologs. What is the equivalent of an interolog if one is investigating the conservation of interactions within a species? In particular, are one-same inferences to count as interologs ? If so, there are considerably more within-species interologs than across-species interologs, as demonstrated below. However, one could consider allowing correct one-same inferences to count as interologs to be unfair in making a comparison to the across-species case. One might instead argue that both-different inferences are a suitable comparison to across-species inferences.

In an influential study, Mika and Rost [216] presented evidence that interactions are more conserved within species than across species. They excluded all one-same inferences on the basis that including them made the comparison to across-species inferences unfair. In comparing the success of both-different within-species transferred interactions to across-species transferred interactions, they found that interactions were more conserved within than across species. They considered this result surprising, as it runs against what is commonly believed about the similarity in function of different types of homologs (orthologs and paralogs): for a gene duplicate (a paralog) to be retained in the genome, it quickly tends to cease functioning identically to its parent gene and hence must diverge in function; however, a gene pair that results from a speciation event (orthologs) normally maintains the ancestral protein's functions and is hence subject to higher functional conservation [193]. This general expectation that paralogs must always change function in order to be maintained has recently been questioned by Kondrashov and Koonin [168] who argued that some genes are retained simply because their protein products are needed in greater quantity in the cell (the 'gene dosage' hypothesis; see also Ref. [167]). The evidence supporting this hypothesis has been questioned by Qian and Zhang [261] who argued that the evidence instead supports the 'dosage balance' effect [17], whereby a duplication of a single member of a protein complex is deleterious. The extent of maintenance of ancestral gene function by orthologs is also under debate [104, 193, 229].

Although we agree that a direct comparison of across-species transferred interactions to one-same transferred interactions does not represent a fair comparison, we argue that Mika and Rost can be considered to have over-counted the number of both-different transferred interactions. Interactions can be inferred multiple times and some inferences can be from considerably more sequence-similar proteins. As we did in Section 4.3.4 we define the 'closest' inference as the one that would be made at the most stringent E-value cut-off. For example, consider the case in which the interaction $A - B$ is predicted both from the interaction $A - B'$ and the interaction

Table 4.17: **Within species: Ratio of correct inferences using the real data compared to randomly rewired interactions.** The one-same inferences perform better than the both-different inferences. The values in this table should be compared to those in Table 4.5. A comparison with Figure 4.14 C illustrates that the choice of how to measure the improvement over random can have large effects on the results.

	$E_{\text{val}} \leq 10^{-10}$		$E_{\text{val}} \leq 10^{-70}$	
	One-same	Both-different	One-same	Both-different
SC	4.6 (0.48)	1.5 (0.26)	7.9 (0.99)	2.4 (0.71)
CE	3.3 (0.26)	1.5 (0.25)	8.2 (0.61)	7.4 (0.54)
DM	3.5 (0.38)	1.1 (0.14)	11 (2.2)	2.4 (2.0)
HS	4.9 (0.33)	1.1 (0.14)	10 (0.95)	2.8 (0.53)

$A'' - B''$. If one of the two homology relationships $[A, A'']$ and $[B, B'']$ is more distant than $[B, B']$, then the first inference is closer than the second.

We observed in practice that for a great many interaction pairs $A - B$ and $A'' - B''$, there was a closer interaction $A - B'$. By parsimony, we treat the interaction $A - B'$ as underlying the observed conservation.

For every inferred interaction, we classify the inference to it as either one-same or both-different, depending on which type of inference would be made at the most stringent definition of homology.

We conduct an investigation similar to the across-species case for both types of within-species inference. See Figure 4.14 and Table 4.17; additionally we provide a version of Figure 4.14 using percentage sequence identity instead of E-value (Figure D.4) and for joint sequence-similarity measure (Figure D.5). The number of correct one-same interactions is large in comparison with both across-species and both-different interactions. Indeed, one-same interactions represent a sizeable fraction of the aggregate interaction lists (compare Figure 4.14 A and Table 4.1). However, a comparison to Figure 4.3 shows that the observed fraction of correct one-same inferences is comparable to and sometimes lower than that for across-species inferences (depending on the species pair).

To make a fair comparison to the across-species case, we compare to both-different

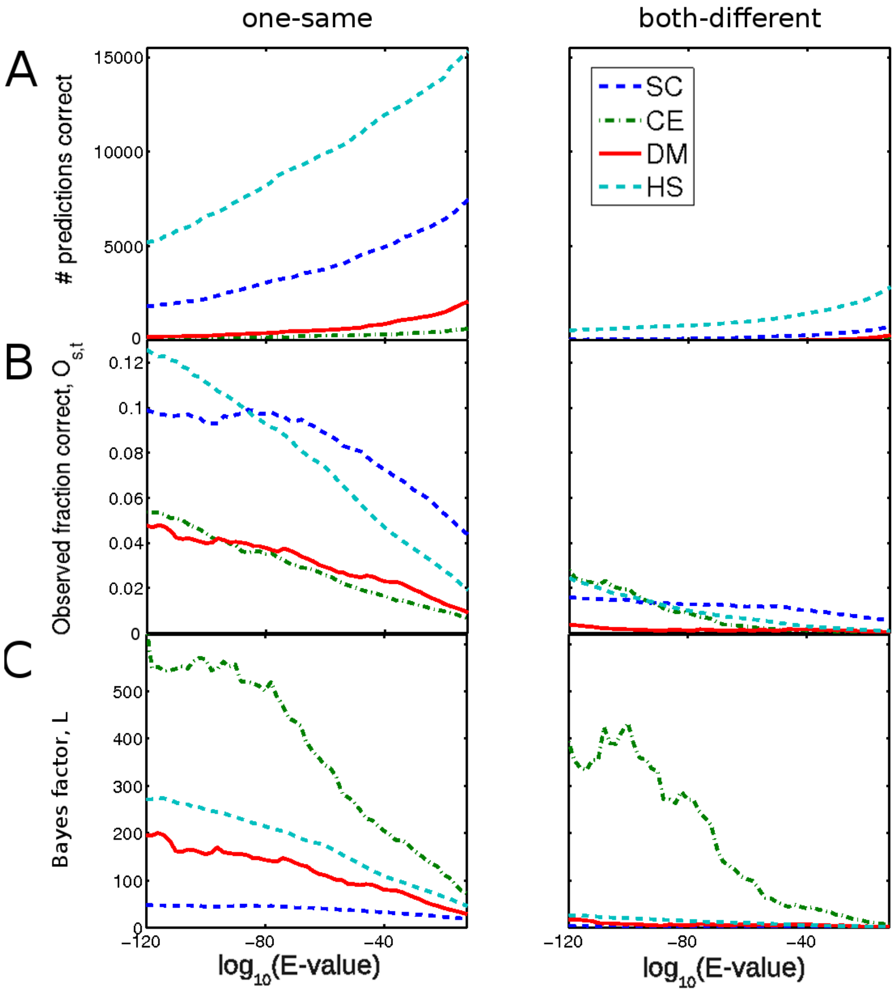


Figure 4.14: **Inferences within a species: 'one-same' inferences (left) dominate 'both-different' inferences (right).** For inferences within *S. cerevisiae* (SC), *C. elegans* (CE), *D. melanogaster* (DM), and *H. sapiens* (HS), one-same inferences dominate for (A) the number of correct inferences, (B) the fraction of inferences observed to be correct $O_{s,t}$, and (C) the Bayes Factor L. We consider a given inferred interaction to be inferred from the 'closest' interaction (see the main text for a definition and discussion). The very large Bayes Factors for *C. elegans*, particularly for the one-same case, are due to small-number effects.

Table 4.18: **Fraction of observed correct inferences $O_{s,t}$ at blastp E-value cutoffs of 10^{-10} and 10^{-70} for across-species and both-different within-species transferred interactions.** This table compares values also shown in Figure 4.3 B and Figure 4.14 B. If investigating only the 'closest' homologous inference (see text), one can see by comparing the diagonal to off-diagonal entries that when considered this way within species inferences are neither more accurate nor result in more correct inferences than across-species inferences.

Fraction of correct inferences, $E_{\text{val}} \leq 10^{-10}$					
target species		SC	CE	DM	HS
source species	SC	0.0055	0.0006	0.0018	0.0041
	CE	0.0207	0.0002	0.0029	0.0041
	DM	0.0157	0.0007	0.0006	0.0024
	HS	0.0175	0.0006	0.0017	0.0009
Fraction of correct inferences, $E_{\text{val}} \leq 10^{-70}$					
source species	SC	0.0128	0.0054	0.0066	0.0221
	CE	0.2201	0.0046	0.0258	0.0464
	DM	0.1285	0.0113	0.0013	0.0373
	HS	0.1092	0.0076	0.0138	0.0079

interactions. We show the values at two different E-value thresholds in Table 4.18. As we have shown, if one is not considering one-same inferences, then taking into account that one-same inferences may provide better evidence for conserved interactions suggests that protein-protein interactions are no more conserved within species than they are across species. This conclusion is opposite to that of Ref. [216], and it arises from our observation that one-same inferences underlie much of the observed conservation of interactions within a species.

4.5 Conclusions

Using six species, a mixture of low-throughput and high-throughput binary protein-protein interaction data and three different sets of homology definitions, we have investigated the conservation of interactions across and within species. Several factors mean that observed conservation rates do not reflect true evolutionary conservation rates. We argue that the data is biased, such that observed conservation rates will be inflated due to preferential investigation of homologous interactions. We develop a framework that takes interaction incompleteness into account – in contrast to previous studies, which have side-stepped the question of interactome errors. Using this framework, we are able to estimate interactome sizes with a method that is different from others in the literature.

Our estimates for the fraction of conserved interactions, which will be too high due to the bias in the data, are very low for definitions of homology that are often associated with the transfer of functional annotations across species.

We used our results on the conservation of interactions to estimate the rate at which protein-protein interactions are lost through evolution, though we stress the caveats involved with such an estimate.

Given that inferred interactions are not accurate unless stringent definitions of homology are used but that few interactions are transferable when such definitions are in place, we considered the possibility that certain types of inference were substantially less likely to yield conserved interactions. For example, we considered it possible that inferences from proteins in large protein families were substantially less accurate. Despite investigating a range of properties that might influence the conservation of interactions, we found no properties that, when taken into account, gave much improvement in conservation rates.

A previous study that compared the conservation of interactions within and across species found that interactions within a species are more conserved than those across species [216]. In contrast, we have shown that if 'one-same' inferences are taken into account and considered as potentially more parsimonious explanations for conservation, then one obtains the opposite conclusion. Moreover, our result is in line with the general expectation that orthologs retain more functional similarity than paralogs, and we thereby contribute to a current debate on whether this is a valid expectation [229]. These results could be developed through an investigation of the conservation of interactions of paralogs with different evolutionary histories. *S. cerevisiae* underwent a whole-genome duplication event [198] and comparison of its genome to that of other yeasts that diverged prior to this event (e.g. *Zygosaccharomyces rouxii*) has revealed the presence of ancient paralogous gene families as well as novel, lineage specific ones [61]. It would also be possible to compare duplicates that were retained after the whole-genome duplication to those in which one of the copies was lost.

The present study concentrates on the success of interolog inferences, which is the basis for a large number of widely-used methods to predict interactions [41, 42, 77, 102, 137, 138, 150, 177, 185, 255, 346, 356]. We urge caution in interpreting interactions transferred across species unless the definition of homology employed is a strict one, and we believe that interactome incompleteness is not solely responsible for the lack of observed conservation of interactions.

Chapter 5

Conclusions and future directions

5.1 Communities

Systems biologists stress that it is one thing to enumerate parts of a cell – genes and their products – and another thing entirely to understand how these act together to bring about biological function. A network perspective holds promise for bridging these gaps in our understanding. However, establishing connections between network structure and biological function has proven difficult, with many initial findings subject to debate (see Chapter 1).

The community structure of protein-protein interactions is of considerable interest because there is a strong *a priori* hypothesis about its relevance: communities are hypothesised to be good candidates for functional modules. As functional modules are themselves believed to be present at multiple different scales within biological systems, it makes sense to probe the community structure of protein-protein interaction networks (PINs) at multiple scales.

The literature that applies community detection algorithms to PINs is considerable, but the biological relevance of this community structure has been probed insufficiently: previous work overwhelmingly relies on one measure of functional enrichment that is insufficient to test the hypothesis that communities are good candidates for functional modules. The results presented in Chapter 3 represent an attempt at making a robust connection between PIN community structure and biological function. Our study represents the first systematic attempt to investigate community structure in PINs at multiple scales. We design new tests of the functional homogeneity of detected communities, finding that many communities are indeed good candidates for functional modules and that almost every protein is found within a functionally homogeneous community at some scale. We demonstrate how the community membership of an individual protein of interest can be traced. Finally, we show that different functional types of proteins are organised in different ways – for example, the scales at which different types of proteins are most concentrated within communities can vary dramatically.

There are several directions this work can be taken in. A recently introduced method detects communities in multiple networks simultaneously [226], through incorporating edges between the same node in different networks. The different networks, called different 'slices', could be networks involving different types of relationship between the same objects, views of the network at different resolutions/scales, or the network as it existed at different points in time. Other data types could be incorporated using this method: for example, data on gene co-expression, genetic interactions, correlated-phenotypes, and functional similarity could each be represented as a 'slice'.

It is always possible to incorporate new data sources, but I feel that in order for the relevance of community structure to truly be tested, it is important to demonstrate its utility for particular cases, whether this be for biological processes or for individual proteins. This would perhaps be best facilitated by a mechanism (e.g. a website) that allowed querying of particular proteins or GO terms. Another tack could be to use some of the results concerning the heterogeneous behaviour of the different protein functional classes to develop more sophisticated protein function prediction methods.

5.2 Homology

Existing protein-protein interaction data is concentrated within a few model organisms, and the coverage of this data for all species save *S. Cerevisiae* remains limited. This situation is by no means restricted to protein-protein interactions: in general, knowledge about particular biological processes is gleaned from the study of one or a few model organisms. It is widely assumed that knowledge gained in one species can be transferred to another species, even among species that are widely separated on the tree of life.

In Chapter 4, we investigated the validity of this transfer of knowledge for the case of protein-protein interactions. This common procedure is known to have shortcomings, which are generally ascribed to the incompleteness of protein interaction data. We show, however, that the procedure is unreliable even when the incompleteness of the data is taken into account. Our results imply that, unless very stringent definitions of homology are in place, interactions rewire at a rate too fast to allow reliable transfer between species that are well separated on the tree of life. We thus urge caution in interpreting the results of such transfers.

Transfer of interactions can also be performed *within* a species. We find that, when controlling for factors that favour within-species interaction transfer, this type of interaction transfer is even less reliable than that between species. Our result, though counter to previous studies, agrees with the general belief that duplicated proteins in the same species tend to diverge to be maintained, whereas the same protein in different species tends to stay the same in order to preserve ancestral functionality.

We can use our estimates of the reliability of interaction transfer to model the speed at which interactions are lost in evolution. These estimates are preliminary ones and can be improved upon in many ways, as suggested in Section 4.3.3.

We find that properties of proteins, such as their age, the number of domains that constitute them, and the size of the protein family to which they belong, are not very helpful for identifying which interactions are more likely to be conserved across species. This could be a limitation of the simple modelling framework that we employed. Our observations in Chapter 3 concerning the heterogeneity of behaviour of proteins of different types suggest that investigating mixture models (where sub-populations are represented differently) could be a promising line of enquiry. Different sub-populations that could be modelled include classifications by protein structure (e.g. using the classes from the Structural classification of protein folds, SCOP [6]) and GO slim classes.

5.3 Communities and homology of protein-protein interactions

The main finding from Chapter 4, the low rates of conservation of protein-protein interactions, has three primary potential explanations:

- Protein-protein interactions are not very constrained evolutionarily. This could be because they are not adaptive in the first place [84].

- Protein-protein interactions are a large source of phenotypic diversity. Small sequence changes can lead to large changes in phenotype via interaction rewiring. That is, in the case of protein-protein interactions, sequence is not very informative about function.

- Protein-protein interactions are not conserved at the level of individual interactions but rather at a 'fuzzier' level. This would suggest the hypothesis that interactions are not that conserved because of rewiring internal to communities rather than between.

This third possibility could be investigated by looking not for patterns of precise wiring but for communities of evolutionary conserved interactions. There are several questions that a combination of approaches from Chapters 3 and 4 suggest:

1. What evidence is there for the homology of protein-protein interaction communities across species?

2. Where are the successful interolog inferences in a network? Do they tend to be within communities, or between communities? Are they randomly distributed throughout the communities, or are they concentrated in particular communities? If they are concentrated in particular communities, which functional categories dominate in these?

3. Are communities in inferred interaction networks comparable to the communities based on real data in terms of their constituent proteins? Although individual inferences may be bad (see Chapter 4), perhaps taken in the aggregate they are better.

4. How do communities inferred in an individual species compare to communities inferred in all species simultaneously? The simultaneous detection of communities in multiple species could be achieved using the multi-slice community detection method discussed above [226]: the PIN of each species would be a 'slice', and either homology or functional similarity relationships could link the proteins in each slice. The success of the technique could also be benchmarked against local network alignment tools (see Section 1.4.5.2).

These research questions connect closely to the debate about whether functional modules are also evolutionary modules. There seem to be at least two main senses in which they could be. Wagner [337] made the case that neutral mutations (i.e. mutations with no phenotypic effect) can be hidden – possibly inside modules – until a changing environment yields a phenotypic effect. Recently, Navlakha and Kingsford [228] suggested that new edges indeed tend to form within existing complexes. This could enable overall 'fuzzy' homology of interaction patterns without conservation of individual interactions, as discussed above. In another sense of a functional module being an evolutionary module, nature could settle on functioning modules, and then link them up in different ways, much as engineers do with components. Recently, Zinman et al. [364] reported that interactions within modules are more conserved than interactions between modules. The study [299] mentioned in Section 1.5 reported little evidence for the functional cohesiveness of evolutionary modules, but this can be contrasted with examples such as that of signalling pathways that can be co-opted in different developmental contexts. As Pereira-Leal et al. [254] pointed out this indicates that a functional module can be re-used in different contexts. The results in the literature appear difficult to reconcile, but it seems that modules may well play a crucial role in the accommodation of evolutionary robustness and evolutionary flexibility [125], though whether and how this is the case is still very much open to debate.

Distinguishing between these different theories, or proposing new ones based on the results of a carefully designed comparative analysis, could go a long way toward increasing understanding of the evolution and organisation of biological parts into the whole.

Appendix A

Methods for predicting protein interactions

This Appendix is in large part reproduced from a paper published jointly with Ramazan Saeed and Charlotte Deane [182].

An increasing amount of protein interaction data is available, but it is error-prone and focuses on a few model organisms (see Section 1.2). Predicting protein interactions is thus a key challenge.

Protein interactions can be predicted computationally by employing various sources of information – including protein features, evolutionary knowledge, and network information. In this Appendix, I explain the ideas behind many of the most popular interaction prediction methods. Rather than giving algorithmic details, the evolutionary ideas behind the methods are highlighted. Figure A.1 illustrates some of the methods I will discuss.

The main category of prediction methods not covered here are ones based on machine-learning approaches. Two reviews of protein interaction prediction that include these approaches are given in Shoemaker and Panchenko [295] and Pitre et al. [257].

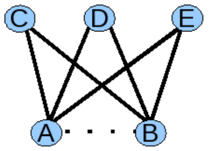
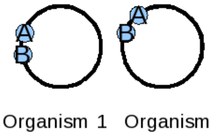
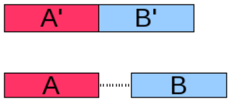
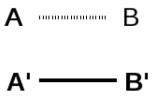
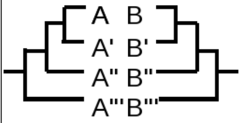
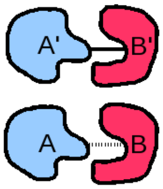
Method (refs)	**Prediction of interaction between A and B**	**Interaction predicted if:**
Network based (40)		A and B have many interacting partners in common
Gene Neighbourhood (70, 111)		Homologs of A and B occur in the same neighbourhood across multiple genomes
Gene Fusion (110, 112)		Homologs of A and B (A', B') occur fused together in another organism
Phylogenetic Profile (113)	1 2 3 4 A 1 1 0 0 B 1 1 0 0 C 1 0 1 0 D 0 1 1 1	Homologs of A and B have the same pattern of occurrence across multiple organisms (1, 2, 3, 4 etc)
Interologs (114)		A and B interact if their homologs, A' and B', in another species interact
Phylogenetic tree (71, 121)		A and B interact if their phylogenetic trees are similar
Structure based (122)		A and B interact if A' and B', which have similar structures to A and B, interact

Figure A.1: Different interaction prediction methods.

A.1 Assessing prediction methods

Most commonly, interaction prediction methods are judged by using the reference methodology in which an overlap between a true-positive gold standard set and a true-negative gold standard set is employed in a receiver operating characteristic (ROC) analysis (see Section 2.3 and [83, 150, 249]). The larger the area under the ROC curve for a binary classifier, the better that classifier is. As discussed in Section 2.3, ROC curves are unsuitable when the size of the two sets are very different.

A common way of obtaining true positive gold-standard sets is to select interactions observed in multiple assays [14] or manually curated datasets [115, 269]. However, such sets do not exist for all species due to low experimental coverage, and they can lead to bias in results [26, 283].

A.2 Network-based methods

An early example of using the topology of a network to assess the quality of interactions was proposed by Saito et al. [284], who suggested that the greater the number of isolated interaction partners two proteins had, the more unreliable the interaction. This observation has been employed to predict putative protein interactions in an existing network [51]. In particular they were able to rank the reliability of a predicted interaction using a measure of the shortest alternative pathway between the two interactors. Chen et al. [52] also exploited the high level of clustering in the network to predict interactions based upon triangular motifs in the network. This work demonstrated that there was information encoded in triangles of interactions in the PIN, which suggests that it is necessary to look beyond pairwise relationships to understand the evolution of the PIN as a whole. This claim has been surprisingly hard to demonstrate [52].

Clauset et al. [56] used the hierarchical structure (that is, structures present at different scales) of networks to predict 'missing' edges between nodes. Interaction probabilities were assigned between hierarchical groups, and a pair of nodes was suggested to be possibly linked if they possessed a high mean probability of connection within these hierarchical groups but were observed as unconnected. Whilst this method was tested on various networks, including a metabolic network, it has yet to be applied to PINs. It has the potential to be successful if PINs are found to have biologically meaningful hierarchical structure (Yook et al. [358] claimed that this is the case). This, in turn, rests on our understanding of the evolution of hierarchical modularity.

A common shortcoming of such network-based methodologies is that the results are likely to be sensitive to errors in the network data (see Sections 1.2.3 and 1.2.5).

A.3 Genome-based methods

Genomic information methods rely on the context of the gene/protein in an organism's genome, as well as the context of its homologs in other genomes. For a review, see Marcotte et al. [201]. *Homologs* are proteins that are believed to share characteristics because of shared ancestry. They are usually detected through sequence alignment. Proteins that are homologs due to a speciation events are called *orthologs*. Proteins that are homologs due to a gene duplication event are known as *paralogs*.

A.3.1 Gene neighbourhood

Gene neighbourhood methods are based on the observation mentioned in Section 1.4.3 that products of genes that are co-regulated have a higher chance of interacting physically, and on the fact that co-regulated genes – particularly in bacteria – tend to be close together in the genome. Gene neighbourhood methods exploit this relationship by searching for genes that are conserved and remain in the same neighbourhood across genomes. This adjacency is used to predict possible functional association [141, 242].

A.3.2 Gene fusion

The gene fusion method, also known as the Rosetta Stone method, assumes that the two interacting proteins depend on each other such that at some point in evolution, the two proteins were fused into one. The fused protein, dubbed the 'Rosetta Stone', is used to predict interactions in species in which the two proteins remain separate [80, 201]. This assumes that there is an evolutionary pressure for proteins that always interact to be transcribed as a single protein.

A.3.3 Phylogenetic profile

In the phylogenetic profile method, interactions are predicted based on the presence or absence of genes in related species. Utilising the recent dramatic increase in fully sequenced genomes, a phylogenetic profile is constructed for each gene. (A phylogenetic profile is a binary vector showing whether a protein is present in a genome or not). Similarities between the phylogenetic profiles of any two genes can be taken to indicate that the genes have some functional interdependencies on each other, thereby explaining their co-conservation across different species [252].

Here it is assumed that there is enough selective pressure that if two proteins interact to perform their cellular function and one protein is lost, then the other protein will also be lost (for example, this would be the case for some complexes). In the case of horizontal gene transfer in bacteria, genes will only be kept if they are transferred with other genes with which they need to interact to perform a fitness enhancing function.

A.4 Sequence-based methods

A.4.1 Interologs

I discuss interologs in Section 1.4.5.1 and Chapter 4. As discussed in Chapter 4, inferring interactions on the basis of homology is inaccurate unless strict definitions of homology are employed. At such strict definitions, few inferences are made.

A.4.2 Phylogenetic tree methods

Sequence homology can be used to build the phylogenetic trees of protein families [109]. These can then be compared. Pazos and Valencia [250] developed a method called *mirrortree* that utilises phylogenetic trees to make predictions. This explicitly uses the idea that co-evolving proteins are likely to be functionally associated. Valencia and Pazos [329] based predictions on the co-evolution of interacting interfaces only.

As discussed in Section 1.4.3, the main issue is in the inference from functional association to physical interaction (from co-evolution to co-adaptation).

A.5 Structure-based methods

A further category of interaction prediction methods consists of approaches that exploit structural similarities and make predictions based upon structural models.

In analogy with sequence-based interologs, structure-based interologs have been investigated. Aloy et al. [4] found that proteins with the same folds or structural domains tended to participate in similar interactions if the sequence identity of the proteins was above approximately 30%. Below this percentage, there is a 'twilight zone' where proteins may or may not share similar interactions. The evolutionary assumptions and potential problems are similar to those for sequence-based interologs (see Section A.4.1). Their advantage is that they are more accurate compared to sequence-based methods, their disadvantage is the relative paucity of structural information.

Protein-docking methods [58, 298], in which two structural proteins are rigidly combined and then refined, can also be used to predict interactions. However, due to the computational cost of such methods, they are in general more useful for providing information on the interacting interface of two structurally defined subunits [281]. Other methods predict interactions based on surface patch comparison [47] and oligomeric protein structure networks [40]. Carugo and Franzot [47] divided atoms on the surface of each protein into small, partially overlapping sets called 'patches'. They compared the shapes of each pair of patches belonging to different proteins, and they used a statistical analysis of the shape complementarity values to discriminate interacting and non-interacting protein pairs with an accuracy of up to 80%. Brinda and Vishveshwara [40] attempted to understand the factors involved in protein interactions by analysing interactions between amino acids based on the number of non-covalent bonds, which are known to play a role in mediating protein interactions [189].

One lesson from the field of protein docking is that supplementing physical and chemical considerations with information deduced from sequence and structural databases can improve predictions greatly [209]. The use of these databases rests on the assumption that similar structures imply similar interactions because the proteins are related to each other through evolution.

A.6 Domain-based methods

A range of methods have been devised that attempt to predict which of the domains in a protein interact. The methods annotate protein sequences with domains defined by Pfam, SCOP, CDD, and other domain databases [6, 86, 200].

Association methods are a group of prediction methods that look for blocks of sequence or structural motifs that distinguish interacting proteins from non-interacting proteins. In one such study, Sprinzak and Margalit [304] looked for sequence domains that were found to interact more often than expected by chance. They used such domains as signatures to predict new interactions.

A problem with association models is that they only consider one domain at a time and ignore the effect of other domains on the interaction. This was addressed by Deng et al. [71], who estimated the probabilities of interactions between every pair of domains and used them to predict interactions between proteins. Rare interactions between two domains can be missed by this method. To compensate for this, Riley et al. [275] developed measures based on the reduction in the likelihood of the PIN, caused by disallowing a given domain-domain interaction. This can give some indication of which domain-domain interaction is more likely to be responsible for a given protein-protein interaction.

In all these approaches, domains are assumed to interact independently, though this can depend on other domains within a protein pair and remains a severe limitation of these methods.

These methods could potentially be powerful in predicting entire PINs of organisms for which interaction data is not available [257]. This is because surprisingly few domains have been duplicated and recombined to form proteins across the tree of life: 50% of domain structure annotations in each organism are to fewer than 200 domain families common to all kingdoms. There is cause for caution, however, as one needs to ascertain that domain interactions are not organism-specific, remembering that domain combinations tend to be specific to organisms [240]. Basu et al. [22] demonstrated both that domains that occur in diverse domain architectures tend to have more interactions and that which domains end up in diverse architectures is organism-specific.

A.7 Summary

There are two main areas in which caution is needed in inferring protein interactions from evolutionary assumptions. The first is the use of interologs. There is widespread belief in the bioinformatics community that transfer of biological function, including protein interactions, should be possible from sequence homology. The evolutionary assumption is that close sequence similarity implies little functional divergence. This fundamental assumption has been bought under suspicion in the case of protein interactions (see Chapter 4). The second is separating the effects of actual molecular co-adaptation from the observation of co-evolution.

Key to the reliable use of interologs and co-evolution is a better understanding of the molecular mechanisms underpinning the evolution of interactions. There is evolutionary knowledge that has yet to be exploited in this regard. Notably, these include the large documented differences between transient and obligate interactions. (Proteins that must interact in order to carry out their cellular role are said to partake in an obligate interaction.) These differences can potentially be detected on the basis of sequence alone [239]. An investigation of the different sequence cutoffs that should be employed for interolog prediction for transient and obligate interactions would be a useful starting point.

Another step for improving interolog-based and co-evolution-based predictions is to use the comparative protein-protein interaction data that is now becoming available. Alignments of entire interaction networks (see Section 1.4.5.2), rather than just pairs of proteins, can give additional information as to when an interolog inference is acceptable. If proteins A, B, and C all interact in species S and proteins A' and B' and A' and C' interact in species S', then there is better evidence to predict that B'

and C' also interact.

There is potential for models of network evolution to be used in the prediction of protein interactions. At present, there is no clear consensus as to which are the good models of network evolution. It is clear that there is evolutionary information that can be incorporated into these models to make them more realistic. For example, no model proposed (to our knowledge) distinguishes between transient and obligate interactions, which is perhaps surprising given the known differences between the evolution of these different interaction types. Good models of protein-protein interaction network growth will be helpful in predicting new interactions. Such models could be used to generate ensembles of interaction networks matching observed statistics in empirical interaction networks. The frequency with which a given pair of proteins interacts in an ensemble can be used to predict likely interactors [56], although this has yet to be put into practice.

An important observation is that it is not easy to assess the relative success of different prediction methods. It is likely that different methods are more successful on certain types of proteins, which would be related to the different evolutionary assumptions underlying different methods. Through direct comparison of which methods do best on which proteins, interaction prediction could be tailored to what else is known about the proteins involved.

Appendix B

Examples of community membership

The proteins found in the communities given in Section 3.5. Protein numbers are the SGD identification numbers (Saccharomyces Genome Database, `www.yeastgenome.org`, [53]), the short descriptions are given on the SGD website.

Table B.1: **Example given in Figure 3.4 a). Of all those proteins found in the community at $\log(\lambda) = 0$, the proteins found in the red community at the partition at $\log(\lambda) = 0.5$. Continues on next page.**

100	Component of the small-subunit (SSU) processome, which is involved in the biogenesis of the 18S rRNA
451	Protein associated with U3 and U14 snoRNAs, required for pre-rRNA processing and 40S ribosomal subunit synthesis; localized in the nucleus and concentrated in the nucleolus
541	Protein of unknown function, identified as a high-copy suppressor of a dbp5 mutation
564	Essential nucleolar protein required for the synthesis of 18S rRNA and for the assembly of 40S ribosomal subunit
609	
643	Protein involved in bud-site selection; diploid mutants display a random budding pattern instead of the wild-type bipolar pattern
653	Conserved 90S pre-ribosomal component essential for proper endonucleolytic cleavage of the 35 S rRNA precursor at A0, A1, and A2 sites; contains eight WD-repeats; PWP2 deletion leads to defects in cell cycle and bud morphogenesis
752	RNA binding protein, part of U3 snoRNP involved in rRNA processing, part of U4/U6-U5 tri-snRNP involved in mRNA splicing, similar to human 15.5K protein
781	DNA Polymerase phi; has sequence similarity to the human MybBP1A and weak sequence similarity to B-type DNA polymerases, not required for chromosomal DNA replication; required for the synthesis of rRNA
837	RNA-binding protein, activates mRNA decapping directly by binding to the mRNA substrate and enhancing the activity of the decapping proteins Dcp1p and Dcp2p
884	Nucleolar protein, component of the small subunit (SSU) processome containing the U3 snoRNA that is involved in processing of pre-18S rRNA
929	Essential protein involved in maturation of 18S rRNA; depletion leads to inhibited pre-rRNA processing and reduced polysome levels; localizes primarily to the nucleolus
964	Protein that recognizes and binds damaged DNA (with Rad23p) during nucleotide excision repair; subunit of Nuclear Excision Repair Factor 2 (NEF2); homolog of human XPC protein
1030	Mitochondrial protein required for splicing of the group I intron aI5 of the COB pre-mRNA, binds to the RNA to promote splicing; also involved in but not essential for splicing of the COB bI2 intron and the intron in the 21S rRNA gene
1081	Protein of unknown function, green fluorescent protein (GFP)-fusion protein localizes to the endoplasmic reticulum; msc7 mutants are defective in directing meiotic recombination events to homologous chromatids
1107	Protein involved in rRNA processing; required for maturation of the 35S primary transcript of pre-rRNA and for cleavage leading to mature 18S rRNA; homologous to eIF-4a, which is a DEAD box RNA-dependent ATPase with helicase activity
1131	Protein component of the H/ACA snoRNP pseudouridylase complex, involved in the modification and cleavage of the 18S pre-rRNA
1191	Component of the SSU processome, which is required for pre-18S rRNA processing, essential protein that interacts with Mpp10p and mediates interactions of Imp4p and Mpp10p with U3 snoRNA
1239	Nucleolar protein, component of the small subunit (SSU) processome containing the U3 snoRNA that is involved in processing of pre-18S rRNA
1281	Protein required for pre-rRNA processing and 40S ribosomal subunit assembly
1498	Transcriptional activator of proline utilization genes, constitutively binds PUT1 and PUT2 promoter sequences and undergoes a conformational change to form the active state; has a Zn(2)-Cys(6) binuclear cluster domain
1518	UDP-glucose pyrophosphorylase (UGPase), catalyses the reversible formation of UDP-Glc from glucose 1-phosphate and UTP, involved in a wide variety of metabolic pathways, expression modulated by Pho85p through Pho4p
1561	Predominantly nucleolar DEAH-box ATP-dependent RNA helicase, required for 18S rRNA synthesis
1582	Subunit of U3-containing Small Subunit (SSU) processome complex involved in production of 18S rRNA and assembly of small ribosomal subunit
1768	Subunit of U3-containing 90S preribosome complex involved in production of 18S rRNA and assembly of small ribosomal subunit
2172	Nucleolar protein, component of the small subunit processome complex, which is required for processing of pre-18S rRNA; has similarity to mammalian fibrillarin
2307	Nucleolar protein, forms a complex with Noc4p that mediates maturation and nuclear export of 40S ribosomal subunits; also present in the small subunit processome complex, which is required for processing of pre-18S rRNA
2367	Nuclear protein related to mammalian high mobility group (HMG) proteins, essential for function of H/ACA-type snoRNPs, which are involved in 18S rRNA processing
2372	Putative RNA-binding protein implicated in ribosome biogenesis; contains an RNA recognition motif (RRM) and has similarity to hydrophilins; NOP6 may be a fungal-specific gene as no homologs have been yet identified in higher eukaryotes
2509	
2707	Essential protein that is a component of 90S preribosomes; may be involved in rRNA processing; multicopy suppressor of sensitivity to Brefeldin A; expression is induced during lag phase and also by cold shock
2732	Subunit of U3-containing 90S preribosome and Small Subunit (SSU) processome complexes involved in production of 18S rRNA and assembly of small ribosomal subunit; member of t-Utp subcomplex involved with transcription of 35S rRNA transcript
2747	Putative PINc domain nuclease required for early cleavages of 35S pre-rRNA and maturation of 18S rRNA; component of the SSU (small subunit) processome involved in 40S ribosomal subunit biogenesis; copurifies with Faf1p
2773	Nucleolar protein involved in pre-rRNA processing; depletion causes severely decreased 18S rRNA levels
2806	Subunit of U3-containing Small Subunit (SSU) processome complex involved in production of 18S rRNA and assembly of small ribosomal subunit

Continued from previous page.

2857	Nucleolar protein, component of the small subunit (SSU) processome containing the U3 snoRNA that is involved in processing of pre-18S rRNA
2860	Vacuolar endopolyphosphatase with a role in phosphate metabolism; functions as a homodimer
2953	Helicase encoded by the Y' element of subtelomeric regions, highly expressed in the mutants lacking the telomerase component TLC1; potentially phosphorylated by Cdc28p
3036	Protein associated with the mitochondrial nucleoid; putative mitochondrial ribosomal protein with similarity to E. coli L7/L12 ribosomal protein; required for normal respiratory growth
3046	Putative ATP-dependent RNA helicase of the DEAD-box family involved in ribosomal biogenesis
3139	ATP-dependent RNA helicase of the DEAD box family; required for 18S rRNA synthesis
3181	Ski complex component and WD-repeat protein, mediates 3'-5' RNA degradation by the cytoplasmic exosome; also required for meiotic double-strand break recombination; null mutants have superkiller phenotype
3322	Possible U3 snoRNP protein involved in maturation of pre-18S rRNA, based on computational analysis of large-scale protein-protein interaction data
3332	Cytoplasmic GTPase-activating protein for Ypt/Rab transport GTPases Ypt6p, Ypt31p and Sec4p; involved in recycling of internalized proteins and regulation of Golgi secretory function
3360	Nucleolar protein required for export of tRNAs from the nucleus; also copurifies with the small subunit (SSU) processome containing the U3 snoRNA that is involved in processing of pre-18S rRNA
3377	Essential nucleolar protein of unknown function; contains WD repeats, interacts with Mpp10p and Bfr2p, and has homology to Spb1p
3391	Nucleolar protein that binds nuclear localization sequences, required for pre-rRNA processing and ribosome biogenesis
3430	Cargo-transport protein involved in endocytosis; interacts with phosphatidylinositol-4-kinase Stt4; GFP-fusion protein localizes to the cytoplasm; YGR198W is an essential gene
3515	
3547	Essential subunit of U3-containing 90S preribosome involved in production of 18S rRNA and assembly of small ribosomal subunit; also part of pre-40S ribosome and required for its export into cytoplasm; binds RNA and contains pumilio domain
3605	Possible U3 snoRNP protein involved in maturation of pre-18S rRNA, based on computational analysis of large-scale protein-protein interaction data
3645	Nucleolar protein, component of the small subunit (SSU) processome containing the U3 snoRNA that is involved in processing of pre-18S rRNA
3683	
3762	Component of the SSU processome and 90S preribosome, required for pre-18S rRNA processing, interacts with and controls the stability of Imp3p and Imp4p, essential for viability; similar to human Mpp10p
3934	Essential protein required for biogenesis of 40S (small) ribosomal subunit; has similarity to the beta subunit of trimeric G-proteins and the splicing factor Prp4p
4041	Essential nucleolar protein involved in the early steps of 35S rRNA processing; interacts with Faf1p; member of a transcriptionally co-regulated set of genes called the RRB regulon
4053	
4119	Nucleolar protein, specifically associated with the U3 snoRNA, part of the large ribonucleoprotein complex known as the small subunit (SSU) processome, required for 18S rRNA biogenesis, part of the active pre-rRNA processing complex
4165	Pseudouridine synthase catalytic subunit of box H/ACA small nucleolar ribonucleoprotein particles (snoRNPs), acts on both large and small rRNAs and on snRNA U2; mutations in human ortholog dyskerin cause the disorder dyskeratosis congenita
4176	Member of the alpha/beta knot fold methyltransferase superfamily; required for maturation of 18S rRNA and for 40S ribosome production; interacts with RNA and with S-adenosylmethionine; associates with spindle/microtubules; forms homodimers
4187	Essential evolutionarily-conserved nucleolar protein component of the box C/D snoRNP complexes that direct 2'-O-methylation of pre-rRNA during its maturation; overexpression causes spindle orientation defects
4212	Nucleolar protein, component of the small subunit (SSU) processome containing the U3 snoRNA that is involved in processing of pre-18S rRNA
4372	Phosphatidylinositol transfer protein with a potential role in regulating lipid and fatty acid metabolism under heme-depleted conditions; interacts specifically with thioredoxin peroxidase; may have a role in oxidative stress resistance
4401	Subunit of U3-containing 90S preribosome and Small Subunit (SSU) processome complexes involved in production of 18S rRNA and assembly of small ribosomal subunit; synthetic defect with STI1 Hsp90 cochaperone; human homolog linked to glaucoma
4526	DNA helicase involved in telomere formation and elongation; acts as a catalytic inhibitor of telomerase; also plays a role in repair and recombination of mitochondrial DNA
4558	Subunit of U3-containing Small Subunit (SSU) processome complex involved in production of 18S rRNA and assembly of small ribosomal subunit
4616	Protein involved in bud-site selection; diploid mutants display a random budding pattern instead of the wild-type bipolar pattern
4699	Nucleolar protein, component of the small subunit (SSU) processome containing the U3 snoRNA that is involved in processing of pre-18S rRNA
4735	Essential DEAH-box ATP-dependent RNA helicase specific to the U3 snoRNP, predominantly nucleolar in distribution, required for 18S rRNA synthesis
4751	Protein component of the small (40S) ribosomal subunit; identical to Rps16Bp and has similarity to E. coli S9 and rat S16 ribosomal proteins
4842	RNA binding protein with preference for single stranded tracts of U's involved in synthesis of both 18S and 5.8S rRNAs; component of both the ribosomal small subunit (SSU) processosome and the 90S preribosome
4927	
5019	Component of the SSU processome, which is required for pre-18S rRNA processing; interacts with Mpp10p; member of a superfamily of proteins that contain a sigma(70)-like motif and associate with RNAs

Continued from previous page.

5076	Essential protein of unknown function; heterozygous mutant shows haploinsufficiency in K1 killer toxin resistance
5199	Protein with seven cysteine-rich CCHC zinc-finger motifs, similar to human CNBP, proposed to be involved in the RAS/cAMP signaling pathway
5252	Essential nucleolar protein required for 40S ribosome biogenesis; physically and functionally interacts with Krr1p
5337	Essential nucleolar protein involved in pre-18S rRNA processing; binds to RNA and stimulates ATPase activity of Dbp8; involved in assembly of the small subunit (SSU) processome
5368	Coenzyme Q (ubiquinone) binding protein, functions in the delivery of Q¡sub¿6¡/sub¿ to its proper location for electron transport during respiration; START domain protein with homologs in bacteria and eukaryotes
5370	Subunit of U3-containing 90S preribosome processome complex involved in 18S rRNA biogenesis and small ribosomal subunit assembly; stimulates Bms1p GTPase and U3 binding activity; similar to RNA cyclase-like proteins but no activity detected
5462	tRNA 2'-phosphotransferase, catalyzes the final step in yeast tRNA splicing
5530	Essential nucleolar protein that is a component of the SSU (small subunit) processome involved in 40S ribosomal subunit biogenesis; has homology to PINc domain protein Fcf1p, although the PINc domain of Utp23p is not required for function
5604	Component of small ribosomal subunit (SSU) processosome that contains U3 snoRNA; originally isolated as bud-site selection mutant that displays a random budding pattern
5671	Essential nucleolar protein required for pre-18S rRNA processing, interacts with Dim1p, an 18S rRNA dimethyltransferase, and also with Nob1p, which is involved in proteasome biogenesis; contains a KH domain
5837	Protein involved in pre-rRNA processing, 18S rRNA synthesis, and snoRNA synthesis; component of the small subunit processome complex, which is required for processing of pre-18S rRNA
5908	Ferric reductase, reduces siderophore-bound iron prior to uptake by transporters; expression induced by low iron levels
5933	Protein required for export of the ribosomal subunits; associates with the RNA components of the pre-ribosomes; contains HEAT-repeats
6047	U3 snoRNP protein, component of the small (ribosomal) subunit (SSU) processosome containing U3 snoRNA; required for the biogenesis of18S rRNA
6138	GTPase required for synthesis of 40S ribosomal subunits and for processing the 35S pre-rRNA at sites A0, A1, and A2; interacts with Rcl1p, which stimulates its GTPase and U3 snoRNA binding activities; has similarity to Tsr1p
6187	Essential 18S rRNA dimethylase (dimethyladenosine transferase), responsible for conserved m6(2)Am6(2)A dimethylation in 3'-terminal loop of 18S rRNA, part of 90S and 40S pre-particles in nucleolus, involved in pre-ribosomal RNA processing
6316	Essential conserved protein that is part of the 90S preribosome; required for production of 18S rRNA and small ribosomal subunit; contains five consensus RNA-binding domains
6341	Protein involved in pre-rRNA processing, associated with U3 snRNP; component of small ribosomal subunit (SSU) processosome; ortholog of the human U3-55k protein
6348	Nucleolar protein, forms a complex with Nop14p that mediates maturation and nuclear export of 40S ribosomal subunits
7455	Constituent of small nucleolar ribonucleoprotein particles containing H/ACA-type snoRNAs, which are required for pseudouridylation and processing of pre-18S rRNA
7608	Essential protein required for maturation of 18S rRNA; null mutant is sensitive to hydroxyurea and is delayed in recovering from alpha-factor arrest; green fluorescent protein (GFP)-fusion protein localizes to the nucleolus
7650	

Table B.2: **Example given in Figure 3.4 a). Of all those proteins found in the community at $\log(\lambda) = 0$, the proteins found in the yellow community at the partition at $\log(\lambda) = 0.5$. Continues on next page.**

7	Putative regulatory subunit of Nem1p-Spo7p phosphatase holoenzyme, regulates nuclear growth by controlling phospholipid biosynthesis, required for normal nuclear envelope morphology, premeiotic replication, and sporulation
14	Regulatory subunit A of the heterotrimeric protein phosphatase 2A, which also contains regulatory subunit Cdc55p and either catalytic subunit Pph21p or Pph22p; required for cell morphogenesis and for transcription by RNA polymerase III
33	GTPase, required for general translation initiation by promoting Met-tRNAiMet binding to ribosomes and ribosomal subunit joining; homolog of bacterial IF2
69	Protein involved in bud-site selection, Bud14p-Glc7p complex is a cortical regulator of dynein; inhibitor of the actin assembly factor Bnr1p (formin); diploid mutants display a random budding pattern instead of the wild-type bipolar pattern
103	Cytoskeletal protein binding protein required for assembly of the cortical actin cytoskeleton; interacts with proteins regulating actin dynamics and proteins required for endocytosis; found in the nucleus and cell cortex; has 3 SH3 domains
123	Protein component of the large (60S) ribosomal subunit, nearly identical to Rpl19Ap and has similarity to rat L19 ribosomal protein; rpl19a and rpl19b single null mutations result in slow growth, while the double null mutation is lethal
131	B subunit of DNA polymerase alpha-primase complex, required for initiation of DNA replication during mitotic and premeiotic DNA synthesis; also functions in telomere capping and length regulation
133	Alpha-adaptin, large subunit of the clathrin associated protein complex (AP-2); involved in vesicle mediated transport
135	Major CTP synthase isozyme (see also URA8), catalyzes the ATP-dependent transfer of the amide nitrogen from glutamine to UTP, forming CTP, the final step in de novo biosynthesis of pyrimidines; involved in phospholipid biosynthesis
147	Protein involved in G2/M phase progression and response to DNA damage, interacts with Rad53p; contains an RNA recognition motif, a nuclear localization signal, and several SQ/TQ cluster domains; hyperphosphorylated in response to DNA damage
168	Protein component of the small (40S) ribosomal subunit; identical to Rps8Bp and has similarity to rat S8 ribosomal protein
188	Protein component of the large (60S) ribosomal subunit, has similarity to rat L32 ribosomal protein; overexpression disrupts telomeric silencing
235	N-terminally acetylated protein component of the large (60S) ribosomal subunit, nearly identical to Rpl4Bp and has similarity to E. coli L4 and rat L4 ribosomal proteins
252	Protein component of the small (40S) ribosomal subunit; identical to Rps11Ap and has similarity to E. coli S17 and rat S11 ribosomal proteins
288	Mitochondrial C1-tetrahydrofolate synthase, involved in interconversion between different oxidation states of tetrahydrofolate (THF); provides activities of formyl-THF synthetase, methenyl-THF cyclohydrolase, and methylene-THF dehydrogenase
370	Prephenate dehydrogenase involved in tyrosine biosynthesis, expression is dependent on phenylalanine levels
376	Protein of unknown function involved in COPII vesicle formation; interacts with the Sec23p/Sec24p subcomplex; overexpression suppresses the temperature sensitivity of a myo2 mutant; has similarity to S. pombe Mpd2
385	Protein component of the small (40S) ribosomal subunit; identical to Rps6Ap and has similarity to rat S6 ribosomal protein
393	Protein component of the small (40S) ribosomal subunit; nearly identical to Rps9Ap and has similarity to E. coli S4 and rat S9 ribosomal proteins
395	Protein component of the large (60S) ribosomal subunit, nearly identical to Rpl21Bp and has similarity to rat L21 ribosomal protein
420	Protein required for oxidation of specific cysteine residues of the transcription factor Yap1p, resulting in the nuclear localization of Yap1p in response to stress
426	Peroxisomal AMP-binding protein, localizes to both the peroxisomal peripheral membrane and matrix, expression is highly inducible by oleic acid, similar to E. coli long chain acyl-CoA synthetase
464	GTPase-activating protein (RhoGAP) for Rho3p and Rho4p, possibly involved in control of actin cytoskeleton organization
467	Mitochondrial serine hydroxymethyltransferase, converts serine to glycine plus 5,10 methylenetetrahydrofolate; involved in generating precursors for purine, pyrimidine, amino acid, and lipid biosynthesis; reverse reaction generates serine
471	Cytoplasmic pre-60S factor; required for the correct recycling of shuttling factors Alb1, Arx1 and Tif6 at the end of the ribosomal large subunit biogenesis; involved in bud growth in the mitotic signaling network
482	Third-largest subunit of DNA polymerase II (DNA polymerase epsilon), required to maintain fidelity of chromosomal replication and also for inheritance of telomeric silencing; mRNA abundance peaks at the G1/S boundary of the cell cycle
627	Ribosomal protein 59 of the small subunit, required for ribosome assembly and 20S pre-rRNA processing; mutations confer cryptopleurine resistance; nearly identical to Rps14Bp and similar to E. coli S11 and rat S14 ribosomal proteins
780	Protein component of the large (60S) ribosomal subunit, nearly identical to Rpl12Bp; rpl12a rpl12b double mutant exhibits slow growth and slow translation; has similarity to E. coli L11 and rat L12 ribosomal proteins
838	ATPase of the ATP-binding cassette (ABC) family involved in 40S and 60S ribosome biogenesis, has similarity to Gcn20p; shuttles from nucleus to cytoplasm, physically interacts with Tif6p, Lsg1p
876	Protein component of the small (40S) ribosomal subunit; identical to Rps24Bp and has similarity to rat S24 ribosomal protein
904	Protein component of the small (40S) ribosomal subunit; identical to Rps8Ap and has similarity to rat S8 ribosomal protein
943	Protein required for the hydroxylation of heme O to form heme A, which is an essential prosthetic group for cytochrome c oxidase
993	Protein component of the large (60S) ribosomal subunit, nearly identical to Rpl14Ap and has similarity to rat L14 ribosomal protein
1007	Protein component of the small (40S) ribosomal subunit; overproduction suppresses mutations affecting RNA polymerase III-dependent transcription; has similarity to E. coli S10 and rat S20 ribosomal proteins
1025	Ribosomal protein L4 of the large (60S) ribosomal subunit, nearly identical to Rpl8Bp and has similarity to rat L7a ribosomal protein; mutation results in decreased amounts of free 60S subunits

Continued from previous page.

1026	Putative RNA binding protein; involved in translational repression and found in cytoplasmic P bodies; found associated with small nucleolar RNAs snR10 and snR11
1052	Protein component of the large (60S) ribosomal subunit, nearly identical to Rpl27Bp and has similarity to rat L27 ribosomal protein
1055	Subunit of N-terminal acetyltransferase NatA (Nat1p, Ard1p, Nat5p); acetylates many proteins and thus affects telomeric silencing, cell cycle, heat-shock resistance, mating, and sporulation; human Ard1p levels are elevated in cancer cells
1106	Hsp70 protein that interacts with Zuo1p (a DnaJ homolog) to form a ribosome-associated complex that binds the ribosome via the Zuo1p subunit; also involved in pleiotropic drug resistance via sequential activation of PDR1 and PDR5; binds ATP
1183	Protein component of the large (60S) ribosomal subunit, identical to Rpl42Ap and has similarity to rat L44; required for propagation of the killer toxin-encoding M1 double-stranded RNA satellite of the L-A double-stranded RNA virus
1213	Protein involved in nuclear export of the large ribosomal subunit; acts as a Crm1p-dependent adapter protein for export of nascent ribosomal subunits through the nuclear pore complex
1236	Alpha subunit of the heteromeric nascent polypeptide-associated complex (NAC) involved in protein sorting and translocation, associated with cytoplasmic ribosomes
1246	Protein component of the small (40S) ribosomal subunit; identical to Rps4Ap and has similarity to rat S4 ribosomal protein
1278	Protein of unknown function proposed to be involved in nuclear pore complex biogenesis and maintenance as well as protein folding; has similarity to the mammalian BAG-1 protein
1280	Protein component of the large (60S) ribosomal subunit, identical to Rpl2Ap and has similarity to E. coli L2 and rat L8 ribosomal proteins; expression is upregulated at low temperatures
1331	Protein component of the small (40S) ribosomal subunit; identical to Rps24Ap and has similarity to rat S24 ribosomal protein
1387	Component of the mitochondrial alpha-ketoglutarate dehydrogenase complex, which catalyzes a key step in the tricarboxylic acid (TCA) cycle, the oxidative decarboxylation of alpha-ketoglutarate to form succinyl-CoA
1395	N-terminally acetylated protein component of the large (60S) ribosomal subunit, binds to 5.8 S rRNA; has similarity to Rpl16Bp, E. coli L13 and rat L13a ribosomal proteins; transcriptionally regulated by Rap1p
1423	
1465	Protein phosphatase involved in vegetative growth at low temperatures, sporulation, and glycogen accumulation; transcription induced by low temperature and nitrogen starvation; member of the dual-specificity family of protein phosphatases
1550	Nucleoside diphosphate kinase, catalyzes the transfer of gamma phosphates from nucleoside triphosphates, usually ATP, to nucleoside diphosphates by a mechanism that involves formation of an autophosphorylated enzyme intermediate
1618	Beta-adaptin, large subunit of the clathrin-associated protein (AP-1) complex; binds clathrin; involved in clathrin-dependent Golgi protein sorting
1626	Component of the GSE complex, which is required for proper sorting of amino acid permease Gap1p; required for ribosomal small subunit export from nucleus; required for growth at low temperature
1663	Protein component of the large (60S) ribosomal subunit, nearly identical to Rpl17Bp and has similarity to E. coli L22 and rat L17 ribosomal proteins; copurifies with the Dam1 complex (aka DASH complex)
1695	Phosphatidylinositol phosphate (PtdInsP) phosphatase involved in hydrolysis of PtdIns[4]P; transmembrane protein localizes to ER and Golgi; involved in protein trafficking and processing, secretion, and cell wall maintenance
1765	Protein component of the small (40S) ribosomal subunit; nearly identical to Rps21Bp and has similarity to rat S21 ribosomal protein
2104	Protein component of the large (60S) ribosomal subunit, identical to Rpl2Bp and has similarity to E. coli L2 and rat L8 ribosomal proteins
2129	Protein that forms a heterotrimeric complex with Erp2p, Emp24p, and Erv25p; member, along with Emp24p and Erv25p, of the p24 family involved in ER to Golgi transport and localized to COPII-coated vesicles
2134	Non-essential protein of unknown function; expression induced in response to heat stress
2156	Protein component of the large (60S) ribosomal subunit, nearly identical to Rpl19Bp and has similarity to rat L19 ribosomal protein; rpl19a and rpl19b single null mutations result in slow growth, while the double null mutation is lethal
2218	Protein required for processing of 20S pre-rRNA in the cytoplasm, associates with pre-40S ribosomal particles
2233	Protein component of the large (60S) ribosomal subunit, nearly identical to Rpl31Bp and has similarity to rat L31 ribosomal protein; associates with the karyopherin Sxm1p; loss of both Rpl31p and Rpl39p confers lethality
2240	Protein component of the large (60S) ribosomal subunit, nearly identical to Rpl13Bp; not essential for viability; has similarity to rat L13 ribosomal protein
2241	Protein component of the small (40S) ribosomal subunit; identical to Rps16Ap and has similarity to E. coli S9 and rat S16 ribosomal proteins
2288	Ribosomal protein P1 beta, component of the ribosomal stalk, which is involved in interaction of translational elongation factors with ribosome; accumulation is regulated by phosphorylation and interaction with the P2 stalk component
2295	Protein component of the large (60S) ribosomal subunit, identical to Rpl35Ap and has similarity to rat L35 ribosomal protein
2296	ADP-ribosylation factor, GTPase of the Ras superfamily involved in regulation of coated formation vesicles in intracellular trafficking within the Golgi; functionally interchangeable with Arf1p
2354	Essential phosphoprotein component (p150) of the COPII coat of secretory pathway vesicles, in complex with Sec13p; required for ER-derived transport vesicle formation
2419	Protein component of the large (60S) ribosomal subunit, nearly identical to Rpl4Ap and has similarity to E. coli L4 and rat L4 ribosomal proteins

Continued from previous page.

2428	Nucleolar protein required for maturation of 18S rRNA, member of the eIF4A subfamily of DEAD-box ATP-dependent RNA helicases
2432	Protein component of the small (40S) ribosomal subunit; identical to Rps11Bp and has similarity to E. coli S17 and rat S11 ribosomal proteins
2471	Protein component of the small (40S) ribosomal subunit; has similarity to E. coli S15 and rat S13 ribosomal proteins
2489	Telomere end-binding and capping protein, plays a key role with Pol12p in linking telomerase action with completion of lagging strand synthesis, and in a regulatory step required for telomere capping
2498	Essential iron-sulfur protein required for ribosome biogenesis and translation initiation; facilitates binding of a multifactor complex (MFC) of translation initiation factors to the small ribosomal subunit; predicted ABC family ATPase
2524	Protein of unknown function that associates with ribosomes; has a putative RNA binding domain
2602	DEAD-box protein required for efficient splicing of mitochondrial Group I and II introns; non-polar RNA helicase that also facilities strand annealing
2604	Probable dephospho-CoA kinase (DPCK) that catalyzes the last step in coenzyme A biosynthesis; null mutant lethality is complemented by E. coli coaE (encoding DPCK); detected in purified mitochondria in high-throughput studies
2612	Protein with a role in ubiquinone (Coenzyme Q) biosynthesis, possibly functioning in stabilization of Coq7p; located on the matrix face of the mitochondrial inner membrane; component of a mitochondrial ubiquinone-synthesizing complex
2790	Ribosomal protein P2 beta, a component of the ribosomal stalk, which is involved in the interaction between translational elongation factors and the ribosome; regulates the accumulation of P1 (Rpp1Ap and Rpp1Bp) in the cytoplasm
2826	Protein component of the large (60S) ribosomal subunit, nearly identical to Rpl12Ap; rpl12a rpl12b double mutant exhibits slow growth and slow translation; has similarity to E. coli L11 and rat L12 ribosomal proteins
2837	eIF3g subunit of the core complex of translation initiation factor 3 (eIF3), which is essential for translation
2855	Ribosomal protein 51 (rp51) of the small (40s) subunit; nearly identical to Rps17Ap and has similarity to rat S17 ribosomal protein
2858	Protein component of the small (40S) ribosomal subunit; nearly identical to Rps18Bp and has similarity to E. coli S13 and rat S18 ribosomal proteins
2879	Protein component of the large (60S) ribosomal subunit, nearly identical to Rpl27Ap and has similarity to rat L27 ribosomal protein
2957	Protein of unknown function that associates with ribosomes and has a putative RNA binding domain; interacts with Tma22p; null mutant exhibits translation defects; has homology to human oncogene MCT-1
2982	Member of the PUF protein family, which is defined by the presence of Pumilio homology domains that confer RNA binding activity; preferentially binds mRNAs encoding nucleolar ribosomal RNA-processing factors
2997	Protein involved in nucleolar integrity and processing of the pre-rRNA for the 60S ribosome subunit; transcript is induced in response to cytotoxic stress but not genotoxic stress
2998	Protein component of the large (60S) ribosomal subunit, has similarity to rat L30 ribosomal protein; involved in pre-rRNA processing in the nucleolus; autoregulates splicing of its transcript
2999	Ribosomal protein L30 of the large (60S) ribosomal subunit, nearly identical to Rpl24Bp and has similarity to rat L24 ribosomal protein; not essential for translation but may be required for normal translation rate
3044	Protein component of the large (60S) ribosomal subunit, nearly identical to Rpl7Bp and has similarity to E. coli L30 and rat L7 ribosomal proteins; contains a conserved C-terminal Nucleic acid Binding Domain (NDB2)
3067	Putative GTPase involved in 60S ribosomal subunit biogenesis; required for the release of Nmd3p from 60S subunits in the cytoplasm
3071	Ribosomal protein of the large (60S) ribosomal subunit, has similarity to E. coli L15 and rat L27a ribosomal proteins; may have peptidyl transferase activity; can mutate to cycloheximide resistance
3091	Protein component of the small (40S) subunit, essential for control of translational accuracy; phosphorylation by C-terminal domain kinase I (CTDK-I) enhances translational accuracy; similar to E. coli S5 and rat S2 ribosomal proteins
3103	N-terminally acetylated protein component of the large (60S) ribosomal subunit, nearly identical to Rpl1Ap and has similarity to E. coli L1 and rat L10a ribosomal proteins; rpl1a rpl1b double null mutation is lethal
3115	Protein component of the large (60S) ribosomal subunit, nearly identical to Rpl9Bp and has similarity to E. coli L6 and rat L9 ribosomal proteins
3215	Nuclear protein that binds to and stabilizes the exoribonuclease Rat1p, required for pre-rRNA processing
3258	
3259	Protein component of the small (40S) ribosomal subunit; nearly identical to Rps25Bp and has similarity to rat S25 ribosomal protein
3266	Protein component of the large (60S) ribosomal subunit, nearly identical to Rpl26Ap and has similarity to E. coli L24 and rat L26 ribosomal proteins; binds to 5.8S rRNA
3286	
3313	Protein required for pre-rRNA processing; associated with the 90S pre-ribosome and 43S small ribosomal subunit precursor; interacts with U3 snoRNA; deletion mutant has synthetic fitness defect with an sgs1 deletion mutant
3317	Protein component of the large (60S) ribosomal subunit, nearly identical to Rpl11Ap; involved in ribosomal assembly; depletion causes degradation of proteins and RNA of the 60S subunit; has similarity to E. coli L5 and rat L11
3334	
3350	Ribosomal protein 28 (rp28) of the small (40S) ribosomal subunit, required for translational accuracy; nearly identical to Rps23Bp and similar to E. coli S12 and rat S23 ribosomal proteins; deletion of both RPS23A and RPS23B is lethal

Continued from previous page.

3384	GTP-binding protein of the ras superfamily required for bud site selection, morphological changes in response to mating pheromone, and efficient cell fusion; localized to the plasma membrane; significantly similar to mammalian Rap GTPases
3434	Cholinephosphate cytidylyltransferase, also known as CTP
3513	Plasma membrane ATP-binding cassette (ABC) transporter, multidrug transporter mediates export of many different organic anions including oligomycin; similar to human cystic fibrosis transmembrane receptor (CFTR)
3517	Cytosolic ribosome-associated chaperone that acts, together with Ssz1p and the Ssb proteins, as a chaperone for nascent polypeptide chains; contains a DnaJ domain and functions as a J-protein partner for Ssb1p and Ssb2p
3541	Integral membrane protein of the Golgi required for targeting of the Arf-like GTPase Arl3p to the Golgi; multicopy suppressor of ypt6 null mutation
3616	Essential RNA-binding G protein effector of mating response pathway, mainly associated with nuclear envelope and ER, interacts in mRNA-dependent manner with translating ribosomes via multiple KH domains, similar to vertebrate vigilins
3661	Subunit of tRNA (1-methyladenosine) methyltransferase, with Gcd10p, required for the modification of the adenine at position 58 in tRNAs, especially tRNAi-Met; first identified as a negative regulator of GCN4 expression
3713	Protein component of the large (60S) ribosomal subunit, nearly identical to Rpl17Ap and has similarity to E. coli L22 and rat L17 ribosomal proteins
3726	Protein component of the small (40S) ribosomal subunit; nearly identical to Rps22Bp and has similarity to E. coli S8 and rat S15a ribosomal proteins
3727	Ribosomal protein 59 of the small subunit, required for ribosome assembly and 20S pre-rRNA processing; mutations confer cryptopleurine resistance; nearly identical to Rps14Ap and similar to E. coli S11 and rat S14 ribosomal proteins
3786	3-hydroxyanthranilic acid dioxygenase, required for the de novo biosynthesis of NAD from tryptophan via kynurenine; expression regulated by Hst1p
3884	Protein component of the small (40S) ribosomal subunit, the least basic of the non-acidic ribosomal proteins; phosphorylated in vivo; essential for viability; has similarity to E. coli S7 and rat S5 ribosomal proteins
3906	Protein component of the small (40S) ribosomal subunit; mutation affects 20S pre-rRNA processing; identical to Rps4Bp and has similarity to rat S4 ribosomal protein
3958	Protein of unknown function, required for cell growth and possibly involved in rRNA processing; mRNA is cell cycle regulated
3968	Ribosomal protein L4 of the large (60S) ribosomal subunit, nearly identical to Rpl8Ap and has similarity to rat L7a ribosomal protein; mutation results in decreased amounts of free 60S subunits
4000	Protein that regulates telomeric length; protects telomeric ends in a complex with Cdc13p and Stn1p
4019	Protein component of the large (60S) ribosomal subunit, nearly identical to Rpl15Bp and has similarity to rat L15 ribosomal protein; binds to 5.8 S rRNA
4065	Protein component of the large (60S) ribosomal subunit, responsible for joining the 40S and 60S subunits; regulates translation initiation; has similarity to rat L10 ribosomal protein and to members of the QM gene family
4254	Protein component of the small (40S) ribosomal subunit; nearly identical to Rps28Ap and has similarity to rat S28 ribosomal protein
4278	Protein component of the small (40S) ribosomal subunit; nearly identical to Rps30Bp and has similarity to rat S30 ribosomal protein
4332	Conserved ribosomal protein P0 similar to rat P0, human P0, and E. coli L10e; shown to be phosphorylated on serine 302
4336	Protein component of the large (60S) ribosomal subunit, nearly identical to Rpl26Bp and has similarity to E. coli L24 and rat L26 ribosomal proteins; binds to 5.8S rRNA
4359	Protein component of the small (40S) ribosomal subunit; nearly identical to Rps22Ap and has similarity to E. coli S8 and rat S15a ribosomal proteins
4364	Elongase, involved in fatty acid and sphingolipid biosynthesis; synthesizes very long chain 20-26-carbon fatty acids from C18-CoA primers; involved in regulation of sphingolipid biosynthesis
4398	Protein component of the large (60S) ribosomal subunit, nearly identical to Rpl31Ap and has similarity to rat L31 ribosomal protein; associates with the karyopherin Sxm1p; loss of both Rpl31p and Rpl39p confers lethality
4433	Ribosomal protein 10 (rp10) of the small (40S) subunit; nearly identical to Rps1Bp and has similarity to rat S3a ribosomal protein
4440	Protein component of the large (60S) ribosomal subunit, has similarity to Rpl6Ap and to rat L6 ribosomal protein; binds to 5.8S rRNA
4486	Ribosomal protein 51 (rp51) of the small (40s) subunit; nearly identical to Rps17Bp and has similarity to rat S17 ribosomal protein
4488	Protein component of the small (40S) ribosomal subunit; nearly identical to Rps18Ap and has similarity to E. coli S13 and rat S18 ribosomal proteins
4524	Serine esterase that deacylates exogenous lysophospholipids, homolog of human neuropathy target esterase (NTE); mammalian NTE1 deacylates phosphatidylcholine to glycerophosphocholine
4528	Ribosomal protein 10 (rp10) of the small (40S) subunit; nearly identical to Rps1Ap and has similarity to rat S3a ribosomal protein
4538	N-terminally acetylated protein component of the large (60S) ribosomal subunit, has similarity to Rpl6Bp and to rat L6 ribosomal protein; binds to 5.8S rRNA
4722	G-protein beta subunit and guanine nucleotide dissociation inhibitor for Gpa2p; ortholog of RACK1 that inhibits translation; core component of the small (40S) ribosomal subunit; represses Gcn4p in the absence of amino acid starvation
4750	Protein component of the large (60S) ribosomal subunit, nearly identical to Rpl13Ap; not essential for viability; has similarity to rat L13 ribosomal protein
4807	N-terminally acetylated protein component of the large (60S) ribosomal subunit, nearly identical to Rpl36Bp and has similarity to rat L36 ribosomal protein; binds to 5.8 S rRNA

Continued from previous page.

4843	Protein component of the small (40S) ribosomal subunit; nearly identical to Rps10Ap and has similarity to rat ribosomal protein S10
4855	Protein component of the large (60S) ribosomal subunit, nearly identical to Rpl20Bp and has similarity to rat L18a ribosomal protein
4917	Integral inner mitochondrial membrane protein with a role in maintaining mitochondrial nucleoid structure and number; mutants exhibit an increased rate of mitochondrial DNA escape; shows some sequence similarity to exonucleases
5013	N-terminally acetylated protein component of the large (60S) ribosomal subunit, binds to 5.8 S rRNA; has similarity to Rpl16Ap, E. coli L13 and rat L13a ribosomal proteins; transcriptionally regulated by Rap1p
5040	Protein component of the small (40S) ribosomal subunit, nearly identical to Rps7Ap; interacts with Kti11p; deletion causes hypersensitivity to zymocin; has similarity to rat S7 and Xenopus S8 ribosomal proteins
5056	Essential ATP-dependent RNA helicase of the DEAD-box protein family, involved in nonsense-mediated mRNA decay and rRNA processing
5068	RNA-binding protein required for the assembly of box H/ACA snoRNPs and thus for pre-rRNA processing, forms a complex with Shq1p and interacts with H/ACA snoRNP components Nhp2p and Cbf5p; similar to Gar1p
5122	Protein component of the small (40S) ribosomal subunit, has apurinic/apyrimidinic (AP) endonuclease activity; essential for viability; has similarity to E. coli S3 and rat S3 ribosomal proteins
5151	Essential serine kinase involved in the processing of the 20S pre-rRNA into mature 18S rRNA; has similarity to Rio1p
5171	Co-chaperone that stimulates the ATPase activity of Ssa1p, required for a late step of ribosome biogenesis; associated with the cytosolic large ribosomal subunit; contains a J-domain; mutation causes defects in fluid-phase endocytosis
5175	Phosphatidylinositol transfer protein (PITP) controlled by the multiple drug resistance regulator Pdr1p, localizes to lipid particles and microsomes, controls levels of various lipids, may regulate lipid synthesis, homologous to Pdr17p
5245	Protein component of the large (60S) ribosomal subunit, identical to Rpl18Ap and has similarity to rat L18 ribosomal protein
5246	Protein component of the small (40S) ribosomal subunit, required for assembly and maturation of pre-40 S particles; mutations in human RPS19 are associated with Diamond Blackfan anemia; nearly identical to Rps19Ap
5399	Ribosomal protein P2 alpha, a component of the ribosomal stalk, which is involved in the interaction between translational elongation factors and the ribosome; regulates the accumulation of P1 (Rpp1Ap and Rpp1Bp) in the cytoplasm
5400	Protein component of the small (40S) ribosomal subunit; has similarity to E. coli S19 and rat S15 ribosomal proteins
5480	Protein component of the large (60S) ribosomal subunit, identical to Rpl18Bp and has similarity to rat L18 ribosomal protein; intron of RPL18A pre-mRNA forms stem-loop structures that are a target for Rnt1p cleavage leading to degradation
5481	Protein component of the small (40S) ribosomal subunit, required for assembly and maturation of pre-40 S particles; mutations in human RPS19 are associated with Diamond Blackfan anemia; nearly identical to Rps19Bp
5487	Primary rRNA-binding ribosomal protein component of the large (60S) ribosomal subunit, has similarity to E. coli L23 and rat L23a ribosomal proteins; binds to 26S rRNA via a conserved C-terminal motif
5540	B-type regulatory subunit of protein phosphatase 2A (PP2A); homolog of the mammalian B' subunit of PP2A
5574	Nuclear 5' to 3' single-stranded RNA exonuclease, involved in RNA metabolism, including rRNA and snRNA processing as well as mRNA transcription termination
5582	Essential nuclear protein involved in proteasome maturation and synthesis of 40S ribosomal subunits; required for cleavage of the 20S pre-rRNA to generate the mature 18S rRNA
5589	Protein component of the large (60S) ribosomal subunit, has similarity to E. coli L3 and rat L3 ribosomal proteins; involved in the replication and maintenance of killer double stranded RNA virus
5622	Protein component of the small (40S) ribosomal subunit, nearly identical to Rps7Bp; interacts with Kti11p; deletion causes hypersensitivity to zymocin; has similarity to rat S7 and Xenopus S8 ribosomal proteins
5705	Subunit of the APT subcomplex of cleavage and polyadenylation factor, may have a role in 3' end formation of both polyadenylated and non-polyadenylated RNAs
5724	Component of mRNP complexes associated with polyribosomes; implicated in secretion and nuclear segregation; multicopy suppressor of BFA (Brefeldin A) sensitivity
5730	ATP-dependent DEAD (Asp-Glu-Ala-Asp)-box RNA helicase, required for translation initiation of all yeast mRNAs; mutations in human DEAD-box DBY are a frequent cause of male infertility
5760	Ribosomal protein L37 of the large (60S) ribosomal subunit, nearly identical to Rpl33Ap and has similarity to rat L35a; rpl33b null mutant exhibits normal growth while rpl33a rpl33b double null mutant is inviable
5819	Protein component of the small (40S) ribosomal subunit; nearly identical to Rps10Bp and has similarity to rat ribosomal protein S10
5839	Protein component of the large (60S) ribosomal subunit, nearly identical to Rpl20Ap and has similarity to rat L18a ribosomal protein
5930	
6002	Protein component of the small (40S) ribosomal subunit; nearly identical to Rps9Bp and has similarity to E. coli S4 and rat S9 ribosomal proteins
6006	COPII vesicle coat protein required for ER transport vesicle budding and autophagosome formation; Sec16p is bound to the periphery of ER membranes and may act to stabilize initial COPII complexes; interacts with Sec23p, Sec24p and Sec31p
6011	Protein component of the small (40S) ribosomal subunit; identical to Rps6Bp and has similarity to rat S6 ribosomal protein
6020	Protein of unknown function; the authentic, non-tagged protein is detected in purified mitochondria in high-throughput studies; null mutant displays elevated frequency of mitochondrial genome loss

Continued from previous page.

6064	N-terminally acetylated ribosomal protein L37 of the large (60S) ribosomal subunit, nearly identical to Rpl33Bp and has similarity to rat L35a; rpl33a null mutant exhibits slow growth while rpl33a rpl33b double null mutant is inviable
6119	Protein component of the large (60S) ribosomal subunit, nearly identical to Rpl7Ap and has similarity to E. coli L30 and rat L7 ribosomal proteins; contains a conserved C-terminal Nucleic acid Binding Domain (NDB2)
6133	tRNA
6141	N-terminally acetylated protein component of the large (60S) ribosomal subunit, nearly identical to Rpl1Bp and has similarity to E. coli L1 and rat L10a ribosomal proteins; rpl1a rpl1b double null mutation is lethal
6208	Protein with similarity to mammalian electron transfer flavoprotein complex subunit ETF-alpha; interacts with frataxin, Yfh1p; null mutant displays elevated frequency of mitochondrial genome loss
6229	Cyclin associated with protein kinase Kin28p, which is the TFIIH-associated carboxy-terminal domain (CTD) kinase involved in transcription initiation at RNA polymerase II promoters
6245	Translation initiation factor eIF-5; N-terminal domain functions as a GTPase-activating protein to mediate hydrolysis of ribosome-bound GTP; C-terminal domain is the core of ribosomal preinitiation complex formation
6271	Protein required for maturation of mitochondrial and cytosolic Fe/S proteins, localizes to the mitochondrial intermembrane space, overexpression of ISA2 suppresses grx5 mutations
6290	Transcription factor TFIIB, a general transcription factor required for transcription initiation and start site selection by RNA polymerase II
6299	Guanine nucleotide exchange factor (GEF) for Arf proteins; involved in vesicular transport; suppressor of ypt3 mutations; member of the Sec7-domain family
6306	Protein component of the large (60S) ribosomal subunit, nearly identical to Rpl11Bp; involved in ribosomal assembly; depletion causes degradation of proteins and RNA of the 60S subunit; has similarity to E. coli L5 and rat L11
6318	
6393	Ski complex component and TPR protein, mediates 3'-5' RNA degradation by the cytoplasmic exosome; null mutants have superkiller phenotype of increased viral dsRNAs and are synthetic lethal with mutations in 5'-3' mRNA decay
6437	Protein component of the large (60S) ribosomal subunit, has similarity to rat L29 ribosomal protein; not essential for translation, but required for proper joining of the large and small ribosomal subunits and for normal translation rate
6438	Protein component of the large (60S) ribosomal subunit, nearly identical to Rpl36Ap and has similarity to rat L36 ribosomal protein; binds to 5.8 S rRNA
7246	Protein of unknown that associates with ribosomes; null mutant exhibits translation defects, altered polyribosome profiles, and resistance to the translation inhibitor anisomcyin
7376	

Table B.3: **Example given in Figure 3.4 a). Of all those proteins found in the community at $\log(\lambda) = 0$, the proteins found in the blue community at the partition at $\log(\lambda) = 0.5$. Continued on next page.**

23	Essential nuclear protein, constituent of 66S pre-ribosomal particles; required for maturation of 25S and 5.8S rRNAs; required for maintenance of M1 satellite double-stranded RNA of the L-A virus
48	Oleate-activated transcription factor, acts alone and as a heterodimer with Pip2p; activates genes involved in beta-oxidation of fatty acids and peroxisome organization and biogenesis
183	Protein component of the large (60S) ribosomal subunit, identical to Rpl23Bp and has similarity to E. coli L14 and rat L23 ribosomal proteins
346	Essential nucleolar protein, putative DEAD-box RNA helicase required for maintenance of M1 dsRNA virus; involved in biogenesis of large (60S) ribosomal subunits
391	Putative protein of unknown function; expression is reduced in a gcr1 null mutant; GFP-fusion protein localizes to the vacuole; expression pattern and physical interactions suggest a possible role in ribosome biogenesis
446	
559	AdoMet-dependent methyltransferase involved in rRNA processing and 60S ribosomal subunit maturation; methylates G2922 in the tRNA docking site of the large subunit rRNA and in the absence of snR52, U2921; suppressor of PAB1 mutants
569	Catabolic L-serine (L-threonine) deaminase, catalyzes the degradation of both L-serine and L-threonine; required to use serine or threonine as the sole nitrogen source, transcriptionally induced by serine and threonine
644	Acyl-CoA
668	WD-repeat protein involved in ribosome biogenesis; may interact with ribosomes; required for maturation and efficient intra-nuclear transport or pre-60S ribosomal subunits, localizes to the nucleolus
804	Constituent of 66S pre-ribosomal particles, involved in 60S ribosomal subunit biogenesis
808	GTPase that associates with nuclear 60S pre-ribosomes, required for export of 60S ribosomal subunits from the nucleus
879	
928	Protein constituent of 66S pre-ribosomal particles, contributes to processing of the 27S pre-rRNA
1080	Mitochondrial ribosome recycling factor, essential for mitochondrial protein synthesis and for the maintenance of the respiratory function of mitochondria
1094	Essential protein that interacts with proteasome components and has a potential role in proteasome substrate specificity; also copurifies with 66S pre-ribosomal particles
1108	Constituent of 66S pre-ribosomal particles, required for ribosomal large subunit maturation; functionally redundant with Ssf2p; member of the Brix family
1127	Essential component of the Rix1 complex (with Rix1p and Ipi3p) that is required for processing of ITS2 sequences from 35S pre-rRNA; Rix1 complex associates with Mdn1p in pre-60S ribosomal particles
1130	Nucleolar protein involved in the assembly and export of the large ribosomal subunit; constituent of 66S pre-ribosomal particles; contains a sigma(70)-like motif, which is thought to bind RNA
1180	
1240	Essential component of the Rix1 complex (Rix1p, Ipi1p, Ipi3p) that is required for processing of ITS2 sequences from 35S pre-rRNA; Rix1 complex associates with Mdn1p in pre-60S ribosomal particles
1492	Protein involved in mRNA turnover and ribosome assembly, localizes to the nucleolus
1497	Nucleolar protein required for the normal accumulation of 25S and 5.8S rRNAs, associated with the 27SA2 pre-ribosomal particle; proposed to be involved in the biogenesis of the 60S ribosomal subunit
1504	Protein involved in an early, nucleolar step of 60S ribosomal subunit biogenesis; essential for cell growth and replication of killer M1 dsRNA virus; contains four beta-transducin repeats
1565	Essential protein, constituent of 66S pre-ribosomal particles; interacts with proteins involved in ribosomal biogenesis and cell polarity; member of the SURF-6 family
1655	Essential protein required for the maturation of 25S rRNA and 60S ribosomal subunit assembly, localizes to the nucleolus; constituent of 66S pre-ribosomal particles
1659	Protein possibly involved in a post-Golgi secretory pathway; required for the transport of nitrogen-regulated amino acid permease Gap1p from the Golgi to the cell surface
1732	Putative ATP-dependent RNA helicase of the DEAD-box family involved in ribosomal biogenesis; essential for growth under anaerobic conditions
1789	Essential protein involved in the processing of pre-rRNA and the assembly of the 60S ribosomal subunit; interacts with ribosomal protein L11; localizes predominantly to the nucleolus; constituent of 66S pre-ribosomal particles
1839	Low-affinity amino acid permease, may act to supply the cell with amino acids as nitrogen source in nitrogen-poor conditions; transcription is induced under conditions of sulfur limitation; plays a role in regulating Ty1 transposition
1894	Putative ATP-dependent RNA helicase, nucleolar protein required for synthesis of 60S ribosomal subunits at a late step in the pathway; sediments with 66S pre-ribosomes in sucrose gradients
1897	Nuclear protein involved in asymmetric localization of ASH1 mRNA; binds double-stranded RNA in vitro; constituent of 66S pre-ribosomal particles
2189	Putative ATP-dependent RNA helicase of the DEAD-box protein family, constituent of 66S pre-ribosomal particles; essential protein involved in ribosome biogenesis
2327	Bifunctional enzyme containing both alcohol dehydrogenase and glutathione-dependent formaldehyde dehydrogenase activities, functions in formaldehyde detoxification and formation of long chain and complex alcohols, regulated by Hog1p-Sko1p
2467	Constituent of 66S pre-ribosomal particles, required for large (60S) ribosomal subunit biogenesis; involved in nuclear export of pre-ribosomes; required for maintenance of dsRNA virus; homolog of human CAATT-binding protein

Continued from previous page.

2494	Essential evolutionarily conserved nucleolar protein necessary for biogenesis of 60S ribosomal subunits and processing of pre-rRNAs to mature rRNAs, associated with several distinct 66S pre-ribosomal particles
2508	Shuttling pre-60S factor; involved in the biogenesis of ribosomal large subunit biogenesis; interacts directly with Alb1; responsible for Tif6 recycling defects in absence of Rei1; associated with the ribosomal export complex
2720	Protein required for ribosomal large subunit maturation, functionally redundant with Ssf1p; member of the Brix family
2769	Essential protein involved in nuclear export of Mss4p, which is a lipid kinase that generates phosphatidylinositol 4,5-biphosphate and plays a role in actin cytoskeleton organization and vesicular transport
2794	Subunit of the structure-specific Mms4p-Mus81p endonuclease that cleaves branched DNA; involved in DNA repair, replication fork stability, and joint molecule formation/resolution during meiotic recombination; helix-hairpin-helix protein
2820	Component of the pre-60S pre-ribosomal particle; required for cell viability under standard (aerobic) conditions but not under anaerobic conditions
2904	Pumilio-homology domain protein that binds ASH1 mRNA at PUF consensus sequences in the 3' UTR and represses its translation, resulting in proper asymmetric localization of ASH1 mRNA
3079	Constituent of 66S pre-ribosomal particles, involved in 60S ribosomal subunit biogenesis
3088	RNA helicase in the DEAH-box family, functions in both RNA polymerase I and polymerase II transcript metabolism, involved in release of the lariat-intron from the spliceosome
3335	Component of several different pre-ribosomal particles; forms a complex with Ytm1p and Erb1p that is required for maturation of the large ribosomal subunit; required for exit from G¡sub¿0¡/sub¿ and the initiation of cell proliferation
3440	Phosphoserine phosphatase of the phosphoglycerate pathway, involved in serine and glycine biosynthesis, expression is regulated by the available nitrogen source
3477	Highly conserved nuclear protein required for actin cytoskeleton organization and passage through Start, plays a critical role in G1 events, binds Nap1p, also involved in 60S ribosome biogenesis
3586	ATP-dependent 3'-5' RNA helicase, involved in nuclear RNA quality control both as a component of the TRAMP complex and in TRAMP independent processes; member of the Dead-box family of helicases
3658	Shuttling pre-60S factor; involved in the biogenesis of ribosomal large subunit; interacts directly with Arx1p; responsible for Tif6p recycling defects in absence of Rei1p
3667	Putative protein of unknown function; the authentic non-tagged protein is detected in highly purified mitochondria; null mutant is viable, displays severe respiratory growth defect and elevated frequency of mitochondrial genome loss
3802	Nucleolar protein required for normal metabolism of the rRNA primary transcript, proposed to be involved in ribosome biogenesis
3931	Nucleolar DEAD-box protein required for ribosome assembly and function, including synthesis of 60S ribosomal subunits; constituent of 66S pre-ribosomal particles
3957	Putative ATPase of the AAA family, required for export of pre-ribosomal large subunits from the nucleus; distributed between the nucleolus, nucleoplasm, and nuclear periphery depending on growth conditions
3992	Protein that forms a nuclear complex with Noc2p that binds to 66S ribosomal precursors to mediate their intranuclear transport; also binds to chromatin to promote the association of DNA replication factors and replication initiation
3999	Essential protein with similarity to Rpl24Ap and Rpl24Bp, associated with pre-60S ribosomal subunits and required for ribosomal large subunit biogenesis
4064	Protein involved in bud-site selection; diploid mutants display a random budding pattern instead of the wild-type bipolar pattern
4096	Huge dynein-related AAA-type ATPase (midasin), forms extended pre-60S particle with the Rix1 complex (Rix1p-Ipi1p-Ipi3p), may mediate ATP-dependent remodeling of 60S subunits and subsequent export from nucleoplasm to cytoplasm
4126	mRNA-binding protein expressed during iron starvation; binds to a sequence element in the 3'-untranslated regions of specific mRNAs to mediate their degradation; involved in iron homeostasis
4186	Protein with WD-40 repeats involved in rRNA processing; associates with trans-acting ribosome biogenesis factors; similar to beta-transducin superfamily
4211	Protein with a likely role in ribosomal maturation, required for accumulation of wild-type levels of large (60S) ribosomal subunits; binds to the helicase Dbp6p in pre-60S ribosomal particles in the nucleolus
4223	TLC1 RNA-associated factor involved in telomere length regulation as the recruitment subunit of the telomerase holoenzyme, has a possible role in activating Est2p-TLC1-RNA bound to the telomere
4266	ATP-dependent RNA helicase of the DEAD-box family involved in biogenesis of the 60S ribosomal subunit
4317	Protein component of the large (60S) ribosomal subunit, has similarity to rat L38 ribosomal protein
4389	ATPase of the CDC48/PAS1/SEC18 (AAA) family, forms a hexameric complex; may be involved in degradation of aberrant mRNAs
4419	Cytoplasmic protein of unknown function; ubiquitinated protein with similarity to the human ring finger motif protein, RNF10; predicted to encode a DNA-3-methyladenine glycosidase II that catalyzes hydrolysis of alkylated DNA
4441	Peptidyl-prolyl cis-trans isomerase (PPIase) (proline isomerase) localized to the nucleus; catalyzes isomerization of proline residues in histones H3 and H4, which affects lysine methylation of those histones
4539	Nucleolar peptidyl-prolyl cis-trans isomerase (PPIase); FK506 binding protein; phosphorylated by casein kinase II (Cka1p-Cka2p-Ckb1p-Ckb2p) and dephosphorylated by Ptp1p
4637	
4652	Constituent of 66S pre-ribosomal particles, forms a complex with Nop7p and Ytm1p that is required for maturation of the large ribosomal subunit; required for maturation of the 25S and 5.8S ribosomal RNAs; homologous to mammalian Bop1
4703	Peripheral GTPase of the mitochondrial inner membrane, essential for respiratory competence, likely functions in assembly of the large ribosomal subunit, has homologs in plants and animals

Continued from previous page.

4728	Protein component of the large (60S) ribosomal subunit, nearly identical to Rpl15Ap and has similarity to rat L15 ribosomal protein; binds to 5.8 S rRNA
4738	Essential nuclear protein involved in early steps of ribosome biogenesis; physically interacts with the ribosomal protein Rpl3p
4903	ATP-dependent RNA helicase; localizes to both the nuclear periphery and nucleolus; highly enriched in nuclear pore complex fractions; constituent of 66S pre-ribosomal particles
4947	Nucleolar protein with similarity to large ribosomal subunit L7 proteins; constituent of 66S pre-ribosomal particles; plays an essential role in processing of precursors to the large ribosomal subunit RNAs
4983	Protein involved in the synthesis of N-acetylglucosaminyl phosphatidylinositol (GlcNAc-PI), the first intermediate in the synthesis of glycosylphosphatidylinositol (GPI) anchors; homologous to the human PIG-H protein
5005	Probable RNA m(5)C methyltransferase, essential for processing and maturation of 27S pre-rRNA and large ribosomal subunit biogenesis; localized to the nucleolus; constituent of 66S pre-ribosomal particles
5054	Constituent of 66S pre-ribosomal particles, involved in 60S ribosomal subunit biogenesis; localizes to both nucleolus and cytoplasm
5119	Nucleolar protein found in preribosomal complexes; contains an RNA recognition motif (RRM)
5126	Essential component of the Rix1 complex (Rix1p, Ipi1p, Ipi3p) that is required for processing of ITS2 sequences from 35S pre-rRNA; highly conserved and contains WD40 motifs; Rix1 complex associates with Mdn1p in pre-60S ribosomal particles
5174	Elongin A, F-box protein that forms a heterodimer with Elc1p and is required for ubiquitin-dependent degredation of the RNA Polymerase II subunit RPO21; subunit of the Elongin-Cullin-Socs (ECS) ligase complex
5336	Putative GTPase that associates with pre-60S ribosomal subunits in the nucleolus and is required for their nuclear export and maturation
5401	Nucleolar protein involved in pre-25S rRNA processing and biogenesis of large 60S ribosomal subunit; contains an RNA recognition motif (RRM); binds to Ebp2; similar to Nop13p and Nsr1p
5437	Nucleolar protein, constituent of 66S pre-ribosomal particles; depletion leads to defects in rRNA processing and a block in the assembly of large ribosomal subunits; possesses a sigma(70)-like RNA-binding motif
5504	Nucleolar protein required for 60S ribosomal subunit biogenesis
5531	DNA ligase required for nonhomologous end-joining (NHEJ), forms stable heterodimer with required cofactor Lif1p, interacts with Nej1p; involved in meiosis, not essential for vegetative growth
5606	Origin-binding F-box protein that forms an SCF ubiquitin ligase complex with Skp1p and Cdc53p; plays a role in DNA replication, involved in invasive and pseudohyphal growth
5693	Protein component of the small (40S) ribosomal subunit; nearly identical to Rps28Bp and has similarity to rat S28 ribosomal protein
5732	Protein that forms a nucleolar complex with Mak21p that binds to 90S and 66S pre-ribosomes, as well as a nuclear complex with Noc3p that binds to 66S pre-ribosomes; both complexes mediate intranuclear transport of ribosomal precursors
5767	Folylpolyglutamate synthetase, catalyzes extension of the glutamate chains of the folate coenzymes, required for methionine synthesis and for maintenance of mitochondrial DNA
5778	Protein of unknown function that associates with ribosomes
5798	Constituent of 66S pre-ribosomal particles, forms a complex with Nop7p and Erb1p that is required for maturation of the large ribosomal subunit; has seven C-terminal WD repeats
5820	Essential protein that binds ribosomal protein L11 and is required for nuclear export of the 60S pre-ribosomal subunit during ribosome biogenesis; mouse homolog shows altered expression in Huntington's disease model mice
5964	Nucleolar protein, essential for processing and maturation of 27S pre-rRNA and large ribosomal subunit biogenesis; constituent of 66S pre-ribosomal particles; contains four RNA recognition motifs (RRMs)
6014	Putative GTPase that associates with free 60S ribosomal subunits in the nucleolus and is required for 60S ribosomal subunit biogenesis; constituent of 66S pre-ribosomal particles; member of the ODN family of nucleolar G-proteins
6052	Protein component of the large (60S) ribosomal subunit with similarity to E. coli L18 and rat L5 ribosomal proteins; binds 5S rRNA and is required for 60S subunit assembly
6067	Nucleolar protein; involved in biogenesis of the 60S subunit of the ribosome; interacts with rRNA processing factors Cbf5p and Nop2p; null mutant is viable but growth is severely impaired
6129	SET-domain lysine-N-methyltransferase, catalyzes the formation of dimethyllysine residues on the large ribsomal subunit protein L23a (RPL23A and RPL23B)
6132	Nucleolar protein required for 60S ribosome subunit biogenesis, constituent of 66S pre-ribosomal particles; physically interacts with Nop8p and the exosome subunit Rrp43p
6180	Mu1-like medium subunit of the clathrin-associated protein complex (AP-1); binds clathrin; involved in clathrin-dependent Golgi protein sorting
6220	Constituent of 66S pre-ribosomal particles, has similarity to human translation initiation factor 6 (eIF6); may be involved in the biogenesis and or stability of 60S ribosomal subunits
6221	Guanine nucleotide dissociation stimulator for Sec4p, functions in the post-Golgi secretory pathway; binds zinc, found both on membranes and in the cytosol
6249	
6347	Nucleolar protein, constituent of pre-60S ribosomal particles; required for proper processing of the 27S pre-rRNA at the A3 and B1 sites to yield mature 5.8S and 25S rRNAs
6365	Cyclin (Bur2p)-dependent protein kinase that functions in transcriptional regulation; phosphorylates the carboxy-terminal domain of Rpo21p and the C-terminal repeat domain of Spt5p; regulated by Cak1p
6373	Essential protein of unknown function; interacts with proteins involved in RNA processing, ribosome biogenesis, ubiquitination and demethylation; tagged protein localizes to nucleus and nucleolus; similar to WDR55, a human WD repeat protein
7259	
28857	

Table B.4: **Example given in Figure 3.4 a). Of all those proteins found in the community at** $\log(\lambda) = 0$**, the proteins not found in the red, blue or yellow communities at the partition at** $\log(\lambda) = 0.5$**. Continues on next page.**

6	Mitochondrial protein of unknown function
95	Nonfunctional protein with homology to IMP dehydrogenase; probable pseudogene, located close to the telomere; is not expressed at detectable levels; YAR073W and YAR075W comprise a continuous reading frame in some strains of S. cerevisiae
167	
215	Cytoplasmic inorganic pyrophosphatase (PPase), homodimer that catalyzes the rapid exchange of oxygens from Pi with water, highly expressed and essential for viability, active-site residues show identity to those from E. coli PPase
224	Galactokinase, phosphorylates alpha-D-galactose to alpha-D-galactose-1-phosphate in the first step of galactose catabolism; expression regulated by Gal4p
225	Uracil permease, localized to the plasma membrane; expression is tightly regulated by uracil levels and environmental cues
238	Nuclear SAM-dependent mono- and asymmetric arginine dimethylating methyltransferase that modifies hnRNPs, including Npl3p and Hrp1p, thus facilitating nuclear export of these proteins; required for viability of npl3 mutants
242	Chitin synthase II, requires activation from zymogenic form in order to catalyze the transfer of N-acetylglucosamine (GlcNAc) to chitin; required for the synthesis of chitin in the primary septum during cytokinesis
248	Protein involved in the assembly of the mitochondrial succinate dehydrogenase complex; putative chaperone
253	RNA polymerase I enhancer binding protein; DNA binding protein which binds to genes transcribed by both RNA polymerase I and RNA polymerase II; required for termination of RNA polymerase I transcription
262	Ubiquitin-specific protease that specifically disassembles unanchored ubiquitin chains; involved in fructose-1,6-bisphosphatase (Fbp1p) degradation; similar to human isopeptidase T
283	eIF3a subunit of the core complex of translation initiation factor 3 (eIF3), essential for translation; part of a subcomplex (Prt1p-Rpg1p-Nip1p) that stimulates binding of mRNA and tRNA(i)Met to ribosomes
347	Polypeptide release factor (eRF1) in translation termination; mutant form acts as a recessive omnipotent suppressor; methylated by Mtq2p-Trm112p in ternary complex eRF1-eRF3-GTP; mutation of methylation site confers resistance to zymocin
352	Protein required for normal prospore membrane formation; interacts with Gip1p, which is the meiosis-specific regulatory subunit of the Glc7p protein phosphatase; expressed specifically in spores and localizes to the prospore membrane
357	Diaminohydroxyphoshoribosylaminopyrimidine deaminase; catalyzes the second step of the riboflavin biosynthesis pathway
363	Microsomal beta-keto-reductase; contains oleate response element (ORE) sequence in the promoter region; mutants exhibit reduced VLCFA synthesis, accumulate high levels of dihydrosphingosine, phytosphingosine and medium-chain ceramides
517	Poly(A+) RNA-binding protein, involved in the export of mRNAs from the nucleus to the cytoplasm; similar to Hrb1p and Npl3p; also binds single-stranded telomeric repeat sequence in vitro
542	Cytoplasmic RNA-binding protein that associates with translating ribosomes; involved in heme regulation of Hap1p as a component of the HMC complex, also involved in the organization of actin filaments; contains a La motif
742	Nucleotide pyrophosphatase/phosphodiesterase family member; mediates extracellular nucleotide phosphate hydrolysis along with Npp1p and Pho5p; activity and expression enhanced during conditions of phosphate starvation
760	Translation elongation factor eIF-5A, previously thought to function in translation initiation; similar to and functionally redundant with Anb1p; structural homolog of bacterial EF-P; undergoes an essential hypusination modification
818	Microtubule-binding protein that together with Kar9p makes up the cortical microtubule capture site and delays the exit from mitosis when the spindle is oriented abnormally
827	Gamma subunit of the translation initiation factor eIF2, involved in the identification of the start codon; binds GTP when forming the ternary complex with GTP and tRNAi-Met
832	Histone chaperone for Htz1p/H2A-H2B dimer; required for the stabilization of the Chz1p-Htz1-H2B complex; has overlapping function with Nap1p; null mutant displays weak sensitivity to MMS and benomyl; contains a highly conserved CHZ motif
865	Conserved nuclear RNA-binding protein; specifically binds to transcribed chromatin in a THO- and RNA-dependent manner, genetically interacts with shuttling hnRNP NAB2; overproduction suppresses transcriptional defect caused by hpr1 mutation
902	Ubiquitin-conjugating enzyme involved in ER-associated protein degradation; located at the cytosolic side of the ER membrane; tail region contains a transmembrane segment at the C-terminus; substrate of the ubiquitin-proteasome pathway
949	Subunit of cohesin loading factor (Scc2p-Scc4p), a complex required for the loading of cohesin complexes onto chromosomes; involved in establishing sister chromatid cohesion during double-strand break repair via phosphorylated histone H2AX
956	Mitochondrial inner membrane insertase, mediates the insertion of both mitochondrial- and nuclear-encoded proteins from the matrix into the inner membrane, interacts with mitochondrial ribosomes; conserved from bacteria to animals
968	Aminophospholipid translocase (flippase) that localizes primarily to the plasma membrane; contributes to endocytosis, protein transport and cell polarity; type 4 P-type ATPase
1073	DNA helicase involved in rDNA replication and Ty1 transposition; relieves replication fork pauses at telomeric regions; structurally and functionally related to Pif1p
1151	Cytochrome c lysine methyltransferase, trimethylates residue 72 of apo-cytochrome c (Cyc1p) in the cytosol; not required for normal respiratory growth
1159	Mitochondrial outer membrane protein with similarity to Tom70p; probable minor component of the TOM (translocase of outer membrane) complex responsible for recognition and import of mitochondrially directed proteins
1197	Protein implicated in Mms22-dependent DNA repair during S phase, DNA damage induces phosphorylation by Mec1p at one or more SQ/TQ motifs; interacts with Mms22p and Slx4p; has four BRCT domains; has a role in regulation of Ty1 transposition
1259	Inosine monophosphate dehydrogenase, catalyzes the first step of GMP biosynthesis, expression is induced by mycophenolic acid resulting in resistance to the drug, expression is repressed by nutrient limitation
1300	Subunit of the CCR4-NOT complex, which is a global transcriptional regulator with roles in transcription initiation and elongation and in mRNA degradation
1305	Microsomal cytochrome b reductase, not essential for viability; also detected in mitochondria; mutation in conserved NADH binding domain of the human ortholog results in type I methemoglobinemia

Continued from previous page.

1341	Zinc knuckle protein, involved in nuclear RNA quality control as a component of the TRAMP complex; stimulates the poly(A) polymerase activity of Pap2p in vitro; functionally redundant with Air2p
1391	Component of the RAM signaling network that is involved in regulation of Ace2p activity and cellular morphogenesis, interacts with protein kinase Cbk1p and also with Kic1p
1399	Protein that associates with ribosomes; putative metalloprotease
1440	Cytoplasmic RNA-binding protein, contains an RNA recognition motif (RRM); may have a role in mRNA translation, as suggested by genetic interactions with genes encoding proteins involved in translational initiation
1447	Subunit of DNA primase, which is required for DNA synthesis and double-strand break repair
1489	N-terminally acetylated protein component of the large (60S) ribosomal subunit, nearly identical to Rpl14Bp and has similarity to rat L14 ribosomal protein; rpl14a csh5 double null mutant exhibits synthetic slow growth
1523	Protein involved in iron metabolism in mitochondria; similar to NifU, which is a protein required for the maturation of the Fe/S clusters of nitrogenase in nitrogen-fixing bacteria
1539	Protein that associates with ribosomes; homolog of translationally controlled tumor protein; green fluorescent protein (GFP)-fusion protein localizes to the cytoplasm and relocates to the mitochondrial outer surface upon oxidative stress
1558	
1848	Putative protein of unknown function; the authentic, non-tagged protein is detected in highly purified mitochondria in high-throughput studies
1930	Basic helix-loop-helix (bHLH) transcription factor of the myc-family; binds cooperatively with Pho2p to the PHO5 promoter; function is regulated by phosphorylation at multiple sites and by phosphate availability
2145	
2161	Essential subunit of the cohesin complex required for sister chromatid cohesion in mitosis and meiosis; apoptosis induces cleavage and translocation of a C-terminal fragment to mitochondria; expression peaks in S phase
2224	Mitochondrial NADP-specific isocitrate dehydrogenase, catalyzes the oxidation of isocitrate to alpha-ketoglutarate; not required for mitochondrial respiration and may function to divert alpha-ketoglutarate to biosynthetic processes
2227	Mitochondrial translational activator of the COB mRNA; membrane protein that interacts with translating ribosomes, acts on the COB mRNA 5'-untranslated leader
2239	Ribosomal stalk protein P1 alpha, involved in the interaction between translational elongation factors and the ribosome; accumulation of P1 in the cytoplasm is regulated by phosphorylation and interaction with the P2 stalk component
2263	Nuclear protein that plays a role in the function of the Smc5p-Rhc18p complex
2334	Zinc knuckle protein, involved in nuclear RNA quality control as a component of the TRAMP complex; stimulates the poly(A) polymerase activity of Pap2p in vitro; functionally redundant with Air1p
2341	Homocitrate synthase isozyme, catalyzes the condensation of acetyl-CoA and alpha-ketoglutarate to form homocitrate, which is the first step in the lysine biosynthesis pathway; highly similar to the other isozyme, Lys21p
2350	Protein component of the large (60S) ribosomal subunit, identical to Rpl35Bp and has similarity to rat L35 ribosomal protein
2407	Protein of unknown function, member of the DUP380 subfamily of conserved, often subtelomerically-encoded proteins; the authentic, non-tagged protein is detected in highly purified mitochondria in high-throughput studies
2476	Ubiquitin isopeptidase, required for recycling ubiquitin from proteasome-bound ubiquitinated intermediates, acts at the late endosome/prevacuolar compartment to recover ubiquitin from ubiquitinated membrane proteins en route to the vacuole
2518	Putative alanine transaminase (glutamic pyruvic transaminase)
2642	Homoaconitase, catalyzes the conversion of homocitrate to homoisocitrate, which is a step in the lysine biosynthesis pathway
2757	Putative GPI-anchored aspartic protease, located in the cytoplasm and endoplasmic reticulum
2758	Mitochondrial inner membrane protein required for assembly of the F0 sector of mitochondrial F1F0 ATP synthase, which is a large, evolutionarily conserved enzyme complex required for ATP synthesis
2759	Protein involved in the transport of cell wall components from the Golgi to the cell surface; required for bud growth
2789	Nuclear protein that binds to RNA and to Mex67p, required for export of poly(A)+ mRNA from the nucleus; member of the REF (RNA and export factor binding proteins) family; another family member, Yra2p, can substitute for Yra1p function
2801	Mitochondrial inner membrane protein required for normal mitochondrial morphology, may be involved in fission of the inner membrane; forms a homo-oligomeric complex
2823	
2923	RNA binding protein that associates with polysomes; proposed to be involved in regulating mRNA translation; involved in the copper-dependent mineralization of copper sulfide complexes on cell surface in cells cultured in copper salts
2935	Protein involved in transcription; interacts with RNA polymerase II subunits Rpb2p, Rpb3, and Rpb11p; has similarity to human RPAP1
2985	Arginyl-tRNA-protein transferase, catalyzes post-translational conjugation of arginine to the amino termini of acceptor proteins which are then subject to degradation via the N-end rule pathway
3017	Translation initiation factor eIF4G, subunit of the mRNA cap-binding protein complex (eIF4F) that also contains eIF4E (Cdc33p); associates with the poly(A)-binding protein Pab1p, also interacts with eIF4A (Tif1p); homologous to Tif4631p
3049	

Continued from previous page.

3157	Protein component of the small (40S) ribosomal subunit; nearly identical to Rps26Bp and has similarity to rat S26 ribosomal protein
3206	Subunit of the heme-activated, glucose-repressed Hap2p/3p/4p/5p CCAAT-binding complex, a transcriptional activator and global regulator of respiratory gene expression; contains sequences sufficient for both complex assembly and DNA binding
3356	Asparagine synthetase, isozyme of Asn1p; catalyzes the synthesis of L-asparagine from L-aspartate in the asparagine biosynthetic pathway
3380	Ribosomal protein L30 of the large (60S) ribosomal subunit, nearly identical to Rpl24Ap and has similarity to rat L24 ribosomal protein; not essential for translation but may be required for normal translation rate
3394	Translation initiation factor eIF4G, subunit of the mRNA cap-binding protein complex (eIF4F) that also contains eIF4E (Cdc33p); associates with the poly(A)-binding protein Pab1p, also interacts with eIF4A (Tif1p); homologous to Tif4632p
3417	Cytoplasmic tyrosyl-tRNA synthetase, required for cytoplasmic protein synthesis; interacts with positions 34 and 35 of the tRNATyr anticodon; mutations in human ortholog YARS are associated with Charcot-Marie-Tooth (CMT) neuropathies
3436	Cytoplasmic trifunctional enzyme C1-tetrahydrofolate synthase, involved in single carbon metabolism and required for biosynthesis of purines, thymidylate, methionine, and histidine
3570	Putative nucleolar DEAD box RNA helicase; high-copy number suppression of a U14 snoRNA processing mutant suggests an involvement in 18S rRNA synthesis
3610	Subunit of the multiprotein cohesin complex required for sister chromatid cohesion in mitotic cells; also required, with Rec8p, for cohesion and recombination during meiosis; phylogenetically conserved SMC chromosomal ATPase family member
3653	Endoplasmic reticulum (ER) resident protein required for ER exit of the high-affinity phosphate transporter Pho84p, specifically required for packaging of Pho84p into COPII vesicles
3725	Protein component of the large (60S) ribosomal subunit, has similarity to rat L39 ribosomal protein; required for ribosome biogenesis; loss of both Rpl31p and Rpl39p confers lethality; also exhibits genetic interactions with SIS1 and PAB1
3765	Beta-adaptin, large subunit of the clathrin associated protein complex (AP-2); involved in vesicle mediated transport; similar to mammalian beta-chain of the clathrin associated protein complex
3819	Small subunit of the clathrin-associated adaptor complex AP-2, which is involved in protein sorting at the plasma membrane; related to the sigma subunit of the mammalian plasma membrane clathrin-associated protein (AP-2) complex
3828	Protein of unknown function, essential for growth under standard (aerobic) conditions but not under anaerobic conditions
3855	Protein component of the large (60S) ribosomal subunit, identical to Rpl43Ap and has similarity to rat L37a ribosomal protein
3950	Mitochondrial matrix protein involved in biogenesis of the iron-sulfur (Fe/S) cluster of Fe/S proteins, isa1 deletion causes loss of mitochondrial DNA and respiratory deficiency; depletion reduces growth on nonfermentable carbon sources
4012	Essential protein involved in 60S ribosome maturation; ortholog of the human protein (SBDS) responsible for autosomal recessive Shwachman-Bodian-Diamond Syndrome; highly conserved across archae and eukaryotes
4140	Protein required for optimal translation under nutrient stress; involved in TOR signaling pathway; binds G4 quadruplex and purine motif triplex nucleic acid; acts with Cdc13p to maintain telomere structure
4182	Dual function protein involved in translation initiation as a substoichiometric component (eIF3j) of translation initiation factor 3 (eIF3) and required for processing of 20S pre-rRNA; binds to eIF3 subunits Rpg1p and Prt1p and 18S rRNA
4322	Component of the exomer complex, which also contains Csh6p, Bch1p, Bch2p, and Bud7p and is involved in export of selected proteins, such as chitin synthase Chs3p, from the Golgi to the plasma membrane
4325	Protein component of the small (40S) ribosomal subunit; nearly identical to Rps25Ap and has similarity to rat S25 ribosomal protein
4363	GDP/GTP exchange protein (GEP) for Rho1p and Rho2p; mutations are synthetically lethal with mutations in rom1, which also encodes a GEP
4380	Protein component of the small (40S) ribosomal subunit; nearly identical to Rps29Bp and has similarity to rat S29 and E. coli S14 ribosomal proteins
4422	Presumed helicase required for RNA polymerase II transcription termination and processing of RNAs; homolog of Senataxin which causes Ataxia-Oculomotor Apraxia 2 and a dominant form of amyotrophic lateral sclerosis
4424	Inosine monophosphate dehydrogenase, catalyzes the first step of GMP biosynthesis, member of a four-gene family in S. cerevisiae, constitutively expressed
4427	Protein with a potential role in pre-rRNA processing
4479	Asn rich cytoplasmic protein that contains RGG motifs; high-copy suppressor of group II intron-splicing defects of a mutation in MRS2 and of a conditional mutation in POL1 (DNA polymerase alpha); possible role in mitochondrial mRNA splicing
4520	Inosine monophosphate dehydrogenase, catalyzes the first step of GMP biosynthesis, member of a four-gene family in S. cerevisiae, constitutively expressed
4537	Lipid-binding protein, localized to the bud via specific mRNA transport; non-tagged protein detected in a phosphorylated state in mitochondria; GFP-fusion protein localizes to the cell periphery; C-termini of Tcb1p, Tcb2p and Tcb3p interact
4648	
4725	Putative integral membrane E3 ubiquitin ligase; acts with Asi2p and Asi3p to ensure the fidelity of SPS-sensor signalling by maintaining the dormant repressed state of gene expression in the absence of inducing signals
4754	eIF3i subunit of the core complex of translation initiation factor 3 (eIF3), which is essential for translation
4873	Translation initiation factor eIF1A, essential protein that forms a complex with Sui1p (eIF1) and the 40S ribosomal subunit and scans for the start codon; C-terminus associates with Fun12p (eIF5B); N terminus interacts with eIF2 and eIF3
4880	Mitochondrial inorganic pyrophosphatase, required for mitochondrial function and possibly involved in energy generation from inorganic pyrophosphate
4926	eIF3c subunit of the eukaryotic translation initiation factor 3 (eIF3), involved in the assembly of preinitiation complex and start codon selection

Continued from previous page.

4949	Poly(A+) RNA-binding protein, involved in the export of mRNAs from the nucleus to the cytoplasm; similar to Gbp2p and Npl3p
4953	Putative integral membrane E3 ubiquitin ligase; acts with Asi1p and Asi2p to ensure the fidelity of SPS-sensor signalling by maintaining the dormant repressed state of gene expression in the absence of inducing signals
4961	Poly (A)+ RNA-binding protein, abundant mRNP-component protein that binds mRNA and is required for stability of many mRNAs; component of glucose deprivation induced stress granules, involved in P-body-dependent granule assembly
4968	Protein that binds to Fpr1p, conferring rapamycin resistance by competing with rapamycin for Fpr1p binding; accumulates in the nucleus upon treatment of cells with rapamycin; has similarity to D. melanogaster shuttle craft and human NFX1
5006	Subunit of tRNA (1-methyladenosine) methyltransferase with Gcd14p, required for the modification of the adenine at position 58 in tRNAs, especially tRNAi-Met; first identified as a negative regulator of GCN4 expression
5107	Cytoplasmic GTPase involved in biogenesis of the 60S ribosome; has similarity to translation elongation factor 2 (Eft1p and Eft2p)
5168	Protein of unknown function; overexpression antagonizes the suppression of splicing defects by spp382 mutants; green fluorescent protein (GFP)-fusion protein localizes to both the cytoplasm and the nucleus
5188	Translation initiation factor eIF1; component of a complex involved in recognition of the initiator codon; modulates translation accuracy at the initiation phase
5206	Catalytic subunit of DNA polymerase (II) epsilon, a chromosomal DNA replication polymerase that exhibits processivity and proofreading exonuclease activity; also involved in DNA synthesis during DNA repair; interacts extensively with Mrc1p
5243	Non-canonical poly(A) polymerase, involved in nuclear RNA quality control as a component of the TRAMP complex; catalyzes polyadenylation of rRNA precursors; overlapping functions with Pap2p
5324	Para hydroxybenzoate
5334	Ubiquitin protease cofactor, forms deubiquitination complex with Ubp3p that coregulates anterograde and retrograde transport between the endoplasmic reticulum and Golgi compartments; null is sensitive to brefeldin A
5393	Mitochondrial glutamyl-tRNA synthetase, predicted to be palmitoylated
5475	Non-canonical poly(A) polymerase, involved in nuclear RNA quality control as a component of the TRAMP complex; catalyzes polyadenylation of unmodified tRNAs, and snoRNA and rRNA precursors; overlapping functions with Trf5p
5483	Subunit of cleavage factor I, a five-subunit complex required for the cleavage and polyadenylation of pre-mRNA 3' ends; RRM-containing heteronuclear RNA binding protein and hnRNPA/B family member that binds to poly (A) signal sequences
5499	Cytoplasmic mRNA cap binding protein and translation initiation factor eIF4E; the eIF4E-cap complex is responsible for mediating cap-dependent mRNA translation via interactions with translation initiation factor eIF4G (Tif4631p or Tif4632p)
5543	Protein with a role in 5'-end processing of mitochondrial RNAs, located in the mitochondrial membrane
5741	Putative protein of unknown function; the authentic protein is detected in highly purified mitochondria in high-throughput studies; null mutant displays reduced frequency of mitochondrial genome loss
5887	High-affinity cyclic AMP phosphodiesterase, component of the cAMP-dependent protein kinase signaling system, protects the cell from extracellular cAMP, contains readthrough motif surrounding termination codon
5888	eIF3b subunit of the core complex of translation initiation factor 3 (eIF3), essential for translation; part of a subcomplex (Prt1p-Rpg1p-Nip1p) that stimulates binding of mRNA and tRNA(i)Met to ribosomes
5907	Transcriptional repressor involved in the control of multidrug resistance; negatively regulates expression of the PDR5 gene; member of the Gal4p family of zinc cluster proteins
6125	Protein kinase involved in regulating diverse events including vesicular trafficking, DNA repair, and chromosome segregation; binds the CTD of RNA pol II; homolog of mammalian casein kinase 1delta (CK1delta)
6148	UDP-glucose
6247	Protein component of the large (60S) ribosomal subunit, identical to Rpl43Bp and has similarity to rat L37a ribosomal protein; null mutation confers a dominant lethal phenotype
6289	Putative protein of unknown function; subunit of the ASTRA complex (Rvb1p, Rvb2p, Tra1p, Tti1p, Tti2, Asa1p and Tra1p) which is part of the chromatin remodeling machinery
6336	Ribosomal protein 28 (rp28) of the small (40S) ribosomal subunit, required for translational accuracy; nearly identical to Rps23Ap and similar to E. coli S12 and rat S23 ribosomal proteins; deletion of both RPS23A and RPS23B is lethal
6349	Asparagine synthetase, isozyme of Asn2p; catalyzes the synthesis of L-asparagine from L-aspartate in the asparagine biosynthetic pathway
6367	Translation initiation factor eIF-4B, has RNA annealing activity; contains an RNA recognition motif and binds to single-stranded RNA
6379	Second largest subunit of DNA polymerase II (DNA polymerase epsilon), required for normal yeast chromosomal replication; expression peaks at the G1/S phase boundary; potential Cdc28p substrate
7385	
7603	Component of the RNA polymerase II general transcription and DNA repair factor TFIIH; involved in transcription initiation and in nucleotide-excision repair; homolog of Chlamydomonas reinhardtii REX1-S protein involved in DNA repair

Table B.5: **Example given in Figures 3.4 b). Of the proteins found in the community at $\log(\lambda) = 0.5$, the proteins in the pink community at $\log(\lambda) = 0.75$.**

189	Subunit of the RNA polymerase II mediator complex; associates with core polymerase subunits to form the RNA polymerase II holoenzyme
316	General transcriptional co-repressor, acts together with Tup1p; also acts as part of a transcriptional co-activator complex that recruits the SWI/SNF and SAGA complexes to promoters; can form the prion [OCT+]
397	Subunit of the RNA polymerase II mediator complex; associates with core polymerase subunits to form the RNA polymerase II holoenzyme; essential for transcriptional regulation
457	Subunit of the RNA polymerase II mediator complex; associates with core polymerase subunits to form the RNA polymerase II holoenzyme; essential for transcriptional regulation
677	Subunit of the RNA polymerase II mediator complex; associates with core polymerase subunits to form the RNA polymerase II holoenzyme; essential for transcriptional regulation; involved in glucose repression
680	General repressor of transcription, forms complex with Cyc8p, involved in the establishment of repressive chromatin structure through interactions with histones H3 and H4, appears to enhance expression of some genes
824	Subunit of the RNA polymerase II mediator complex; associates with core polymerase subunits to form the RNA polymerase II holoenzyme; essential for transcriptional regulation
1083	Subunit of the RNA polymerase II mediator complex; associates with core polymerase subunits to form the RNA polymerase II holoenzyme; general transcription factor involved in telomere maintenance
1100	Subunit of the RNA polymerase II mediator complex; associates with core polymerase subunits to form the RNA polymerase II holoenzyme; essential for transcriptional regulation
2163	Subunit of the RNA polymerase II mediator complex; associates with core polymerase subunits to form the RNA polymerase II holoenzyme; essential for transcriptional regulation
2716	Subunit of the RNA polymerase II mediator complex; associates with core polymerase subunits to form the RNA polymerase II holoenzyme; essential for transcriptional regulation; target of the global repressor Tup1p
2851	Subunit of the RNA polymerase II mediator complex; associates with core polymerase subunits to form the RNA polymerase II holoenzyme; required for stable association of Srb10p-Srb11p kinase; essential for transcriptional regulation
2993	Subunit of the RNA polymerase II mediator complex; associates with core polymerase subunits to form the RNA polymerase II holoenzyme; essential for basal and activated transcription; direct target of Cyc8p-Tup1p transcriptional corepressor
3095	Subunit of the RNA polymerase II mediator complex; associates with core polymerase subunits to form the RNA polymerase II holoenzyme; involved in telomere maintenance; conserved with other metazoan MED31 subunits
3119	Component of the RNA polymerase II mediator complex, which is required for transcriptional activation and also has a role in basal transcription
3336	Subunit of the RNA polymerase II mediator complex; associates with core polymerase subunits to form the RNA polymerase II holoenzyme; essential for transcriptional regulation; involved in telomere maintenance
4061	Subunit of the RNA polymerase II mediator complex; associates with core polymerase subunits to form the RNA polymerase II holoenzyme; required for glucose repression, HO repression, RME1 repression and sporulation
4466	Basic leucine zipper (bZIP) transcription factor required for oxidative stress tolerance; activated by H2O2 through the multistep formation of disulfide bonds and transit from the cytoplasm to the nucleus; mediates resistance to cadmium
4718	Subunit of the RNA polymerase II mediator complex; associates with core polymerase subunits to form the RNA polymerase II holoenzyme; essential protein
4970	Cyclin-like component of the RNA polymerase II holoenzyme, involved in phosphorylation of the RNA polymerase II C-terminal domain; involved in glucose repression and telomere maintenance
5180	Subunit of the RNA polymerase II mediator complex; associates with core polymerase subunits to form the RNA polymerase II holoenzyme; contributes to both postive and negative transcriptional regulation; dispensible for basal transcription
5293	Subunit of the RNA polymerase II mediator complex; associates with core polymerase subunits to form the RNA polymerase II holoenzyme; component of the Med9/10 module; required for regulation of RNA polymerase II activity
5411	Subunit of the RNA polymerase II mediator complex; associates with core polymerase subunits to form the RNA polymerase II holoenzyme; affects transcription by acting as target of activators and repressors
5495	Subunit of the RNA polymerase II mediator complex; associates with core polymerase subunits to form the RNA polymerase II holoenzyme; essential for transcriptional regulation
5539	Dubious opening reading frame unlikely to encode a protein, based on available experimental and comparative sequence data; partially overlaps the uncharacterized gene YOR012C; null mutant displays increased levels of spontaneous Rad52 foci
5666	Transcriptional repressor and activator; involved in repression of flocculation-related genes, and activation of stress responsive genes; negatively regulated by cAMP-dependent protein kinase A subunit Tpk2p
5700	Subunit of the RNA polymerase II mediator complex; associates with core polymerase subunits to form the RNA polymerase II holoenzyme; essential for transcriptional regulation
5963	Cyclin-dependent protein kinase, component of RNA polymerase II holoenzyme; involved in phosphorylation of the RNA polymerase II C-terminal domain; involved in glucose repression
6050	Subunit of TFIID, TFIIF, INO80, SWI/SNF, and NuA3 complexes, involved in RNA polymerase II transcription initiation and in chromatin modification; contains a YEATS domain
6169	DNA-binding transcription factor required for the activation of the GAL genes in response to galactose; repressed by Gal80p and activated by Gal3p
6272	Putative class I histone deacetylase (HDAC) with sequence similarity to Hda1p, Rpd3p, Hos2p, and Hos3p; deletion results in increased histone acetylation at rDNA repeats; interacts with the Tup1p-Ssn6p corepressor complex
6274	Subunit of the RNA polymerase II mediator complex; associates with core polymerase subunits to form the RNA polymerase II holoenzyme; essential for transcriptional regulation
6372	Subunit of the RNA polymerase II mediator complex; associates with core polymerase subunits to form the RNA polymerase II holoenzyme; required for transcriptional activation and has a role in basal transcription

Table B.6: **Example given in Figures 3.4 b). Of the proteins found in the community at $\log(\lambda) = 0.5$, the proteins in the green community at $\log(\lambda) = 0.75$.**

285	Subunit of the SAGA transcriptional regulatory complex, involved in proper assembly of the complex; also present as a C-terminally truncated form in the SLIK/SALSA transcriptional regulatory complex
402	Subunit (90 kDa) of TFIID and SAGA complexes, involved in RNA polymerase II transcription initiation and in chromatin modification
516	Probable subunit of SAGA histone acetyltransferase complex
638	TFIID subunit (150 kDa), involved in RNA polymerase II transcription initiation
678	Protein of unknown function, putative transcriptional regulator; proposed to be a Ada Histone acetyltransferase complex component; GFP tagged protein is localized to the cytoplasm and nucleus
735	Basic leucine zipper (bZIP) transcriptional activator of amino acid biosynthetic genes in response to amino acid starvation; expression is tightly regulated at both the transcriptional and translational levels
817	Long chain fatty acyl-CoA synthetase; accepts a wider range of acyl chain lengths than Faa1p, preferring C9
950	TATA-binding protein, general transcription factor that interacts with other factors to form the preinitiation complex at promoters, essential for viability
966	Nucleosome remodeling factor that functions in regulation of transcription elongation; contains a chromo domain, a helicase domain and a DNA-binding domain; component of both the SAGA and SLIK complexes
1141	Subunit of SAGA and NuA4 histone acetyltransferase complexes; interacts with acidic activators (e.g., Gal4p) which leads to transcription activation; similar to human TRRAP, which is a cofactor for c-Myc mediated oncogenic transformation
1221	Transcription factor, involved in regulating multidrug resistance and oxidative stress response; forms a heterodimer with Pdr1p; contains a Zn(II)2Cys6 zinc finger domain that interacts with a pleiotropic drug resistance element in vitro
2552	Subunit (61/68 kDa) of TFIID and SAGA complexes, involved in RNA polymerase II transcription initiation and in chromatin modification, similar to histone H2A
2553	Transcription factor that activates transcription of genes expressed at the M/G1 phase boundary and in G1 phase; localization to the nucleus occurs during G1 and appears to be regulated by phosphorylation by Cdc28p kinase
2574	Subunit (145 kDa) of TFIID and SAGA complexes, involved in RNA polymerase II transcription initiation and in chromatin modification
2583	Transcriptional regulator involved in glucose repression of Gal4p-regulated genes; component of transcriptional adaptor and histone acetyltransferase complexes, the ADA complex, the SAGA complex, and the SLIK complex
2624	Carbon source-responsive zinc-finger transcription factor, required for transcription of the glucose-repressed gene ADH2, of peroxisomal protein genes, and of genes required for ethanol, glycerol, and fatty acid utilization
2800	Subunit of the SAGA and SAGA-like transcriptional regulatory complexes, interacts with Spt15p to activate transcription of some RNA polymerase II-dependent genes, also functions to inhibit transcription at some promoters
2856	Transcription coactivator, component of the ADA and SAGA transcriptional adaptor/HAT (histone acetyltransferase) complexes
2981	Zinc cluster protein that is a master regulator involved in recruiting other zinc cluster proteins to pleiotropic drug response elements (PDREs) to fine tune the regulation of multidrug resistance genes
3034	Subunit of SAGA histone acetyltransferase complex; involved in formation of the preinitiation complex assembly at promoters; null mutant displays defects in premeiotic DNA synthesis
3080	Subunit (60 kDa) of TFIID and SAGA complexes, involved in transcription initiation of RNA polymerase II and in chromatin modification, similar to histone H4
3506	TFIID subunit (145 kDa), involved in RNA polymerase II transcription initiation, has histone acetyltransferase activity, involved in promoter binding and G1/S progression
4045	Subunit of the SAGA transcriptional regulatory complex but not present in SAGA-like complex SLIK/SALSA, required for SAGA-mediated inhibition at some promoters
4477	TFIID subunit (40 kDa), involved in RNA polymerase II transcription initiation, similar to histone H3 with atypical histone fold motif of Spt3-like transcription factors
4564	TFIID subunit (19 kDa), involved in RNA polymerase II transcription initiation, similar to histone H4 with atypical histone fold motif of Spt3-like transcription factors
4582	TFIID subunit (65 kDa), involved in RNA polymerase II transcription initiation
4607	TFIID subunit (48 kDa), involved in RNA polymerase II transcription initiation; potential Cdc28p substrate
4609	
4621	Protein that binds Sin3p in a two-hybrid assay; contains a Zn(II)2Cys6 zinc finger domain characteristic of DNA-binding proteins; computational analysis suggests a role in regulation of expression of genes encoding transporters
4836	Ubiquitin-specific protease that is a component of the SAGA (Spt-Ada-Gcn5-Acetyltransferase) acetylation complex; required for SAGA-mediated deubiquitination of histone H2B
4840	TFIID subunit (67 kDa), involved in RNA polymerase II transcription initiation
4849	Subunit (17 kDa) of TFIID and SAGA complexes, involved in RNA polymerase II transcription initiation and in chromatin modification, similar to histone H3
5433	Nuclear pore-associated protein, forms a complex with Sac3p that is involved in transcription and in mRNA export from the nucleus; contains a PAM domain implicated in protein-protein binding
5508	Subunit of the SAGA transcriptional regulatory complex, involved in maintaining the integrity of the complex
5549	Subunit of the Ada histone acetyltransferase complex, required for structural integrity of the complex
5932	TFIID subunit (47 kDa), involved in promoter binding and RNA polymerase II transcription initiation
5968	Integral subunit of SAGA histone acetyltransferase complex, regulates transcription of a subset of SAGA-regulated genes, required for the Ubp8p association with SAGA and for H2B deubiquitylation
6175	Adaptor protein required for structural integrity of the SAGA complex, a histone acetyltransferase-coactivator complex that is involved in global regulation of gene expression through acetylation and transcription functions
28510	Protein involved in mRNA export coupled transcription activation and elongation; component of both the SAGA histone acetylase complex and TREX-2, and interacts with RNAPII

Table B.7: **Example given in Figures 3.4 b). Of the proteins found in the community at** $\log(\lambda) = 0.5$**, the proteins in neither the pink nor the green community at** $\log(\lambda) = 0.75$**.**

223	UDP-glucose-4-epimerase, catalyzes the interconversion of UDP-galactose and UDP-D-glucose in galactose metabolism; also catalyzes the conversion of alpha-D-glucose or alpha-D-galactose to their beta-anomers
598	Citrate synthase, catalyzes the condensation of acetyl coenzyme A and oxaloacetate to form citrate, peroxisomal isozyme involved in glyoxylate cycle; expression is controlled by Rtg1p and Rtg2p transcription factors
1770	TFIIE small subunit, involved in RNA polymerase II transcription initiation
2266	Serine/threonine protein kinase, subunit of the transcription factor TFIIH; involved in transcription initiation at RNA polymerase II promoters
2779	Protein similar to Ashbya gossypii sporulation-specific chitinase
3221	Sensor of mitochondrial dysfunction; regulates the subcellular location of Rtg1p and Rtg3p, transcriptional activators of the retrograde (RTG) and TOR pathways; Rtg2p is inhibited by the phosphorylated form of Mks1p
3484	Histone acetyltransferase, acetylates N-terminal lysines on histones H2B and H3; catalytic subunit of the ADA and SAGA histone acetyltransferase complexes; founding member of the Gcn5p-related N-acetyltransferase superfamily
3629	Outward-rectifier potassium channel of the plasma membrane with two pore domains in tandem, each of which forms a functional channel permeable to potassium; carboxy tail functions to prevent inner gate closures; target of K1 toxin
4103	Mitogen-activated protein kinase involved in osmoregulation via three independent osmosensors; mediates the recruitment and activation of RNA Pol II at Hot1p-dependent promoters; localization regulated by Ptp2p and Ptp3p
4305	Protein involved in shmoo formation and bipolar bud site selection; homologous to Spa2p, localizes to sites of polarized growth in a cell cycle dependent- and Spa2p-dependent manner, interacts with MAPKKs Mkk1p, Mkk2p, and Ste7p
5011	Protein component of the large (60S) ribosomal subunit, nearly identical to Rpl9Ap and has similarity to E. coli L6 and rat L9 ribosomal proteins
5111	Basic leucine zipper (bZIP) transcription factor of the ATF/CREB family, forms a complex with Tup1p and Ssn6p to both activate and repress transcription; cytosolic and nuclear protein involved in osmotic and oxidative stress responses
5452	
6037	Trichostatin A-insensitive homodimeric histone deacetylase (HDAC) with specificity in vitro for histones H3, H4, H2A, and H2B; similar to Hda1p, Rpd3p, Hos1p, and Hos2p; deletion results in increased histone acetylation at rDNA repeats
6043	Subunit of TFIIH and nucleotide excision repair factor 3 complexes, involved in transcription initiation, required for nucleotide excision repair, similar to 52 kDa subunit of human TFIIH
6260	Subunit of TFIIH complex, involved in transcription initiation, similar to 34 kDa subunit of human TFIIH; interacts with Ssl1p
7251	Ubiquitin-like protein modifier, may function in modification of Sph1p and Hbt1p, functionally complemented by the human or S. pombe ortholog; mechanism of Hub1p adduct formation not yet clear
7603	Component of the RNA polymerase II general transcription and DNA repair factor TFIIH; involved in transcription initiation and in nucleotide-excision repair; homolog of Chlamydomonas reinhardtii REX1-S protein involved in DNA repair

Table B.8: **Example given in Figures 3.4 b). Of the proteins found in the community at** $\log(\lambda) = 0.5$**, the proteins in the pink community at** $\log(\lambda) = 1.6$**.**

189	Subunit of the RNA polymerase II mediator complex; associates with core polymerase subunits to form the RNA polymerase II holoenzyme
397	Subunit of the RNA polymerase II mediator complex; associates with core polymerase subunits to form the RNA polymerase II holoenzyme; essential for transcriptional regulation
677	Subunit of the RNA polymerase II mediator complex; associates with core polymerase subunits to form the RNA polymerase II holoenzyme; essential for transcriptional regulation; involved in glucose repression
824	Subunit of the RNA polymerase II mediator complex; associates with core polymerase subunits to form the RNA polymerase II holoenzyme; essential for transcriptional regulation
1083	Subunit of the RNA polymerase II mediator complex; associates with core polymerase subunits to form the RNA polymerase II holoenzyme; general transcription factor involved in telomere maintenance
1100	Subunit of the RNA polymerase II mediator complex; associates with core polymerase subunits to form the RNA polymerase II holoenzyme; essential for transcriptional regulation
2163	Subunit of the RNA polymerase II mediator complex; associates with core polymerase subunits to form the RNA polymerase II holoenzyme; essential for transcriptional regulation
2716	Subunit of the RNA polymerase II mediator complex; associates with core polymerase subunits to form the RNA polymerase II holoenzyme; essential for transcriptional regulation; target of the global repressor Tup1p
2851	Subunit of the RNA polymerase II mediator complex; associates with core polymerase subunits to form the RNA polymerase II holoenzyme; required for stable association of Srb10p-Srb11p kinase; essential for transcriptional regulation
2993	Subunit of the RNA polymerase II mediator complex; associates with core polymerase subunits to form the RNA polymerase II holoenzyme; essential for basal and activated transcription; direct target of Cyc8p-Tup1p transcriptional corepressor
3095	Subunit of the RNA polymerase II mediator complex; associates with core polymerase subunits to form the RNA polymerase II holoenzyme; involved in telomere maintenance; conserved with other metazoan MED31 subunits
3119	Component of the RNA polymerase II mediator complex, which is required for transcriptional activation and also has a role in basal transcription
3336	Subunit of the RNA polymerase II mediator complex; associates with core polymerase subunits to form the RNA polymerase II holoenzyme; essential for transcriptional regulation; involved in telomere maintenance
4061	Subunit of the RNA polymerase II mediator complex; associates with core polymerase subunits to form the RNA polymerase II holoenzyme; required for glucose repression, HO repression, RME1 repression and sporulation
4466	Basic leucine zipper (bZIP) transcription factor required for oxidative stress tolerance; activated by H2O2 through the multistep formation of disulfide bonds and transit from the cytoplasm to the nucleus; mediates resistance to cadmium
4718	Subunit of the RNA polymerase II mediator complex; associates with core polymerase subunits to form the RNA polymerase II holoenzyme; essential protein
5180	Subunit of the RNA polymerase II mediator complex; associates with core polymerase subunits to form the RNA polymerase II holoenzyme; contributes to both postive and negative transcriptional regulation; dispensible for basal transcription
5293	Subunit of the RNA polymerase II mediator complex; associates with core polymerase subunits to form the RNA polymerase II holoenzyme; component of the Med9/10 module; required for regulation of RNA polymerase II activity
5411	Subunit of the RNA polymerase II mediator complex; associates with core polymerase subunits to form the RNA polymerase II holoenzyme; affects transcription by acting as target of activators and repressors
5495	Subunit of the RNA polymerase II mediator complex; associates with core polymerase subunits to form the RNA polymerase II holoenzyme; essential for transcriptional regulation
5539	Dubious opening reading frame unlikely to encode a protein, based on available experimental and comparative sequence data; partially overlaps the uncharacterized gene YOR012C; null mutant displays increased levels of spontaneous Rad52 foci
5700	Subunit of the RNA polymerase II mediator complex; associates with core polymerase subunits to form the RNA polymerase II holoenzyme; essential for transcriptional regulation
5963	Cyclin-dependent protein kinase, component of RNA polymerase II holoenzyme; involved in phosphorylation of the RNA polymerase II C-terminal domain; involved in glucose repression
6274	Subunit of the RNA polymerase II mediator complex; associates with core polymerase subunits to form the RNA polymerase II holoenzyme; essential for transcriptional regulation
6372	Subunit of the RNA polymerase II mediator complex; associates with core polymerase subunits to form the RNA polymerase II holoenzyme; required for transcriptional activation and has a role in basal transcription

Table B.9: **Example given in Figures 3.4 b). Of the proteins found in the community at $\log(\lambda) = 0.5$, the proteins in the green community at $\log(\lambda) = 1.6$.**

285	Subunit of the SAGA transcriptional regulatory complex, involved in proper assembly of the complex; also present as a C-terminally truncated form in the SLIK/SALSA transcriptional regulatory complex
516	Probable subunit of SAGA histone acetyltransferase complex
966	Nucleosome remodeling factor that functions in regulation of transcription elongation; contains a chromo domain, a helicase domain and a DNA-binding domain; component of both the SAGA and SLIK complexes
2574	Subunit (145 kDa) of TFIID and SAGA complexes, involved in RNA polymerase II transcription initiation and in chromatin modification
2583	Transcriptional regulator involved in glucose repression of Gal4p-regulated genes; component of transcriptional adaptor and histone acetyltransferase complexes, the ADA complex, the SAGA complex, and the SLIK complex
2800	Subunit of the SAGA and SAGA-like transcriptional regulatory complexes, interacts with Spt15p to activate transcription of some RNA polymerase II-dependent genes, also functions to inhibit transcription at some promoters
2856	Transcription coactivator, component of the ADA and SAGA transcriptional adaptor/HAT (histone acetyltransferase) complexes
3034	Subunit of SAGA histone acetyltransferase complex; involved in formation of the preinitiation complex assembly at promoters; null mutant displays defects in premeiotic DNA synthesis
4045	Subunit of the SAGA transcriptional regulatory complex but not present in SAGA-like complex SLIK/SALSA, required for SAGA-mediated inhibition at some promoters
4609	
4836	Ubiquitin-specific protease that is a component of the SAGA (Spt-Ada-Gcn5-Acetyltransferase) acetylation complex; required for SAGA-mediated deubiquitination of histone H2B
4849	Subunit (17 kDa) of TFIID and SAGA complexes, involved in RNA polymerase II transcription initiation and in chromatin modification, similar to histone H3
5508	Subunit of the SAGA transcriptional regulatory complex, involved in maintaining the integrity of the complex
5968	Integral subunit of SAGA histone acetyltransferase complex, regulates transcription of a subset of SAGA-regulated genes, required for the Ubp8p association with SAGA and for H2B deubiquitylation
6175	Adaptor protein required for structural integrity of the SAGA complex, a histone acetyltransferase-coactivator complex that is involved in global regulation of gene expression through acetylation and transcription functions
28510	Protein involved in mRNA export coupled transcription activation and elongation; component of both the SAGA histone acetylase complex and TREX-2, and interacts with RNAPII

Table B.10: **Example given in Figures 3.4 b). Of the proteins found in the community at $\log(\lambda) = 0.5$, the proteins in the orange community at $\log(\lambda) = 1.6$.**

402	Subunit (90 kDa) of TFIID and SAGA complexes, involved in RNA polymerase II transcription initiation and in chromatin modification
638	TFIID subunit (150 kDa), involved in RNA polymerase II transcription initiation
3506	TFIID subunit (145 kDa), involved in RNA polymerase II transcription initiation, has histone acetyltransferase activity, involved in promoter binding and G1/S progression
4477	TFIID subunit (40 kDa), involved in RNA polymerase II transcription initiation, similar to histone H3 with atypical histone fold motif of Spt3-like transcription factors
4564	TFIID subunit (19 kDa), involved in RNA polymerase II transcription initiation, similar to histone H4 with atypical histone fold motif of Spt3-like transcription factors
4582	TFIID subunit (65 kDa), involved in RNA polymerase II transcription initiation
4607	TFIID subunit (48 kDa), involved in RNA polymerase II transcription initiation; potential Cdc28p substrate
4840	TFIID subunit (67 kDa), involved in RNA polymerase II transcription initiation
5932	TFIID subunit (47 kDa), involved in promoter binding and RNA polymerase II transcription initiation

Table B.11: **Example given in Figures 3.4 b). Of the proteins found in the community at $\log(\lambda) = 0.5$, the proteins in neither the pink, green nor orange communities at $\log(\lambda) = 1.6$.**

223	UDP-glucose-4-epimerase, catalyzes the interconversion of UDP-galactose and UDP-D-glucose in galactose metabolism; also catalyzes the conversion of alpha-D-glucose or alpha-D-galactose to their beta-anomers
316	General transcriptional co-repressor, acts together with Tup1p; also acts as part of a transcriptional co-activator complex that recruits the SWI/SNF and SAGA complexes to promoters; can form the prion [OCT+]
457	Subunit of the RNA polymerase II mediator complex; associates with core polymerase subunits to form the RNA polymerase II holoenzyme; essential for transcriptional regulation
598	Citrate synthase, catalyzes the condensation of acetyl coenzyme A and oxaloacetate to form citrate, peroxisomal isozyme involved in glyoxylate cycle; expression is controlled by Rtg1p and Rtg2p transcription factors
678	Protein of unknown function, putative transcriptional regulator; proposed to be a Ada Histone acetyltransferase complex component; GFP tagged protein is localized to the cytoplasm and nucleus
680	General repressor of transcription, forms complex with Cyc8p, involved in the establishment of repressive chromatin structure through interactions with histones H3 and H4, appears to enhance expression of some genes
735	Basic leucine zipper (bZIP) transcriptional activator of amino acid biosynthetic genes in response to amino acid starvation; expression is tightly regulated at both the transcriptional and translational levels
817	Long chain fatty acyl-CoA synthetase; accepts a wider range of acyl chain lengths than Faa1p, preferring C9
950	TATA-binding protein, general transcription factor that interacts with other factors to form the preinitiation complex at promoters, essential for viability
1141	Subunit of SAGA and NuA4 histone acetyltransferase complexes; interacts with acidic activators (e.g., Gal4p) which leads to transcription activation; similar to human TRRAP, which is a cofactor for c-Myc mediated oncogenic transformation
1221	Transcription factor, involved in regulating multidrug resistance and oxidative stress response; forms a heterodimer with Pdr1p; contains a Zn(II)2Cys6 zinc finger domain that interacts with a pleiotropic drug resistance element in vitro
1770	TFIIE small subunit, involved in RNA polymerase II transcription initiation
2266	Serine/threonine protein kinase, subunit of the transcription factor TFIIH; involved in transcription initiation at RNA polymerase II promoters
2552	Subunit (61/68 kDa) of TFIID and SAGA complexes, involved in RNA polymerase II transcription initiation and in chromatin modification, similar to histone H2A
2553	Transcription factor that activates transcription of genes expressed at the M/G1 phase boundary and in G1 phase; localization to the nucleus occurs during G1 and appears to be regulated by phosphorylation by Cdc28p kinase
2624	Carbon source-responsive zinc-finger transcription factor, required for transcription of the glucose-repressed gene ADH2, of peroxisomal protein genes, and of genes required for ethanol, glycerol, and fatty acid utilization
2779	Protein similar to Ashbya gossypii sporulation-specific chitinase
2981	Zinc cluster protein that is a master regulator involved in recruiting other zinc cluster proteins to pleiotropic drug response elements (PDREs) to fine tune the regulation of multidrug resistance genes
3080	Subunit (60 kDa) of TFIID and SAGA complexes, involved in transcription initiation of RNA polymerase II and in chromatin modification, similar to histone H4
3221	Sensor of mitochondrial dysfunction; regulates the subcellular location of Rtg1p and Rtg3p, transcriptional activators of the retrograde (RTG) and TOR pathways; Rtg2p is inhibited by the phosphorylated form of Mks1p
3484	Histone acetyltransferase, acetylates N-terminal lysines on histones H2B and H3; catalytic subunit of the ADA and SAGA histone acetyltransferase complexes; founding member of the Gcn5p-related N-acetyltransferase superfamily
3629	Outward-rectifier potassium channel of the plasma membrane with two pore domains in tandem, each of which forms a functional channel permeable to potassium; carboxy tail functions to prevent inner gate closures; target of K1 toxin
4103	Mitogen-activated protein kinase involved in osmoregulation via three independent osmosensors; mediates the recruitment and activation of RNA Pol II at Hot1p-dependent promoters; localization regulated by Ptp2p and Ptp3p
4305	Protein involved in shmoo formation and bipolar bud site selection; homologous to Spa2p, localizes to sites of polarized growth in a cell cycle dependent- and Spa2p-dependent manner, interacts with MAPKKs Mkk1p, Mkk2p, and Ste7p
4621	Protein that binds Sin3p in a two-hybrid assay; contains a Zn(II)2Cys6 zinc finger domain characteristic of DNA-binding proteins; computational analysis suggests a role in regulation of expression of genes encoding transporters
4970	Cyclin-like component of the RNA polymerase II holoenzyme, involved in phosphorylation of the RNA polymerase II C-terminal domain; involved in glucose repression and telomere maintenance
5011	Protein component of the large (60S) ribosomal subunit, nearly identical to Rpl9Ap and has similarity to E. coli L6 and rat L9 ribosomal proteins
5111	Basic leucine zipper (bZIP) transcription factor of the ATF/CREB family, forms a complex with Tup1p and Ssn6p to both activate and repress transcription; cytosolic and nuclear protein involved in osmotic and oxidative stress responses
5433	Nuclear pore-associated protein, forms a complex with Sac3p that is involved in transcription and in mRNA export from the nucleus; contains a PAM domain implicated in protein-protein binding
5452	
5549	Subunit of the Ada histone acetyltransferase complex, required for structural integrity of the complex
5666	Transcriptional repressor and activator; involved in repression of flocculation-related genes, and activation of stress responsive genes; negatively regulated by cAMP-dependent protein kinase A subunit Tpk2p
6037	Trichostatin A-insensitive homodimeric histone deacetylase (HDAC) with specificity in vitro for histones H3, H4, H2A, and H2B; similar to Hda1p, Rpd3p, Hos1p, and Hos2p; deletion results in increased histone acetylation at rDNA repeats
6043	Subunit of TFIIH and nucleotide excision repair factor 3 complexes, involved in transcription initiation, required for nucleotide excision repair, similar to 52 kDa subunit of human TFIIH
6050	Subunit of TFIID, TFIIF, INO80, SWI/SNF, and NuA3 complexes, involved in RNA polymerase II transcription initiation and in chromatin modification; contains a YEATS domain
6169	DNA-binding transcription factor required for the activation of the GAL genes in response to galactose; repressed by Gal80p and activated by Gal3p
6260	Subunit of TFIIH complex, involved in transcription initiation, similar to 34 kDa subunit of human TFIIH; interacts with Ssl1p
6272	Putative class I histone deacetylase (HDAC) with sequence similarity to Hda1p, Rpd3p, Hos2p, and Hos3p; deletion results in increased histone acetylation at rDNA repeats; interacts with the Tup1p-Ssn6p corepressor complex
7251	Ubiquitin-like protein modifier, may function in modification of Sph1p and Hbt1p, functionally complemented by the human or S. pombe ortholog; mechanism of Hub1p adduct formation not yet clear
7603	Component of the RNA polymerase II general transcription and DNA repair factor TFIIH; involved in transcription initiation and in nucleotide-excision repair; homolog of Chlamydomonas reinhardtii REX1-S protein involved in DNA repair

Appendix C

Examples of community membership following individual proteins

The proteins found in some of the communities discussed in Section 3.9. For each of the five proteins selected as examples in that section, we illustrate the proteins they co-occur with in some example communities. Protein numbers are the SGD identification numbers (Saccharomyces Genome Database, `www.yeastgenome.org`, [53]), the short descriptions are given on the SGD website.

Table C.1: **From Figure 3.13: Protein 514 (YCL008C), component of the ESCRT-I complex, which is involved in ubiquitin-dependent sorting of proteins into the endosome,** $\log(\lambda) = 2$ **in the** A **network**

177	
514	Component of the ESCRT-I complex, which is involved in ubiquitin-dependent sorting of proteins into the endosome; homologous to the mouse and human Tsg101 tumor susceptibility gene; mutants exhibit a Class E Vps phenotype
661	Forkhead transcription factor that drives S-phase specific expression of genes involved in chromosome segregation, spindle dynamics, and budding; suppressor of calmodulin mutants with specific SPB assembly defects; telomere maintenance role
1485	Class E Vps protein of the ESCRT-III complex, required for sorting of integral membrane proteins into lumenal vesicles of multivesicular bodies, and for delivery of newly synthesized vacuolar enzymes to the vacuole, involved in endocytosis
1524	One of four subunits of the endosomal sorting complex required for transport III (ESCRT-III); forms an ESCRT-III subcomplex with Did4p; involved in the sorting of transmembrane proteins into the multivesicular body (MVB) pathway
2894	Cytoplasmic and vacuolar membrane protein involved in late endosome to vacuole transport; required for normal filament maturation during pseudohyphal growth; may function in targeting cargo proteins for degradation; interacts with Vta1p
3438	ESCRT-I subunit required to stabilize oligomers of the ESCRT-I core complex (Stp22p, Vps28p, Srn2p), which is involved in ubiquitin-dependent sorting of proteins into the endosome; deletion mutant is sensitive to rapamycin and nystatin
3708	Vacuolar carboxypeptidase yscS; expression is induced under low-nitrogen conditions
3863	Component of the ESCRT-II complex, which is involved in ubiquitin-dependent sorting of proteins into the endosome
4015	One of four subunits of the endosomal sorting complex required for transport III (ESCRT-III); involved in the sorting of transmembrane proteins into the multivesicular body (MVB) pathway; recruited from the cytoplasm to endosomal membranes
4109	Component of the ESCRT-I complex, which is involved in ubiquitin-dependent sorting of proteins into the endosome; suppressor of rna1-1 mutation; may be involved in RNA export from nucleus
4171	Multivesicular body (MVB) protein involved in endosomal protein sorting; regulates Vps4p activity by promoting its oligomerization; has an N-terminal Vps60- and Did2- binding domain, a linker region, and a C-terminal Vps4p binding domain
4409	Component of the ESCRT-II complex; contains the GLUE (GRAM Like Ubiquitin binding in EAP45) domain which is involved in interactions with ESCRT-I and ubiquitin-dependent sorting of proteins into the endosome
4682	Myristoylated subunit of ESCRTIII, the endosomal sorting complex required for transport of transmembrane proteins into the multivesicular body pathway to the lysosomal/vacuolar lumen; cytoplasmic protein recruited to endosomal membranes
5923	Component of the ESCRT-II complex, which is involved in ubiquitin-dependent sorting of proteins into the endosome; appears to be functionally related to SNF7; involved in glucose derepression
5986	Component of the ESCRT-I complex (Stp22p, Srn2p, Vps28p, and Mvb12p), which is involved in ubiquitin-dependent sorting of proteins into the endosome; conserved C-terminal domain interacts with ESCRT-III subunit Vps20p
6005	Cytoplasmic class E vacuolar protein sorting (VPS) factor that coordinates deubiquitination in the multivesicular body (MVB) pathway by recruiting Doa4p to endosomes
6377	AAA-ATPase involved in multivesicular body (MVB) protein sorting, ATP-bound Vps4p localizes to endosomes and catalyzes ESCRT-III disassembly and membrane release; ATPase activity is activated by Vta1p; regulates cellular sterol metabolism
6435	Class E protein of the vacuolar protein-sorting (Vps) pathway; binds Vps4p and directs it to dissociate ESCRT-III complexes; forms a functional and physical complex with Ist1p; human ortholog may be altered in breast tumors

Table C.2: **From Figure 3.13: Protein 514 (YCL008C),** $\log(\lambda) = 1$ **in the** *P* **network**

401	
466	Protein of unknown function; the authentic, non-tagged protein is detected in purified mitochondria in high-throughput studies; null mutant displays elevated frequency of mitochondrial genome loss
514	Component of the ESCRT-I complex, which is involved in ubiquitin-dependent sorting of proteins into the endosome; homologous to the mouse and human Tsg101 tumor susceptibility gene; mutants exhibit a Class E Vps phenotype
554	
783	
800	
894	Protein that associates with the INO80 chromatin remodeling complex under low-salt conditions
923	
930	
1019	Transcriptional repressor involved in response to pH and in cell wall construction; required for alkaline pH-stimulated haploid invasive growth and sporulation; activated by proteolytic processing; similar to A. nidulans PacC
1485	Class E Vps protein of the ESCRT-III complex, required for sorting of integral membrane proteins into lumenal vesicles of multivesicular bodies, and for delivery of newly synthesized vacuolar enzymes to the vacuole, involved in endocytosis
1524	One of four subunits of the endosomal sorting complex required for transport III (ESCRT-III); forms an ESCRT-III subcomplex with Did4p; involved in the sorting of transmembrane proteins into the multivesicular body (MVB) pathway
2476	Ubiquitin isopeptidase, required for recycling ubiquitin from proteasome-bound ubiquitinated intermediates, acts at the late endosome/prevacuolar compartment to recover ubiquitin from ubiquitinated membrane proteins en route to the vacuole
2698	
2894	Cytoplasmic and vacuolar membrane protein involved in late endosome to vacuole transport; required for normal filament maturation during pseudohyphal growth; may function in targeting cargo proteins for degradation; interacts with Vta1p
2949	
3227	Protein of unknown function; highly induced in zinc-depleted conditions and has increased expression in NAP1 deletion mutants
3354	
3438	ESCRT-I subunit required to stabilize oligomers of the ESCRT-I core complex (Stp22p, Vps28p, Srn2p), which is involved in ubiquitin-dependent sorting of proteins into the endosome; deletion mutant is sensitive to rapamycin and nystatin
3592	Zinc-regulated transcription factor, binds to zinc-responsive promoter elements to induce transcription of certain genes in the presence of zinc; regulates its own transcription; contains seven zinc-finger domains
3863	Component of the ESCRT-II complex, which is involved in ubiquitin-dependent sorting of proteins into the endosome
4015	One of four subunits of the endosomal sorting complex required for transport III (ESCRT-III); involved in the sorting of transmembrane proteins into the multivesicular body (MVB) pathway; recruited from the cytoplasm to endosomal membranes
4063	Protein that inhibits Doa4p deubiquitinating activity; contributes to ubiquitin homeostasis by regulating the conversion of free ubiquitin chains to ubiquitin monomers by Doa4p; GFP-fusion protein localizes to endosomes
4109	Component of the ESCRT-I complex, which is involved in ubiquitin-dependent sorting of proteins into the endosome; suppressor of rna1-1 mutation; may be involved in RNA export from nucleus
4171	Multivesicular body (MVB) protein involved in endosomal protein sorting; regulates Vps4p activity by promoting its oligomerization; has an N-terminal Vps60- and Did2- binding domain, a linker region, and a C-terminal Vps4p binding domain
4409	Component of the ESCRT-II complex; contains the GLUE (GRAM Like Ubiquitin binding in EAP45) domain which is involved in interactions with ESCRT-I and ubiquitin-dependent sorting of proteins into the endosome
4682	Myristoylated subunit of ESCRTIII, the endosomal sorting complex required for transport of transmembrane proteins into the multivesicular body pathway to the lysosomal/vacuolar lumen; cytoplasmic protein recruited to endosomal membranes
5209	Protein with a positive role in the multivesicular body sorting pathway; functions and forms a complex with Did2p; recruitment to endosomes is mediated by the Vps2p-Vps24p subcomplex of ESCRT-III; also interacts with Vps4p
5801	Protein involved in proteolytic activation of Rim101p in response to alkaline pH; PalA/AIP1/Alix family member; interaction with the ESCRT-III subunit Snf7p suggests a relationship between pH response and multivesicular body formation
5923	Component of the ESCRT-II complex, which is involved in ubiquitin-dependent sorting of proteins into the endosome; appears to be functionally related to SNF7; involved in glucose derepression
5986	Component of the ESCRT-I complex (Stp22p, Srn2p, Vps28p, and Mvb12p), which is involved in ubiquitin-dependent sorting of proteins into the endosome; conserved C-terminal domain interacts with ESCRT-III subunit Vps20p
6005	Cytoplasmic class E vacuolar protein sorting (VPS) factor that coordinates deubiquitination in the multivesicular body (MVB) pathway by recruiting Doa4p to endosomes
6377	AAA-ATPase involved in multivesicular body (MVB) protein sorting, ATP-bound Vps4p localizes to endosomes and catalyzes ESCRT-III disassembly and membrane release; ATPase activity is activated by Vta1p; regulates cellular sterol metabolism
6435	Class E protein of the vacuolar protein-sorting (Vps) pathway; binds Vps4p and directs it to dissociate ESCRT-III complexes; forms a functional and physical complex with Ist1p; human ortholog may be altered in breast tumors

Table C.3: **From Figure 3.13: Protein 514 (YCL008C),** $\log(\lambda) = 2$ **in the** P **network**

.

514	Component of the ESCRT-I complex, which is involved in ubiquitin-dependent sorting of proteins into the endosome; homologous to the mouse and human Tsg101 tumor susceptibility gene; mutants exhibit a Class E Vps phenotype
3438	ESCRT-I subunit required to stabilize oligomers of the ESCRT-I core complex (Stp22p, Vps28p, Srn2p), which is involved in ubiquitin-dependent sorting of proteins into the endosome; deletion mutant is sensitive to rapamycin and nystatin
3863	Component of the ESCRT-II complex, which is involved in ubiquitin-dependent sorting of proteins into the endosome
4109	Component of the ESCRT-I complex, which is involved in ubiquitin-dependent sorting of proteins into the endosome; suppressor of rna1-1 mutation; may be involved in RNA export from nucleus
4409	Component of the ESCRT-II complex; contains the GLUE (GRAM Like Ubiquitin binding in EAP45) domain which is involved in interactions with ESCRT-I and ubiquitin-dependent sorting of proteins into the endosome
4682	Myristoylated subunit of ESCRTIII, the endosomal sorting complex required for transport of transmembrane proteins into the multivesicular body pathway to the lysosomal/vacuolar lumen; cytoplasmic protein recruited to endosomal membranes
5923	Component of the ESCRT-II complex, which is involved in ubiquitin-dependent sorting of proteins into the endosome; appears to be functionally related to SNF7; involved in glucose derepression
5986	Component of the ESCRT-I complex (Stp22p, Srn2p, Vps28p, and Mvb12p), which is involved in ubiquitin-dependent sorting of proteins into the endosome; conserved C-terminal domain interacts with ESCRT-III subunit Vps20p

Table C.4: **From Figure 3.14 a): Protein 2 (YAL002W), Membrane-associated protein that interacts with Vps21p to facilitate soluble vacuolar protein localization; component of the CORVET complex; required for localization and trafficking of the CPY sorting receptor; contains RING finger motif, $\log(\lambda) = 1$ in the A network**

2	Membrane-associated protein that interacts with Vps21p to facilitate soluble vacuolar protein localization; component of the CORVET complex; required for localization and trafficking of the CPY sorting receptor; contains RING finger motif
12	Endosomal SNARE related to mammalian syntaxin 8
146	Peripheral membrane protein required for vesicular transport between ER and Golgi and for the 'priming' step in homotypic vacuole fusion, part of the cis-SNARE complex; has similarity to alpha-SNAP
284	ATPase required for the release of Sec17p during the 'priming' step in homotypic vacuole fusion and for ER to Golgi transport; homolog of the mammalian NSF
335	Protein involved in vacuolar assembly, essential for autophagy and the cytoplasm-to-vacuole pathway
404	Protein containing SH3-domains, involved in establishing cell polarity and morphogenesis; functions as a scaffold protein for complexes that include Cdc24p, Ste5p, Ste20p, and Rsr1p
492	Mu3-like subunit of the clathrin associated protein complex (AP-3); functions in transport of alkaline phosphatase to the vacuole via the alternate pathway
1003	5-phospho-ribosyl-1(alpha)-pyrophosphate synthetase, synthesizes PRPP, which is required for nucleotide, histidine, and tryptophan biosynthesis; one of five related enzymes, which are active as heteromultimeric complexes
1679	Vesicle membrane protein (v-SNARE) with acyltransferase activity; involved in trafficking to and within the Golgi, endocytic trafficking to the vacuole, and vacuolar fusion; membrane localization due to prenylation at the carboxy-terminus
2235	Vacuolar protein that plays a critical role in the tethering steps of vacuolar membrane fusion by facilitating guanine nucleotide exchange on small guanosine triphosphatase Ypt7p
2487	Vacuolar membrane protein that is a subunit of the homotypic vacuole fusion and vacuole protein sorting (HOPS) complex; essential for membrane docking and fusion at the Golgi-to-endosome and endosome-to-vacuole stages of protein transport
2597	Hydrophilic protein involved in vesicle trafficking between the ER and Golgi; SM (Sec1/Munc-18) family protein that binds the tSNARE Sed5p and stimulates its assembly into a trans-SNARE membrane-protein complex
2672	Palmitoyl transferase involved in protein palmitoylation; acts as a negative regulator of pheromone response pathway; required for endocytosis of pheromone receptors; involved in cell shape control; contains ankyrin repeats
2731	Multivalent adaptor protein that facilitates vesicle-mediated vacuolar protein sorting by ensuring high-fidelity vesicle docking and fusion, which are essential for targeting of vesicles to the endosome; required for vacuole inheritance
2903	Component of CORVET tethering complex; cytoplasmic protein required for the sorting and processing of soluble vacuolar proteins, acidification of the vacuolar lumen, and assembly of the vacuolar H+-ATPase
2906	Membrane glycoprotein v-SNARE involved in retrograde transport from the Golgi to the ER; required for N- and O-glycosylation in the Golgi but not in the ER; forms a complex with the cytosolic Tip20p
3063	Protein of the Sec1p/Munc-18 family, essential for vacuolar protein sorting; required for the function of Pep12p and the early endosome/late Golgi SNARE Tlg2p; essential for fusion of Golgi-derived vesicles with the prevacuolar compartment
3066	Essential SNARE protein localized to the ER, involved in retrograde traffic from the Golgi to the ER; forms a complex with the SNAREs Sec22p, Sec20p and Ufe1p
3092	Protein required for fusion of cvt-vesicles and autophagosomes with the vacuole; associates, as a complex with Ccz1p, with a perivacuolar compartment; potential Cdc28p substrate
3113	Peripheral membrane protein required for fusion of COPI vesicles with the ER, prohibits back-fusion of COPII vesicles with the ER, may act as a sensor for vesicles at the ER membrane; interacts with Sec20p
3180	Component of the vacuole SNARE complex involved in vacuolar morphogenesis; SNAP-25 homolog; functions with a syntaxin homolog Vam3p in vacuolar protein trafficking
3493	Beta3-like subunit of the yeast AP-3 complex; functions in transport of alkaline phosphatase to the vacuole via the alternate pathway; exists in both cytosolic and peripherally associated membrane-bound pools
3561	Small subunit of the clathrin-associated adaptor complex AP-3, which is involved in vacuolar protein sorting; related to the sigma subunit of the mammalian clathrin AP-3 complex; suppressor of loss of casein kinase 1 function
4083	v-SNARE component of the vacuolar SNARE complex involved in vesicle fusion; inhibits ATP-dependent Ca(2+) transport activity of Pmc1p in the vacuolar membrane
4138	Component of CORVET tethering complex; vacuolar peripheral membrane protein that promotes vesicular docking/fusion reactions in conjunction with SNARE proteins, required for vacuolar biogenesis
4388	ATP-binding protein that is a subunit of the HOPS complex and the CORVET tethering complex; essential for membrane docking and fusion at both the Golgi-to-endosome and endosome-to-vacuole stages of protein transport
4432	Protein of unknown function proposed to be involved in protein secretion; interacts with Dsl1p and localizes to the ER and nuclear envelope
4460	GTPase; GTP-binding protein of the rab family; required for homotypic fusion event in vacuole inheritance, for endosome-endosome fusion, similar to mammalian Rab7
4810	Protein involved in cis-Golgi membrane traffic; v-SNARE that interacts with two t-SNARES, Sed5p and Pep12p; required for multiple vacuolar sorting pathways
4844	Component of CORVET tethering complex; peripheral vacuolar membrane protein required for protein trafficking and vacuole biogenesis; interacts with Pep7p
5202	Peripheral membrane protein required for Golgi-to-ER retrograde traffic; component of the ER target site that interacts with coatomer, the major component of the COPI vesicle protein coat; also interacts with Cin5p and Sec39p
5560	Ankyrin repeat-containing protein similar to Akr1p; member of a family of putative palmitoyltransferases containing an Asp-His-His-Cys-cysteine rich (DHHC-CRD) domain; possibly involved in constitutive endocytosis of Ste3p
5562	Target membrane receptor (t-SNARE) for vesicular intermediates traveling between the Golgi apparatus and the vacuole; controls entry of biosynthetic, endocytic, and retrograde traffic into the prevacuolar compartment; syntaxin
5601	t-SNARE required for ER membrane fusion and vesicular traffic, integral membrane protein that constitutes with Sec20p and Use1p the trimeric acceptor for R/v-SNAREs on Golgi-derived vesicles at the ER; part of Dsl1p complex
5632	Syntaxin-related protein required for vacuolar assembly; functions with Vam7p in vacuolar protein trafficking; member of the syntaxin family of proteins
5697	Sphingoid long-chain base kinase, responsible for synthesis of long-chain base phosphates, which function as signaling molecules, regulates synthesis of ceramide from exogenous long-chain bases, localizes to the Golgi and late endosomes
5966	Subunit of the vacuole fusion and protein sorting HOPS complex and the CORVET tethering complex; part of the Class C Vps complex essential for membrane docking and fusion at Golgi-to-endosome and endosome-to-vacuole protein transport stages
6116	Delta adaptin-like subunit of the clathrin associated protein complex (AP-3); functions in transport of alkaline phosphatase to the vacuole via the alternate pathway, suppressor of loss of casein kinase 1 function

Table C.5: **From Figure 3.14 a): Protein 2 (YAL002W),** $\log(\lambda) = 2$ **in the** A **network**

2	Membrane-associated protein that interacts with Vps21p to facilitate soluble vacuolar protein localization; component of the CORVET complex; required for localization and trafficking of the CPY sorting receptor; contains RING finger motif
335	Protein involved in vacuolar assembly, essential for autophagy and the cytoplasm-to-vacuole pathway
2235	Vacuolar protein that plays a critical role in the tethering steps of vacuolar membrane fusion by facilitating guanine nucleotide exchange on small guanosine triphosphatase Ypt7p
2487	Vacuolar membrane protein that is a subunit of the homotypic vacuole fusion and vacuole protein sorting (HOPS) complex; essential for membrane docking and fusion at the Golgi-to-endosome and endosome-to-vacuole stages of protein transport
2903	Component of CORVET tethering complex; cytoplasmic protein required for the sorting and processing of soluble vacuolar proteins, acidification of the vacuolar lumen, and assembly of the vacuolar H+-ATPase
3180	Component of the vacuole SNARE complex involved in vacuolar morphogenesis; SNAP-25 homolog; functions with a syntaxin homolog Vam3p in vacuolar protein trafficking
4083	v-SNARE component of the vacuolar SNARE complex involved in vesicle fusion; inhibits ATP-dependent Ca(2+) transport activity of Pmc1p in the vacuolar membrane
4138	Component of CORVET tethering complex; vacuolar peripheral membrane protein that promotes vesicular docking/fusion reactions in conjunction with SNARE proteins, required for vacuolar biogenesis
4388	ATP-binding protein that is a subunit of the HOPS complex and the CORVET tethering complex; essential for membrane docking and fusion at both the Golgi-to-endosome and endosome-to-vacuole stages of protein transport
4460	GTPase; GTP-binding protein of the rab family; required for homotypic fusion event in vacuole inheritance, for endosome-endosome fusion, similar to mammalian Rab7
4810	Protein involved in cis-Golgi membrane traffic; v-SNARE that interacts with two t-SNARES, Sed5p and Pep12p; required for multiple vacuolar sorting pathways
4844	Component of CORVET tethering complex; peripheral vacuolar membrane protein required for protein trafficking and vacuole biogenesis; interacts with Pep7p
5632	Syntaxin-related protein required for vacuolar assembly; functions with Vam7p in vacuolar protein trafficking; member of the syntaxin family of proteins
5966	Subunit of the vacuole fusion and protein sorting HOPS complex and the CORVET tethering complex; part of the Class C Vps complex essential for membrane docking and fusion at Golgi-to-endosome and endosome-to-vacuole protein transport stages

.

Table C.6: **From Figure 3.14 b): Protein 9 (YAL011W), Protein of unknown function, component of the SWR1 complex, which exchanges histone variant H2AZ (Htz1p) for chromatin-bound histone H2A; required for formation of nuclear-associated array of smooth endoplasmic reticulum known as karmellae,** $\log(\lambda) = 1.75$ **in the** A **network**

9	Protein of unknown function, component of the SWR1 complex, which exchanges histone variant H2AZ (Htz1p) for chromatin-bound histone H2A; required for formation of nuclear-associated array of smooth endoplasmic reticulum known as karmellae
435	Protein of unknown function, component of the SWR1 complex, which exchanges histone variant H2AZ (Htz1p) for chromatin-bound histone H2A
2742	Swi2/Snf2-related ATPase that is the structural component of the SWR1 complex, which exchanges histone variant H2AZ (Htz1p) for chromatin-bound histone H2A
2893	Htz1p-binding component of the SWR1 complex, which exchanges histone variant H2AZ (Htz1p) for chromatin-bound histone H2A; required for vacuolar protein sorting
3234	Component of the Swr1p complex that incorporates Htz1p into chromatin; component of the NuA4 histone acetyltransferase complex
3617	Nuclear actin-related protein involved in chromatin remodeling, component of chromatin-remodeling enzyme complexes
4075	Actin-related protein that binds nucleosomes; a component of the SWR1 complex, which exchanges histone variant H2AZ (Htz1p) for chromatin-bound histone H2A
4377	Protein of unknown function, component of the Swr1p complex that incorporates Htz1p into chromatin
4391	Protein involved in transcription initiation at TATA-containing promoters; associates with the basal transcription factor TFIID; contains two bromodomains; corresponds to the C-terminal region of mammalian TAF1; redundant with Bdf2p
4505	Nucleosome-binding component of the SWR1 complex, which exchanges histone variant H2AZ (Htz1p) for chromatin-bound histone H2A; required for vacuolar protein sorting
5051	Subunit of both the NuA4 histone H4 acetyltransferase complex and the SWR1 complex, may function to antagonize silencing near telomeres; interacts directly with Swc4p, has homology to human leukemogenic protein AF9, contains a YEATS domain
5372	Histone variant H2AZ, exchanged for histone H2A in nucleosomes by the SWR1 complex; involved in transcriptional regulation through prevention of the spread of silent heterochromatin

.

Table C.7: **From Figure 3.14 b): Protein 9 (YAL011W),** $\log(\lambda) = 2.25$ **in the** A **network**

.

9	Protein of unknown function, component of the SWR1 complex, which exchanges histone variant H2AZ (Htz1p) for chromatin-bound histone H2A; required for formation of nuclear-associated array of smooth endoplasmic reticulum known as karmellae
435	Protein of unknown function, component of the SWR1 complex, which exchanges histone variant H2AZ (Htz1p) for chromatin-bound histone H2A
2742	Swi2/Snf2-related ATPase that is the structural component of the SWR1 complex, which exchanges histone variant H2AZ (Htz1p) for chromatin-bound histone H2A
2893	Htz1p-binding component of the SWR1 complex, which exchanges histone variant H2AZ (Htz1p) for chromatin-bound histone H2A; required for vacuolar protein sorting
4075	Actin-related protein that binds nucleosomes; a component of the SWR1 complex, which exchanges histone variant H2AZ (Htz1p) for chromatin-bound histone H2A
4377	Protein of unknown function, component of the Swr1p complex that incorporates Htz1p into chromatin
4505	Nucleosome-binding component of the SWR1 complex, which exchanges histone variant H2AZ (Htz1p) for chromatin-bound histone H2A; required for vacuolar protein sorting
5051	Subunit of both the NuA4 histone H4 acetyltransferase complex and the SWR1 complex, may function to antagonize silencing near telomeres; interacts directly with Swc4p, has homology to human leukemogenic protein AF9, contains a YEATS domain

Table C.8: **From Figure 3.14 c): Protein 14 (YAL016W), Regulatory subunit A of the heterotrimeric protein phosphatase 2A, which also contains regulatory subunit Cdc55p and either catalytic subunit Pph21p or Pph22p; required for cell morphogenesis and for transcription by RNA polymerase III,** $\log(\lambda) = 1.5$ **in the** A **network**

.

14	Regulatory subunit A of the heterotrimeric protein phosphatase 2A, which also contains regulatory subunit Cdc55p and either catalytic subunit Pph21p or Pph22p; required for cell morphogenesis and for transcription by RNA polymerase III
1117	Protein with carboxyl methyl esterase activity that may have a role in demethylation of the phosphoprotein phosphatase catalytic subunit; also identified as a small subunit mitochondrial ribosomal protein
2292	Catalytic subunit of protein phosphatase 2A, functionally redundant with Pph22p; methylated at C terminus; forms alternate complexes with several regulatory subunits; involved in signal transduction and regulation of mitosis
2347	Catalytic subunit of protein phosphatase 2A, functionally redundant with Pph21p; methylated at C terminus; forms alternate complexes with several regulatory subunits; involved in signal transduction and regulation of mitosis
3158	Non-essential regulatory subunit B of protein phosphatase 2A, which has multiple roles in mitosis and protein biosynthesis; involved in regulation of mitotic exit; found in the nucleus of most cells, also at the bud neck and at the bud tip
3393	Putative component of the protein phosphatase type 2A complex
4577	Protein that interacts with silencing proteins at the telomere, involved in transcriptional silencing; implicated in the mitotic exit network through regulation of Cdc14p localization; paralog of Zds1p
5540	B-type regulatory subunit of protein phosphatase 2A (PP2A); homolog of the mammalian B' subunit of PP2A
5599	Component of the spindle checkpoint, involved in sensing lack of tension on mitotic chromosomes; protects centromeric Rec8p at meiosis I; required for accurate chromosomal segregation at meiosis II and for mitotic chromosome stability
5688	Zn2-Cys6 zinc-finger transcription factor that activates genes involved in multidrug resistance; paralog of Yrm1p, acting on an overlapping set of target genes
6073	Activator of the phosphotyrosyl phosphatase activity of PP2A,peptidyl-prolyl cis/trans-isomerase; regulates G1 phase progression, the osmoresponse, microtubule dynamics; subunit of the Tap42p-Pph21p-Rrd2p complex

Table C.9: **From Figure 3.14 c): Protein 14 (YAL016W),** $\log(\lambda) = 1.3$ **in the** P **network**

.

14	Regulatory subunit A of the heterotrimeric protein phosphatase 2A, which also contains regulatory subunit Cdc55p and either catalytic subunit Pph21p or Pph22p; required for cell morphogenesis and for transcription by RNA polymerase III
22	Putative GDP/GTP exchange factor required for mitotic exit at low temperatures; acts as a guanine nucleotide exchange factor (GEF) for Tem1p, which is a key regulator of mitotic exit; physically associates with Ras2p-GTP
200	
1201	Protein required for proper cell fusion and cell morphology; functions in a complex with Kel2p to negatively regulate mitotic exit, interacts with Tem1p and Lte1p; localizes to regions of polarized growth; potential Cdc28p substrate
3470	Protein that functions in a complex with Kel1p to negatively regulate mitotic exit, interacts with Tem1p and Lte1p; localizes to regions of polarized growth; potential Cdc28p substrate
3899	Protein of unknown function, green fluorescent protein (GFP)-fusion protein localizes to the vacuolar membrane
5540	B-type regulatory subunit of protein phosphatase 2A (PP2A); homolog of the mammalian B' subunit of PP2A
5876	Tubulin folding factor D involved in beta-tubulin (Tub2p) folding; isolated as mutant with increased chromosome loss and sensitivity to benomyl

Table C.10: **From Figure 3.14 d): Protein 19 (YAL021C), component of the CCR4-NOT transcriptional complex, which is involved in regulation of gene expression; component of the major cytoplasmic deadenylase, which is involved in mRNA poly(A) tail shortening,** $\log(\lambda) = 1$ **in the** A **network**

.

19	Component of the CCR4-NOT transcriptional complex, which is involved in regulation of gene expression; component of the major cytoplasmic deadenylase, which is involved in mRNA poly(A) tail shortening
433	Glucosidase II catalytic subunit required for normal cell wall synthesis; mutations in rot2 suppress tor2 mutations, and are synthetically lethal with rot1 mutations
689	Component of the CCR4-NOT complex, which has multiple roles in regulating mRNA levels including regulation of transcription and destabilizing mRNAs by deadenylation; basal transcription factor
870	Subunit of the CCR4-NOT complex, which has roles in transcription regulation, mRNA degradation, and post-transcriptional modifications; with Ubc4p, ubiquitinates nascent polypeptide-associated complex subunits and histone demethyase Jhd2p
1236	Alpha subunit of the heteromeric nascent polypeptide-associated complex (NAC) involved in protein sorting and translocation, associated with cytoplasmic ribosomes
1300	Subunit of the CCR4-NOT complex, which is a global transcriptional regulator with roles in transcription initiation and elongation and in mRNA degradation
2324	Component of the CCR4-NOT complex, which has multiple roles in regulating mRNA levels including regulation of transcription and destabilizing mRNAs by deadenylation; basal transcription factor
2629	Glucosidase II beta subunit, forms a complex with alpha subunit Rot2p, involved in removal of two glucose residues from N-linked glycans during glycoprotein biogenesis in the ER
2660	Beta3 subunit of the heterotrimeric nascent polypeptide-associated complex which binds ribosomes via its beta-subunits in close proximity to nascent polypeptides; interacts with Caf130p of the CCR4-NOT complex; similar to human BTF3
3146	Member of the Puf family of RNA-binding proteins; binds to mRNAs encoding chromatin modifiers and spindle pole body components; involved in longevity, maintenance of cell wall integrity, and sensitivity to and recovery from pheromone arrest
3366	Part of the evolutionarily-conserved CCR4-NOT transcriptional regulatory complex involved in controlling mRNA initiation, elongation, and degradation
3772	
4748	Protein of unknown function; interacts with both the Reg1p/Glc7p phosphatase and the Snf1p kinase
5232	Evolutionarily conserved subunit of the CCR4-NOT complex involved in controlling mRNA initiation, elongation and degradation; binds Cdc39p
5335	RNase of the DEDD superfamily, subunit of the Ccr4-Not complex that mediates 3' to 5' mRNA deadenylation
5958	Subunit beta1 of the nascent polypeptide-associated complex (NAC) involved in protein targeting, associated with cytoplasmic ribosomes; enhances DNA binding of the Gal4p activator; homolog of human BTF3b
6276	Subunit of the CCR4-NOT complex, which is a global transcriptional regulator with roles in transcription initiation and elongation and in mRNA degradation

Table C.11: **From Figure 3.14 d): Protein 19 (YAL021C),** $\log(\lambda) = 1.5$ **in the** A **network**

.

19	Component of the CCR4-NOT transcriptional complex, which is involved in regulation of gene expression; component of the major cytoplasmic deadenylase, which is involved in mRNA poly(A) tail shortening
689	Component of the CCR4-NOT complex, which has multiple roles in regulating mRNA levels including regulation of transcription and destabilizing mRNAs by deadenylation; basal transcription factor
870	Subunit of the CCR4-NOT complex, which has roles in transcription regulation, mRNA degradation, and post-transcriptional modifications; with Ubc4p, ubiquitinates nascent polypeptide-associated complex subunits and histone demethyase Jhd2p
1300	Subunit of the CCR4-NOT complex, which is a global transcriptional regulator with roles in transcription initiation and elongation and in mRNA degradation
2324	Component of the CCR4-NOT complex, which has multiple roles in regulating mRNA levels including regulation of transcription and destabilizing mRNAs by deadenylation; basal transcription factor
2660	Beta3 subunit of the heterotrimeric nascent polypeptide-associated complex which binds ribosomes via its beta-subunits in close proximity to nascent polypeptides; interacts with Caf130p of the CCR4-NOT complex; similar to human BTF3
3146	Member of the Puf family of RNA-binding proteins; binds to mRNAs encoding chromatin modifiers and spindle pole body components; involved in longevity, maintenance of cell wall integrity, and sensitivity to and recovery from pheromone arrest
3366	Part of the evolutionarily-conserved CCR4-NOT transcriptional regulatory complex involved in controlling mRNA initiation, elongation, and degradation
3772	
5232	Evolutionarily conserved subunit of the CCR4-NOT complex involved in controlling mRNA initiation, elongation and degradation; binds Cdc39p
5335	RNase of the DEDD superfamily, subunit of the Ccr4-Not complex that mediates 3' to 5' mRNA deadenylation
6276	Subunit of the CCR4-NOT complex, which is a global transcriptional regulator with roles in transcription initiation and elongation and in mRNA degradation

Table C.12: **From Figure 3.14 d): Protein 19 (YAL021C),** $\log(\lambda) = 1.5$ **in the** P **network**

19	Component of the CCR4-NOT transcriptional complex, which is involved in regulation of gene expression; component of the major cytoplasmic deadenylase, which is involved in mRNA poly(A) tail shortening
147	Protein involved in G2/M phase progression and response to DNA damage, interacts with Rad53p; contains an RNA recognition motif, a nuclear localization signal, and several SQ/TQ cluster domains; hyperphosphorylated in response to DNA damage
689	Component of the CCR4-NOT complex, which has multiple roles in regulating mRNA levels including regulation of transcription and destabilizing mRNAs by deadenylation; basal transcription factor
870	Subunit of the CCR4-NOT complex, which has roles in transcription regulation, mRNA degradation, and post-transcriptional modifications; with Ubc4p, ubiquitinates nascent polypeptide-associated complex subunits and histone demethyase Jhd2p
1236	Alpha subunit of the heteromeric nascent polypeptide-associated complex (NAC) involved in protein sorting and translocation, associated with cytoplasmic ribosomes
1300	Subunit of the CCR4-NOT complex, which is a global transcriptional regulator with roles in transcription initiation and elongation and in mRNA degradation
2324	Component of the CCR4-NOT complex, which has multiple roles in regulating mRNA levels including regulation of transcription and destabilizing mRNAs by deadenylation; basal transcription factor
2383	Putative RNA binding protein and partially redundant Whi3p homolog that regulates the cell size requirement for passage through Start and commitment to cell division
2466	Ubiquitin-conjugating enzyme that mediates selective degradation of short-lived, abnormal, or excess proteins, including histone H3; central component of the cellular stress response; expression is heat inducible
2660	Beta3 subunit of the heterotrimeric nascent polypeptide-associated complex which binds ribosomes via its beta-subunits in close proximity to nascent polypeptides; interacts with Caf130p of the CCR4-NOT complex; similar to human BTF3
2851	Subunit of the RNA polymerase II mediator complex; associates with core polymerase subunits to form the RNA polymerase II holoenzyme; required for stable association of Srb10p-Srb11p kinase; essential for transcriptional regulation
3324	Ser/Thr kinase involved in transcription and stress response; functions as part of a network of genes in exit from mitosis; localization is cell cycle regulated; activated by Cdc15p during the exit from mitosis
3366	Part of the evolutionarily-conserved CCR4-NOT transcriptional regulatory complex involved in controlling mRNA initiation, elongation, and degradation
5035	Protein of unknown function, mediates sensitivity to salt stress; interacts physically with the splicing factor Msl1p and also displays genetic interaction with MSL1
5141	RNA binding protein that sequesters CLN3 mRNA in cytoplasmic foci; cytoplasmic retention factor for Cdc28p and associated cyclins; regulates cell fate and dose-dependently regulates the critical cell size required for passage through Start
5232	Evolutionarily conserved subunit of the CCR4-NOT complex involved in controlling mRNA initiation, elongation and degradation; binds Cdc39p
5335	RNase of the DEDD superfamily, subunit of the Ccr4-Not complex that mediates 3' to 5' mRNA deadenylation
5958	Subunit beta1 of the nascent polypeptide-associated complex (NAC) involved in protein targeting, associated with cytoplasmic ribosomes; enhances DNA binding of the Gal4p activator; homolog of human BTF3b
6276	Subunit of the CCR4-NOT complex, which is a global transcriptional regulator with roles in transcription initiation and elongation and in mRNA degradation

.

Table C.13: **From Figure 3.14 d): Protein 19 (YAL021C),** $\log(\lambda) = 2.2$ **in the** P **network**

19	Component of the CCR4-NOT transcriptional complex, which is involved in regulation of gene expression; component of the major cytoplasmic deadenylase, which is involved in mRNA poly(A) tail shortening
1866	Part of the evolutionarily-conserved CCR4-NOT transcriptional regulatory complex involved in controlling mRNA initiation, elongation, and degradation; putative ABC ATPase; interacts with Ssn2p, Ssn3p, and Ssn8p
2851	Subunit of the RNA polymerase II mediator complex; associates with core polymerase subunits to form the RNA polymerase II holoenzyme; required for stable association of Srb10p-Srb11p kinase; essential for transcriptional regulation
3011	General transcription elongation factor TFIIS, enables RNA polymerase II to read through blocks to elongation by stimulating cleavage of nascent transcripts stalled at transcription arrest sites

.

Appendix D

Alternative sequence similarity scores for inferring interactions

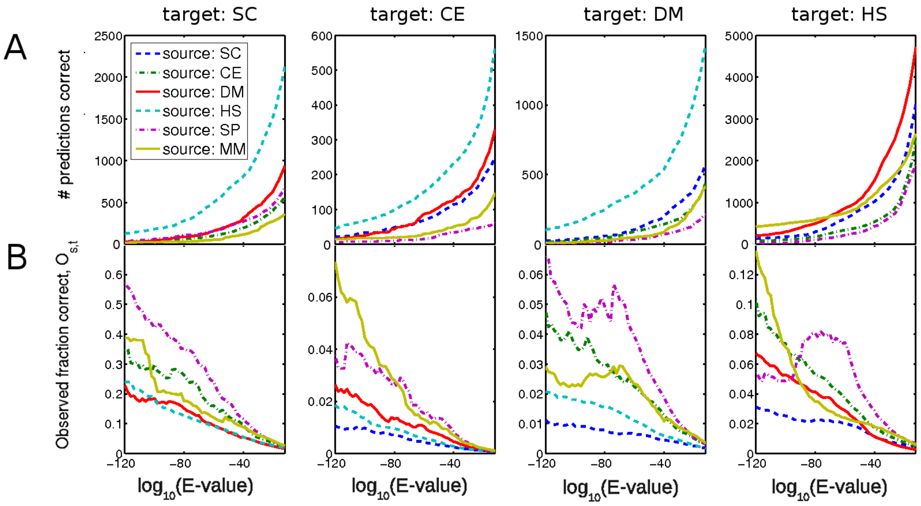


Figure D.1: **As for Figure 4.3 A and B, but with different scales for the y-axes.** We show the results of inferring interactions from *S. cerevisiae* (SC), *C. elegans*, *D. melanogaster* (DM), *H. sapiens* (HS), *S, Pombe* (SP), and *M. musculus* (MM) to the first four of those species. (A) Number of correct interolog inferences across species, for different `blastp` E-value cut-offs. (B) Fraction of all inferences that are observed in the interactions of the target species, $O_{s,t}$.

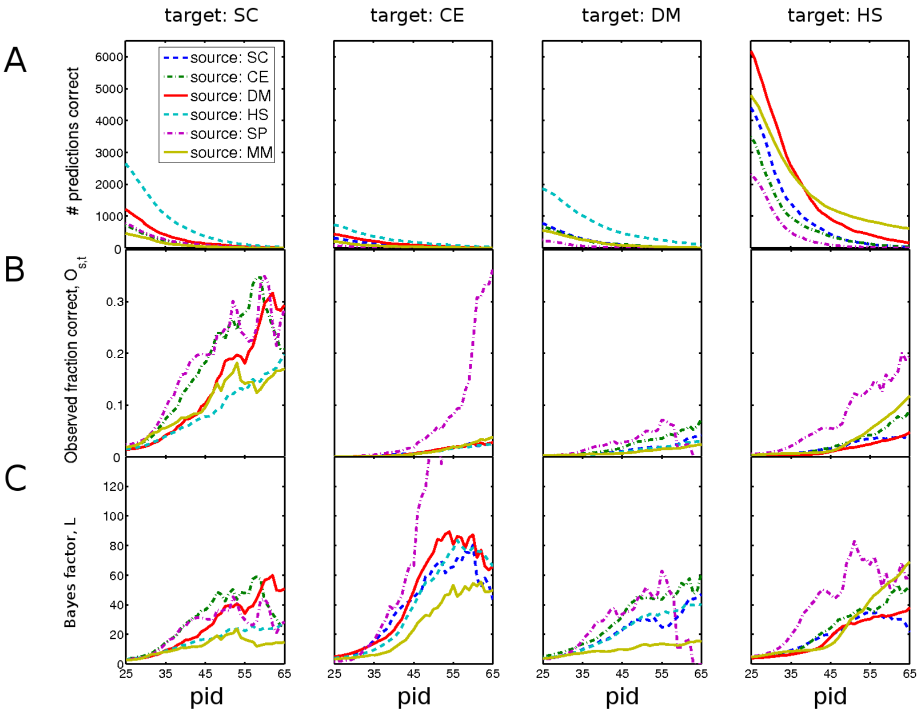


Figure D.2: **As for Figure 4.3, but using thresholds of percentage sequence identity (pid) rather than thresholds on *E*-value.** We show the results of inferring interactions from *S. cerevisiae* (SC), *C. elegans*, *D. melanogaster* (DM), *H. sapiens* (HS), *S, Pombe* (SP), and *M. musculus* (MM) to the first four of those species. (A) Number of correct interolog inferences across species. (B) Fraction of all inferences that are observed in the interactions of the target species, $O_{s,t}$. (C) The Bayes Factor L that an inference is correct.

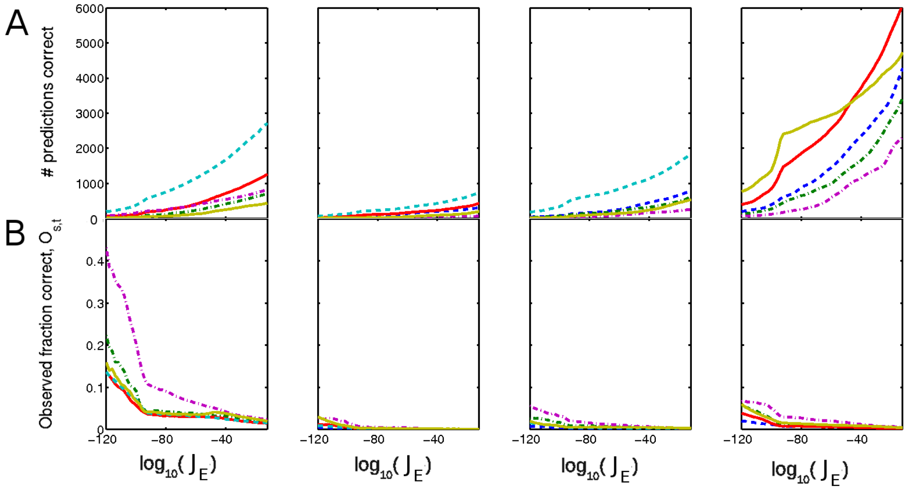


Figure D.3: **As for Figure 4.3 A and B, but using thresholds of the joint similarity measure $J_E = \sqrt(E_{\textbf{val}}(A, A')E_{\textbf{val}}(B, B'))$, rather than a threshold on the definition of homology.** We show the results of inferring interactions from *S. cerevisiae* (SC), *C. elegans*, *D. melanogaster* (DM), *H. sapiens* (HS), *S, Pombe* (SP), and *M. musculus* (MM) to the first four of those species. (A) Number of correct interolog inferences across species. (B) Fraction of all inferences that are observed in the interactions of the target species, $O_{s,t}$. The Bayes Factor measure L does not generalise straightforwardly for joint sequence-similarity measures. Note that `blastp` rounds all E-values of 10^{-180} and below down to 0. In order to see the behaviour of these hits, we replace all hits with an E-value of 0 with an E-value of 10^{-180}, and this explains the 'kink' seen in this figure.

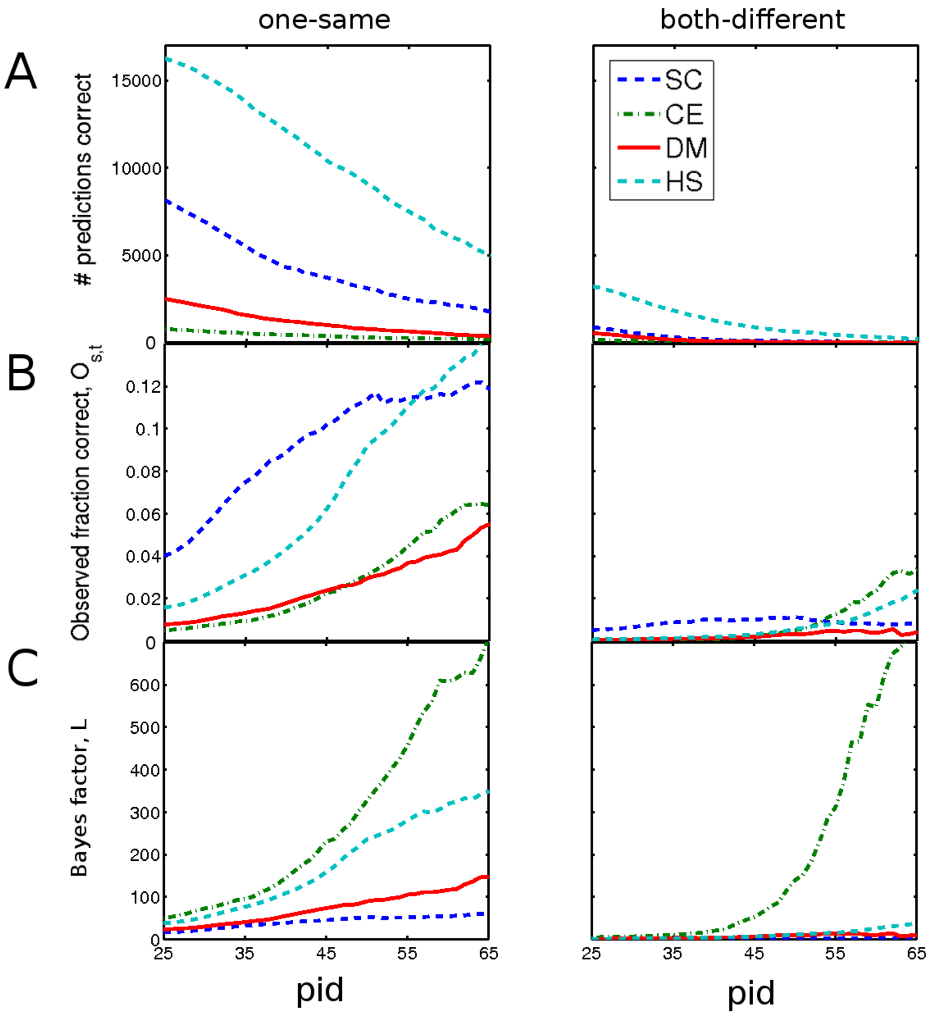


Figure D.4: **As for Figure 4.14, but for using percentage sequence identity (**pid**) rather than** ***E*****-value.** Inferences within a species: 'one-same' inferences (left) dominate 'both-different' inferences (right). For inferences within *S. cerevisiae* (SC), *C. elegans* (CE), *D. melanogaster* (DM), and *H. sapiens* (HS), (A) the number of correct inferences, (B) the fraction of inferences observed to be correct $O_{s,t}$, and (C) the Bayes Factor L that the inferences are correct. We consider a given inferred interaction to be inferred from the 'closest' interaction (see the main text for a definition and discussion).

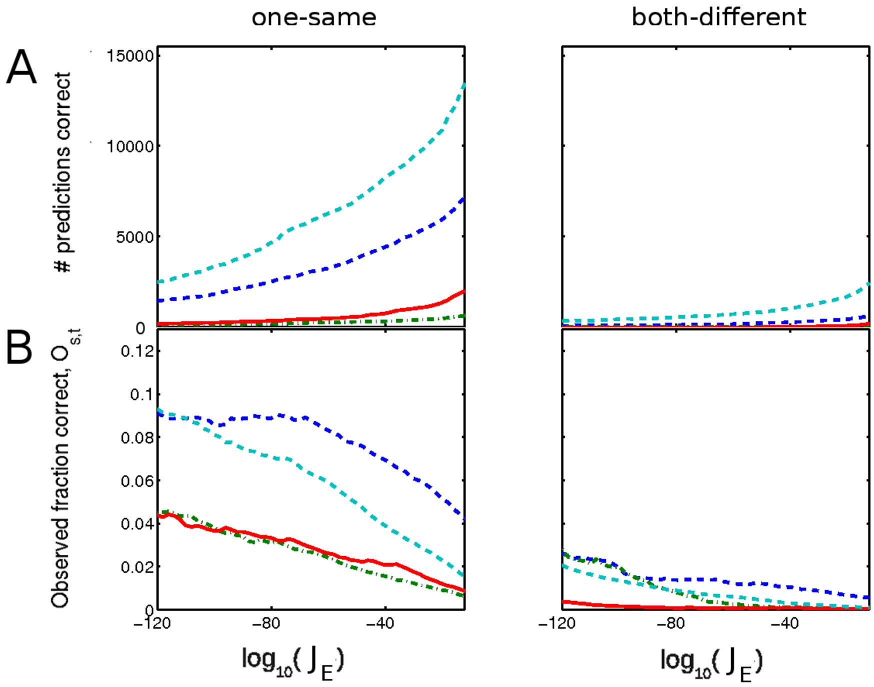


Figure D.5: **As for Figure 4.14 A and B, but using thresholds of the joint similarity measure $J_E = \sqrt{(E_{\textbf{val}}(A, A')E_{\textbf{val}}(B, B'))}$, rather than a threshold on the definition of homology.** Inferences within a species: 'one-same' inferences (left) dominate 'both-different' inferences (right). For inferences within *S. cerevisiae* (SC), *C. elegans* (CE), *D. melanogaster* (DM), and *H. sapiens* (HS), (A) the number of correct inferences, (B) the fraction of inferences observed to be correct $O_{s,t}$.

Bibliography

[1] S. Agarwal, C. M. Deane, M. A. Porter, and N. S. Jones. Revisiting date and party hubs: Novel approaches to role assignment in protein interaction networks. *PLoS Computational Biology*, 6(6):e1000817, 2010.

[2] I. Agrafioti, J. Swire, J. Abbott, D. Huntley, S. Butcher, and M. P. H. Stumpf. Comparative analysis of the Saccharomyces cerevisiae and Caenorhabditis elegans protein interaction networks. *BMC Evolutionary Biology*, 5(23), 2005.

[3] U. Alon. *An Introduction to Systems Biology: Design Principles of Biological Circuits*. Chapman & Hall/CRC, 2007.

[4] P. Aloy, H. Ceulemans, A. Stark, and R. B. Russell. The relationship between sequence and interaction divergence in proteins. *Journal of Molecular Biology*, 332(5):989–998, 2003.

[5] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. Basic local alignment search tool. *Journal of Molecular Biology*, 215(3):403–410, 1990.

[6] A. Andreeva, D. Howorth, S. E. Brenner, T. J. P. Hubbard, C. Chothia, and A. G. Murzin. SCOP database in 2004: refinements integrate structure and sequence family data. *Nucleic Acids Research*, 32:D226–D229, 2004.

[7] B. Aranda, P. Achuthan, Y. Alam-Faruque, I. Armean, A. Bridge, C. Derow, M. Feuermann, A. T. Ghanbarian, S. Kerrien, J. Khadake, J. Kerssemakers,

C. Leroy, M. Menden, M. Michaut, L. Montecchi-Palazzi, S. N. Neuhauser, S. Orchard, V. Perreau, B. Roechert, K. van Eijk, and H. Hermjakob. The IntAct molecular interaction database in 2010. *Nucleic Acids Research*, 38 (suppl 1):D525–D531, 2010.

[8] A. Arenas, A. Fernández, and S. Gómez. Analysis of the structure of complex networks at different resolution levels. *New Journal of Physics*, 10:053039, 2008.

[9] M. Arifuzzaman, M. Maeda, A. Itoh, K. Nishikata, C. Takita, R. Saito, T. Ara, K. Nakahigashi, H.-C. Huang, A. Hirai, K. Tsuzuki, S. Nakamura, M. Altaf-Ul-Amin, T. Oshima, T. Baba, N. Yamamoto, T. Kawamura, T. Ioka-Nakamichi, M. Kitagawa, M. Tomita, S. Kanaya, C. Wada, and H. Mori. Large-scale identification of protein-protein interaction of Escherichia coli K-12. *Genome Research*, 16(5):686–691, 2006.

[10] Y. Artzy-Randrup, S. J. Fleishman, N. Ben-Tal, and L. Stone. Comment on ‘Network motifs: Simple building blocks of complex networks’ and ‘Superfamilies of evolved and designed networks’. *Science*, 305(5687):1107, 2004.

[11] G. E. Ascoli. *Computational Neuroanatomy – Principles and Methods*. Humana Press, 2003.

[12] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock. Gene Ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genetics*, 25(1):25–29, 2000.

[13] P. Bachman and Y. Liu. Structure discovery in PPI networks using pattern-based network decomposition. *Bioinformatics*, 25(14):1814–1821, 2009.

[14] G. D. Bader and C. W. Hogue. Analyzing yeast protein-protein interaction data obtained from different sources. *Nature Biotechnology*, 20(10):991–997, 2002.

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

[16] J. S. Bader, A. Chaudhuri, J. M. Rothberg, and J. Chant. Gaining confidence in high-throughput protein interaction networks. *Nature Biotechnology*, 22(1): 78–85, 2004.

[17] P. Balazs, P. Csaba, and H. L. D. Dosage sensitivity and the evolution of gene families in yeast. *Nature*, 424(6945):194–197, 2003. 10.1038/nature01771.

[18] S. Bandyopadhyay, R. Sharan, and T. Ideker. Systematic identification of functional orthologs based on protein network comparison. *Genome Research*, 16 (3):428–435, 2006.

[19] A.-L. Barabási and R. Albert. Emergence of scaling in random networks. *Science*, 286(5439):509–512, 1999.

[20] A.-L. Barabási and Z. N. Oltvai. Network biology: understanding the cell's functional organization. *Nature Reviews Genetics*, 5(2):101–113, 2004.

[21] D. S. Bassett, N. F. Wymbs, M. A. Porter, P. J. Mucha, J. M. Carlson, and S. T. Grafton. Dynamic reconfiguration of human brain networks during learning. *Proceedings of the National Academy of Sciences*, 2011.

[22] M. K. Basu, L. Carmel, I. B. Rogozin, and E. V. Koonin. Evolution of protein domain promiscuity in eukaryotes. *Genome Research*, 18:449–461, 2008.

[23] N. N. Batada, T. Reguly, A. Breitkreutz, L. Boucher, B. J. Breitkreutz, L. D. Hurst, and M. Tyers. Stratus not altocumulus: A new view of the yeast protein interaction network. *PLoS Biology*, 4(10):e317, 2006.

[24] N. N. Batada, T. Reguly, A. Breitkreutz, L. Boucher, B. J. Breitkreutz, L. D. Hurst, and M. Tyers. Still stratus not altocumulus: further evidence against the date/party hub distinction. *PLoS Biology*, 5(6):e154, 2007.

[25] P. Beltrao and L. Serrano. Specificity and evolvability in eukaryotic protein interaction networks. *PLoS Computational Biology*, 3(2):e25, 2007.

[26] A. Ben-Hur and W. S. Noble. Choosing negative examples for the prediction of protein-protein interactions. *BMC Bioinformatics*, 7 Suppl 1(S2), 2006.

[27] A. Ben-Hur, A. Elisseeff, and I. Guyon. A stability based method for discovering structure in clustered data. *Pacific Symposium on Biocomputing.*, pages 6–17, 2002.

[28] S. A. Benner, M. A. Cohen, and G. H. Gonnet. Empirical and structural models for insertions and deletions in the divergent evolution of proteins. *Journal of Molecular Biology*, 229(4):1065–1082, 1993.

[29] J. Berg, J. L. Tymoczko, and L. Stryer. *Biochemistry, 5th edition.* W. H. Freeman, 2002.

[30] J. Berg, M. Lässig, and A. Wagner. Structure and evolution of protein interaction networks: a statistical model for link dynamics and gene duplications. *BMC Evolutionary Biology*, 4(51), 2004.

[31] N. Bertin, N. Simonis, D. Dupuy, M. E. Cusick, J.-D. J. Han, H. B. Fraser, F. P. Roth, and M. Vidal. Confirmation of organized modularity in the yeast interactome. *PLoS Biology*, 5(6):e153, 2007.

[32] N. Bhardwaj and H. Lu. Correlation between gene expression profiles and proteinprotein interactions within and across genomes. *Bioinformatics*, 21(11): 2730–2738, 2005.

[33] G. Bianconi, P. Pin, and M. Marsili. Assessing the relevance of node features for network structure. *Proceedings of the National Academy of Sciences*, 106 (28):11433–11438, 2009.

[34] V. D. Blondel, J. L. Guillaume, and R. Lambiotte. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, page P10008, 2008.

[35] J. D. Bloom and C. Adami. Apparent dependence of protein evolutionary rate on number of interactions is linked to biases in protein–protein interactions data sets. *BMC Evolutionary Biology*, 3(21), 2003.

[36] M. Borodovsky and S. Ekisheva. *Problems and Solutions in Biological Sequence Analysis*. Cambridge University Press, 2008.

[37] P. M. Bowers, S. J. Cokus, D. Eisenberg, and T. O. Yeates. Use of logic relationships to decipher protein network organization. *Science*, 306(5705): 2246–2249, 2004.

[38] E. I. Boyle, S. Weng, J. Gollub, H. Jin, D. Botstein, J. M. Cherry, and G. Sherlock. GO::TermFinder–open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes. *Bioinformatics*, 20(18):3710, 2004.

[39] U. Brandes, D. Delling, M. Gaertler, R. Goerke, M. Hoefer, Z. Nikoloski, and D. Wagner. On modularity clustering. *IEEE Transactions on Knowledge and Data Engineering*, 20(2):172–188, 2008.

[40] K. V. Brinda and S. Vishveshwara. Oligomeric protein structure networks: insights into protein-protein interactions. *BMC Bioinformatics*, 6(296), 2005.

[41] K. Brown and I. Jurisica. Unequal evolutionary conservation of human protein interactions in interologous networks. *Genome Biology*, 8(5):R95, 2007.

[42] K. R. Brown and I. Jurisica. Online Predicted Human Interaction Database. *Bioinformatics*, 21(9):2076–2082, 2005.

[43] A. Brückner, C. Polge, N. Lentze, D. Auerbach, and U. Schlattner. Yeast two-hybrid, a powerful tool for systems biology. *International Journal of Molecular Sciences*, 10(6):2763–2788, 2009.

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

[45] M. Buckland and F. Gey. The relationship between recall and precision. *Journal of the American Society for Information Science*, 45(1):12–19, 1994.

[46] D. R. Caffrey, S. Somaroo, J. D. Hughes, J. Mintseris, and E. S. Huang. Are protein-protein interfaces more conserved in sequence than the rest of the protein surface? *Protein Science*, 13(1):190–202, 2004.

[47] O. Carugo and G. Franzot. Prediction of protein-protein interactions based on surface patch comparison. *Proteomics*, 4(6):1727–1736, 2004.

[48] A. Ceol, A. Chatr Aryamontri, L. Licata, D. Peluso, L. Briganti, L. Perfetto, L. Castagnoli, and G. Cesareni. MINT, the molecular interaction database: 2009 update. *Nucleic Acids Research*, 38(suppl 1):D532–D539, 2010.

[49] A. Chatr-aryamontri, A. Ceol, L. M. Palazzi, G. Nardelli, M. V. Schneider, L. Castagnoli, and G. Cesareni. MINT: the Molecular INTeraction database. *Nucleic Acids Research*, 35(Database issue):D572, 2007.

[50] J. Chen and B. Yuan. Detecting functional modules in the yeast protein-protein interaction network. *Bioinformatics*, 22(18):2283–2290, 2006.

[51] J. Chen, W. Hsu, M. L. Lee, and S. K. Ng. Increasing confidence of protein interactomes using network topological metrics. *Bioinformatics*, 22(16):1998–2004, 2006.

[52] P. Y. Chen, C. M. Deane, and G. Reinert. A statistical approach using network structure in the prediction of protein characteristics. *Bioinformatics*, 23(17): 2314–2321, 2007.

[53] J. M. Cherry, C. Adler, C. Ball, S. A. Chervitz, S. S. Dwight, E. T. Hester, Y. Jia, G. Juvik, T. Roe, M. Schroeder, et al. SGD: Saccharomyces genome database. *Nucleic Acids Research*, 26(1):73, 1998.

[54] H. N. Chua, W.-K. Sung, and L. Wong. Exploiting indirect neighbours and topological weight to predict protein function from protein-protein interactions. *Bioinformatics*, 22(13):1623–1630, 2006.

[55] T. Clackson and J. A. Wells. A hot spot of binding energy in a hormone-receptor interface. *Science*, 267(5196):383–386, 1995.

[56] A. Clauset, C. Moore, and M. E. J. Newman. Hierarchical structure and the prediction of missing links in networks. *Nature*, 453(7191):98–101, 2008.

[57] A. Clauset, C. R. Shalizi, and M. E. J. Newman. Power-law distributions in empirical data. *SIAM Review*, 51(4):661–703, 2009.

[58] S. J. Cockell, B. Oliva, and R. M. Jackson. Structure-based evaluation of in silico predictions of protein-protein interactions using Comparative Docking. *Bioinformatics*, 23(5):573–581, 2007.

[59] V. Colizza, A. Flammini, M. Serrano, and A. Vespignani. Detecting rich-club ordering in complex networks. *Nature Physics*, 2:110–115, 2006.

[60] M. O. Collins and J. S. Choudhary. Mapping multiprotein complexes by affinity purification and mass spectrometry. *Current Opinion in Biotechnology*, 19(4): 324–330, 2008.

[61] T. G. Consortium, J.-L. Souciet, B. Dujon, C. Gaillardin, M. Johnston, P. V. Baret, P. Cliften, D. J. Sherman, J. Weissenbach, E. Westhof, P. Wincker, C. Jubin, J. Poulain, V. Barbe, B. Sgurens, F. Artiguenave, V. Anthouard, B. Vacherie, M.-E. Val, R. S. Fulton, P. Minx, R. Wilson, P. Durrens, G. Jean, C. Marck, T. Martin, M. Nikolski, T. Rolland, M.-L. Seret, S. Casargola, L. Despons, C. Fairhead, G. Fischer, I. Lafontaine, V. Leh, M. Lemaire, J. de Montigny, C. Neuvglise, A. Thierry, I. Blanc-Lenfle, C. Bleykasten, J. Diffels, E. Fritsch, L. Frangeul, A. Goffon, N. Jauniaux, R. Kachouri-Lafond, C. Payen, S. Potier, L. Pribylova, C. Ozanne, G.-F. Richard, C. Sacerdot, M.-L. Straub, and E. Talla. Comparative genomics of protoploid *Saccharomycetaceae*. *Genome Research*, 19(10):1696–1709, 2009.

[62] L. d. F. Costa, F. A. Rodrigues, G. Travieso, and P. R. Villas Boas. Characterization of complex networks: A survey of measurements. *Advances in Physics*, 56(1):167–242, 2007.

[63] M. E. Cusick, H. Yu, A. Smolyar, K. Venkatesan, A.-R. R. Carvunis, N. Simonis, J.-F. F. Rual, H. Borick, P. Braun, M. Dreze, J. Vandenhaute, M. Galli, J. Yazaki, D. E. Hill, J. R. Ecker, F. P. Roth, and M. Vidal. Literature-curated protein interaction datasets. *Nature Methods*, 6(1):39–46, 2009.

[64] L. Danon, A. Daz-Guilera, J. Duch, and A. Arenas. Comparing community

structure identification. *Journal of Statistical Mechanics: Theory and Experiment*, 09:P09008, 2005.

[65] S. V. Date and E. M. Marcotte. Discovery of uncharacterized cellular systems by genome-wide analysis of functional linkages. *Nature Biotechnology*, 21(9): 1055–1062, 2003.

[66] L. David, R. Oliver, and O. Christine. Predicting protein function from sequence and structure. *Nat Rev Mol Cell Biol*, 8(12):995–1005, dec 2007. 10.1038/nrm2281.

[67] J. Davis and M. Goadrich. The relationship between Precision-Recall and ROC curves. In *Proceedings of the 23rd international conference on Machine learning*, ICML '06, pages 233–240, New York, NY, USA, 2006. ACM.

[68] E. de Silva, T. Thorne, P. Ingram, I. Agrafioti, J. Swire, C. Wiuf, and M. P. H. Stumpf. The effects of incomplete protein interaction data on structural and evolutionary inferences. *BMC Biology*, 4(39), 2006.

[69] D. J. de Solla Price. A general theory of bibliometric and other cumulative advantage processes. *Journal of the American Society for Information Science*, 27:292–306, 1976.

[70] C. M. Deane, L. Salwiski, I. Xenarios, and D. Eisenberg. Protein Interactions: Two methods for assessment of the reliability of high-throughput observations. *Molecular & Cellular Proteomics*, 1(5):349–356, 2002.

[71] M. Deng, S. Mehta, F. Sun, and T. Chen. Inferring domain-domain interactions from protein-protein interactions. *Genome Research*, 12(10):1540–1548, 2002.

[72] M. Deng, F. Sun, and T. Chen. Assessment of the reliability of protein-protein

interactions and protein function prediction. *Pacific Symposium on Biocomputing*, pages 140–151, 2003.

[73] P. D'Haeseleer and G. Church. Estimating and improving protein interaction error rates. *Proceedings of the IEEE Computational Society Bioinformatics Conference*, pages 216–223, 2004.

[74] D. A. Drummond, A. Raval, and C. O. Wilke. A single determinant dominates the rate of yeast protein evolution. *Molecular Biology and Evolution*, 23(2): 327–337, 2006.

[75] R. Dunn, F. Dudbridge, and C. Sanderson. The use of edge-betweenness clustering to investigate biological function in protein interaction networks. *BMC Bioinformatics*, 6(39), 2005.

[76] J. Dutkowski and J. Tiuryn. Identification of functional modules from conserved ancestral proteinprotein interactions. *Bioinformatics*, 23(13):i149–i158, 2007.

[77] J. Dutkowski and J. Tiuryn. Phylogeny-guided interaction mapping in seven eukaryotes. *BMC Bioinformatics*, 10(393), 2009.

[78] A. M. Edwards, B. Kus, R. Jansen, D. Greenbaum, J. Greenblatt, and M. Gerstein. Bridging structural biology and genomics: assessing protein interaction data with known complexes. *Trends in Genetics*, 18(10):529–536, 2002.

[79] E. Eisenberg and E. Y. Levanon. Preferential attachment in the protein network evolution. *Physical Review Letters*, 91(13):138701, 2003.

[80] A. J. Enright, I. Iliopoulos, N. C. Kyrpides, and C. A. Ouzounis. Protein interaction maps for complete genomes based on genome fusion events. *Nature*, 402(6757):86–90, 1999.

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

[82] R. M. Ewing, P. Chu, F. Elisma, H. Li, P. Taylor, S. Climie, L. McBroom-Cerajewski, M. D. Robinson, L. O'Connor, M. Li, R. Taylor, M. Dharsee, Y. Ho, A. Heilbut, L. Moore, S. Zhang, O. Ornatsky, Y. V. Bukhman, M. Ethier, Y. Sheng, J. Vasilescu, M. Abu-Farha, J.-P. Lambert, H. S. Duewel, I. I. Stewart, B. Kuehl, K. Hogue, K. Colwill, K. Gladwish, B. Muskat, R. Kinach, S.-L. Adams, M. F. Moran, G. B. Morin, T. Topaloglou, and D. Figeys. Large-scale mapping of human protein-protein interactions by mass spectrometry. *Molecular Systems Biology*, 3(89), 2007.

[83] T. Fawcett. An introduction to ROC analysis. *Pattern Recognition Letters*, 27 (8):861–874, 2006.

[84] A. Fernandez and M. Lynch. Non-adaptive origins of interactome complexity. *Nature*, 474(7352):502–505, 2011.

[85] S. Fields and O. Song. A novel genetic system to detect protein-protein interactions. *Nature*, 340(6230):245–246, 1989.

[86] R. D. Finn, J. Mistry, B. Schuster-Böckler, S. Griffiths-Jones, V. Hollich, T. Lassmann, S. Moxon, M. Marshall, A. Khanna, R. Durbin, S. R. Eddy, E. L. Sonnhammer, and A. Bateman. Pfam: clans, web tools and services. *Nucleic Acids Research*, 34(Database issue):247–251, 2006.

[87] R. A. Fisher. *The Genetical Theory of Natural Selection.* Dover, New York, 1958.

[88] J. Flannick, A. Novak, C. B. Do, B. S. Srinivasan, and S. Batzoglou. Automatic parameter learning for multiple network alignment. In *Proceedings of the 12th*

annual international conference on Research in computational molecular biology, RECOMB'08, pages 214–231, Berlin, Heidelberg, 2008. Springer-Verlag.

[89] P. Fontana, A. Cestaro, R. Velasco, E. Formentin, and S. Toppo. Rapid annotation of anonymous sequences from genome projects using semantic similarities and a weighting scheme in gene ontology. *PLoS ONE*, 4(2):e4619, 02 2009.

[90] E. Formstecher, S. Aresta, V. Collura, A. Hamburger, A. Meil, A. Trehin, C. Reverdy, V. Betin, S. Maire, C. Brun, B. Jacq, M. Arpin, Y. Bellaiche, S. Bellusci, P. Benaroch, M. Bornens, R. Chanet, P. Chavrier, O. Delattre, V. Doye, R. Fehon, G. Faye, T. Galli, J.-A. Girault, B. Goud, J. de Gunzburg, L. Johannes, M.-P. Junier, V. Mirouse, A. Mukherjee, D. Papadopoulo, F. Perez, A. Plessis, C. Ross, S. Saule, D. Stoppa-Lyonnet, A. Vincent, M. White, P. Legrain, J. Wojcik, J. Camonis, and L. Daviet. Protein interaction mapping: A Drosophila case study. *Genome Research*, 15(3):376–384, 2005.

[91] S. Fortunato. Community detection in graphs. *Physics Reports*, 486:75–174, 2010.

[92] S. Fortunato and M. Barthelemy. Resolution limit in community detection. *Proceedings of the National Academy of Sciences*, 104(1):36–41, 2007.

[93] E. B. Fowlkes and C. L. Mallows. A method for comparing two hierarchical clusterings. *Journal of the American Statistical Association*, 78(383):553–569, 1983.

[94] E. Fox Keller. Revisiting "scale-free" networks. *BioEssays*, 27(10):1060–1068, 2005.

[95] H. B. Fraser, A. E. Hirsh, L. M. Steinmetz, C. Scharfe, and M. W. Feldman. Evolutionary rate in the protein interaction network. *Science*, 296(5568):750–752, 2002.

[96] A. L. N. Fred and A. K. Jain. Robust data clustering. In *Proceedings IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages II 128–133, 2003.

[97] G. Gallone, T. Simpson, J. Armstrong, and A. Jarman. Bio::Homology::InterologWalk - A Perl module to build putative protein-protein interaction networks through interolog mapping. *BMC Bioinformatics*, 12(289), 2011.

[98] T. K. B. Gandhi, J. Zhong, S. Mathivanan, L. Karthick, K. N. Chandrika, M. S. Sujatha, S. Salil, P. Stefan, N. Shilpa, P. Balamurugan, M. Goparani, N. Kannabiran, S. Beiyi, D. Nandan, N. Rashmi, S. Malabika, B. J. D, P. Giovanni, S. Jorg, B. J. S, and P. Akhilesh. Analysis of the human protein interactome and comparison with yeast, worm and fly interaction datasets. *Nature Genetics*, 38(3):285–293, 2006.

[99] B. Gareth, P.-A. J. Manuel, L. Joyce, Y. Wehong, Y. Xiaochun, C. Veronica, S. Andrei, R. Dawn, B. Bryan, K. Nevan, D. Michael, P. John, G. Jack, and E. Andrew. Interaction network containing conserved and essential protein complexes in *Escherichia coli*. *Nature*, 433(7025):531–537, 2005.

[100] A. C. Gavin, M. Bosche, R. Krause, and P. Grandi. Functional organisation of the yeast proteome by systematic analysis of protein complexes. *Nature*, 415 (6868):141–147, 2002.

[101] A. C. Gavin, P. Aloy, P. Grandi, R. Krause, M. Boesche, M. Marzioch, C. Rau, L. J. Jensen, S. Bastuck, B. Dümpelfeld, et al. Proteome survey reveals modularity of the yeast cell machinery. *Nature*, 440(7084):631–636, 2006.

[102] J. Geisler-Lee, N. O'Toole, R. Ammar, N. J. Provart, A. H. Millar, and

M. Geisler. A predicted interactome for *Arabidopsis*. *Plant Physiology*, 145 (2):317–329, 2007.

[103] D. Gfeller, J.-C. Chappelier, and P. De Los Rios. Finding instabilities in the community structure of complex networks. *Physical Review E*, 72:056135, 2005.

[104] T. A. Gibson and D. S. Goldberg. Questioning the ubiquity of neofunctionalization. *PLoS Computational Biology*, 5(1):1–11, 2009.

[105] T. A. Gibson and D. S. Goldberg. Improving evolutionary models of protein interaction networks. *Bioinformatics*, 27(3):376–382, 2010.

[106] J. Gillis and P. Pavlidis. The role of indirect connections in gene networks in predicting function. *Bioinformatics*, 27(13):1860–1866, 2011.

[107] L. Giot, J. S. Bader, C. Brouwer, A. Chaudhuri, B. Kuang, Y. Li, Y. L. Hao, C. E. Ooi, B. Godwin, E. Vitols, G. Vijayadamodar, P. Pochart, H. Machineni, M. Welsh, Y. Kong, B. Zerhusen, R. Malcolm, Z. Varrone, A. Collis, M. Minto, S. Burgess, L. McDaniel, E. Stimpson, F. Spriggs, J. Williams, K. Neurath, N. Ioime, M. Agee, E. Voss, K. Furtak, R. Renzulli, N. Aanensen, S. Carrolla, E. Bickelhaupt, Y. Lazovatsky, A. DaSilva, J. Zhong, C. A. Stanyon, R. L. Finley, K. P. White, M. Braverman, T. Jarvie, S. Gold, M. Leach, J. Knight, R. A. Shimkets, M. P. McKenna, J. Chant, and J. M. Rothberg. A protein interaction map of *Drosophila melanogaster*. *Science*, 302(5651):1727–1736, 2003.

[108] M. Girvan and M. E. J. Newman. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*, 99(12):7821–7826, 2002.

[109] C. S. Goh, A. A. Bogan, M. Joachimiak, D. Walther, and F. E. Cohen. Co-

evolution of proteins with their interaction partners. *Journal of Molecular Biology*, 299(2):283–293, 2000.

[110] K.-I. Goh, M. E. Cusick, D. Valle, B. Childs, M. Vidal, and A.-L. Barabsi. The human disease network. *Proceedings of the National Academy of Sciences*, 104 (21):8685–8690, 2007.

[111] B. H. Good, Y.-A. de Montjoye, and A. Clauset. Performance of modularity maximization in practical contexts. *Physical Review E*, 81(4):046106, 2010.

[112] S. Götz, J. M. García-Gómez, J. Terol, T. D. Williams, S. H. Nagaraj, M. J. Nueda, M. Robles, M. Talón, J. Dopazo, and A. Conesa. High-throughput functional annotation and data mining with the Blast2GO suite. *Nucleic Acids Research*, 36(10):3420–3435, 2008.

[113] A. Grigoriev. On the number of protein-protein interactions in the yeast proteome. *Nucleic Acids Research*, 31(14):4157–4161, 2003.

[114] R. Guimera and L. A. N. Amaral. Functional cartography of complex metabolic networks. *Nature*, 433(7028):895–900, 2005.

[115] U. Güldener, M. Münsterkötter, G. Kastenmüller, N. Strack, J. van Helden, C. Lemer, J. Richelles, S. J. Wodak, J. García-Martínez, J. E. Pérez-Ortín, H. Michael, A. Kaps, E. Talla, B. Dujon, B. André, J. L. Souciet, J. De Montigny, E. Bon, C. Gaillardin, and H. W. Mewes. CYGD: the Comprehensive Yeast Genome Database. *Nucleic Acids Research*, 33(Database issue):364–368, 2005.

[116] M. W. Hahn and A. D. Kern. Comparative genomics of centrality and essentiality in three eukaryotic protein-interaction networks. *Molecular Biology and Evolution*, 22(4):803–806, 2005.

[117] M. W. Hahn, G. C. Conant, and A. Wagner. Molecular evolution in large genetic networks: does connectivity equal constraint? *Journal Molecular Evolution*, 58 (2):203–211, 2004.

[118] L. Hakes, S. C. Lovell, S. G. Oliver, and D. L. Robertson. Specificity in protein interactions and its relationship with sequence diversity and coevolution. *Proceedings of the National Academy of Sciences*, 104(19):7999–8004, 2007.

[119] L. Hakes, J. W. Pinney, D. L. Robertson, and S. C. Lovell. Protein-protein interaction networks and biology–what's the connection? *Nature Biotechnology*, 26:69–72, 2008.

[120] J.-D. Han, N. Bertin, T. Hao, D. Goldberg, G. F. Berriz, L. Zhang, D. Dupuy, A. J. Walhout, M. E. Cusick, F. P. Roth, and M. Vidal. Evidence for dynamically organized modularity in the yeast protein-protein interaction network. *Nature*, 430(6997):88–93, 2004.

[121] J.-D. Han, D. Dupuy, N. Bertin, M. E. Cusick, and M. Vidal. Effect of sampling on topology predictions of protein-protein interaction networks. *Nature Biotechnology*, 23(7):839–844, 2005.

[122] B. Hanczar, J. Hua, C. Sima, J. Weinstein, M. Bittner, and E. R. Dougherty. Small-sample precision of ROC-related estimates. *Bioinformatics*, 26(6):822–830, 2010.

[123] D. Hand. Measuring classifier performance: acoherent alternative to the area under the ROC curve. *Machine Learning*, 77:103–123, 2009.

[124] G. T. Hart, A. Ramani, and E. Marcotte. How complete are current yeast and human protein-interaction networks? *Genome Biology*, 7(11):120, 2006.

[125] L. H. Hartwell, J. J. Hopfield, S. Leibler, and A. W. Murray. From molecular to modular cell biology. *Nature*, 402(6761):C4–C52, 1999.

[126] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* Springer, 2001.

[127] M. B. Hastings. Community detection as an inference problem. *Physical Review E*, 74(3):035102, 2006.

[128] H. Hermjakob, L. Montecchi-Palazzi, G. Bader, J. Wojcik, L. Salwinski, A. Ceol, S. Moore, S. Orchard, U. Sarkans, C. von Mering, et al. The HUPO PSI's molecular interaction formata community standard for the representation of protein interaction data. *Nature Biotechnology*, 22(2):177–183, 2004.

[129] M. E. Hillenmeyer, E. Fung, J. Wildenhain, S. E. Pierce, S. Hoon, W. Lee, M. Proctor, R. P. St Onge, M. Tyers, D. Koller, et al. The chemical genomic portrait of yeast: uncovering a phenotype for all genes. *Science*, 320(5874): 362–365, 2008.

[130] E. Hirsh and R. Sharan. Identification of conserved protein complexes based on a model of protein network evolution. *Bioinformatics*, 23(2):e170–e176, 2007.

[131] Y. Ho, A. Gruhler, A. Heilbut, G. Bader, L. Moore, S. Adams, A. Millar, P. Taylor, K. Bennett, K. Boutilier, et al. Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectometry. *Nature*, 415 (6868):180–183, 2002.

[132] E. L. Hong, R. Balakrishnan, Q. Dong, K. R. Christie, J. Park, G. Binkley, M. C. Costanzo, S. S. Dwight, S. R. Engel, D. G. Fisk, et al. Gene Ontology annotations at SGD: new data sources and annotation methods. *Nucleic Acids Research*, 36(Database issue):D577, 2008.

[133] F. Hormozdiari, R. Salari, M. Hsing, A. Schönhuth, S. K. Chan, S. C. Sahinalp, and A. Cherkasov. The effect of insertions and deletions on wirings in protein-protein interaction networks: a large-scale study. *Journal of Computational Biology*, 16(2):159–167, 2009.

[134] D. W. Hosmer and S. Lemeshow. *Applied Logistic Regression, 2nd ed.* Wiley, 2000.

[135] M. Hsing, K. G. Byler, and A. Cherkasov. The use of Gene Ontology terms for predicting highly-connected 'hub' nodes in protein-protein interaction networks. *BMC Systems Biology*, 2(80), 2008.

[136] H. Huang and J. S. Bader. Precision and recall estimates for two-hybrid screens. *Bioinformatics*, 25(3):372–378, 2009.

[137] H. Huang, B. M. Jedynak, and J. S. Bader. Where have all the interactions gone? Estimating the coverage of two-hybrid protein interaction maps. *PLoS Computational Biology*, 3(11):e214, 2007.

[138] T.-W. Huang, A.-C. Tien, W.-S. Huang, Y.-C. G. Lee, C.-L. Peng, H.-H. Tseng, C.-Y. Kao, and C.-Y. F. Huang. POINT: a database for the prediction of protein-protein interactions based on the orthologous interactome. *Bioinformatics*, 20(17):3273–3276, 2004.

[139] T.-W. Huang, C.-Y. Lin, and C.-Y. Kao. Reconstruction of human protein interolog network using evolutionary conserved network. *BMC Bioinformatics*, 8(152), 2007.

[140] L. Hubert and P. Arabie. Comparing partitions. *Journal of Classification*, 2: 193–218, 1985.

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

[142] T. Ideker and R. Sharan. Protein networks in disease. *Genome Research*, 18 (4):644–652, 2008.

[143] P. Ingram, M. P. H. Stumpf, and J. Stark. Network motifs: structure does not determine function. *BMC Genomics*, 7(108), 2006.

[144] I. Ispolatov, P. L. Krapivsky, and A. Yuryev. Duplication-divergence model of protein interaction network. *Physical Review E*, 71(6):061911, 2005.

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

[146] R. Jansen, H. Yu, D. Greenbaum, Y. Kluger, N. J. Krogan, S. Chung, A. Emili, M. Snyder, J. F. Greenblatt, and M. Gerstein. A Bayesian networks approach for predicting protein-protein interactions from genomic data. *Science*, 302 (5644):449–453, 2003.

[147] L. J. Jensen, M. Kuhn, M. Stark, S. Chaffron, C. Creevey, J. Muller, T. Doerks, P. Julien, A. Roth, M. Simonovic, P. Bork, and C. von Mering. STRING 8a global view on proteins and their functional interactions in 630 organisms. *Nucleic Acids Research*, 37(suppl 1):D412–D416, 2009.

[148] H. Jeong, R. Tombor, R. Albert, A.-L. Barabási, and Z. N. Oltvai. The large-scale organization of metabolic networks. *Nature*, 407(6804):651–654, 2000.

[149] H. Jeong, S. P. Mason, A.-L. Barabási, and Z. N. Oltvai. Lethality and centrality in protein networks. *Nature*, 411(6833):41–42, 2001.

[150] P. Jonsson, T. Cavanna, D. Zicha, and P. Bates. Cluster analysis of networks generated through homology: automatic identification of important protein communities involved in cancer metastasis. *BMC Bioinformatics*, 7(2), 2006.

[151] I. K. Jordan, Y. I. Wolf, and E. V. Koonin. No simple dependence between protein evolution rate and the number of protein-protein interactions: only the most prolific interactors tend to evolve slowly. *BMC Evolutionary Biology*, 3 (5), 2003.

[152] R. Jothi, P. F. Cherukuri, A. Tasneem, and T. M. Przytycka. Co-evolutionary analysis of domains in interacting proteins reveals insights into domain-domain interactions mediating protein-protein interactions. *Journal of Molecular Biology*, 362(4):861–875, 2006.

[153] D. Juan, F. Pazos, and A. Valencia. High-confidence prediction of global interactomes based on genome-wide coevolutionary networks. *Proceedings of the National Academy of Sciences*, 105(3):934–939, 2008.

[154] T. Kamada and S. Kawai. An algorithm for drawing general undirected graphs. *Information Processing Letters*, 31(1):7–15, 1989.

[155] M. G. Kann, B. A. Shoemaker, A. R. Panchenko, and T. M. Przytycka. Correlated evolution of interacting proteins: looking behind the mirrortree. *Journal of Molecular Biology*, 385(1):91–98, 2009.

[156] B. Karrer, E. Levina, and M. E. J. Newman. Robustness of community structure in networks. *Physical Review E*, 77:046119, 2008.

[157] M. Kasahara. The 2R hypothesis: an update. *Current Opinion in Immunology*, 19(5):547–552, 2007.

[158] A. K. Kenworthy. Imaging protein-protein interactions using Fluorescence Resonance Energy Transfer microscopy. *Methods*, 24(3):289–296, 2001.

[159] S. Kerrien, Y. Alam-Faruque, B. Aranda, I. Bancarz, A. Bridge, C. Derow, E. Dimmer, M. Feuermann, A. Friedrichsen, R. Huntley, et al. IntAct–open source resource for molecular interaction data. *Nucleic Acids Research*, 35 (Database issue):D561, 2007.

[160] S. Kerrien, S. Orchard, L. Montecchi-Palazzi, B. Aranda, A. Quinn, N. Vinod, G. Bader, I. Xenarios, J. Wojcik, D. Sherman, M. Tyers, J. Salama, S. Moore, A. Ceol, A. Chatr-aryamontri, M. Oesterheld, V. Stumpflen, L. Salwinski, J. Nerothin, E. Cerami, M. Cusick, M. Vidal, M. Gilson, J. Armstrong, P. Woollard, C. Hogue, D. Eisenberg, G. Cesareni, R. Apweiler, and H. Hermjakob. Broadening the horizon - level 2.5 of the HUPO-PSI format for molecular interactions. *BMC Biology*, 5(44), 2007.

[161] T. S. Keshava Prasad, R. Goel, K. Kandasamy, S. Keerthikumar, S. Kumar, S. Mathivanan, D. Telikicherla, R. Raju, B. Shafreen, A. Venugopal, L. Balakrishnan, A. Marimuthu, S. Banerjee, D. S. Somanathan, A. Sebastian, S. Rani, S. Ray, C. J. Harrys Kishore, S. Kanth, M. Ahmed, M. K. Kashyap, R. Mohmood, Y. L. Ramachandra, V. Krishna, B. A. Rahiman, S. Mohan, P. Ranganathan, S. Ramabadran, R. Chaerkady, and A. Pandey. Human Protein Reference Database – 2009 update. *Nucleic Acids Research*, 37(suppl 1): D767–D772, 2009.

[162] O. Keskin, B. Ma, and R. Nussinov. Hot regions in protein–protein interactions: the organization and contribution of structurally conserved hot spot residues. *Journal of Molecular Biology*, 345(5):1281–1294, 2005.

[163] J. Kim and T. Wilhelm. What is a complex graph? *Physica A*, 387:2637–2652, 2008.

[164] P. M. Kim, L. J. Lu, Y. Xia, and M. B. Gerstein. Relating three-dimensional structures to protein networks provides evolutionary insights. *Science*, 314 (5807):1938–1941, 2006.

[165] W. K. Kim and E. M. Marcotte. Age-dependent evolution of the yeast protein interaction network suggests a limited role of gene duplication and divergence. *PLoS Computational Biology*, 4:e1000232, 2008.

[166] H. Kitano. Systems biology: A brief overview. *Science*, 295(5560):1662–1664, 2002.

[167] F. A. Kondrashov and A. S. Kondrashov. Role of selection in fixation of gene duplications. *Journal of Theoretical Biology*, 239(2):141–151, 2006.

[168] F. A. Kondrashov and E. V. Koonin. A common framework for understanding the origin of genetic dominance and evolutionary fates of gene duplications. *Trends in Genetics*, 20(7):287–290, 2004.

[169] E. Kotelnikova, A. Kalinin, A. Yuryev, and S. Maslov. Prediction of protein-protein interactions on the basis of evolutionary conservation of protein functions. *Evolutionary Bioinformatics*, 3:197–206, 2007.

[170] N. J. Krogan, G. Cagney, H. Yu, G. Zhong, X. Guo, A. Ignatchenko, J. Li, S. Pu, N. Datta, A. P. Tikuisis, T. Punna, J. M. Peregrín-Alvarez, M. Shales, X. Zhang, M. Davey, M. D. Robinson, A. Paccanaro, J. E. Bray, A. Sheung, B. Beattie, D. P. Richards, V. Canadien, A. Lalev, F. Mena, P. Wong, A. Starostine, M. M. Canete, J. Vlasblom, S. Wu, C. Orsi, S. R. Collins, S. Chandran, R. Haw, J. J. Rilstone, K. Gandi, N. J. Thompson, G. Musso, P. St Onge, S. Ghanny, M. H.

Lam, G. Butland, A. M. Altaf-Ul, S. Kanaya, A. Shilatifard, E. O'Shea, J. S. Weissman, C. J. Ingles, T. R. Hughes, J. Parkinson, M. Gerstein, S. J. Wodak, A. Emili, and J. F. Greenblatt. Global landscape of protein complexes in the yeast *Saccharomyces cerevisiae*. *Nature*, 440(7084):637–643, 2006.

[171] R. Lambiotte. Multi-scale modularity in complex networks. In *Modeling and Optimization in Mobile, Ad Hoc and Wireless Networks (WiOpt), 2010 Proceedings of the 8th International Symposium on*, pages 546–553, 2010.

[172] R. Lambiotte, R. Sinatra, J.-C. Delvenne, T. S. Evans, M. Barahona, and V. Latora. Flow graphs: Interweaving dynamics and structure. *Physical Review E*, 84:017102, 2011.

[173] A. Lancichinetti and S. Fortunato. Community detection algorithms: a comparative analysis. *Physical Review E*, 80(5):056117, 2009.

[174] A. Lancichinetti, S. Fortunato, and F. Radicchi. Benchmark graphs for testing community detection algorithms. *Physical Review E*, 78(4):46110, 2008.

[175] A. Lancichinetti, S. Fortunato, and J. Kertsz. Detecting the overlapping and hierarchical community structure in complex networks. *New Journal of Physics*, 11:033015, 2009.

[176] V. Latora and M. Marchiori. Efficient behavior of small-world networks. *Physical Review Letters*, 87(19):198701, 2001.

[177] S.-A. Lee, C.-h. Chan, C.-H. Tsai, J.-M. Lai, F.-S. Wang, C.-Y. Kao, and C.-Y. Huang. Ortholog-based protein-protein interaction prediction and its application to inter-species interactions. *BMC Bioinformatics*, 9 (Suppl 12) (S11), 2008.

[178] I. Lemmens, S. Eyckerman, L. Zabeau, D. Catteeuw, E. Vertenten, K. Verschueren, D. Huylebroeck, J. Vandekerckhove, and J. Tavernier. Heteromeric MAPPIT: a novel strategy to study modificationdependent proteinprotein interactions in mammalian cells. *Nucleic Acids Research*, 31(14):e75, 2003.

[179] E. D. Levy and J. B. Pereira-Leal. Evolution and dynamics of protein interactions and networks. *Current Opinion in Structural Biology*, 18(3):349–357, 2008.

[180] A. C. F. Lewis, N. S. Jones, M. A. Porter, and C. M. Deane. What evidence is there for the homology of protein interactions? *submitted*.

[181] A. C. F. Lewis, N. S. Jones, M. A. Porter, and C. M. Deane. The function of communities in protein interaction networks at multiple scales. *BMC Systems Biology*, 4(100), 2010.

[182] A. C. F. Lewis, R. Saeed, and C. M. Deane. Predicting protein-protein interactions in the context of protein evolution. *Molecular Biosystems*, 6:55–64, 2010.

[183] C. Li and P. K. Maini. An evolving network model with community structure. *Journal of Physics A: Mathematical and General*, 38:9741–9749, 2005.

[184] M. Li, J. Wang, and J. Chen. A graph-theoretic method for mining overlapping functional modules in protein interaction networks. *Lecture Notes in Bioinformatics*, 4983:208–219, 2008.

[185] S. Li, C. M. Armstrong, N. Bertin, H. Ge, S. Milstein, M. Boxem, P. O. Vidalain, J. D. Han, A. Chesneau, T. Hao, D. S. Goldberg, N. Li, M. Martinez, J. F. Rual, P. Lamesch, L. Xu, M. Tewari, S. L. Wong, L. V. Zhang, G. F. Berriz, L. Jacotot, P. Vaglio, J. Reboul, T. Hirozane-Kishikawa, Q. Li, H. W.

Gabel, A. Elewa, B. Baumgartner, D. J. Rose, H. Yu, S. Bosak, R. Sequerra, A. Fraser, S. E. Mango, W. M. Saxton, S. Strome, S. Van Den Heuvel, F. Piano, J. Vandenhaute, C. Sardet, M. Gerstein, L. Doucette-Stamm, K. C. Gunsalus, J. W. Harper, M. E. Cusick, F. P. Roth, D. E. Hill, and M. Vidal. A map of the interactome network of the metazoan *C. elegans*. *Science*, 303(5657):540–543, 2004.

[186] Z. Liang, M. Xu, M. Teng, and L. Niu. Comparison of protein interaction networks reveals species conservation and divergence. *BMC Bioinformatics*, 7 (457), 2006.

[187] C.-S. Liao, K. Lu, M. Baym, R. Singh, and B. Berger. IsoRankN: spectral methods for global alignment of multiple protein networks. *Bioinformatics*, 25 (12):i253–i258, 2009.

[188] D. Lin. An information-theoretic definition of similarity. In *In Proceedings of the 15th International Conference on Machine Learning*, pages 296–304. Morgan Kaufmann, 1998.

[189] J. A. Loo. Studying noncovalent protein complexes by electrospray ionization mass spectrometry. *Mass Spectrometry Review*, 16(1):1–23, 1997.

[190] P. W. Lord, R. D. Stevens, A. Brass, and C. A. Goble. Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation. *Bioinformatics*, 19(10):1275–1283, 2003.

[191] F. Luo, Y. Yang, C. F. Chen, R. Chang, J. Zhou, and R. H. Scheuermann. Modular organization of protein interaction networks. *Bioinformatics*, 23(2): 207–214, 2007.

[192] M. Lynch. The evolution of genetic networks by non-adaptive processes. *Nature Reviews Genetics*, 8(10):803–813, 2007.

[193] M. Lynch. *The Origins of Genome Architecture*. Sinauer Associates Inc., 2007.

[194] D. Lynn, C. Chan, M. Naseer, M. Yau, R. Lo, A. Sribnaia, G. Ring, J. Que, K. Wee, G. Winsor, M. Laird, K. Breuer, A. Foroushani, F. Brinkman, and R. Hancock. Curating the innate immunity interactome. *BMC Systems Biology*, 4(1):117, 2010.

[195] D. J. C. MacKay. *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press, 2003.

[196] J. P. MacKay, M. Sunde, J. A. Lowry, M. Crossley, and J. M. Matthews. Protein interactions: is seeing believing? *Trends in Biochemical Sciences*, 32:530–531, 2007.

[197] J. Macqueen. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley Symposium on Mathematical Statistics and Probability*, pages 281–297, 1967.

[198] K. Manolis, B. B. W., and L. E. S. Proof and evolutionary analysis of ancient genome duplication in the yeast saccharomyces cerevisiae. *Nature*, 428(6983): 617–624, 2004. 10.1038/nature02424.

[199] I. Maraziotis, K. Dimitrakopoulou, and A. Bezerianos. An *in silico* method for detecting overlapping functional modules from composite biological networks. *BMC Systems Biology*, 2(93), 2008.

[200] A. Marchler-Bauer, J. B. Anderson, M. K. Derbyshire, C. DeWeese-Scott, N. R. Gonzales, M. Gwadz, L. Hao, S. He, D. I. Hurwitz, J. D. Jackson, Z. Ke, D. Krylov, C. J. Lanczycki, C. A. Liebert, C. Liu, F. Lu, S. Lu, G. H. Marchler, M. Mullokandov, J. S. Song, N. Thanki, R. A. Yamashita, J. J. Yin, D. Zhang, and S. H. Bryant. CDD: a conserved domain database for interactive domain family analysis. *Nucleic Acids Research*, 35(Database issue):237–240, 2007.

[201] C. J. Marcotte, M. Pellegrini, H. L. Ng, D. W. Rice, T. O. Yeates, and D. Eisenberg. Detecting protein function and protein-protein interactions from genome sequences. *Science*, 285(5428):751–753, 1999.

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

[203] C. P. Massen and J. P. K. Doye. Thermodynamics of community structure. *arXiv:cond-mat/0610077v1*, 2006.

[204] L. R. Matthews, P. Vaglio, J. Reboul, H. Ge, B. P. Davis, J. Garrels, S. Vincent, and M. Vidal. Identification of potential interaction networks using sequence-based searches for conserved protein-protein interactions or 'interologs'. *Genome Research*, 11(12):2120–2126, 2001.

[205] K. Mehmet, K. Yohan, T. Umut, S. Shankar, S. Wojciech, and G. Ananth. Pairwise alignment of protein interaction networks. *Journal of Computational Biology*, 13(2):182–199, 2006.

[206] M. Meila and D. Heckerman. An experimental comparison of model-based clustering methods. *Machine Learning*, 42(1/2):9–29, 2001.

[207] M. Meilă. Comparing clusterings–an information based distance. *Journal of Multivariate Analysis*, 98(5):873–895, 2007.

[208] W. Mendenhall, R. J. Beaver, and B. M. Beaver. *Introduction to Probability and Statistics*. Brooks/Cole, 2008.

[209] R. Mendez, R. Leplae, M. F. Lensink, and S. J. Wodak. Assessment of CAPRI predictions in rounds 3–5 shows progress in docking procedures. *Proteins*, 60 (2):150–169, 2005.

[210] M. Mete, F. Tang, X. Xu, and N. Yuruk. A structural approach for finding functional modules from large biological networks. *BMC Bioinformatics*, 9 (S19), 2008.

[211] H. W. Mewes, D. Frishman, U. Guldener, G. Mannhaupt, K. Mayer, M. Mokrejs, B. Morgenstern, M. Munsterkotter, S. Rudd, and B. Weil. MIPS: A database for genomes and protein sequences. *Nucleic Acids Research*, 30(1): 31–34, 2002.

[212] H. W. Mewes, C. Amid, R. Arnold, D. Frishman, U. Güldener, G. Mannhaupt, M. Münsterkötter, P. Pagel, N. Strack, V. Stümpflen, J. Warfsmann, and A. Ruepp. MIPS: analysis and annotation of proteins from whole genomes. *Nucleic Acids Research*, 32(Database issue):41–44, 2004.

[213] H. W. Mewes, D. Frishman, K. F. Mayer, M. Münsterkötter, O. Noubibou, P. Pagel, T. Rattei, M. Oesterheld, A. Ruepp, and V. Stümpflen. MIPS: analysis and annotation of proteins from whole genomes in 2005. *Nucleic Acids Research*, 34(Database issue):169–172, 2006.

[214] M. Michaut, S. Kerrien, L. Montecchi-Palazzi, C. Cassier-Chauvat, F. Chauvat, J.-C. Aude, P. Legrain, and H. Hermjakob. InteroPORC: an automated tool to predict highly conserved protein interaction networks. *BMC Bioinformatics*, 9 (Suppl 10)(P1), 2008.

[215] M. Middendorf, E. Ziv, and C. Wiggins. Inferring network mechanisms: The Drosophila melanogaster protein interaction network. *Proceedings of the National Academy of Sciences*, 102(9):3192–3197, 2004.

[216] S. Mika and B. Rost. Protein-protein interactions more conserved within species than across species. *PLoS Computational Biology*, 2(7):e79, 2006.

[217] J. P. Miller, R. S. Lo, A. Ben-Hur, C. Desmarais, I. Stagljar, W. S. Noble, and S. Fields. Large-scale identification of yeast integral membrane protein interactions. *Proceedings of the National Academy of Sciences*, 102(34):12123–12128, 2005.

[218] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon. Network motifs: Simple building blocks of complex networks. *Science*, 298 (5594):824–827, 2002.

[219] R. Milo, S. Itzkovitz, N. Kashtan, R. Levitt, and U. Alon. Response to Comment on “Network Motifs: Simple Building Blocks of Complex Networks” and “Superfamilies of Evolved and Designed Networks”. *Science*, 305(5687):1107, 2004.

[220] J. Mintseris and Z. Weng. Structure, function, and evolution of transient and obligate protein-protein interactions. *Proceedings of the National Academy of Sciences*, 102(31):10930–10935, 2005.

[221] I. S. Moreira, P. A. Fernandes, and M. J. Ramos. Hot spots–a review of the protein-protein interface determinant amino-acid residues. *Proteins*, 68(4):803–812, 2007.

[222] E. Morett, J. O. Korbel, E. Rajan, G. Saab-Rincon, L. Olvera, M. Olvera, S. Schmidt, B. Snel, and P. Bork. Systematic discovery of analogous enzymes in thiamin biosynthesis. *Nature Biotechnology*, 21(7):790–795, 2003.

[223] D. A. Morrison. The Timetree of Life. *Systematic Biology*, 58(4):461–462, 2009.

[224] S. Mostafavi, D. Ray, D. Warde-Farley, C. Grouios, and Q. Morris. GeneMANIA: a real-time multiple association network integration algorithm for predicting gene function. *Genome Biology*, 9(Suppl 1):S4, 2008.

[225] F. Mousson, A. Kolkman, W. W. M. P. Pijnappel, H. T. M. Timmers, and A. J. R. Heck. Quantitative proteomics reveals regulation of dynamic components within TATA-binding protein (TBP) transcription complexes. *Molecular & Cellular Proteomics*, 7(5):845–852, 2008.

[226] P. J. Mucha, T. Richardson, K. Macon, M. A. Porter, and J.-P. Onnela. Community structure in time-dependent, multiscale, and multiplex networks. *Science*, 328(5980):876–878, 2010.

[227] A. G. Murzin, S. E. Brenner, T. Hubbard, and C. Chothia. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *Journal of Molecular Biology*, 247(4):536–540, 1995.

[228] S. Navlakha and C. Kingsford. Network archaeology: Uncovering ancient networks from present-day interactions. *PLoS Computational Biology*, 7(4): e1001119, 2011.

[229] N. L. Nehrt, W. T. Clark, P. Radivojac, and M. W. Hahn. Testing the ortholog conjecture with comparative functional genomic data from mammals. *PLoS Computational Biology*, 7(6):e1002073, 2011.

[230] M. E. J. Newman. Assortative mixing in networks. *Physical Review Letters*, 89 (20):208701, 2002.

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

[232] M. E. J. Newman. Detecting community structure in networks. *The European Physical Journal B*, 38:321–330, 2004.

[233] M. E. J. Newman. Finding community structure in networks using the eigenvectors of matrices. *Physical Review E*, 74(3):36104, 2006.

[234] M. E. J. Newman. The physics of networks. *Physics Today*, 61(11):33–38, 2008.

[235] M. E. J. Newman. *Networks: An Introduction*. Oxford University Press, 2010.

[236] M. E. J. Newman and M. Girvan. Finding and evaluating community structure in networks. *Physical Review E*, 69(2):26113, 2004.

[237] D. Noble. Modeling the heart–from genes to cells to the whole organ. *Science*, 295(5560):1678–1682, 2002.

[238] K. P. O'Brien, M. Remm, and E. L. L. Sonnhammer. Inparanoid: a comprehensive database of eukaryotic orthologs. *Nucleic Acids Research*, 33(suppl 1): D476–D480, 2005.

[239] Y. Ofran and B. Rost. Analysing six types of protein-protein interfaces. *Journal of Molecular Biology*, 325(2):377–387, 2003.

[240] C. A. Orengo and J. M. Thornton. Protein families and their evolution-a structural perspective. *Annual Review Biochemistry*, 74:867–900, 2005.

[241] A. A. Orr. The genetic theory of adaptation: a brief history. *Nature Reviews Genetics*, 6(2):119–127, 2005.

[242] R. Overbeek, M. Fonstein, M. D'Souza, G. D. Pusch, and N. Maltsev. The use of gene clusters to infer functional coupling. *Proceedings of the National Academy of Sciences*, 96(6):2896–2901, 1999.

[243] P. Pagel, P. Wong, and D. Frishman. A domain interaction map based on phylogenetic profiling. *Journal of Molecular Biology*, 344(5):1331–1346, 2004.

[244] C. Pál, B. Papp, and M. J. Lercher. An integrated view of protein evolution. *Nature Reviews Genetics*, 7(5):337–348, 2006.

[245] G. Palla, I. Derényi, I. Farkas, and T. Vicsek. Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, 435 (7043):814–818, 2005.

[246] G. Palla, A.-L. Barabási, and T. Vicsek. Quantifying social group evolution. *Nature*, 446(7136):664–667, 2007.

[247] J. Pandey, M. Koyutrk, S. Subramaniam, and A. Grama. Functional coherence in domain interaction networks. *Bioinformatics*, 24(16):i28–i34, 2008.

[248] B. Papp, C. Pál, and L. D. Hurst. Dosage sensitivity and the evolution of gene families in yeast. *Nature*, 424(6945):194–197, 2003.

[249] A. Patil and H. Nakamura. Filtering high-throughput protein-protein interaction data using a combination of genomic features. *BMC Bioinformatics*, 6 (100), 2005.

[250] F. Pazos and A. Valencia. Similarity of phylogenetic trees as indicator of protein-protein interaction. *Protein Engineering*, 14(9):609–614, 2001.

[251] F. Pazos and A. Valencia. Protein co-evolution, co-adaptation and interactions. *EMBO Journal*, 27(20):2648–2655, 2008.

[252] M. Pellegrini, E. Marcotte, M. J. Thompson, D. Eisenberg, and T. O. Yeates. Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proceedings of the National Academy of Sciences*, 96(8):4285–4288, 1999.

[253] J. B. Pereira-Leal, A. J. Enright, and C. A. Ouzounis. Detection of functional modules from protein interaction networks. *Proteins: Structure, Function and Genetics*, 54(1):49–57, 2004.

[254] J. B. Pereira-Leal, E. D. Levy, and S. A. Teichmann. The origins and evolution of functional modules: lessons from protein complexes. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 361(1467):507–517, 2006.

[255] M. Persico, A. Ceol, C. Gavrila, R. Hoffmann, A. Florio, and G. Cesareni. HomoMINT: an inferred human network based on orthology mapping of protein interactions discovered in model organisms. *BMC Bioinformatics*, 6 (Suppl 4) (S21), 2005.

[256] Pinkert, S. and Schultz, J. and Reichardt, J. Protein interaction networks - more than mere modules. *PLoS Computational Biology*, 6(1):e1000659, 2010.

[257] S. Pitre, M. Alamgir, J. R. Green, M. Dumontier, F. Dehne, and A. Golshani. Computational methods for predicting protein-protein interactions. *Advances in Biochemical Engineering / Biotechnology*, 110:247–267, 2008.

[258] P. Pons and M. Latapy. Post-processing hierarchical community structures: Quality improvements and multi-scale view. *Theoretical Computer Science*, 412 (8-10):892–900, 2011.

[259] M. A. Porter, J.-P. Onnela, and P. J. Mucha. Communities in networks. *Notices of the American Mathematical Society*, 56(9):1082–1097, 1164–1166, 2009.

[260] O. Puig, F. Caspary, G. Rigaut, B. Rutz, E. Bouveret, E. Bragado-Nilsson, M. Wilm, and B. Séraphin. The Tandem Affinity Purification (TAP) method: A general procedure of protein complex purification. *Methods*, 24(3):218–229, 2001.

[261] W. Qian and J. Zhang. Gene dosage and gene duplicability. *Genetics*, 179(4): 2319–2324, 2008.

[262] W. Qian, X. He, E. Chan, H. Xu, and J. Zhang. Measuring the evolutionary rate of protein-protein interaction. *Proceedings of the National Academy of Sciences*, 108(21):8725–8730, 2011.

[263] H. Qin, H. H. S. Lu, W. B. Wu, and W.-H. Li. Evolution of the yeast protein interaction network. *Proceedings of the National Academy of Sciences*, 100(22): 12820–12824, 2003.

[264] F. Radicchi, A. Lancichinetti, and J. J. Ramasco. Combinatorial approach to modularity. *Physical Review E*, 82(2):026102, 2010.

[265] W. M. Rand. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, 66(336):846–850, 1971.

[266] O. Ratmann, O. Jorgensen, T. Hinkley, M. M. P. Stumpf, S. Richardson, and C. Wiuf. Using likelihood-free inference to compare evolutionary dynamics of the protein networks of *H. pylori* and *P. falciparum*. *PLoS Computational Biology*, 3(11):2266–2278, 2007.

[267] T. Ravasi, H. Suzuki, C. V. Cannistraci, S. Katayama, V. B. Bajic, T. Kai, A. Altuna, S. Sebastian, K.-K. Mutsumi, B. Nicolas, C. Piero, D. C. O., F. A. R.R., G. Julian, G. Sean, H. Jung-Hoon, H. Takehiro, H. Winston, H. Oliver, K. Atanas, K. Mandeep, K. Hideya, K. Atsutaka, L. Timo, van Nimwegen Erik, M. C. Ross, O. Chihiro, R. Aleksandar, S. Ariel, T. R. D., T. Jesper, L. Boris, T. S. A., A. Takahiro, N. Noriko, M. Kayoko, T. Michihira, F. Shiro, I. Kengo, K. Chikatoshi, I. Ryoko, K. Yayoi, K. Jun, H. D. A., I. Trey, and H. Yoshihide. An atlas of combinatorial transcriptional regulation in mouse and man. *Cell*, 140(5):744–752, 2010.

[268] E. Ravasz, A. L. Somera, D. A. Mongru, Z. N. Oltvai, and A.-L. Barabási.

Hierarchical organization of modularity in metabolic networks. *Science*, 297 (5586):1551–1555, 2002.

[269] T. Reguly, A. Breitkreutz, L. Boucher, B. J. Breitkreutz, G. C. Hon, C. L. Myers, A. Parsons, H. Friesen, R. Oughtred, A. Tong, C. Stark, Y. Ho, D. Botstein, B. Andrews, C. Boone, O. G. Troyanskya, T. Ideker, K. Dolinski, N. N. Batada, and M. Tyers. Comprehensive curation and analysis of global interaction networks in Saccharomyces cerevisiae. *Journal of Biology*, 5(4):11, 2006.

[270] J. Reichardt and S. Bornholdt. Statistical mechanics of community detection. *Physical Review E*, 74(1):16110, 2006.

[271] J. Reichardt and S. Bornholdt. When are networks truly modular? *Physica D*, 224:20–26, 2006.

[272] J. Reichardt and S. Bornholdt. Partitioning and modularity of graphs with arbitrary degree distribution. *Physical Review E*, 76:015102, 2007.

[273] P. Resnik. Using information content to evaluate semantic similarity in a taxonomy. In *International Joint Conference on Artificial Intelligence*, volume 14, pages 448–453. Lawrence Erlbaum Associates Ltd, 1995.

[274] S. A. Rice. The identication of blocs in small political bodies. *American Political Science Review*, 21:619–627, 1927.

[275] R. Riley, C. Lee, C. Sabatti, and D. Eisenberg. Inferring protein domain interactions from databases of interacting proteins. *Genome Biology*, 6(R89), 2005.

[276] A. W. Rives and T. Galitski. Modular organization of cellular networks. *Proceedings of the National Academy of Sciences*, 100(3):1128–1133, 2003.

[277] S. Roded and I. Trey. Modeling cellular machinery through biological network comparison. *Nature Biotechnology*, 24(4):427–433, 2006.

[278] P. Ronhovde and Z. Nussinov. Local resolution-limit-free potts model for community detection. *Physical Review E*, 81:046114, 2010.

[279] J. F. Rual, K. Venkatesan, T. Hao, T. Hirozane-Kishikawa, A. Dricot, N. Li, G. F. Berriz, F. D. Gibbons, M. Dreze, N. Ayivi-Guedehoussou, et al. Towards a proteome-scale map of the human protein-protein interaction network. *Nature*, 437(7062):1173–1178, 2005.

[280] J. Ruan, H. Li, Z. Chen, A. Coghlan, L. J. M. Coin, Y. Guo, J.-K. Hrich, Y. Hu, K. Kristiansen, R. Li, T. Liu, A. Moses, J. Qin, S. Vang, A. J. Vilella, A. Ureta-Vidal, L. Bolund, J. Wang, and R. Durbin. Treefam: 2008 update. *Nucleic Acids Research*, 36(suppl 1):D735–D740, 2008.

[281] R. B. Russell, F. Alber, P. Aloy, F. P. Davis, D. Korkin, M. Pichaud, M. Topf, and A. Sali. A structural perspective on protein-protein interactions. *Current Opinion in Structural Biology*, 14(3):313–324, 2004.

[282] R. Saeed and C. M. Deane. Protein protein interactions, evolutionary rate, abundance and age. *BMC Bioinformatics*, 7(128), 2006.

[283] R. Saeed and C. M. Deane. An assessment of the uses of homologous interactions. *Bioinformatics*, 24:689–695, 2008.

[284] R. Saito, H. Suzuki, and Y. Hayashizaki. Interaction generality, a measurement to assess the reliability of a protein-protein interaction. *Nucleic Acids Research*, 30(5):1163–1168, 2002.

[285] L. Salwinski, C. S. Miller, A. J. Smith, F. K. Pettit, J. U. Bowie, and D. Eisen-

berg. The database of interacting proteins: 2004 update. *Nucleic Acids Research*, 32(suppl 1):D449–D451, 2004.

[286] L. Salwinski, L. Licata, A. Winter, D. Thorneycroft, J. Khadake, A. Ceol, A. C. Aryamontri, R. Oughtred, M. Livstone, L. Boucher, D. Botstein, K. Dolinski, T. Berardini, E. Huala, M. Tyers, D. Eisenberg, G. Cesareni, and H. Hermjakob. Recurated protein interaction datasets. *Nature Methods*, 6(12):860–861, 2009.

[287] L. Sambourg and N. Thierry-Mieg. New insights into protein-protein interaction data lead to increased estimates of the S. cerevisiae interactome size. *BMC Bioinformatics*, 11(605), 2010.

[288] V. Sangar, D. Blankenberg, N. Altman, and A. Lesk. Quantitative sequence-function relationships in proteins based on gene ontology. *BMC Bioinformatics*, 8(1):294, 2007.

[289] D. R. Scannell, G. Butler, and K. H. Wolfe. Yeast genome evolution–the origin of the species. *Yeast*, 24(11):929–942, 2007.

[290] A. Schlicker, F. S. Domingues, J. Rahnenfuhrer, and T. Lengauer. A new measure for functional similarity of gene products based on Gene Ontology. *BMC Bioinformatics*, 7(302), 2006.

[291] J. L. Sevilla, V. Segura, A. Podhorski, E. Guruceaga, J. M. Mato, L. A. Martinez-Cruz, F. J. Corrales, and A. Rubio. Correlation between gene expression and GO semantic similarity. *Computational Biology and Bioinformatics, IEEE/ACM Transactions on*, 2(4):330–338, 2005.

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

[293] R. Sharan, I. Ulitsky, and R. Shamir. Network-based prediction of protein function. *Molecular Systems Biology*, 3(88), 2007.

[294] B. A. Shoemaker and A. R. Panchenko. Deciphering protein–protein interactions. Part I. Experimental techniques and databases. *PLoS Computational Biology*, 3(3):337–334, 2007.

[295] B. A. Shoemaker and A. R. Panchenko. Deciphering protein-protein interactions. Part II. Computational methods to predict protein and domain interaction partners. *PLoS Computational Biology*, 3(4):595–601, 2007.

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

[297] R. Singh, J. Xu, and B. Berger. Global alignment of multiple protein interaction networks with application to functional orthology detection. *Proceedings of the National Academy of Sciences*, 105(35):12763–12768, 2008.

[298] G. R. Smith and M. J. Sternberg. Prediction of protein-protein interactions by docking methods. *Current Opinion in Structural Biology*, 12(1):28–35, 2002.

[299] B. Snel and M. A. Huynen. Quantifying modularity in the evolution of biomolecular systems. *Genome Research*, 14(3):391–397, 2004.

[300] S. N. Soffer and A. Vázquez. Network clustering coefficient without degree-correlation biases. *Physical Review E*, 71(5):57101, 2005.

[301] R. R. Sokal and F. J. Rohlf. *Biometry: The Principles and Practice of Statistics in Biological Research.* Freeman, San Francisco, 1995.

[302] R. V. Solé and S. Valverde. Are network motifs the spandrels of cellular complexity? *Trends in Ecology & Evolution*, 21(8):419–422, 2006.

[303] J. Song and M. Singh. How and when should interactome-derived clusters be used to predict functional modules and protein function? *Bioinformatics*, 25 (23):3143–3150, 2009.

[304] E. Sprinzak and H. Margalit. Correlated sequence-signatures as markers of protein-protein interaction. *Journal of Molecular Biology*, 311(4):681–692, 2001.

[305] E. Sprinzak, S. Sattath, and H. Margalit. How reliable are experimental protein-protein interaction data? *Journal of Molecular Biology*, 327(5):919–923, 2003.

[306] C. Stark, B. J. Breitkreutz, T. Reguly, L. Boucher, A. Breitkreutz, and M. Tyers. BioGRID: a general repository for interaction datasets. *Nucleic Acids Research*, 34:D535–539, 2006.

[307] C. Stark, B.-J. Breitkreutz, A. Chatr-aryamontri, L. Boucher, R. Oughtred, M. S. Livstone, J. Nixon, K. Van Auken, X. Wang, X. Shi, T. Reguly, J. M. Rust, A. Winter, K. Dolinski, and M. Tyers. The BioGRID Interaction Database: 2011 update. *Nucleic Acids Research*, 39(suppl 1):D698–D704, 2011.

[308] L. D. Stein. Human genome: End of the beginning. *Nature*, 431(7011):915–916, 2004.

[309] O. Stephen. Proteomics: Guilt-by-association goes global. *Nature*, 403(6770): 601–603, 2000.

[310] M. P. H. Stumpf, P. Ingram, I. Nouvel, and C. Wiuf. Statistical model selection methods applied to biological networks. In C. Priami, E. Merelli, P. Gonzalez, and A. Omicini, editors, *Transactions on Computational Systems Biology III*, volume 3737 of *Lecture Notes in Computer Science*, pages 65–77. Springer Berlin / Heidelberg, 2005.

[311] M. P. H. Stumpf, C. Wiuf, and R. M. May. Subnets of scale-free networks are not scale-free: sampling properties of networks. *Proceedings of the National Academy of Sciences*, 102(12):4221–4224, 2005.

[312] M. P. H. Stumpf, W. P. Kelly, T. Thorne, and C. Wiuf. Evolution at the system level: the natural history of protein interaction networks. *Trends in Ecology and Evolution*, 22(7):366–373, 2007.

[313] M. P. H. Stumpf, T. Thorne, E. de Silva, R. Stewart, H. J. An, M. Lappe, and C. Wiuf. Estimating the size of the human interactome. *Proceedings of the National Academy of Sciences*, 105(19):6959–6964, 2008.

[314] K. Tarassov, V. Messier, C. R. Landry, S. Radinovic, M. M. Molina, I. Shames, Y. Malitskaya, J. Vogel, H. Bussey, and S. W. Michnick. An *in vivo* map of the yeast protein interactome. *Science*, 320(5882):1465–1470, 2008.

[315] R. Tarjan. Depth-first search and linear graph algorithms. In *Switching and Automata Theory, 1971., 12th Annual Symposium on*, pages 114 –121, 1971.

[316] R. L. Tatusov, E. V. Koonin, and D. J. Lipman. A genomic perspective on protein families. *Science*, 278(5338):631–637, 1997.

[317] J. Tavernier, S. Eyckerman, I. Lemmens, J. Van der Heyden, J. Vandekerckhove, and X. Van Ostade. MAPPIT: a cytokine receptor-based two-hybrid method in mammalian cells. *Clinical & Experimental Allergy*, 32(10):1397–1404, 2002.

[318] S. A. Teichmann. The constraints protein-protein interactions place on sequence divergence. *Journal of Molecular Biology*, 324(3):399–407, 2002.

[319] M. L. Thompson and W. Zucchini. On the statistical analysis of ROC curves. *Statistics in Medicine*, 8(10):1277–1290, 1989.

[320] V. A. Traag and J. Bruggeman. Community detection in networks with positive and negative links. *Physical Review E*, 80:036115, 2009.

[321] K. Trachana, T. A. Larsson, S. Powell, W.-H. Chen, T. Doerks, J. Muller, and P. Bork. Orthology prediction methods: A quality assessment using curated protein families. *BioEssays*, 33(10):769–780, 2011.

[322] A. L. Traud, C. Frost, P. J. Mucha, and M. A. Porter. Visualization of communities in networks. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 19(4):041104, 2009.

[323] A. L. Traud, E. D. Kelsic, P. J. Mucha, and M. A. Porter. Comparing community structure to characteristics in online collegiate social networks. *SIAM Review*, 53(3):526–543, 2011.

[324] S. Tsukiyama, I. Shirakawa, H. Ozaki, and H. Ariyoshi. An algorithm to enumerate all cutsets of a graph in linear time per cutset. *Journal of the ACM*, 27 (4):619–632, 1980.

[325] M. Turanalp and T. Can. Discovering functional interaction patterns in protein-protein interaction networks. *BMC Bioinformatics*, 9(276), 2008.

[326] P. Uetz, L. Giot, G. Cagney, T. A. Mansfield, R. S. Judson, J. R. Knight, D. Lockshon, V. Narayan, M. Srinivasan, P. Pochart, et al. A comprehensive analysis of protein–protein interactions in *Saccharomyces cerevisiae*. *Nature*, 403(6770):623–627, 2000.

[327] S. Ulrich, W. Uwe, L. Maciej, H. Christian, B. F. H., G. Heike, S. Martin, Z. Martina, S. Anke, K. Susanne, T. Jan, M. Sascha, A. Claudia, B. Nicole, K. Silvia, G. Astrid, T. Engin, D. Anja, K. Sylvia, K. Bernhard, B. Walter, L. Hans, and W. E. E. A human protein-protein interaction network: A resource for annotating the proteome. *Cell*, 122(6):957–968, 2005.

[328] W. S. Valdar and J. M. Thornton. Protein-protein interfaces: analysis of amino acid conservation in homodimers. *Proteins*, 42(1):108–124, 2001.

[329] A. Valencia and F. Pazos. Computational methods for the prediction of protein interactions. *Current Opinion Structural Biology*, 12(3):368–373, 2002.

[330] S. van Dongen. *Graph Clustering by Flow Simulation*. PhD thesis, University of Utrecht, May 2000.

[331] A. Vazquez, A. Flammini, A. Maritan, and A. Vespignani. Modeling of protein interaction networks. *ComPlexUs*, 1:38–44, 2002.

[332] K. Venkatesan, J.-F. Rual, A. Vazquez, U. Stelzl, I. Lemmens, T. Hirozane-Kishikawa, T. Hao, M. Zenkner, X. Xin, K.-I. Goh, M. A. Yildirim, N. Simonis, K. Heinzmann, F. Gebreab, J. M. Sahalie, S. Cevik, C. Simon, A.-S. de Smet, E. Dann, A. Smolyar, A. Vinayagam, H. Yu, D. Szeto, H. Borick, A. Dricot, N. Klitgord, R. R. Murray, C. Lin, M. Lalowski, J. Timm, K. Rau, C. Boone, P. Braun, M. E. Cusick, F. P. Roth, D. E. Hill, J. Tavernier, E. E. Wanker, A.-L. Barabási, and M. Vidal. An empirical framework for binary interactome mapping. *Nature Methods*, 6(1):83–90, 2009.

[333] A. J. Vilella, J. Severin, A. Ureta-Vidal, L. Heng, R. Durbin, and E. Birney. Ensemblcompara genetrees: Complete, duplication-aware phylogenetic trees in vertebrates. *Genome Research*, 19(2):327–335, 2009.

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

[335] C. von Mering, L. J. Jensen, B. Snel, S. D. Hooper, M. Krupp, M. Foglierini, N. Jouffre, M. A. Huynen, and P. Bork. STRING: known and predicted protein-

protein associations, integrated and transferred across organisms. *Nucleic Acids Research*, 33(Database issue):433–437, 2005.

[336] A. Wagner. The yeast protein interaction network evolves rapidly and contains few redundant duplicate genes. *Molecular Biology and Evolution*, 18(7):1283–1292, 2001.

[337] A. Wagner. Robustness, evolvability, and neutrality. *FEBS Letters*, 579(8): 1772–1778, 2005.

[338] A. J. M. Walhout, R. Sordella, X. Lu, J. L. Hartley, G. F. Temple, M. A. Brasch, N. Thierry-Mieg, and M. Vidal. Protein interaction mapping in *C. elegans* using proteins involved in vulval development. *Science*, 287(5450):116–122, 2000.

[339] D. L. Wallace. A method for comparing two hierarchical clusterings: Comment. *Journal of the American Statistical Association*, 78(383):569–576, 1983.

[340] P. I. Wang and E. M. Marcotte. It's the machine that matters: Predicting gene function and phenotype from protein networks. *Journal of Proteomics*, 73(11): 2277–2289, 2010.

[341] X. Wang and L. Huang. Identifying dynamic interactors of protein complexes by quantitative Mass Spectrometry. *Molecular & Cellular Proteomics*, 7(1): 46–57, 2008.

[342] S. Wasserman and K. Faust. *Social Network Analysis: Methods and Applications*. Cambridge University Press, 1994.

[343] J. D. Watson, R. A. Laskowski, and J. M. Thornton. Predicting protein function from sequence and structural data. *Current Opinion in Structural Biology*, 15 (3):275–284, 2005.

[344] D. J. Watts and S. H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393(6684):440–442, 1998.

[345] R. S. Weiss and E. Jacobson. A method for the analysis of the structure of complex organizations. *American Sociological Review*, 20(6):661–668, 1955.

[346] A. Wiles, M. Doderer, J. Ruan, T.-T. Gu, D. Ravi, B. Blackman, and A. Bishop. Building and analyzing protein interactome networks by cross-species comparisons. *BMC Systems Biology*, 4(36), 2010.

[347] M. R. Wilkins and S. K. Kummerfeld. Sticking together? Falling apart? Exploring the dynamics of the interactome. *Trends in Biochemical Sciences*, 33 (5):195–200, 2008.

[348] S. G. Williams and S. C. Lovell. The effect of sequence evolution on protein structural divergence. *Molecular Biology and Evolution*, 26:1055–1065, 2009.

[349] H. F. Winstanley, S. Abeln, and C. M. Deane. How old is your fold? *Bioinformatics*, 21(suppl 1):i449–i458, 2005.

[350] C. Wiuf, M. Brameier, O. Hagberg, and M. P. H. Stumpf. A likelihood approach to analysis of network data. *Proceedings of the National Academy of Sciences*, 103(20):7566–7570, 2006.

[351] S. J. Wodak, S. Pu, J. Vlasblom, and B. Sraphin. Challenges and rewards of interaction proteomics. *Molecular & Cellular Proteomics*, 8(1):3–18, 2009.

[352] V. Wood. *Schizosaccharomyces pombe*; comparative genomics; from sequence to systems. In P. Sunnerhagen and J. Piskur, editors, *Comparative Genomics*, volume 15 of *Topics in Current Genetics*, pages 233–285. Springer Berlin - Heidelberg, 2006.

[353] P.-Y. J. Wu, C. Ruhlmann, F. Winston, and P. Schultz. Molecular Architecture of the *S. cerevisiae* SAGA Complex. *Molecular Cell*, 15(2):199–208, 2004.

[354] S. Wuchty. Evolution and topology in the yeast protein interaction network. *Genome Research*, 14(7):1310–1314, 2004.

[355] S. Wuchty, Z. N. Oltvai, and A.-L. Barabási. Evolutionary conservation of motif constituents in the yeast protein interaction network. *Nature Genetics*, 35(2): 176–179, 2003.

[356] S. Yellaboina, D. Dudekula, and M. Ko. Prediction of evolutionarily conserved interologs in *Mus musculus*. *BMC Genomics*, 9(465), 2008.

[357] J. M. Yeomans. *Statistical mechanics of phase transitions.* 1992.

[358] S. H. Yook, Z. N. Oltvai, and A.-L. Barabási. Functional and topological characterization of protein interaction networks. *Proteomics*, 4(4):928–942, 2004.

[359] H. Yu, N. M. Luscombe, H. X. Lu, X. Zhu, Y. Xia, J. D. Han, N. Bertin, S. Chung, M. Vidal, and M. Gerstein. Annotation transfer between genomes: protein-protein interologs and protein-DNA regulogs. *Genome Research*, 14(6): 1107–1118, 2004.

[360] H. Yu, P. Braun, M. A. Yildirim, I. Lemmens, K. Venkatesan, J. Sahalie, T. Hirozane-Kishikawa, F. Gebreab, N. Li, N. Simonis, T. Hao, J. F. Rual, A. Dricot, A. Vazquez, R. R. Murray, C. Simon, L. Tardivo, S. Tam, N. Svrzikapa, C. Fan, A. S. de Smet, A. Motyl, M. E. Hudson, J. Park, X. Xin, M. E. Cusick, T. Moore, C. Boone, M. Snyder, F. P. Roth, A.-L. Barabási, J. Tavernier, D. E. Hill, and M. Vidal. High-quality binary protein interaction map of the yeast interactome network. *Science*, 322(5898):104–110, 2008.

[361] A. Zanzoni, L. Montecchi-Palazzi, G. Quondam, M. Helmer-Citterich, and G. Cesareni. MINT: a Molecular INTeraction database. *FEBS Letters*, 513: 135–140, 2002.

[362] M. Zaslavskiy, F. Bach, and J.-P. Vert. Global alignment of protein-protein interaction networks by graph matching methods. *Bioinformatics*, 25(12):i259–1267, 2009.

[363] J. Zhang. Evolution by gene duplication: an update. *Trends in Ecology & Evolution*, 18(6):292–298, 2003.

[364] G. Zinman, S. Zhong, and Z. Bar-Joseph. Biological interaction networks are conserved at the module level. *BMC Systems Biology*, 5(134), 2011.

[365] E. Zuckerkandl. Evolutionary processes and evolutionary noise at the molecular level. I. Functional density in proteins. *Journal of Molecular Evolution*, 7(3): 167–183, 1976.