

Knowledge Derivation and Data Mining Strategies for Probabilistic Functional Integrated Networks

Katherine James

*Submitted for the degree of Doctor of
Philosophy in the School of Computing
Science, Newcastle University*

July 2011

Acknowledgements

First and foremost I am indebted my supervisors Neil Wipat and Jen Hallinan for their continuous encouragement, support and advice throughout this project.

I also owe thanks to everyone in the CISBAN, Ondex and Integrative Bioinformatics groups, both past and present, for their assistance and feedback during the past few years. In particular, special thanks go to Matt Pocock for his valuable coding advice, Phil Lord for guidance regarding the Gene Ontology and the writing group for their indispensable proof-reading.

Additional thanks go to Dr Julian Rutherford, Debbie Thomas, Wendy Smith and Orr Yarkoni for their considerable help and advice during the laboratory portion of the project.

A special thank you goes to all my friends and family for supporting me during this time and listening to my endless lectures about how interesting PFINs are.

Finally, a huge thank you to Dave for letting me put The Thesis first for three years.

Dedicated in memory of
Kenny J. and Peter C.

Declaration

I declare that this thesis is my own work unless otherwise stated. No part of this thesis has previously been submitted for a degree or any other qualification at Newcastle University or any other institution.

Katherine James

July 2011

Publications

The following publications have been produced by, or in conjunction with, the author during her PhD candidacy:

- James, K., Wipat, A., Hallinan, J. Is newer better? An evaluation of the effects of data curation on integrated analyses in *Saccharomyces cerevisiae*. *Integrative Biology*, 2012 (in press).
- Hallinan, J.S., James, K., Wipat, A. Network approaches to the functional analysis of microbial proteins. *Advances in Microbial Physiology* 2011; 59: pp. 101-133.
- James, K., Lycett, S.J., Wipat, A., Hallinan, J.S. Multiple Gold Standards address bias in functional network integration. *Newcastle University School of Computing Science Technical Report Series* 2011; CS-TR-1302.
- James, K., Wipat, A., Hallinan, J. Integration of full-coverage probabilistic functional networks with relevance to specific biological processes. *Data Integration in the Life Sciences: 6th International Workshop (DILS)* 2009; 5647 LNBI, pp. 31-46.
- Lister A., Charoensawan V., De S., James K., Janga SC., Huppert J. Interfacing systems biology and synthetic biology. *Genome Biology* 2009; 10(6): 309.
- Greenall A., Lei G., Swan DC., James K., Wang L., Peters H., Wipat A., Wilkinson DJ., Lydall D. A genome wide analysis of the response to uncapped telomeres in budding yeast reveals a novel role for the NAD⁺ biosynthetic gene BNA2 in chromosome end protection. *Genome Biology* 2008; 9(10): R146.
- Addinall SG., Downey M., Yu M., Zubko MK., Dewar J., Leake A., Hallinan J., Shaw O., James K., Wilkinson DJ., Wipat A., Durocher D., Lydall D. A genomewide suppressor and enhancer analysis of *cdc13-1* reveals varied cellular processes influencing telomere capping in *Saccharomyces cerevisiae*. *Genetics* 2008 Dec ; 180(4): 2251-66.

Abstract

One of the fundamental goals of systems biology is the experimental verification of the interactome: the entire complement of molecular interactions occurring in the cell. Vast amounts of high-throughput data have been produced to aid this effort. However these data are incomplete and contain high levels of both false positives and false negatives. In order to combat these limitations in data quality, computational techniques have been developed to evaluate the datasets and integrate them in a systematic fashion using graph theory. The result is an integrated network which can be analysed using a variety of network analysis techniques to draw new inferences about biological questions and to guide laboratory experiments.

Individual research groups are interested in specific biological problems and, consequently, network analyses are normally performed with regard to a specific question. However, the majority of existing data integration techniques are global and do not focus on specific areas of biology. Currently this issue is addressed by using known annotation data (such as that from the Gene Ontology) to produce process-specific subnetworks. However, this approach discards useful information and is of limited use in poorly annotated areas of the interactome. Therefore, there is a need for network integration techniques that produce process-specific networks without loss of data. The work described here addresses this requirement by extending one of the most powerful integration techniques, probabilistic functional integrated networks (PFINs), to incorporate a concept of biological *relevance*.

Initially, the available functional data for the baker's yeast *Saccharomyces cerevisiae* was evaluated to identify areas of bias and specificity which could be exploited during network integration. This information was used to develop an integration technique which emphasises interactions relevant to specific biological questions, using yeast ageing as an exemplar. The integration method improves performance during network-based protein functional prediction in relation to this process. Further, the process-relevant networks complement classical network integration techniques and significantly improve network analysis in a wide range of biological processes.

The method developed has been used to produce novel predictions for 505 Gene Ontology biological processes. Of these predictions 41,610 are consistent with existing computational annotations, and 906 are consistent with known expert-curated annotations. The approach significantly reduces the hypothesis space for experimental validation of genes hypothesised to be involved in the oxidative stress response. Therefore, incorporation of biological relevance into network integration can significantly improve network analysis with regard to individual biological questions.

Contents

1	Introduction	1
1.1	Integrative Bioinformatics	2
1.2	Biological Networks and Graph Theory	3
1.3	Probabilistic Functional Integrated Networks	4
1.4	Motivation for this Work	5
1.5	Project Aims and Objectives	6
1.6	Thesis Structure	7
2	Background	8
2.1	Functional Data	8
2.1.1	Physical Data	8
2.1.2	Genetic Data	16
2.1.3	Other Data Types	20
2.1.4	Biological Databases	21
2.2	Deciphering the Interactome	24
2.2.1	The Interactome	24
2.2.2	Computational Protein-Protein Interaction Prediction	25
2.3	Graph Theoretic Analysis	28
2.3.1	Application of Graph Theory to Biological Networks	30
2.3.2	Network Properties and Statistics	31
2.3.3	Network Modularity	36
2.3.4	Alignment and Comparison	43
2.3.5	Network Tools	44
2.4	Network Integration	47
2.4.1	Dataset Noise	51
2.4.2	Gold Standard Data	53
2.4.3	Probabilistic Functional Integrated Networks	54
2.4.4	Dataset Bias	57
2.5	Beyond the Interactome	61

2.5.1	Protein Function	61
2.5.2	Cellular Location	64
2.5.3	Human Disease	65
2.5.4	Annotation Data	65
2.5.5	Network-Based Prediction of Annotation	74
2.6	Yeast as a Model Organism to Study Human Ageing	78
2.6.1	Telomere Maintenance	79
2.6.2	Oxidative Stress	81
2.7	Summary	88
3	Methods	90
3.1	Computational Techniques	90
3.1.1	Data Sources	90
3.1.2	Gene Ontology Analysis	92
3.1.3	Hierarchical Clustering	93
3.1.4	Network Integration	93
3.1.5	Network Visualisation and Evaluation	97
3.2	Laboratory Techniques	98
3.2.1	Strains and Growth Conditions	98
3.2.2	DNA Extraction	98
3.2.3	Polymerase Chain Reaction	99
3.2.4	Stress Sensitivity Tests	100
4	Harnessing Process-Relevance During Network Integration	102
4.1	Harnessing Process Relevance	103
4.1.1	Source Data	103
4.1.2	Evaluation of Dataset Bias	104
4.1.3	Results	105
4.1.4	Discussion	118
4.2	The Integration RelCID Schema	119
4.2.1	Evaluation Strategy	120
4.3	Results	122
4.3.1	Network Integration	122
4.3.2	Network Evaluation	125
4.3.3	Discussion	139

5	Evaluation of the Effect of Database Curation on PFIN Performance in <i>Saccharomyces cerevisiae</i>	143
5.1	Source Data	144
5.2	Database Changes	146
5.2.1	BioGRID	146
5.2.2	KEGG	149
5.2.3	Gene Ontology	153
5.3	Evaluation Strategy	154
5.4	Results	156
5.4.1	Combined Data Source Changes	156
5.4.2	Individual Data Source Changes	159
5.4.3	Relevance Networks	169
5.4.4	Cut-Off Networks	170
5.5	Discussion	175
6	Assessment of GO Biological Processes as POIs and RelCID Performance Optimisation	180
6.1	Datasets	181
6.2	A Full GOBP Sweep	181
6.3	GO Term Choice	182
6.3.1	Term Properties	184
6.3.2	Network Topology	190
6.3.3	Dataset Ranking	190
6.3.4	Dataset Topology	193
6.4	Extending the Relevance Integration Schema	196
6.4.1	Network Performance	197
6.4.2	Combining Relevance	198
6.5	Discussion	203
7	Computational and Laboratory Analysis of Network-Generated Predictions	207
7.1	Functional Prediction	207
7.1.1	Datasets	207
7.1.2	GO Term Selection	208
7.2	Prediction Results	208
7.3	Computational Evaluation	210
7.3.1	Consistency with Existing Annotations	210
7.3.2	Consistency with Previous Computational Predictions	213
7.3.3	Multiple Functional Predictions	214
7.3.4	Discussion	214

7.4	Laboratory Evaluation of a Functional Prediction	218
7.4.1	Choice of Prediction	218
7.4.2	Comparison with Traditional Database Searching	223
7.5	Experimental Results	227
7.5.1	Strain Confirmation	227
7.5.2	Stress Responses	227
7.6	Discussion	229
8	Discussion and Future Work	234
8.1	Introduction	235
8.2	The RelCID Algorithm	236
8.2.1	Dataset Relevance	237
8.2.2	Harnessing Relevance	237
8.2.3	Choice of POI	242
8.2.4	Source data	246
8.2.5	Novel Hypotheses	249
8.3	Contribution	253
8.4	Future Work	254
8.5	Concluding Remarks	257
A	Graph Theoretic Definitions	262
B	Gene Ontology Evidence Types	263
B.1	Experimental Evidence Codes	263
B.2	Author Statement Evidence Codes	264
B.3	Computational Analysis Evidence Codes	264
B.4	Computationally-assigned Evidence Codes	265
C	Relevance Network Production	266
D	BioGRID Evidence Types	267
D.1	Physical interactions	267
D.2	Genetic Interactions	268
E	GO Term Enrichment	270
F	Dataset Integration Rankings	273
G	Dataset Versions	275
H	Functional Predictions	280

Chapter 1

Introduction

The availability of whole genome sequences has spurred a revolution in biological analysis [1, 2]. Several sequenced strain collections have been established [3–6]. Technologies have been developed for the mass production of data in a high-throughput (HTP) manner within a short space of time [7–9]. The mRNA expression of all genes within a genome can be measured under a wide variety of circumstances [10–13], genetic interactions can be screened on a genome-wide scale [14–17], the subcellular location of proteins can be probed [3, 18], large-scale biochemical activity screens can be carried out [19, 20] and techniques have been developed for the proteome-wide detection of both binary protein interactions and protein complex membership [21–24]. Consequently, genes and their products, complexes and pathways are no longer seen as isolated components to be studied solely in a reductionist manner, but as parts of larger, more complex systems, which can now be analysed in their entirety (Figure 1.1) [25, 26].

A wide range of online databases exists to store the resulting data, with new databases continuing to be developed; the 2011 *Nucleic Acids Research* Database Issue reports on 1,330 curated biological databases [27], an increase of 100 from the 2010 edition [28], and 160 from the 2009 edition [29]. Meanwhile, new computational techniques, drawing on knowledge from computer science, statistics and physics, have been developed to study the wealth of data produced by these experimental technologies [30, 31]. The cross-disciplinary field of systems biology encompasses these experimental and computational analyses [32, 33].

Systems biology is the study of biology in terms of whole systems in an iterative fashion, with data analyses guiding experimental design, and experimental results, in turn, forming the basis for further analyses and mathematical modelling [26, 34–36]. One of the fundamental goals of the field is the experimental verification of the interactome: the entire complement of molecular interactions within an organism [37, 38]. Achieving this goal requires the systematic confirmation of all interactions

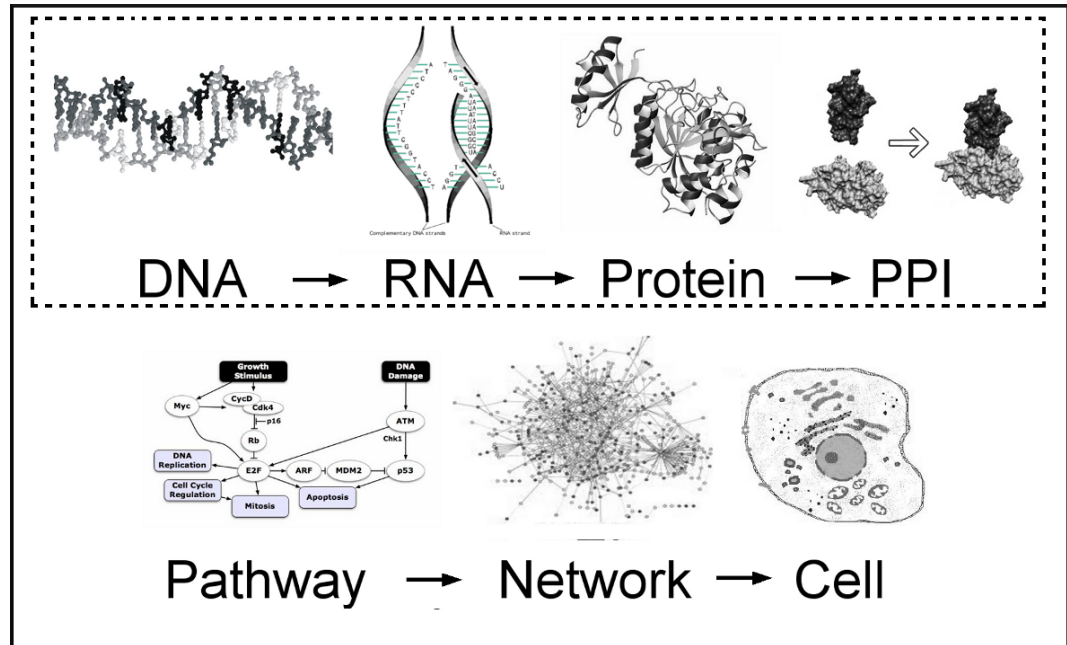


Figure 1.1: Systems biology.

Classical reductionist biology (dashed box) treats the components of the cell, DNA, RNA and proteins, individually. Following the development of high-throughput experimental techniques, systems biology analyses these cellular parts as whole systems in order to model the complex behaviour of the cell (solid box).

occurring between the proteins, genes, ligands and other molecules of the cell [39]. However, even for a relatively simple organism such as the baker's yeast *Saccharomyces cerevisiae*, verification of the interactome is a substantial task, involving the study of thousands of interactions occurring at all stages of the cell cycle, and in response to all possible external stimuli [37].

Beyond the binary interactions of the interactome, a full understanding of the cell's complex behaviour requires detailed knowledge of each interaction. Proteins have individual, and sometimes multiple, functions and act in specific biological processes, form complexes and participate in pathways. These processes each occur at particular locations within the cell. Moreover, the function(s) of each protein, its location(s) and the kinetics of each interaction can differ depending on cellular conditions. In multicellular organisms such as humans, the interactome also encompasses all tissue types, ages and disease states [40, 41].

1.1 Integrative Bioinformatics

A wealth of biological data has been produced in the past decades. These data are spread over hundreds of diverse databases and are heterogeneous in nature. Data relating to interactions between genes and gene products are of particular relevance to the work described in this thesis. However,

the number of interactions common to datasets from different experiments can be surprisingly low [42–45], primarily because each individual experimental type can only provide information about certain aspects of the cell’s behaviour and interactions [46]. For instance, a study by Beyer and co-workers found that approximately 1% of genetically interacting proteins in humans have been found to interact physically [47], indicating that the proteins act in separate but parallel pathways. In order to fully characterise every aspect of highly complex cellular systems and infer new knowledge from the data, diverse data sources must be systematically integrated [48, 49].

Integrative bioinformatics is a field which aims to develop methods for the large-scale integration of heterogeneous data. Combining data sources can provide a more complete view of the cell and reduce the impact of experimental noise [49, 50]. Data integration can also lead to a fuller understanding of cellular interactions by combining several multiple sources of evidence [51] and by revealing global properties not evident in a single data type [52]. Integrative bioinformatics approaches have been applied to a number of biological questions including the study of human disease [40, 53–66].

1.2 Biological Networks and Graph Theory

The majority of experimental datasets can be represented as networks of parts and interactions [67, 68]. A network of protein-protein interactions (PPIs) represents the physical interactions between the proteins of the cell [69], while synthetic genetic arrays (SGAs) produce networks representing shared lethality of gene deletions [14]. Both PPI and SGA networks are undirected since the edges link nodes equally in both directions. Biological networks may also have direction with edges linking source nodes to destination nodes. For instance the kinase-target relationships of phosphorylation data [20, 70, 71] and the pathway directionality in signal transduction and metabolic networks [72–76]. Each type of biological network contains valuable information that can be integrated to infer new hypotheses.

The networks generated by different experimental techniques differ in size and complexity [77] and therefore efficient analysis requires specific tools and techniques. One of the most powerful computational approaches to the interpretation of heterogeneous data is network analysis. Networks can be viewed as a graph [78]. Nodes in a graph correspond to genes or gene products and edges represent the accumulated evidence links between nodes. The simplest graphs include an edge between two nodes if there is evidence of a functional link between them from at least one data source (Figure 1.2 A) [79]. The resulting graph represents all the available evidence of linkage between all pairs of nodes (Figure 1.2 B). More complex graph models can include nodes representing other entities, such as pathways, ligands, annotations and publication references [80]. Graph theory has also been

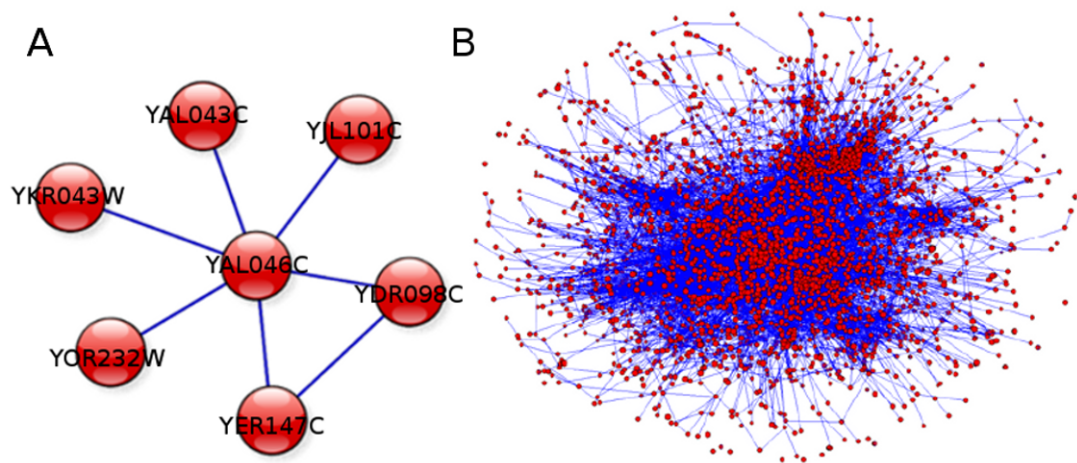


Figure 1.2: Biological data can be visualised using graph theory.

Nodes (depicted as red circles) correspond to genes or gene products and edges (depicted as blue lines) represent a summary of the evidence for interaction between them. **A.** A simple example of a small sub-graph connecting seven yeast genes. **B.** A biological network of several hundred genes.

applied to other aspects of biology, for example the nucleotide interactions in ribosomal RNA [81] and organism interactions in food webs [82].

A network representation allows biological data to be visualised and represented in a manner that is tractable for human visual analysis as well as being computationally amenable [77, 83]. Networks may contain additional information reflecting other aspects of their components' biology such as weights, directionality and types [84]. Many tools have been developed for the visualisation and manipulation of complex networks [85] and several formats have been developed to represent network structure in a standardised manner [86, 87]. These tools allow complex network data to be used for a number of applications, including: detection of protein complexes [88–90]; prediction of protein functions [91, 92]; identification of evolutionary relationships [93, 94]; and inference of novel interactions that were not detected experimentally [95, 96]. Therefore the use of graph theory for the analysis of biological data can add substantially to our understanding of cellular behaviour.

1.3 Probabilistic Functional Integrated Networks

The quality of different datasets, in terms of coverage of the genome and accuracy of the identification of interactions, depends upon the experimental technique used. Consequently, several methods have been employed to assess data quality prior to integration. The most common scoring method involves comparison of the data with a high-quality Gold Standard dataset [59]. A Gold Standard is a high-confidence, often manually-curated, set of interactions believed to be biologically correct [97]. In many cases these data represent a single data type obtained from a human expert-curated database

[49, 98] such as shared metabolic pathway membership in the Kyoto Encyclopedia of Genes and Genomes (KEGG) [99] database or shared biological process in the Gene Ontology (GO) [100]. Composite Gold Standards, derived from multiple data sources, can also be created [101].

Gold Standards can be used in two ways. Experimental datasets may be compared with a Gold Standard using a statistical test that produces a confidence score for each dataset. Scores from multiple datasets can then be integrated using techniques such as Bayesian data fusion [49, 102]. The second use of Gold Standards involves machine learning. Here, the Gold Standard is used to train a classifier, such as a support vector machine (SVM) [103] or Markov random field (MRF)[104, 105], which can then score raw datasets prior to integration. The final integrated networks, termed Probabilistic Functional Integrated Networks (PFINs), are annotated with edge weights representing the level of confidence in the evidence for each association. The weights allow the use of statistical algorithms that take these confidence weightings into account [57]. For many types of analyses PFINs outperform unweighted networks, for example in protein functional inference [49, 103, 105–112].

As functional networks, PFINs are distinct from interactomes since they link pairs of proteins if they have any type of association. Edges may represent a protein’s involvement in whole cellular processes [113]. The greater density of links provided by functional data provides a more informative basis for network analysis and functional discovery than physical interactions alone. PFINs have been used to analyse data from several different species, including yeast [114, 115], mouse [116, 117] and human [118], and to compare patterns of interaction across multiple species [119].

1.4 Motivation for this Work

PFINs, while far more informative than single source and unweighted networks, have several drawbacks. Firstly, high-throughput data are very noisy, with estimates of false positive rates varying from 20% to as high as 91% depending upon the technology used to generate the data [37, 120–122]. Estimates of false negatives range from 17% to 96% [122–124]. Secondly, many studies have shown that the integrated networks, the Gold Standards and the individual data sources may be biased towards specific cellular processes [42, 98]. For instance, co-expression data shows significant bias towards interactions between genes involved in ribosome biogenesis [125]. Current approaches to these problems usually involve attempting to identify, and subsequently remove, the noise [126, 127] or bias [98]. However, these approaches may further complicate the analysis by removing true positive interactions, resulting in a loss of valid data.

More importantly, existing network integration methods do not take into account the relevance of each experimental dataset to specific biological processes. Therefore functional prediction tech-

niques applied to networks are global rather than tailored to a specific area of biology, despite the fact that most network analysis is performed with regard to a specific biological question. Individual research groups address different questions. While the use of PFINs can produce a wealth of novel hypotheses, only a few will be of relevance to each group's interests [128]. The issue of process relevance has been addressed by using protein annotation data either to extract process-specific sub-networks from the PFIN [59, 129] or to build process-specific networks using a subset of the data [46, 130]. These methods also discard potentially useful information and are of limited use in areas of the network that contain large proportions of unannotated proteins.

The inherent biases of experimental data are a valuable source of information. Bias exists in experimental data for several reasons: the type of experimental technique or conditions chosen; the experimental design; or the choice of data for publication [131]. The combination of these factors gives each dataset its own unique set of biases and, consequently, when analysing the data, some datasets will be more informative than others regarding a particular biological process. Further, given the scale of HTP data, the noise in datasets with low relevance to a particular biological question may mask the relevant data contained in more informative datasets.

1.5 Project Aims and Objectives

The aim of this project was to research and develop techniques to exploit, rather than eliminate, data bias in order to optimise network predictions relevant to specific processes without loss of data.

To achieve this aim it was necessary to meet a number of objectives:

1. To investigate the inherent biases of functional data.
2. To use the biases to quantify dataset relevance to specific biological processes.
3. To develop and assess network integration techniques that harness process relevance.
4. To develop and assess process-specific network analysis techniques.
5. To apply the developed techniques to real data in order to produce novel hypotheses about the yeast *Saccharomyces cerevisiae*.
6. To evaluate the hypotheses using computational and laboratory techniques.

1.6 Thesis Structure

The remainder of this thesis is divided into the following sections:

- Chapter 2 provides background information and a literature review of functional data, graph theory, PFINs and network analysis techniques. The basis of cellular ageing is also introduced as the primary biological focus of this project.
- Chapter 3 details the biological datasets and computational methods utilised and developed in this work. Section 3.2 then describes the laboratory techniques used to evaluate the resulting hypotheses.
- Chapter 4 describes the development of a novel process-relevant network integration technique. The technique, RelCID, extends an existing integration method in order to tailor PFINs to answer specific questions. A detailed evaluation of RelCID using *S. cerevisiae* ageing data is then presented.
- Chapter 5 presents a systematic analysis of the effects of data curation on PFIN performance. Source databases are constantly changing over time as new knowledge is gained. It is often assumed that these changes lead to an improvement in PFIN performance over time. This chapter demonstrates that performance in fact fluctuates over time due to bias and noise and that process-tailored techniques, such as RelCID, may overcome some of these effects.
- Chapter 6 describes the optimisation of the RelCID technique and its application to a wide range of biological processes. Several additional aspects of dataset relevance are identified and incorporated into a composite network integration technique in order to produce optimal PFIN performance. The relationship between network performance and biological area of interest is also investigated and found to be closely associated with the structure of GO.
- Chapter 7 demonstrates the power of RelCID in hypothesis generation. Functional predictions are produced for 505 GO terms and computationally evaluated before a single prediction is chosen for laboratory analysis. Importantly, the RelCID technique is shown to significantly reduce analysis time in comparison to traditional methods.
- Chapter 8 discusses the implications of this project and suggests areas for future extension and analysis.

Chapter 2

Background

2.1 Functional Data

Experimental datasets are the building blocks of functional integrated networks. Unlike **PPI** networks, which only include direct physical interactions, a functional network links pairs of proteins if there is any evidence that they are functionally related [69, 113]. Several different types of functional association data exist and each provides information about a different aspect of cellular biology [67]. Functional interactions include any evidence of a functional link between two genes or gene products, such as complex formation, catalysis, genetic interaction, co-localisation and regulatory relationships.

2.1.1 Physical Data

The detection of physical binding between proteins is the basis of the majority of molecular network analyses [67, 132]. **PPIs** can be either binary or protein complex interactions [133]. In binary interactions pairs of proteins have one-to-one physical contact (Figure 2.1 A) [69, 134]. In protein complex interactions a group of proteins is associated as members of the same complex. However, there may or may not be a direct physical interaction between any pair of proteins within the complex. Additionally, the physical interactions that occur in a complex may rely on other complex members and, therefore, do not occur in a binary fashion (Figure 2.1 B). Both binary and complex interactions may be either stable or transient [135, 136].

Several experimental technologies have been developed to detect binary and protein complex interactions, and these methods differ in their methodology, interpretation and the interactions they detect [69, 137–139]. Initially these methods were designed for small-scale analysis [140]. However, recently **HTP** techniques have been developed for the detection of **PPIs** on a genome-wide scale [141],

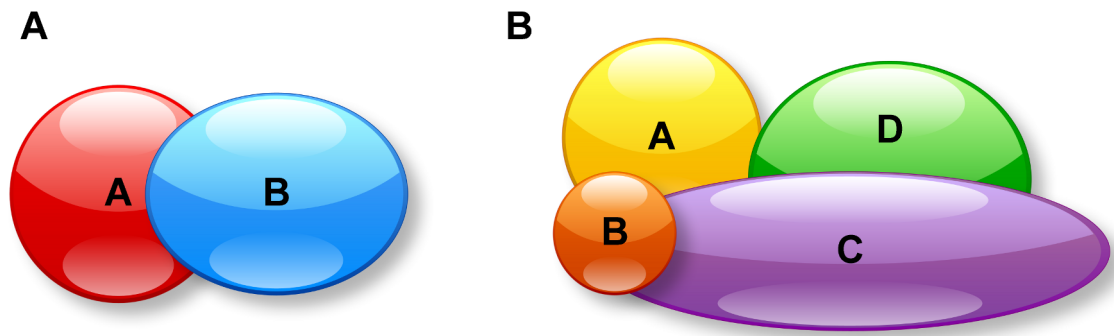


Figure 2.1: Binary and complex protein interactions.

A. In a binary interaction there is a one-to-one direct physical interaction between protein A and protein B. **B.** In protein complexes interactions are inferred between all the members of a detected complex. However the pairs of proteins may not interact physically. In this example protein B and protein D interact as members of the complex but have no physical interaction. Additionally, a direct interaction within a complex may not infer a binary interaction since the complex interaction may rely on other members of the group for stability. For instance while protein A and protein D have a small physical interaction, it may not occur in the absence of protein C and would not be detected by binary detection methods such as Y2H.

allowing more sophisticated analysis of cellular biology [142, 143]. Two of these PPI detection methods have been used to produce genome-wide interaction networks in a number of species; yeast two hybrid (Y2H) for binary interactions and tandem affinity purification (TAP) for complex detection.

2.1.1.1 Binary Interactions

Early methods for PPI detection, such as co-immunoprecipitation, co-fractionation and cross-linking typically required protein purification [140]. Y2H was developed to overcome this requirement by using the *Saccharomyces cerevisiae* Gal4 protein in a bottom-up approach [144]. Gal4 is a transcriptional activator involved in the utilisation of galactose as a carbon source [145]. The Gal4 protein consists of two distinct functional domains: an N-terminal DNA-binding domain; and a C-terminal transcriptional activation domain. Since the two domains function independently it is possible to physically separate them without loss of function [146]. The Y2H system exploits this independence by fusing the separated domains to other query proteins. A *bait* protein is fused to the DNA-binding domain, while a *prey* protein is fused to the activation domain [144]. If the bait and prey interact the Gal4 protein is reconstituted and transcriptional activation occurs (Figure 2.2). By using a reporter gene associated with the *GAL4* transcription activation domain, interaction between the bait and prey can be detected [147]. There are many reporter genes which use colour, fluorescence or selective media growth as indication of gene activation. The Y2H system commonly utilises the *lacZ* reporter gene which, when expressed, causes yeast colonies to turn blue in the presence of bromo-chloro-indolyl-galactopyranoside (X-gal) [148].

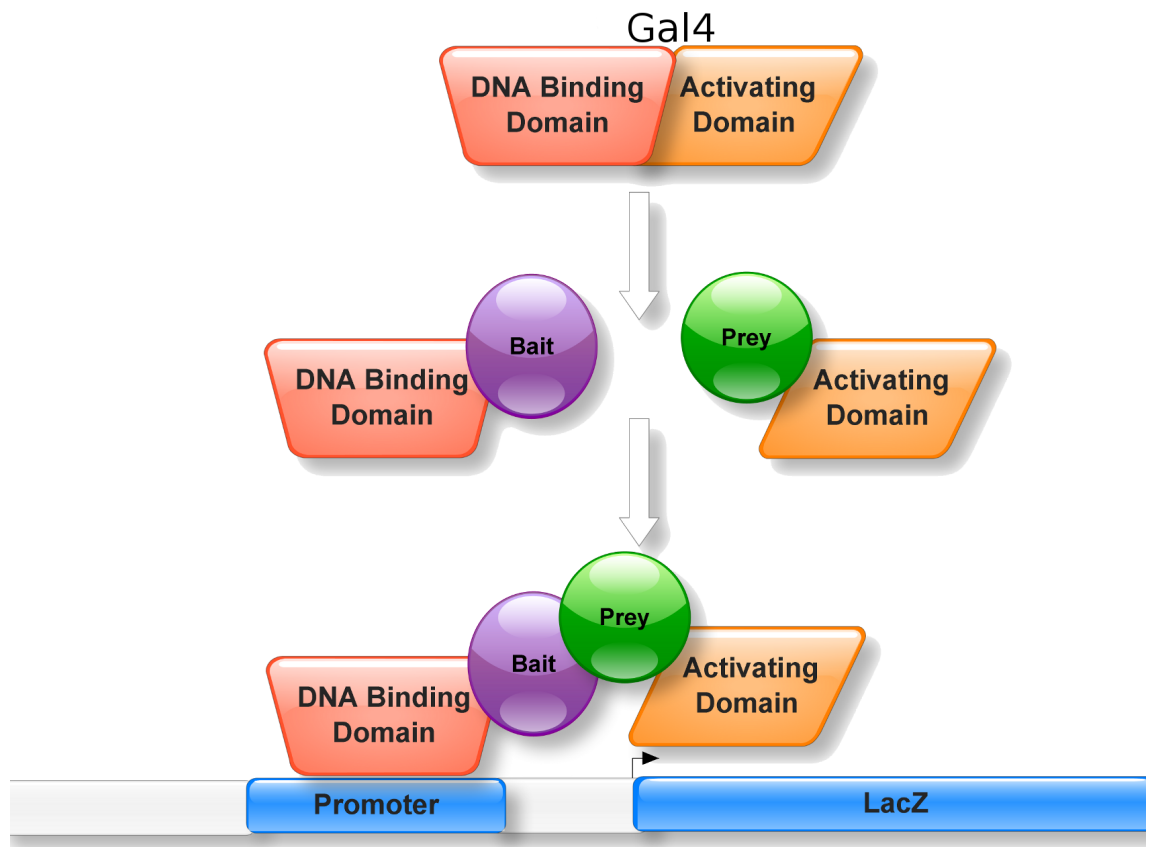


Figure 2.2: The yeast two hybrid (Y2H) method.

The Gal4 protein consists of two domains, the DNA binding domain and the transcription activation domain. In the Yeast Two Hybrid system the binding domain and activation domain are physically separated and fused to bait and prey proteins respectively. If the bait and prey interact, binding to the *GAL4* promoter brings the activation domain into close enough proximity for transcription of a reporter gene to occur. In this example the *lacZ* gene is used as the reporter. When expressed *lacZ* causes yeast colonies to turn blue in the presence of X-gal. Therefore, blue colonies are a positive indicator of interaction between the bait and prey.

Y2H can be implemented at different scales [132]. For the detection of specific interactions low-throughput (**LTP**) **Y2H** is used. This technique involves one-to-one pairwise experiments. On a larger scale a one-vs-all approach can be used to screen a specific protein or group of proteins against the entire proteome [149]. Finally, at the most **HTP** level, **Y2H** can be applied in an all-vs-all manner [150]. Fromont-Racine carried out the first large-scale **Y2H** study in 1997 [151]. Since then large-scale **Y2H** studies have been carried out in several species (Table 2.1). In yeast, two large-scale **Y2H** datasets are of particular note and have been systematically compared [108, 121]. In 2001, Ito and co-workers [21] reported a comprehensive two-hybrid analysis, which identified over 4500 binary interactions among approximately 3000 proteins, using three different reporter genes and multi-copy plasmid constructs. Of these interactions, a *core* high confidence dataset of approximately 800 interactions was selected. In an earlier study by Uetz and colleagues, two separate **Y2H** screens were carried out using a single reporter and low-copy plasmid construct [22]. Unexpectedly, the data of the Uetz and Ito datasets did not have many interactions in common [21]. This lack of overlap was also observed in a later **Y2H** dataset [134] and appears to be a common phenomenon in **Y2H** analysis. Two similar large-scale **Y2H** studies of human proteins [152, 153] showed a similar lack of overlap with only six interactions in common [154, 155].

Several theories have been postulated to account for the lack of overlap between **Y2H** datasets [21, 44, 122]. Since a protein’s function and binding is directly linked to its 3D structure, mutations in one or both of the open reading frames (**ORFs**) may have affected the strength of protein binding. Alternatively, one or both the proteins may have mis-folded due to fusion to the Gal4 pro-

Table 2.1: High-throughput Y2H screens.

Several high-throughput Y2H screens have been carried out in a range of species. Interaction numbers are taken from the BioGRID database (<http://thebiogrid.org/> accessed 20th November 2010) except when marked *, in which case the data are taken from the publication. In the case of the Ito 2001 dataset, only the core high-confidence data are available through BioGRID.

Species	Interactions	Reference
<i>Saccharomyces cerevisiae</i>	167	Fromont-Racine <i>et al.</i> 1997 [151]
<i>Saccharomyces cerevisiae</i>	875	Uetz <i>et al.</i> 2000 [22]
<i>Saccharomyces cerevisiae</i>	848 core (4549* total)	Ito <i>et al.</i> 2001 [21]
<i>Saccharomyces cerevisiae</i>	1778	Yu <i>et al.</i> 2008 [134]
<i>Caenorhabditis elegans</i>	4422	Li <i>et al.</i> 2004 [156]
<i>Drosophila melanogaster</i>	20130	Giot <i>et al.</i> 2003 [157]
<i>Drosophila melanogaster</i>	2185	Formstecher <i>et al.</i> 2005 [158]
<i>Campylobacter jejuni</i>	11687*	Parrish <i>et al.</i> 2007 [159]
<i>Helicobacter pylori</i>	1280*	Rain <i>et al.</i> 2001 [160]
<i>Plasmodium falciparum</i>	2846*	LaCount <i>et al.</i> 2005 [161]
<i>Homo sapiens</i>	755*	Colland <i>et al.</i> 2004 [162]
<i>Homo sapiens</i>	2855	Rual <i>et al.</i> 2005 [152]
<i>Homo sapiens</i>	2527	Stelzl <i>et al.</i> 2005 [153]
Herpesvirus	296*	Uetz <i>et al.</i> 2006 [163]

tein domains. Differences in experimental design, such as the chosen reporter gene or copy number of vectors, could impact the detection of interactions. Additionally, while these studies are large-scale they are non-saturating in that no single study has assessed the entire complement of potential interactions for a species. It has been suggested that false positives may be present in the datasets owing to stochastic activation of the reporter genes. Finally, it has also been suggested that the low overlap is not due to false positives but to poor sensitivity producing false negatives [134, 164]. It is highly likely that the data contains both false positive and negative interactions (See Section 2.4.1). However, the lack of overlap has yet to be fully understood and, consequently, systematic comparison of the datasets with other data types is essential to identify true positive data [165].

Since the development of the Y2H technique, many variations have been produced and the technique has been used to develop further PPI detection techniques [164]. The bacterial transcriptional repressor LexA can be used as an alternative to Gal4 [166]. In this case binding of the bait and prey proteins causes suppression of the reporter gene rather than activation. LexA and Gal4 based systems complement one another and can, therefore, be used together to filter out false results [167]. The protein-fragment complementation assay (PCA) is an *in vivo* technique that uses bait and prey proteins fused to two complementary reporter protein fragments that will only assemble in close proximity. For example the reporter may be an enzyme or fluorescent protein [168–170]. PCA has the advantage of being carried out in the natural cellular environment, thus reducing false positives caused by interactions between proteins that would not naturally meet in the cell. Two recent techniques, fluorescence resonance energy transfer (FRET) and bioluminescence resonance energy transfer (BRET), are real time Y2H variants where the bait and prey are fused to two different fluorescent or bioluminescent molecules with distinct emission factors. Interaction causes energy transfer that changes the signal from the cell [171–173]. The Mammalian protein-protein interaction trap (MAPPIT) is an *in vivo* mammalian variation of Y2H that uses receptors, for instance the cytokine receptor, fused to the bait and prey [174, 175].

Protein chips are also used for the detection of specific protein binding *in vitro* [176]. In this technique large numbers of proteins are immobilised by covalent bonding to a solid surface, for example a glass slide, and then probed with a labelled substrate [177]. Protein chips are produced by high-accuracy spotting robots, allowing a large number of proteins to be immobilised in a small space. The substrate probes can be any type of biological molecule, for instance other proteins, antigens, small molecules, drugs or nucleic acids [178]. Various reporters such as fluorescent proteins are used to detect interaction. Whole-proteome chips are now available allowing genome-wide identification of specific binding partners [4, 177].

2.1.1.2 Complex Detection

Complexes are detected by pull-down proteomics methods. Pull-down experiments have three stages; bait presentation, complex purification and analysis [179]. These techniques have the advantage of being carried out in a protein's natural state and at its normal abundance levels [180]. The tandem affinity purification mass spectrometry (TAP-MS) technique has been developed for HTP detection of protein complexes [181]. In the first stage of TAP-MS a bait protein is modified at its C-terminus by fusion with a TAP tag. The tag consists of a calmodulin binding domain (CBD) and the bacterial immunoglobulin-binding Protein A, separated by a tobacco etch virus (TEV) protease cleavage site [182]. The modified bait is then added to a cellular extract, allowing proteins to bind the bait and form a complex.

The complexed proteins are then purified in two affinity columns (Figure 2.3). The first column consists of beads coated with the immunoglobulin IgG. Protein A binds the IgG beads, and the bound complex is washed before release by protease cleavage at the TEV site. In the second column the CBD binds calmodulin coated beads before further washing of the complex. Finally the bound complex is released and the bait's binding partners are identified by mass spectrometry (MS) [67, 179, 183]. The two-stage purification method of TAP-MS gives high sample purity, however it is thought not to detect weak interactions or those interactions involving low abundance proteins well in comparison to other methods [184, 185].

TAP-MS and related methods identify potential protein complexes. However, due to the nature of complex binding (Figure 2.1 B) analysis of the data is difficult and the results can be interpreted in different ways [186]. There are two major algorithms used to identify binary PPIs from TAP-MS data [42, 184]. In the first, termed the *spoke* model, PPIs are inferred between the bait protein and each of the identified preys (Figure 2.4 A). In the second, termed the *matrix* model, pairwise PPIs are inferred between all pairs of proteins in the complex, including the bait (Figure 2.4 B).

The two models are trade-offs between completeness and accuracy [42]. The spoke model reduces false positives but increases false negatives, while the matrix model increases false positives [187]. Combined models have also been developed for the interpretation of TAP-MS data. Some methods vary the model chosen depending on the complex size [187]. Others calculate probabilities for each individual interaction [45, 98, 188]. For instance, hypergeometric probabilities can be calculated that downweight promiscuous proteins; those proteins which have a larger number of interactions *in vitro* than are statistically likely to occur *in vivo* [98]. This downweighting can also be applied to other data types, such as Y2H datasets, to reduce false positive results.

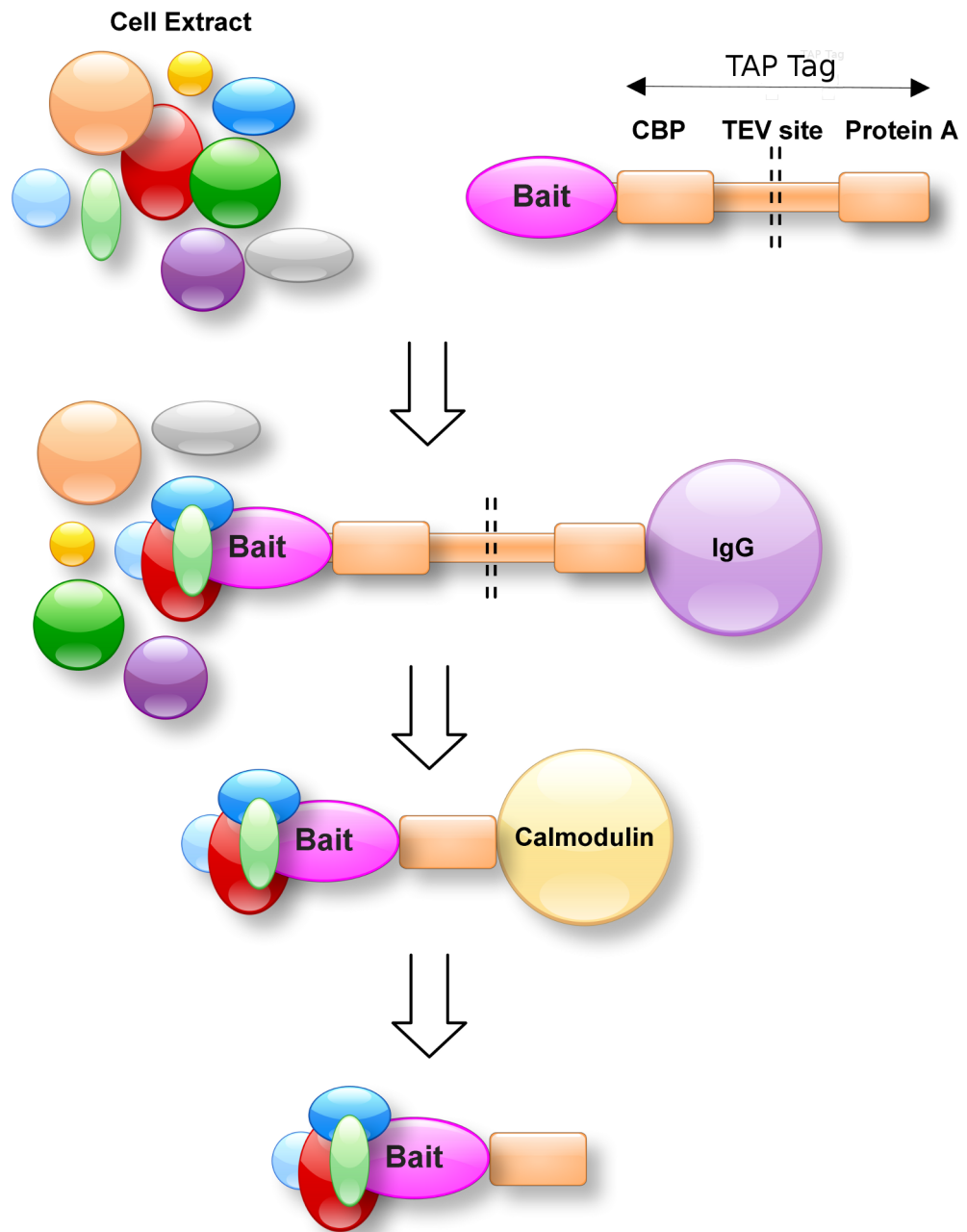


Figure 2.3: Tandem affinity purification.

Cell extract is mixed with modified bait proteins which have been fused to a TAP tag. In the first affinity column Protein A of the tag binds IgG coated beads and the bound proteins are washed to remove un-complexed cell extract. After release of the bait protein by cleavage at the TEV site a second affinity column is utilised. The calmodulin coated beads of the column are bound by the CBD domain of the TAP tag before further washing and release of the bound proteins. The members of the bound complex can then be identified, for instance by mass spectroscopy.

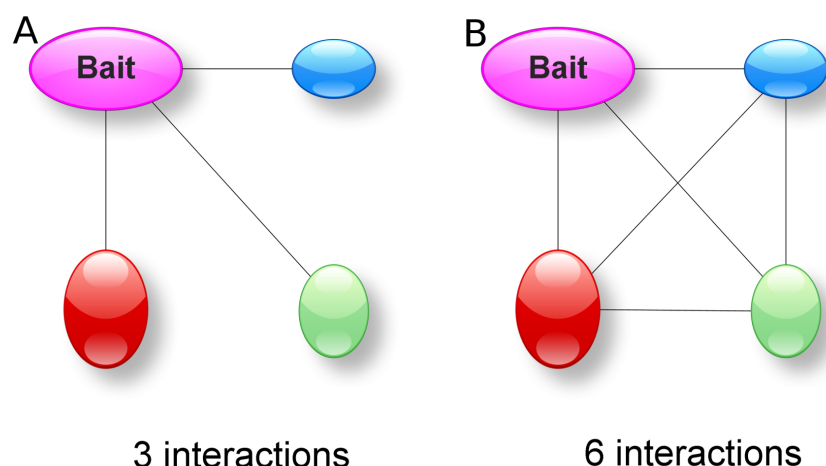


Figure 2.4: Interpretation of TAP-MS data.

There are two methods to interpret the results of a TAP-MS screen. **A.** In the spoke model interactions are inferred between the bait and each of the prey proteins in the complex. **B.** In the matrix model pairwise interactions are inferred between all proteins of the detected complex.

Table 2.2: High-throughput TAP-MS screens.

Several high-throughput TAP-MS screens have been carried out in a number of species. Interaction numbers are taken from the BioGRID database (<http://thebiogrid.org/> accessed 20th November 2010) except when marked *, in which case the data are taken from the publication text. It should be noted that while the publication text of Krogan *et al.* 2006 includes a set of 7,123 interactions, several of the interactions have subsequently been removed from the BioGRID dataset.

Species	Interactions	Reference
<i>Saccharomyces cerevisiae</i>	3400	Gavin <i>et al.</i> 2002 [189]
<i>Saccharomyces cerevisiae</i>	3666	Ho <i>et al.</i> 2002 [190]
<i>Saccharomyces cerevisiae</i>	7079	Krogan <i>et al.</i> 2006 [24]
<i>Saccharomyces cerevisiae</i>	7592	Gavin <i>et al.</i> 2006 [23]
<i>Escherichia coli</i>	5254*	Butland <i>et al.</i> 2005 [191]
<i>Escherichia coli</i>	11511*	Arifuzzman <i>et al.</i> 2006 [192]
<i>Homo sapiens</i>	2068	Ewing <i>et al.</i> 2007 [193]
<i>Homo sapiens</i>	2555	Hutchins <i>et al.</i> 2010 [194]

Currently no model gives a complete set of the physical interactions of the interactome, and analysis of the data in combination with other data types is necessary to accurately identify the protein complex interactions of the interactome [95, 165]. However, all the interactions of both the spoke and matrix models can be considered functional associations.

Affinity purification techniques have been used to detect protein complexes in a number of species (Table 2.2). In yeast, three large-scale TAP-MS datasets [24, 189, 190] have been widely studied [42, 45, 195–197]. Ho and co-workers used a set of 725 baits to capture potential complexes and detected approximately 3500 interactions [190]. While Gavin and colleagues used a significantly larger number of baits (1739), the final number of potential complexed interactions was also approximately 3500 [189]. A later study using approximately 4500 bait proteins applied two distinct

MS-based methods to increase accuracy [24]. The results were then integrated as probabilities using machine learning, producing a final high-confidence set of 7,123 interactions. Surprisingly, the three TAP-MS datasets have low overlap with one another, with known complexes and with the large-scale Y2H datasets (Section 2.1.1.1) [24, 38, 77, 154, 184]. Therefore, in order to increase coverage and reduce noise, a separate combined dataset has been produced by the Krogan group by probabilistic re-analysis of the available data using purification enrichment (PE) scoring [43].

2.1.2 Genetic Data

Genetic data can also be used to infer functional associations. Cells are complex systems with high levels of co-ordination between cellular processes at the genetic level [33]. There are three main types of genetic data:

- **Coexpression.** Genes are each expressed at specific times in response to varying cellular conditions and requirements. Therefore, the expression patterns of genes can reveal underlying cellular biology, since genes expressed at the same times are likely to be functionally related [198].
- **Gene Disruptions.** The disruption of single genes may interfere with cellular processes and can reveal their importance and give clues to their cellular roles [199].
- **Genetic Interactions.** When disruptions are combined, several different types of relationship between pairs of genes can be inferred from the data [200].

2.1.2.1 CoExpression

The coexpression of genes in response to different cellular conditions can be used to infer functional associations. Gene coexpression is measured at the mRNA level using DNA microarrays [13]. Microarrays are similar to protein chips in that they are robotically produced, with a high number of probes immobilised on a small solid surface. However, rather than using proteins as probes, microarrays use synthetic DNA oligonucleotides designed to bind specific RNA sequences [201]. Microarrays can contain tens of thousands of probes, allowing the parallel analysis of gene expression on a genome-wide scale. Large-scale microarrays have been used to study a wide variety of organisms including *S. cerevisiae* [202], *Drosophila melanogaster* [203], *Arabidopsis thaliana* [204] and *H. sapiens* [205]. The National Center for Biotechnology Information (NCBI) Gene Expression Omnibus (GEO¹) database stores HTP gene expression data for over 500 organisms in a standardised format [206].

¹<http://www.ncbi.nlm.nih.gov/geo/> (accessed 14/1/11)

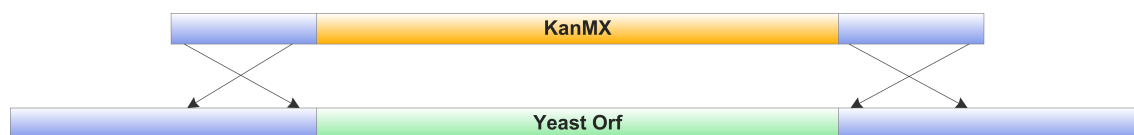


Figure 2.5: Gene deletion in *Sacharomyces cerevisiae*.

An open reading frame may be deleted by replacement with a reporter gene such as the KanMX module. Replacement is accomplished by homologous recombination in the flanking regions of the gene (displayed as ×). The KanMX module consists of the kanamycin resistance gene flanked by unique oligonucleotide sequences which act as a molecular bar-code for the deletion strain. Therefore resistance to the antibiotic kanamycin is a positive indication of deletion.

While coexpression may represent a functional association between a pair of genes, the interpretation of microarray data is non-trivial. Dozens of algorithms have been developed to detect coexpressed genes and cluster the data into functionally-linked groups (for instance [11, 207–217]). However, microarray data are extremely noisy and microarray datasets contain different signal to noise ratios [10]. A large number of coexpressed genes are not functionally related since distinct processes occur at the same time within the cell [10]. Therefore, inferring functional associations between all coexpressed genes can add noise to the data and therefore obscure true positive interactions. Further, functionally linked genes may be expressed sequentially rather than simultaneously [12].

Methods have been developed to detect false negative results [218] and to remove false positives although these methods are unreliable [219]. Often gene expression data are used in combination with other biological data in order to reduce false positives [41, 220–227]. The use of multiple datasets can also improve the detection of co-expression [228]. However, the relationship between microarray data and cellular biology remains difficult to interpret.

2.1.2.2 Gene Disruption

Disruption of a gene can reveal its function [199]. Gene disruption in *S. cerevisiae* can be accomplished in several ways. The overexpression of a gene can disrupt the processes in which it is involved and give clues to its function [4, 229]. Genes may also be rendered functionally inactive by the insertion of DNA, such as a transposon, within their sequences [5]. Finally, genes may be deleted (*knocked out*) from the genome, often by replacement with a marker for the deletion such as the KanMX module used by the Yeast Gene Deletion Project² (Figure 2.5) [230]. The KanMX module consists of the kanamycin resistance gene flanked by unique oligonucleotide sequences which act as a molecular *barcode* for the deletion strain.

²http://www-sequence.stanford.edu/group/yeast_deletion_project/deletions3.html (accessed 11/1/11)

Gene disruption can have several different effects. If a disrupted gene is required for viability the disruption is lethal and the cells will not grow. These genes are termed *essential*, and make up approximately 20% of the yeast genome. Essential genes cannot be permanently disrupted and must be studied using temporary, conditional disruption [14, 231–233].

Cells may also exhibit abnormal growth, termed *sickness*, such as slow growth or small cell size. In many cases the mutant cells will grow normally but show increased sensitivity to particular conditions, such as oxidative stress or low iron [234]. These condition-specific reactions can reveal a gene's cellular role. Disruption mutant strains are available from the Yeast Deletion Project for 90% of the yeast genome.

2.1.2.3 Genetic Interactions

A genetic interaction (GI) occurs when disruption of one gene enhances or suppresses disruption of another [14]. In other words, the two disruptions have a combined effect that is not seen when either gene is disrupted on its own. There are several types of GI which can be classified in different ways. The BioGRID database (see Section 2.1.4.1) classifies GIs into eight types; dosage growth defect, dosage lethality, dosage rescue, synthetic growth defect, synthetic lethality, synthetic rescue, phenotypic enhancement and phenotypic suppression [235].

There are three basic effects of GI; *growth defect* (sickness), *lethality* and *rescue*. In many cases two individual gene disruptions have no effect on the cell but when both genes are disrupted cause growth defect or lethality (Figure 2.6 A-B). In other cases, where the disruption of a single gene may cause growth defect or lethality, a disruption in a second gene will rescue this effect, returning the cell to improved or normal growth (Figure 2.6 C). If the two disruptions are caused by deletion or insertion the effects are termed *synthetic* by BioGRID, while if the second disruption is caused by over-expression these effects are termed *dosage* effects. Finally, genetic interactions can have *phenotypic* effects in which the second disruption causes *enhancement* or *suppression* of the abnormal phenotype produced by the first (termed epistasis [236]).

Originally, GIs were screened in a LTP manner by disruption of specific genes of interest [237–239]. Recently, three HTP methods have been developed to detect GIs on a genome-wide scale; synthetic genetic array (SGA) [14, 16, 123, 240], diploid-based synthetic lethality analysis on microarrays (dSLAM) [15, 241–243] and epistatic miniarray profiles (E-MAPs) [17, 70, 244]. In SGAs gene knockout strains are used to produce haploid double mutants by crossing single gene deletions of interest against a large-scale array of single gene deletions [14]. The resulting mutants are then assessed for synthetic interactions. dSLAM is the diploid version of SLAM [245], a microarray-based screen of competitive mutant growth [15]. An E-MAP is also microarray-based. In this method,

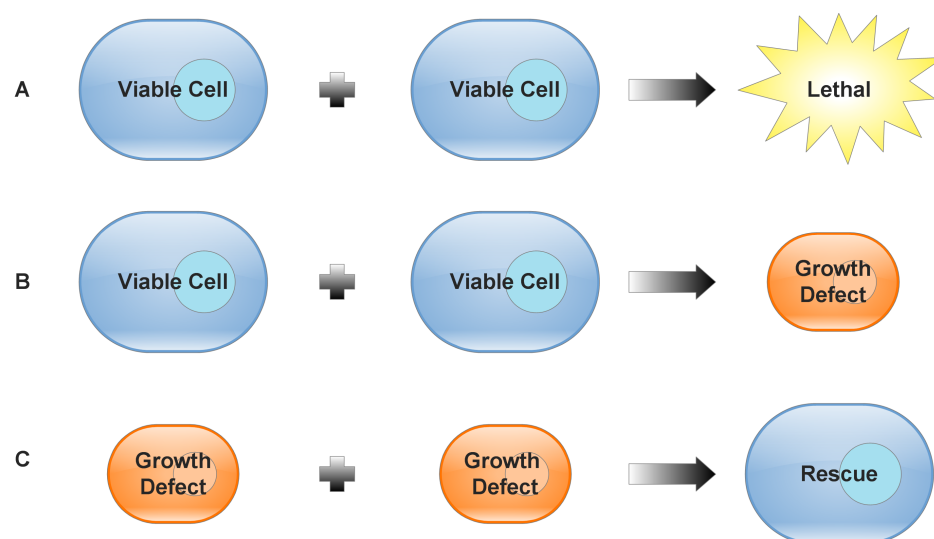


Figure 2.6: Genetic interactions.

A. In lethal interactions two individual genes disruptions have no effect on the cell but when both genes are disrupted the cell is not viable. **B.** In growth defect interactions two individual gene disruptions have no effect on the cell but when disrupted together cause sickness. **C.** In rescue interactions one or both of the individual gene disruptions causes either lethality or sickness which is rescued by the second disruption. In this example the two individual gene deletions cause sickness.

synthetic double mutants are screened for phenotypic responses, by comparison with single mutant strains, to quantify the level of **GI** [17].

Genetically-interacting protein pairs are commonly components of the same pathway or complex and have a relatively high level of conservation across species (approximately 29% between *S. cerevisiae* and *Schizosaccharomyces pombe*) [246]. However, unlike **PPIs**, **GI**s connect genes with related function but which are less likely to have a physical interaction. Further, using condition-specific gene disruption it has been shown that the majority of **GI**s involve an essential gene [232]. Many interacting pairs also have similar structure and, therefore, may share a structural basis for their function [247]. Consequently it may be possible to predict **GI**s based on their conservation, function and/or structural similarity (see Section 2.2.2).

Most **GI**s are thought to be either *between* or *within* pathways in the interactome (Figure 2.7) [248, 249]. In between-pathway interactions, the interacting pair have parallel roles in separate redundant pathways [250]. Therefore, disruption of both genes blocks both of the pathways. In within-pathway interactions, one pathway component can compensate for the disruption of another, however the pathway is blocked if both components are disrupted. Genetically-interacting proteins of this type are more likely to interact physically or be members of the same complex. The between-pathway theory is the favoured explanation for the majority of **GI**s and its principles have been used to predict underlying pathway information [200]. There are several theories as to how genetic redundancy of this type has evolved (discussed in [251]).

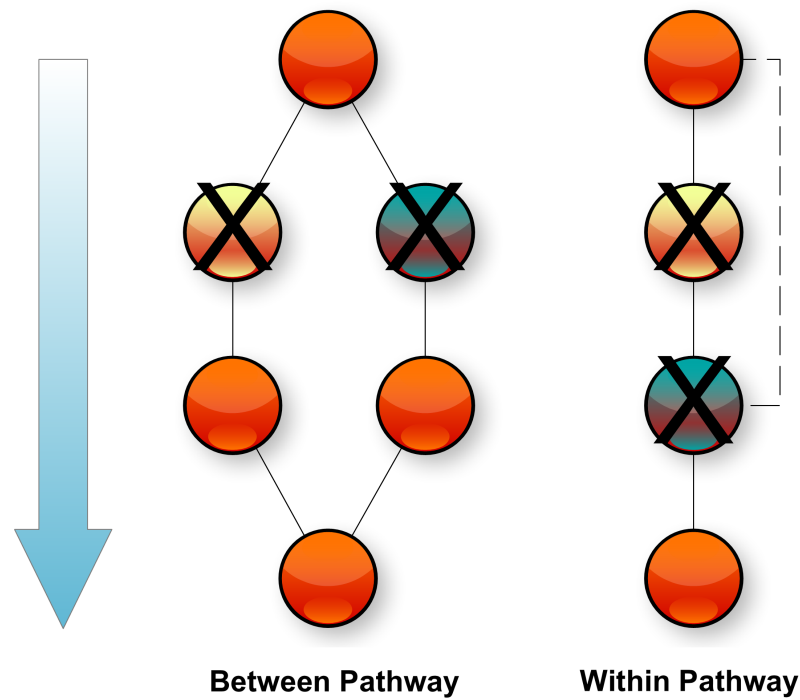


Figure 2.7: Between and within pathway genetic interactions.

A. In between-pathway interactions the yellow and blue components have parallel roles in separate redundant pathways. Therefore, disruption of both genes blocks both of the pathways. **B.** In within-pathway interactions, the blue component can compensate for the disruption of the yellow component, however the pathway is blocked if both components are disrupted.

2.1.3 Other Data Types

Several other types of experimental data exist from which functional associations can be inferred, including:

- **Biochemical Activity:** The modification of one protein by another, for instance by phosphorylation or ubiquitination [19, 20].
- **Co-Localisation:** The co-localisation of a pair of proteins within the same area of the cell [3, 18].
- **Transcription Factor Binding:** The specific binding of proteins to DNA binding sites [252–256].
- **Literature Mining:** The co-occurrence of gene and protein identifiers in scientific literature [257–261].

2.1.4 Biological Databases

A large number of online databases have been developed to store biological data [137, 262]. These resources store a range of data types such as sequences, structures and interactions. The PathGuide resource³ currently lists 325 databases [263]. However, many databases are poorly maintained, or not maintained at all, and can consequently become out of date and contain errors [264].

Database development in bioinformatics has recently been the subject of considerable research [265–269], and has been recognised as a scientific field in its own right [270]. The journal Database [271], dedicated to biological database development and curation strategies, was launched in 2009. Due to the large amount of data being produced, curation efforts have become increasingly important in order to ensure data quality [272–275]. Errors such as inaccurate entry, unintended data duplication or inadequate gene identification can be propagated through multiple databases and can be difficult to identify and remove, particularly from poorly-maintained databases [264].

Nucleic Acid Research maintains an online collection⁴ of selected, high-quality molecular biology databases. Databases are chosen for this collection based on applicability, relevance, coverage, and accuracy [29]. In particular, the included databases all contain up-to-date, curated information. In total 183 of the databases specifically store functional association data [27] (Table 2.3).

Many of these databases cover a wide range of species. For instance, the Biological General Repository for Interaction Datasets (BioGRID) [276], KEGG [277], the Database of Interacting Proteins (DIP) [278] and the Munich Information Center for Protein Sequences (MIPS) [279] all store data for model organisms and several other species. Other databases are area-specific, for example the Proteins Interacting in the Nucleus (PIN) resource [280], the Transcription Factor Database (TRANSFAC) [281] and the Nuclear Protein Database (NPD) [282]. In addition, there are species-specific resources such as the *Saccharomyces* Genome Database (SGD) [283], the Drosophila Interactions Database (DroID) [284], Wormbase [285] and the Human Protein Reference Database (HPRD) [286].

Table 2.3: The Nucleic Acid Research Database Collection as of January 2011.

The NAR maintains a collection of databases that store a wide range of data types including several functional interaction types.

Data Type	No. Databases
Protein-Protein Interactions	84
Metabolic	23
Signalling Pathways	7
Co-expression and Microarrays	69

³<http://www.pathguide.org/> (accessed 11/1/11)

⁴<http://www.oxfordjournals.org/nar/database/c/>

2.1.4.1 BioGRID

BioGRID is a comprehensive and highly-curated resource for functional association data [276]. The database stores interactions of 27 different types, including both physical and genetic interactions⁵. Each interaction is manually curated from the literature in an iterative curation strategy designed to minimise errors⁶. Review articles are excluded from BioGRID, as are interactions from unpublished data. A web-based interaction management system (IMS) allows multiple curators to upload data while avoiding duplication [287]. In addition, the database curators actively encourage community feedback regarding errors and missing data⁷. Due to BioGRID's level of completeness and quality it has been used as the source of data for a large number of studies (for instance [43, 133, 200, 288–292]).

Originally called the GRID and designed to store *S. cerevisiae* data [293], BioGRID has now expanded to store data from 18 different species (Table 2.4). Currently, complete literature coverage is provided for the yeasts *S. cerevisiae* and *S. pombe* and for the thale cress *A. thaliana* [294]. In addition the database provides a high-quality, stand alone literature-curated dataset for *S. cerevisiae* derived from small-scale experimental data alone [235].

Table 2.4: BioGRID interaction statistics.

A summary of the data stored in the BioGRID database based on statistics from <http://wiki.thebiogrid.org/doku.php/statistics> (accessed 13/1/11). Datasets with full literature coverage are marked with an asterisk.

Organism	Total Interactions	Unique Interactions	Proteins	Publications
<i>Arabidopsis thaliana</i> *	5909	4160	2118	848
<i>Bacillus subtilis</i> 168	1	1	2	1
<i>Bos taurus</i>	70	58	84	31
<i>Caenorhabditis elegans</i>	7084	6833	3573	42
<i>Canis familiaris</i>	5	5	8	4
<i>Danio rerio</i>	33	33	38	15
<i>Drosophila melanogaster</i>	34655	26888	7578	1619
<i>Escherichia coli</i> K12 MG1655	43	42	51	9
<i>Gallus gallus</i>	45	37	54	23
<i>Homo sapiens</i>	54578	36737	10213	11188
Human Herpesvirus 1	13	10	12	3
Human Immunodeficiency Virus 1	209	185	187	5
<i>Macaca mulatta</i>	1	1	2	1
<i>Mus musculus</i>	4265	3554	2361	604
<i>Rattus norvegicus</i>	684	496	591	225
<i>Saccharomyces cerevisiae</i> *	244552	163188	6049	9706
<i>Schizosaccharomyces pombe</i> *	16205	13248	2110	1487
<i>Xenopus laevis</i>	121	99	111	39
Total	365574	253138	32475	24876

⁵http://wiki.thebiogrid.org/doku.php/experimental_systems

⁶http://wiki.thebiogrid.org/lib/exe/fetch.php/biogrid_workflow.pdf

⁷http://wiki.thebiogrid.org/doku.php/contribute#send_us_your_interaction_data

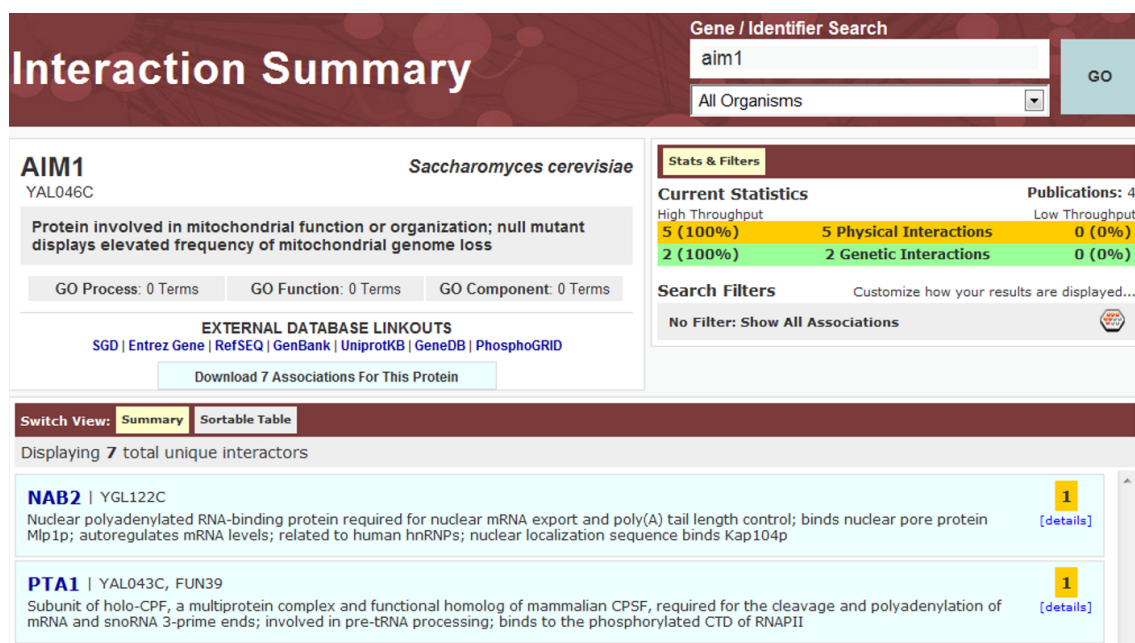


Figure 2.8: The BioGRID database.

A screenshot of the BioGRID web browser output for the yeast gene *AIM1* (accessed 4th March 2011). BioGRID provides curated functional interactions and annotations for each gene, together with experimental details and publication links for each interaction. The interactions for individual genes may also be downloaded as a separate datasets.

The BioGRID website provides comprehensive interaction information for each protein with multiple links to external data sources (Figure 2.8). In addition, the BioGRID datasets are all available for download from the BioGRID website⁸ and through a web service in four standardised formats: PSI-MI XML; Osprey Custom Network; BioGRID TAB 2.0 tab delimited; PSI-MI TAB Version 2.5.

2.1.4.2 The *Saccharomyces* Genome Database

SGD⁹ is an integrated database of molecular biological information about the baker's yeast *S. cerevisiae*. It contains a wide range of data, including sequences [295], annotations [296, 297], phenotypes [298] and publication links for each ORF in the yeast genome and its product (Figure 2.9) [299]. In addition, the database provides various tools, such as BLAST [300], Genome Snapshot [301] and Proteome Browser [302], and has links to other external resources [283].

The SGD community maintains strict guidelines for the curation of new data and provides comprehensive records of all database changes¹⁰. Due to the high level of curation, SGD is one of the most widely used molecular biology resources within the yeast research community.

⁸<http://thebiogrid.org/download.php>

⁹<http://www.yeastgenome.org/>

¹⁰http://wiki.geneontology.org/index.php/SGD_GO_HTTP_guidelines

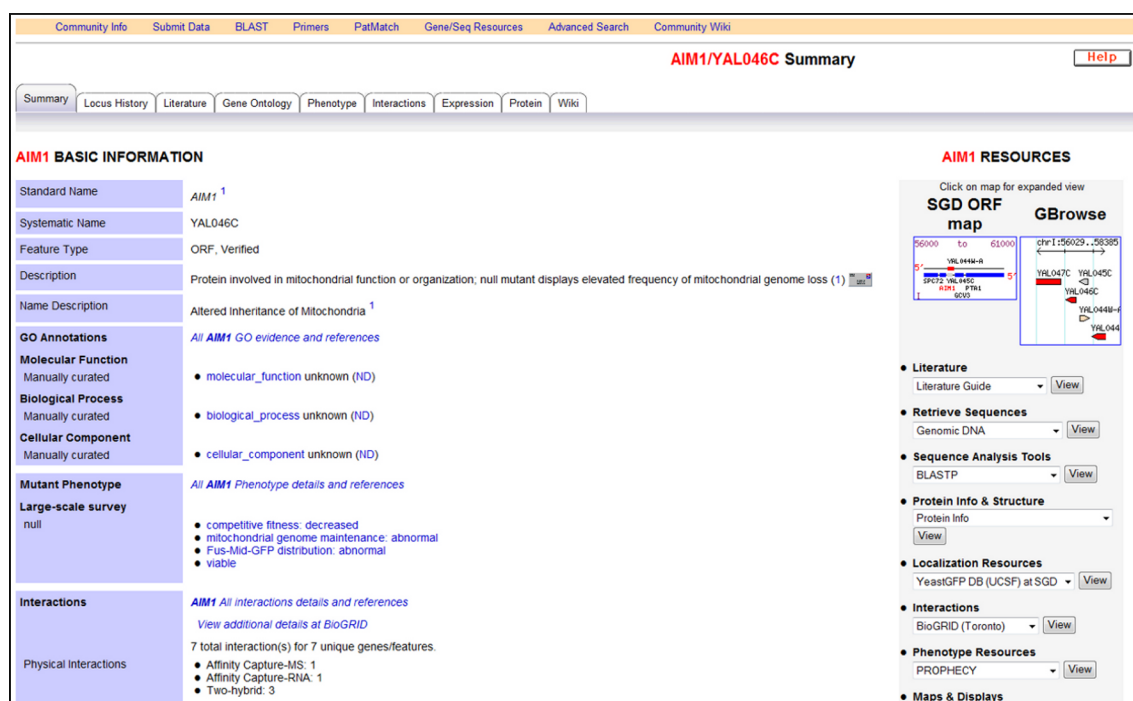


Figure 2.9: The *Saccharomyces* Genome Database.

A screenshot of the SGD web browser displaying the output for the yeast gene *AIM1* (accessed 4th March 2011). SGD provides sequence, annotation and phenotype data for each gene together with publication links. In addition, the database provides a variety of analysis tools and links to other external resources.

2.2 Deciphering the Interactome

2.2.1 The Interactome

The term interactome was first coined in 1999 by Sanchez and co-workers [303] and has since been widely used in scientific literature (Figure 2.10). Initially the concept of the interactome of a species was "*the complete repertoire of interactions potentially encoded by its genome*". This protein-specific definition is commonly used [37, 304–306] and has also been applied to denote subsets of the interactome, such as the microtubule interactome [307], the mitochondrial interactome [308], and the ribonucleoprotein interactome [309].

However, there are several other factors that can be taken into account when defining the interactome. The cell contains other molecules that interact with proteins and contribute to cellular biology [39, 52, 310]. Additionally, complex organisms have different cell types, such as tissues, with a specific subset of interactions occurring in each type. Finally, a full definition of the interactome should allow for differing cellular circumstances, since many molecules vary their function in response to cellular conditions [38, 311].

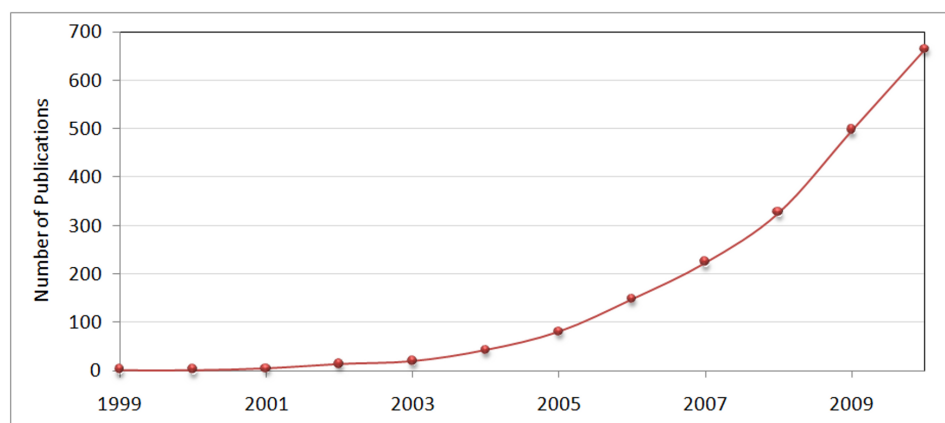


Figure 2.10: The usage of the term "interactome" in PubMed.

(<http://www.ncbi.nlm.nih.gov/pubmed>, accessed 09/04/11).

Systematic identification of the interactome is vital to understanding cellular biology [312]. However, given the complexity of the cell, defining the interactome is not straightforward [37]. Several studies have attempted to define the meaning of the term interactome [303, 304] and estimate its size in different species [42, 134, 184, 313–317]. Additionally, several computational PPI prediction techniques have been developed to complement the available experimental methods [318, 319].

In this thesis the interactome is defined as "*the entire complement of molecular interactions that may occur within an organism under all circumstances and cell types*".

2.2.2 Computational Protein-Protein Interaction Prediction

A full and accurate picture of the interactome requires small-scale LTP confirmation of each interaction. However, there are approximately 6,000 estimated genes in *S. cerevisiae* [320] and therefore a potential of 17,997,000 unique interactions (excluding self-interactions) to be studied. Clearly, many of these interactions would never occur in the cell, since proteins are produced at different times and are localised in different compartments. Therefore, identifying the correct interactions prior to experimental analysis can reduce experimental effort. Several methods have been employed to estimate the number of interactions that may constitute the entire *S. cerevisiae* interactome with a wide range of results (Table 2.5) and it remains unclear how large the true interactome is.

Computational analysis of HTP data can be used to detect potential false negatives and guide the experimental analyses to those interactions most likely to be biologically relevant [321]. Various types of data can be used for PPI prediction including sequence, structures, expression and evolutionary data [319, 322]. A number of databases have been designed to store predicted interactions including PIPS [323], HAPPI [324], OPHID [325], STRING [326], POINT [327], Predictome [328] and UniHI [155].

Table 2.5: Estimates of the size of the yeast interactome.

Various computational methods have been used to estimate the size of the *S. cerevisiae* interactome. Estimates range from 8000* interactions to 75,500** interactions.

Author	Method	Interactions	Ref
Tucker <i>et al.</i> 2001	False positive and negative rates estimated by functional similarity of HTP data and prey interaction rates, followed by estimate reduction for putative and unknown ORFs numbers.	8000* - 12 000	[314]
Legrain <i>et al.</i> 2001	Extrapolation of the interactome size from HTP interaction rates.	15 000 - 20 000	[315]
von Mering <i>et al.</i> 2002	Comparison of known interactions with HTP datasets to estimate minimum interactome size.	30 000+	[184]
Bader & Hogue 2002	Topological estimate based on scaling of the power law distribution in existing biological data.	20 000	[42]
Sprinzak <i>et al.</i> 2003	False positive rate for HTP screens estimated by comparison of HTP and LTP data for co-localisation and shared function.	10 000 - 16 000	[313]
Grigoriev 2003	Estimated based on overlap between datasets coupled with data integration.	16 000 - 26 000	[316]
Hart <i>et al.</i> 2006	Extrapolation from HTP data and estimated false positive rate.	37,800 - 75,500**	[154]
Yu <i>et al.</i> 2008	Extrapolation from HTP binary interaction datasets.	13 500 - 22 500	[134]
Sambourg & Thierry -Mieg 2010	Combination of literature curated data with HTP datasets.	37 600	[317]

Several aspects of genomic sequence can be used to predict **PPIs**. The context of a gene can reveal potential interactions since interacting proteins have increased conservation of their gene order (often as operons in prokaryotes) in comparison to non-interacting proteins [329–331]. Evidence suggests that this prediction method is the most accurate in bacterial genomes [332]. Gene fusion events can also indicate protein interactions by what is termed the Rosetta Stone method. Here, interacting protein pairs have homologs that are fused as one protein in one species, indicating a potential functional link in other species in which they are coded as separate proteins [333–335].

Protein sequence can also be used to predict **PPIs**. Proteins contain specific domains which are crucial to their role and are highly conserved [336, 337]. Interactions between domains are also conserved between species and, therefore, the presence of specific domains and sequence signatures in pairs of proteins can be indicative of **PPI** [338–344]. In addition, the physiochemical properties of a protein, such as charge and hydrophobicity, can be used in combination with sequence to infer interaction [345].

Many interactions are conserved across species and can be identified using orthology and cross-species analysis [327, 346–351]. Conserved interactions, termed interologs, and conserved regulatory interactions, termed regulogs, can be of particular use where there is little interaction data for a species [352, 353]. In addition, the distribution of gene sequences across species, termed the *phylogenetic profile*, is also conserved for many interacting pairs [93, 354–356]. This conservation is thought to occur because interacting protein pairs evolve at the same rate [357]. Therefore, two **ORFs** that have similar profiles are likely to have been co-conserved and, therefore, may interact physically [94, 358–361].

The 3D structure of a protein forms the active site which is essential for its function. Surprisingly, the number of 3D protein structures has been found to be relatively small in comparison to the potential number of sequences [362]. Consequently it may be possible to predict **GIs** and **PPIs** based on a protein's 3D structure, for instance by docking or threading methods [247, 363–369].

Methods that combine sequence data and structural data can improve **PPI** prediction in comparison with sequence or structural data alone [370]. The *in silico* two hybrid (**i2H**) takes advantage of sequence and structural conservation to predict **PPIs**. The algorithm compares multiple sequence alignments for correlated mutations and then calculates interaction scores based on a correlation matrix which can also be used to predict specific residue binding [371]. The **i2H** method has been applied to the bacterium *Escherichia coli* and the predictions are available through the EcID database [372].

Proteins that are expressed at the same time are more likely to interact than those expressed at different times [108]. Moreover, co-expression patterns are conserved between species [373]. Conse-

quently gene expression data can be used to predict functional links [198, 373–376].

An additional source of functional links is the wealth of biological literature stored in databases such as MedLine and PubMed. Many of these publications contain data that is not available in functional databases [377, 378]. Literature mining can be used to extract associations between pairs of proteins by various methods such as pattern matching and natural language processing [258, 379–386].

Machine learning approaches are often used to predict PPIs using a wide variety of data types during classifier training [110, 112, 339, 345, 387–394]. Alternatively, network-based methods, often using heterogeneous data, can be used for PPI prediction [96, 109, 200, 323, 391, 395, 396]. The use of multiple data types by these methods improves prediction accuracy over single-source prediction methods [387, 397]. Network-based prediction of functional associations is discussed in Section 2.4.

2.3 Graph Theoretic Analysis

Graph theory allows biological data to be represented in a manner that is amenable to statistical analysis and manipulation [77, 83, 398]. In graph theory, a *graph* is a mathematical representation of the relationships between entities. A graph consists of a collection of *nodes* connected by a collection of *edges*.

A simple graph G can be represented as:

$$G = (N, E) \quad (2.1)$$

where N is a set of nodes and E is a set of edges. For instance the network depicted in Figure 2.11 consists of:

$$N = \{a, b, c, d, e, f, g\} \quad (2.2)$$

$$E = \{\{a, b\}, \{a, c\}, \{a, d\}, \{a, e\}, \{a, f\}, \{e, f\}, \{a, g\}\}. \quad (2.3)$$

Representation in this manner provides a simple format for study of network structure and its biological implications [312, 321]. Several graph theoretic measures have been used to study network properties and have revealed underlying aspects of network topology such as robustness, connectivity and modularity [76, 399–408].

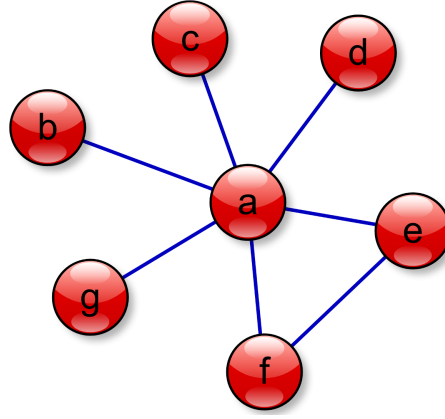


Figure 2.11: A simple undirected graph, G , consisting of seven nodes (red circles labelled a-f) and seven edges (blue lines).

This graph will be used as an exemplar for Section 2.3.

In simple graphs the nodes are of one type. However, graphs may contain multiple node types. Bipartite graphs contain two distinct types of nodes, N_1 and N_2 , and are used to study the relationships between the two groups [409]. At the most complex level, graphs can be multipartite with a range of node types. Large collections of nodes can also be made more tractable for human visual analysis using colour, size and shape, or by attaching attribute labels [77].

Edges can have several types. In simple graphs edges are undirected, with each node having equal importance. Therefore:

$$E \subseteq \binom{N}{2}. \quad (2.4)$$

An example of a simple undirected network of seven nodes and seven edges is shown in Figure 2.11. More complex biological networks can have directed edges where one node is the source and the other is the destination. In this case:

$$E \subseteq N \times N. \quad (2.5)$$

A path is a sequence of nodes within the network. For instance a path of length n is a sequence of nodes $v_0, v_1, v_2, \dots, v_n$, where $(v_i, v_{i+1}) \in E$ for $0 \leq i < n$. Cycles are a specific type of path where $v_0 = v_n$. A directed acyclic graph (DAG) is a specific type of graph, $G = (N, E)$, in which the edges are directed but with no cycles. Edges in a DAG are all directed away from root nodes in parent-to-child relationships such that there are no circular paths from any node back to itself.

In most graphs a single edge connects two nodes. However, edges can be more complex. In some graphs an edge can be connected to the same node at both ends forming a self-loop. In multigraphs a pair of nodes can be connected by several edges, termed multiedges [410]. Finally, hyperedges may connect several nodes in a hypergraph [411]. In visualisations edges may also be coloured to distinguish their type and may be labelled with edge properties such as weights.

Formal graph theoretical definitions are supplied in Appendix A.

2.3.1 Application of Graph Theory to Biological Networks

Biological data can be represented as networks, allowing for visual and graph theoretic analysis of the network structure [357, 412–419]. Biological entities, such as genes, proteins or metabolites, are represented as nodes [420]. In simple networks such as PPI networks or GI networks, nodes are of a single type. Functional networks also have a single node type representing both the gene and its product, since they are derived from several data types.

Many biological datasets consist of two distinct entity types and are represented as bipartite graphs. For instance, regulatory networks have protein nodes and DNA nodes, and metabolic networks have enzymes nodes and metabolite nodes. More complex multipartite graphs can be used to represent multiple biological entities, such as DNA, RNA, proteins, metabolites and ligands, in order to represent cellular biology more accurately [80].

PPI networks are usually undirected, since protein binding is symmetric. Functional interaction networks are also commonly undirected since an edge can represent multiple types of association between a pair of nodes. However, many biological data types can also be represented as more complex networks. DAGs are often used to represent hierarchical data such as annotation data (see Section 2.5.4.3) [100]. Biological interaction data can also be directed. For instance, metabolic networks represent reactions in which the nodes are enzymes, substrates and products, while the edges represent the flux of metabolites [76]. Regulatory and signalling networks are also directed. In regulatory networks the edges represent the binding of transcription factor proteins to their DNA targets [255, 421, 422]. Signalling networks represent the transduction pathways between cellular monitoring components and responses [423, 424]. In combination with PPIs, metabolic, regulatory and signalling networks form a directed functional network in which a cellular signalling response changes gene regulation, resulting in altered metabolic output [425, 426].

Complex edge types can aid in the representation of biological data. For instance, looping edges can represent self-interactions such as dimerisations; multiedges can be represent multiple sources of evidence for protein association [410]; and, hypergraphs can represent complex biological traits such as the formation of protein complexes [411].

For the remainder of this thesis, unless otherwise stated, the term network refers to a functional network of probabilistically-weighted edges where there are no self loops and where a node represents both a gene and its product.

2.3.2 Network Properties and Statistics

2.3.2.1 Node Degree

The *degree* of a node is the number of edges connected to it. Where there are self-loops the looping edges each count as two edges. In directed networks the degree of a node can be subdivided into *out-degree* and *in-degree* to distinguish between edges starting and terminating at the node [412]. The *out/in-degree ratio* can be used to determine a protein's high-level function [427]. In the sample network depicted in Figure 2.11 node **a** has a degree:

$$D(a) = 6. \quad (2.6)$$

A related measure is node *connectivity*. In this case the number of directly connected nodes, termed the node's *neighbourhood*, are counted. Therefore looping edges are ignored [400]. In Figure 2.11 node **a**'s connectivity is also 6 since it has no self-looping edges. Both the *average degree* and *average connectivity* can be measured across the entire network [67].

The *neighbourhood degree* and *neighbourhood connectivity* of a node are the average measures, across the node's neighbourhood, of degree and connectivity respectively. In the example network (Figure 2.11) node **a** has the neighbourhood degree:

$$ND(a) = \frac{D(b) + D(c) + D(d) + D(e) + D(f) + D(g)}{D(a)} = \frac{1 + 1 + 1 + 2 + 2 + 1}{6} = 1.33. \quad (2.7)$$

Random networks, known as Erdős–Rényi networks, can be used to reveal the statistical properties of biological data [428–431]. The *degree distribution* of a network, $p(k)$, is the probability a selected node has k links [432]. Degree distribution can be used to distinguish between different types of network. The degree distribution of random networks follows the Poisson distribution [433]. However, the degree distribution of biological networks is significantly different from random networks, reflecting the high organisation of cellular processes [68].

Many biological networks are thought to be *scale-free* [76, 83, 312, 432, 434–436]. That is they contain many low degree nodes and a small number of high degree nodes, making them highly resistant to random perturbation [312, 414, 437, 438]. In scale-free networks the degree distribution follows

the power law, $p(k) k^{-\gamma}$ [434]. Power law degree distributions have been found in several types of data from a number of species [156, 157, 160, 416, 439].

Scale-free distribution has been observed in several other real world networks such as the world wide web [440, 441] and social networks [442]. However, the scale-free model for biological data has been disputed as the best fitting model in some cases [443–446]. While many biological datasets seem to follow the power law there are other distributions, such as the log-normal and stretched exponential distributions, which may fit some of the data [447]. A geometric random graph model has also been suggested as a better fit for some biological datasets than the scale-free model [407, 448, 449].

Several studies have theorised that the power law distribution may be an artefact of noise in the data or of experimental design, and that the true interactome does not fit this model [68, 430, 450, 451]. Rachlin and colleagues (2006) postulated that this distribution could be plausibly an artefact of the aggregation across multiple process-specific contexts [452]. This view is supported by evidence that in some networks sub-communities have different degree distributions from the network as a whole [453, 454]. Further, some biological networks have been found not to follow the power law [447, 455]. However, the scale-free model remains the most popular model for biological data [450] and is supported by several evolutionary theories [68, 434, 439, 456].

Various models of network evolution have been proposed to account for the scale-free nature of biological data [68]. One of the earliest theories was the *preferential attachment* model which hypothesised that "rich nodes get richer" over time leading to the network hubs [434]. The *link dynamics* model extended the idea of preferential interaction gain to include interaction loss [457]. More recently, new evidence suggests that a node's probability of interaction gain or loss is fixed and does not change through time [458]. Therefore, the scale-free nature of biological networks may have arisen due to selection for robustness and evolvability.

2.3.2.2 Average Path Lengths

The average path lengths within a network can reveal aspects of the underlying network structure. The *shortest path* between two nodes is the smallest number of edges between the two nodes [459]. For instance the shortest path between nodes **b** and **e** in the example network shown in Figure 2.12:

$$SP(b, e) = 2. \quad (2.8)$$

The *diameter* of the network is the longest shortest path within the network. In the example network:

$$D(G) = \text{MAX}(SP(a, b) | a, b \in N) = 2. \quad (2.9)$$

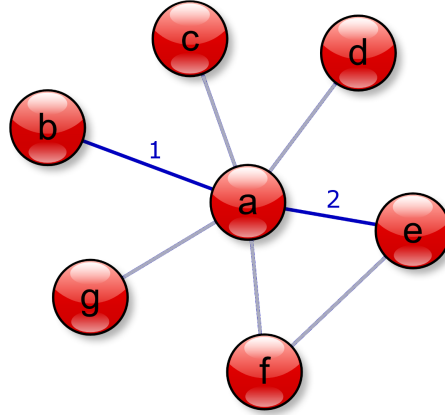


Figure 2.12: Shortest path.

The shortest path between two nodes (red circles) is the shortest number of edges (blue lines) connecting the two nodes. In this example the shortest path between node **b** and node **e** is 2.

The shortest path can be averaged for each pair of nodes in the network to give the network's *characteristic path length* [67]. Biological networks are termed *small world* networks since they have a small characteristic path length and small diameter relative to an equivalently-sized random network [75, 435, 439, 459, 460]. The nodes of small world networks are arranged in local dense regions which are interconnected by a small number of edges [235]. Small world properties can be used in evaluation of network data [461]. The small world phenomenon is also seen in other real world networks such as social networks [462], collaboration networks [459], the world wide web [438] and the internet [401]. Like scale-free networks, small world networks are resistant to perturbations [312, 438]. Scale-free networks are sometimes referred to as *ultra-small world* networks [463, 464].

The *clustering coefficient* of a node is a measure of the degree to which its surrounding nodes cluster together [459]. The clustering coefficient is measured as the ratio of the number of links in the neighbourhood and the maximum possible number of links between them:

$$CC(a) = \frac{2E(a)}{(N(a)(N(a) - 1))} \quad (2.10)$$

where $N(a)$ is the neighbourhood of node **a** and $E(a)$ is the number of connected pairs within $N(a)$. In the example network in Figure 2.11 the clustering coefficient of node **a** is $CC(a) = 2/(6 * 5) = 0.0667$. The distribution of the clustering coefficient in some networks has been observed to follow a power law.

The average clustering coefficient across all nodes of a network can be calculated giving the *network clustering coefficient* [459]. In small world graphs this property is significantly higher than for random networks, while the characteristic path length remains the same.

2.3.2.3 Centralities

Several network measures assess the importance of nodes and edges in network information flow. These measures are termed centralities [465, 466] and can be used to identify essential nodes in the network [467, 468]. Many centrality measures have been developed, for example:

Closeness Centrality [466] of a node is the average of the shortest distance from it to all other nodes in the network N :

$$Ccl(a) = \frac{1}{avgSP(a,t|a,t \in N)} \quad (2.11)$$

where, t ranges over all nodes in the network and $SP(a,t)$ is the shortest path between nodes a and t .

Graph Centrality [469] measures the maximum shortest path from a node to all other nodes in the network:

$$Cg(a) = \frac{1}{maxSP(a,t|a,t \in N)} \quad (2.12)$$

where, t ranges over all nodes in the network and $SP(a,t)$ is the shortest path between nodes a and t .

Stress Centrality [469] is a measure of the number of network shortest paths passing through a node:

$$Cs(a) = \sum_{s \neq a \neq t \in N} \sigma(s,a,t) \quad (2.13)$$

where, s and t are the nodes in the network distinct from node a and $\sigma(s,a,t)$ is the number of shortest paths from s to t on which node a lies. Stress centrality can be considered a measure of a node's importance in network information flow.

Betweenness Centrality or *node betweenness* [469, 470] of a node extends stress centrality to calculate the proportion of shortest paths passing through a node:

$$Cb(a) = \sum_{s \neq a \neq t \in N} (\sigma(s,a,t)/\sigma(s,t)) \quad (2.14)$$

where s and t are the nodes in the network distinct from node a , $\sigma(s,t)$ is the number of shortest paths from s to t and $\sigma(s,a,t)$ is the number of shortest paths on which node a lies. This measure reflects the control a node exerts on the interactions of other nodes in the network and is commonly used as a measure of node essentiality. Nodes with high betweenness centrality and low connectivity are often found linking network modules [471] and are referred to as *bottlenecks* since they restrict the flow of information through the network [470]. *Edge betweenness* can also be measured as the

number of shortest paths an edge lies on [442, 472, 473]. Betweenness centrality measures have been used to identify potential drug targets [438] and to identify disease genes [474].

Bridging Centrality [475] measures the information flow through a node:

$$Cbr(a) = B(a) \times Cb(a) \quad (2.15)$$

where,

$$B(a) = \frac{D(a)^{-1}}{\sum_{i \in N(a)} \frac{1}{D(i)}} \quad (2.16)$$

and, $D(a)$ is the degree, $N(a)$ is the neighbourhood and $Cb(a)$ is the betweenness centrality of node a . This measure can be used to identify nodes situated between highly connected regions of the network that are likely to modulate network information flow. In biological data bridging centrality has been found correlate with gene lethality and may be a good indicator for potential drug targets [77].

Several other more complex centrality measures exist, each measuring different aspects of node importance, including subgraph centrality [476], eigenvector centrality [477] and information centrality [478].

2.3.2.4 Centrality-Lethality

Scale-free networks are extremely robust to random attack (mutation) since the removal of the majority of nodes does not significantly change network structure [479]. However, these networks are vulnerable to attack targeted at the high degree nodes since the removal of these nodes adversely affects network structure by increasing network diameter [438, 480]. In biological networks this property is consistent with yeast experimental data since only a small number of gene mutants (18.2%) are non-viable [5, 6].

The high degree nodes (generally >10 edges) in scale-free networks have been termed *hubs* [413]. If $2 \leq \gamma \leq 3$ in the power law the hubs are considered significant to network structure [434]. Unsurprisingly, node degree correlates with gene lethality and essentiality [416]. Hub proteins in GI networks are also commonly hubs in PPI networks [481]. In functional networks hub proteins are more likely to be essential than non-hubs [482, 483] and take part in more essential PPIs than non-hubs [484]. Additionally, hub essentiality has been found to have a strong correlation with genetic pleiotropy [134] and, in directed networks, with in- and out-degree [485]. Similarly, in protein structure networks hub proteins correspond to active sites and PPI interfaces [486]. However, unlike social network hubs, those of biological networks tend not to interact with each other [402].

Several theories have been proposed to explain hub protein essentiality including interaction dynamics [487] and essential domains [488]. However, it has been shown that the essentiality of a protein is due to its position within the network's topology, a concept known as the *centrality-lethality* rule [416, 489, 490]. Centrality-lethality has been observed in several species [482, 491, 492].

Hubs can be further categorised into two distinct groups [489, 493]. The first, termed *party hubs*, take part in all their interactions at the same time and are thought to occur in functional modules. The second, termed *date hubs*, interact with different proteins at different times, conditions or locations and are thought to have regulatory roles. Using context-specific expression data party hubs have been observed to be interactively conserved across contexts while date hubs are interactively varied with different context-specific roles [452]. A more complex four category system of hub classification has also been proposed based on gene expression characteristics [494]. However, the exact nature and role of many hub proteins remains unclear [493, 495].

2.3.3 Network Modularity

Biological networks are highly complex and, while graph theoretic representation allows visualisation of the data as a network, in most cases large networks remain difficult to study in this format given their scale [412]. Study of biological data has revealed that cellular parts are grouped into node communities of similar function, and many known complexes have been identified as densely connected regions in biological networks [52, 65, 404, 431, 452, 496, 497]. Additionally, it has been observed that many biological networks have a highly-connected hierarchical structure of modules within modules [432, 498]. Consequently, dense network regions are widely believed to relate to the functional units of the cell, each of which performs specific tasks [52, 312, 498–502].

However, it should be noted that there is some evidence this assumption may be incorrect and that the modular structures of PPI networks has less biological significance [451]. Nevertheless, an important aspect of graph theoretic analysis involves identifying network patterns and partitioning large networks into smaller subnetworks in order to identify patterns of connectivity that reveal the underlying mechanisms of cellular biology [442, 503–505]. Since members of these dense areas tend to share common functions, these modules can also be used for functional prediction (see Section 2.5.1) [52, 92, 312, 415, 504, 506–511]. Further, by cross-species comparison of network modularity conserved evolutionary patterns can be identified [405, 512–515].

2.3.3.1 Network Patterns

Various measures of edge density, in particular the centralities and degrees, can reveal topological patterns within networks [432, 459, 461]. A subnetwork, or subgraph, is a subset of the whole network. Several types of subnetwork pattern have been defined, including:

Cliques A clique is a subnetwork, sometimes called a complete subgraph, in which all nodes are connected to one another (Figure 2.13 A) [431]. Often cliques are referred to as *k-cliques* where *k* is the number of nodes in the clique. The simplest *k*-clique is the triangle where *k* = 3.

A *maximal clique* is a clique within a network which cannot be extended by the addition of further nodes. Calculation of a dense network's maximal cliques is considered non-deterministic polynomial-time hard (NP-hard) [516]. However, maximal cliques can be calculated for most PPI datasets since their edges are relatively sparse [517].

The *maximum clique* of a network is its largest fully complete subgraph. However, in large, complex networks calculation of the maximum clique is an non-deterministic polynomial-time complete (NP-complete) problem [516]. Therefore, the maximum clique is often approximated to reduce computation time [518].

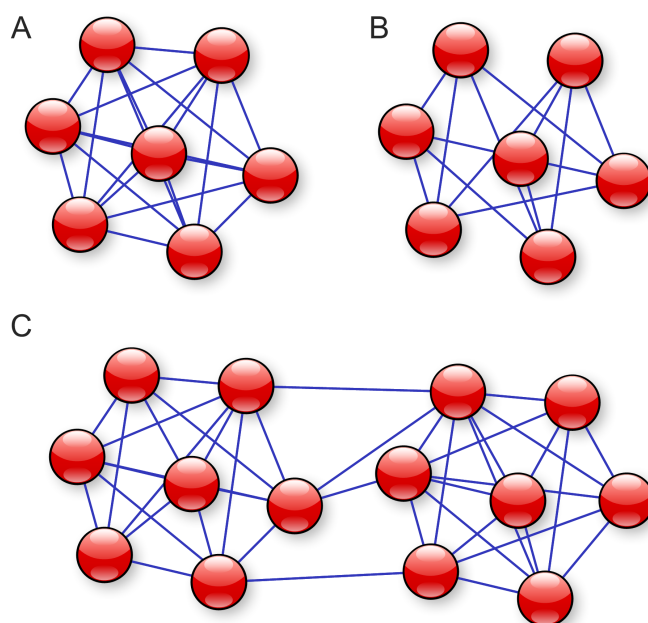


Figure 2.13: Network modularity.

Several types of subnetwork pattern can be defined. **A.** *k*-cliques are complete subgraphs of *k* nodes (red circles) in which all the nodes are connected to one another. Here *k* = 7. **B.** *k*-cores are maximal subgraphs where each node has $\geq k$ edges. Here, *k* = 4. **C.** Modules are a densely connected region with tightly-connected high-degree inner nodes and low-degree outer nodes. Here two densely connected modules are connected by four edges (blue lines).

The k -clique definition of connectivity is very restrictive since many highly-connected network regions do not have the symmetric connectivity required for a k -clique. Further, in terms of cellular biology, it is unlikely that all members of a functional unit will be fully connected in this way. Therefore, a more flexible definition is the n -clique in which the shortest path between all nodes is no higher than n . Where $n = 1$ the clique is equivalent to a k -clique, while higher values of n allow for missing edges within the subgraph [519].

k-Cores Patterns can be identified which have a fixed number of edges while not being fully connected. A k -core is a maximal subgraph within a network in which each node has $\geq k$ edges (Figure 2.13 B) [520]. A k -core, therefore, has higher connectivity than an n -clique but is not as restricted as a k -clique. A related subgraph pattern is the k -plex in which each node has at least degree $|N| - k$, where $|N|$ is the size of the subgraph [521].

Modules A *module* is defined as a densely connected region of a network with tightly-connected, high-degree inner nodes and low-degree outer nodes (Figure 2.13 C) [431, 437]. Each network module is considered a discrete cellular component with a specific task that is separable from other cellular components [52, 432, 522]. In biological data modules can often be found in hierarchical structures in which small modules can be grouped into larger modules [498, 523, 524]. Unlike cliques and cores there is no widely accepted graph theoretic definition of network module connectivity. However, several studies have defined a biological network module using either topological features or additional functional data [429, 431, 472, 525, 526].

2.3.3.2 Clustering

Clustering algorithms use a network's topological properties to search for dense regions in the network or to divide the whole network into distinct parts [527]. There are dozens of network clustering strategies and algorithms which can be applied to biological data (reviewed in [528] and [529]) and the performance of these algorithms differs considerably for different data types [291, 530, 531]. It is beyond the scope of this thesis to discuss each algorithm in depth so this section provides a discussion of the main clustering approaches, concentrating on those algorithms designed for or commonly applied to biological data. Six widely used algorithms, restricted neighbourhood search clustering (RNSC), Girvan-Newman (GN) edge-betweenness, molecular complex detection (MCODE), clique percolation (CP), super paramagnetic clustering (SPC), and Markov clustering algorithm (MCL) are discussed in detail as examples of distinct clustering strategies. These algorithms are summarised in Table 2.6.

Table 2.6: Network clustering algorithms.

A summary of the clustering algorithms discussed in this section. **O** indicates overlapping clusters are produced by the algorithm while **W** indicates networks with weighted edges can be analysed by the algorithm.

Name	Method	O	W	Ref
Restricted Neighbourhood Search Clustering	Iterative network partitioning based on a cost function.	x	x	[532]
Edge Betweenness	Divisive partitioning by iterative removal of high betweenness edges.	x	x	[442]
Molecular Complex Detection	Agglomerative clustering from seed nodes based on neighbourhood densities.	x	x	[89]
Clique Percolation	Clique merging to identify overlapping communities.	✓	x	[453]
Super Paramagnetic Clustering	Ferromagnetic model which assigns spins to network nodes and identifies spin correlation.	x	✓	[533]
Markov Clustering Algorithm	Iterative network flow-based clustering.	x	✓	[534]

Many clustering methods aim to partition networks without loss of network elements. Various parameters are used to select natural partitions in the data. A common partitioning approach is to examine shared interaction partners of the nodes [90, 521, 535–538], or of subgraphs [539], and partition the network to optimise shared interactions within the clusters. Several partitioning algorithms use cost-based functions to assess cluster quality in an iterative fashion. These approaches assume that edge density within a cluster should be significantly higher than between clusters.

The **RNSC** algorithm is a local search clustering algorithm which partitions networks into densely connected regions [532]. The algorithm first partitions the network into random clusters unless an initial clustering is supplied by the user. Single nodes are then iteratively moved between clusters to optimise densely connected regions. The algorithm uses two cost functions based on the number of edges between and within the clusters; a naïve cost and a scaled cost. Higher costs correspond to low density clusters with high between cluster connectivity. The naïve cost acts as a pre-processor. At each iteration the cost function is calculated and nodes are moved between clusters until the cost function is minimised. The process is then repeated until the scaled cost function has also been minimised. **RNSC** runs until the cost functions have not been reduced for a user-specified number of iterations.

Clustering algorithms can exploit the hierarchical modular structures present in biological data [540]. There are two approaches to hierarchical clustering; top-down division and bottom-up agglomeration [541]. In the first approach the network is clustered in a top-down, divisive approach that iteratively

removes network elements based on their topological properties. There are several parameters that can be used to divide the network such as degree, clustering coefficient and centrality measures [472, 475, 525, 526, 542]. One of the most popular measures for divisive hierarchical clustering is edge betweenness.

The Edge-Betweenness Algorithm, often referred to as the Girvan-Newman (GN) algorithm, has been widely used in divisive network clustering [221, 442, 473, 504, 543, 544]. In this approach edges of high betweenness are sequentially removed from the network [442]. Edge betweenness is calculated as the proportion of shortest paths in the network on which an edge lies (see Section 2.3.2.3). After initial calculation the highest scoring edge is removed. Calculation and edge removal is then repeated for a given number of iterations or until a specified cut-off. The final clusters represent areas of dense low-betweenness edges. Unlike partitioning algorithms the clusters may not contain all the original network elements since single nodes are not considered clusters.

Due to the requirement for iterative calculations the original GN algorithm is computationally intensive and does not scale well for large complex networks [545]. A faster implementation of the algorithm has since been developed where partial betweenness scores are calculated for a random subset of the edges, giving an approximation of betweenness across the network [546]. An alternative fast approximation of betweenness can be achieved using a network structure index [547]. More recently a parallel version of the GN algorithm has been developed to improve performance [548].

Networks can also be hierarchically clustered in a bottom-up agglomerative manner. These algorithms begin with seed nodes and grow clusters based on network properties [549, 550]. Seed nodes can be chosen based on several different properties such as distance measures [551, 552], shared neighbours [536, 553] or node degrees [554]. In some cases groups of densely connected proteins are chosen as seed "cores" [555]. Alternatively seed nodes can be specified by the user based on their research interests [292, 541]. Fusion strategies can combine the agglomerative and divisive hierarchical approaches by using average subgraph degree to merge nodes and betweenness to filter them [556]. Like divisive clustering, agglomerative methods do not necessarily cluster all the original network components.

MCODE is a popular agglomerative clustering algorithm that has been incorporated into several network visualisation tools (see Section 2.3.5) [89]. The algorithm first calculates neighbourhood density weightings based on *core clustering coefficient*. The core clustering coefficient of a node is the density of the highest k-core in the neighbourhood of the node, inclusive of the node itself. This measure is then multiplied by the maximum k-core number, k_{MAX} , of the node neighbourhood to give a final density weighting. The highest weighted nodes are selected as seed nodes for agglomerative clustering. In the second stage MCODE finds densely connected regions by recursively building

modules from the seeds based on their neighbourhood weightings. Nodes are added to the core node cluster if they score within a given threshold of the core node's score. Optional post-processing of the clusters allows the user to filter the clusters by connectivity. Finally, the clusters are ranked according to size and edge density.

Several clustering approaches represent a network as a matrix of similarities between its elements. The eigenvectors of the matrix are then used for network clustering [415, 443, 489, 557]. These methods, termed spectral clustering, are popular as they reveal clearer clustering patterns than the network connectivity alone [415, 529].

A drawback of many clustering strategies is the lack of overlap between the final clusters. Many proteins can have multiple functions based on cellular conditions. Consequently, a protein may be a member of several modules within the network, each with its own distinct function [453]. Therefore biological networks contain overlapping communities rather than distinct partitions [96, 453, 498, 558–560]. Several clustering algorithms have been developed which can identify overlapping clusters and allow nodes to be placed in multiple clusters [561]. Many of these methods are extensions of existing clustering approaches such as the GN-based algorithm developed by Wilkinson and colleagues which can assign nodes to several clusters and provide confidence scores for each assignment [544]. Additionally, clustering into communities of edges, rather than nodes, allows for overlapping clusters since a node can be connected to several edges in different clusters [562–564].

The clique percolation (CP) algorithm, also termed clique merging, is a popular agglomerative method which can be used to identify overlapping clusters and locate areas of module cross-talk [453, 499, 558, 561, 565–567]. Cliques are densely connected subgraphs and therefore can be considered network modules (see Section 2.3.3.1) [566]. The CP algorithm uses cliques as seeds for module detection and connects them together if they share $k - 1$ nodes, where k is the number of nodes in the two cliques. The connected cliques, termed communities, can reveal areas of cross-talk at shared nodes. A similar method developed by Spirin and Mirny (2003) identifies maximal cliques as representative of overlapping network modules of different sizes [431].

The majority of clustering algorithms are designed for use with unweighted network data. However, edge weights contain valuable information that can reveal more biologically significant connectivity patterns than topology alone [88, 568]. Several clustering algorithms, many of which extend the principles of existing clustering algorithms, have been designed for use with weighted data. For instance, maximal cliques can be used as the seeds for agglomerative module discovery in weighted networks [517].

A hierarchical clustering algorithm that allows edge weights is [SPC](#) [[431](#), [533](#), [569](#)]. The algorithm, also termed the Pott's Model, simulates a ferromagnetic model subject to temperature dependent fluctuation. [SPC](#) assigns a spin to each node in the network. Spins can be in several different states and, where interactions occur, the spins affect one another. In densely connected regions these spins correlate allowing the algorithm to cluster the network based on correlation patterns. Transitivity clustering also allows for the use of edge weights between protein pairs [[570](#)]. This algorithm clusters the network by addition and removal of edges using a weighted cost function.

Genetic algorithms can be used for cost-based clustering to produce overlapping clusters for weighted or unweighted networks. These algorithms exploit the principles of evolution by simulating natural selection [[560](#)]. At each iteration groups of nodes are moved between clusters, and the resulting clusters are then assessed for fitness based on within-cluster edges. Those clusters that have improved fitness are kept while those changes that reduce fitness are discarded. Evolutionary algorithms run for multiple generations until a required level of fitness is reached. Importantly, since each generation is changed at random, the results of two separate runs on the same network will not necessarily be the same, allowing for several overlapping clustering patterns.

[MCL](#) is a clustering algorithm based on network flow simulation which can be used for weighted and unweighted networks [[534](#)]. The algorithm uses mathematical bootstrapping by simulating random walks across a probabilistic adjacency matrix. [MCL](#) then applies iterative expansion and inflation to find strongly connected regions and weaken sparsely connected ones. In the expansion step flow is simulated across the matrix. The inflation step then strengthens the areas of high flow and weakens those of low flow. The final clusters represent areas of the network with high flow. A single parameter, the inflation value, influences the final number of clusters produced. Clustering comparison studies have shown that [MCL](#) has higher accuracy than many other algorithms [[429](#), [571](#)]. The [MCL](#) algorithm has also been modified and extended for use with several specific data types [[214](#), [564](#), [572–574](#)]. A similar algorithm, repeated random walks ([RRW](#)), produces overlapping clusters on weighted networks based on random walks across the network [[575](#)].

Several clustering strategies use additional biological data to find and evaluate clusters in [PPI](#) networks. These techniques use the additional data either to enhance cluster discovery or in the post-processing of clusters from existing algorithms. Many data types can be used, for instance gene expression data [[215](#), [227](#), [506](#), [576–579](#)], functional annotations [[504](#), [535](#), [580](#), [581](#)], genetic interactions [[582](#)] or domain profiles [[583](#)]. Conversely, [PPI](#) data can be used to enhance the clustering of gene expression data [[225](#)].

2.3.4 Alignment and Comparison

Network comparison can reveal a network's underlying properties, detect noise, predict missing data and reveal conserved areas of interaction [513, 584]. However, network comparison is non-trivial and direct comparison of networks is computationally expensive and often impossible. For instance subgraph isomorphism, which identifies whether a network is an exact subgraph of a larger network, is an **NP-complete** problem. Consequently, heuristics such as global network properties and small local motifs are commonly used for comparison [585], although some non-heuristic algorithms have been developed [586].

Network comparison can be carried out at different levels. At the simplest level a biological network can be compared to a network model [428]. In this context a network model is a randomised network designed to match the biological network's topological properties. Network edges are randomised such that degree distribution, diameter, motif distribution or other topological characteristics are preserved [434, 459, 587–589]. Accurate network models match several aspects of the biological data and can be statistically analysed to predict missing data [448] and identify interesting motifs [590] or used as null hypotheses for prediction algorithms [588]. Network models can also be used to compare biological networks. Here each network's topology is compared with the random model and the resulting network profiles are then compared [512]. Similarity of the profiles can reveal the underlying similarities of the networks.

Networks can also be compared directly by comparison of topological properties such as degree distribution, clustering coefficient and diameter. However, two networks with similar topology can be vastly different [408]. An alternative approach is to analyse the distribution of network motifs [588]. Here, the distribution of small subgraphs, termed graphlets, is used as a network profile. Network profiles can then be compared to reveal the underlying network similarities and differences.

Network alignment allows for a more accurate comparison between network structures [513] and can reveal networks patterns such as those of disease genes [591]. Within-species alignment is relatively straightforward since nodes can be merged based on protein identity [312]. Overlap between networks can be used to identify true positive interactions [44]. Several tools are available for simple comparison of networks from the same species [400, 592].

At a more complex level networks can be compared across multiple species [417, 513, 593]. There are two aspects to these alignments [594]. Locally small conserved regions can be aligned by matching nodes or motifs, while globally entire network structures can be aligned. Cross-species analysis is non-trivial since most biological networks are noisy and incomplete. Alignment complexity increases with the size of the networks and with the number of networks to be aligned [594]. Addition-

ally simple gene matching is not possible for cross-species alignment due to differing genome sizes and naming conventions. Therefore, additional data such as annotation, sequence or topological data are commonly used to aid alignment [595–597].

Additional data can be used for network alignment and comparison since protein function often correlates with network topology [312]. Orthology and sequence similarity scores can be used to identify conserved interactions, termed interologs, and patterns of conserved regulation, termed regulogs [596]. In these algorithms proteins are aligned based on sequence similarity and edges are then merged if they link similar protein pairs in the two networks. Edges may also be inferred in one species where highly similar proteins interact in another species [347, 598]. Similarly, enzyme classification can be used to align metabolic pathways [597, 599]. These cross-species analyses identify areas of network conservation and can give clues into the evolutionary origins of network structure [417, 584, 600–605]. The NetAlign¹¹ tool provides a web-based server for cross-species network alignment based on sequence similarity [606].

Data compression algorithms can also use annotation data to pre-process networks prior to comparison and alignment. Network compression, also termed network simplification, involves the iterative contraction of network components. For instance edges and nodes can be compressed based on node annotations, reducing the complexity of alignment [607–609].

2.3.5 Network Tools

Due to the complexity of biological data a number of tools have been designed for the analysis of network data. These tools range from simple visualisation and layout platforms to more complex data manipulation and analysis platforms (reviewed in [610] and [85]). Dozens of free to use visualisation tools are available both on the web (for example [106, 611–617]) and as stand alone software (for example [80, 618–627]). In some cases these tools are linked to specific databases allowing the automatic import of network data [304, 615, 621, 622]. There are also tools designed for specific types of biological data such as metabolic networks [628, 629], gene regulatory networks [630], protein-small molecule networks [631], microarray data [616] and literature mined data [613].

Network visualisation is relatively straightforward for small, simple networks. However, as network size and connectivity increases efficient visualisation becomes difficult [632]. A significant problem in network visualisation is network layout. Several algorithms have been developed to address this problem. These algorithms aim to arrange the nodes and edges to optimise ease of visualisation. The GEM algorithm uses generalized expectation-maximization to minimise overlapping edges in the network [633]. The force-directed algorithm also aims to reduce edge overlap, while at the same

¹¹<http://www1.ustc.edu.cn/lab/pcrystal/NetAlign>

time keeping edges at approximately equal length [634]. The spring embedded algorithm aims to equally distribute the nodes while minimising the edge length between connected nodes [635].

Network topology parameters can be used to aid network layout. For instance, the BFL algorithm calculates node betweenness and places high scoring nodes in optimal positions. Alternatively biological data, such as annotations (see Section 2.5.4), can be used to arrange network nodes in groups of the same type [636]. Additionally, three-dimensional layouts can also aid efficient visualisation [624]. The majority of network tools provide a number of layout options to suit a variety of data types [610].

Many visualisation tools provide network manipulation options which permit dynamic interaction with the data. Manipulation allows the user to customise the appearance of the network to suit the network analysis problem being addressed. Nodes and edges may be moved, hidden, merged, coloured or labelled in order to aid visualisation and reveal patterns of connectivity. Some tools also link to biological databases, such as SGD and GO, allowing access to further annotation data.

Tools have also been developed for the analysis of network data, for instance the calculation of network topological properties [592], subnetwork prediction [106, 637] and comparison [638], motif analysis [639, 640] and clustering [78]. The most sophisticated network tools combine dynamic visualisation with a wide range of layout and analysis options.

2.3.5.1 Cytoscape

Cytoscape¹² is one of the most comprehensive and widely used network analysis platforms [641]. Cytoscape is an open source platform which provides dynamic visualisation and analysis of network data (Figure 2.14) [619]. A web-based interface is also available [642].

The core Cytoscape program provides a wide range of visualisation and layout options¹³. Nodes and edges may be coloured, annotated and filtered. The platform supports many input types¹⁴ ranging from simple tab delimited text to standardised network formats such as BioPAX [643] and PSI-MI [644].

In addition to its core platform, Cytoscape supports a wide range of plugins and its developers actively encourage plugin development. There are currently 133 Cytoscape plugins available through the Cytoscape website¹⁵, 18 of which were first published during 2010 [553, 645–661].

¹²<http://www.cytoscape.org/>

¹³http://cytoscape.org/manual/Cytoscape2_8Manual.html#AutomaticLayoutAlgorithms

¹⁴http://cytoscape.org/manual/Cytoscape2_8Manual.html#SupportedNetworkFileFormats

¹⁵http://chianti.ucsd.edu/cyto_web/plugins/ (accessed 13th March 2011)

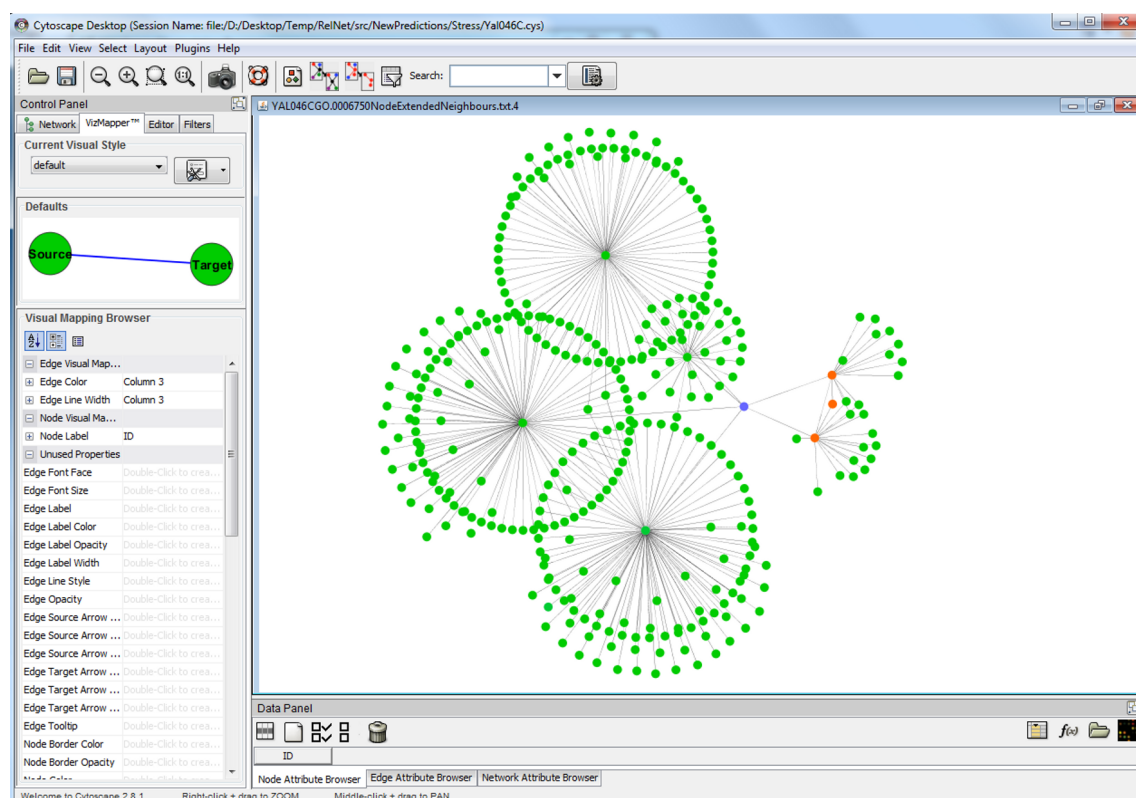


Figure 2.14: Network visualisation in Cytoscape.

A screenshot of the Cytoscape network visualisation platform displaying a network of 339 nodes and 352 edges in the GEM layout.

Table 2.8: Cytoscape plugins.

Summary of the types and number of plugins available from the Cytoscape website.

Type	Number of Plugins
Analysis	56
Network and Attribute Input/Output	30
Network Inference	8
Functional Enrichment	9
Communication/Scripting	8
Miscellaneous	22

Plugins allow a wide range of network analysis and manipulation (Table 2.8). The majority of the plugins provide network analysis functions, such as the NetworkAnalyser plugin which calculates topological parameters [400] and the MCODE network clustering plugin [89]. Additionally, several plugins have been developed to allow access to external databases (for instance [304, 662, 663]). Cytoscape was chosen as the primary network analysis tool for this project since it fulfils all the necessary requirements for functional network manipulation and analysis.

2.3.5.2 Ondex

Ondex¹⁶ is a stand-alone, open-source, semantic data integration platform which allows dynamic visualisation of network data from a wide variety of biological sources [80]. Ondex produces a semantically-rich integrated network which has multiple types of node (referred to as *concepts*) and edge (referred to as *relations*). The concepts are displayed as differently coloured shapes for ease of visualisation (Figure 2.15). The program uses workflow-based parsers to map entities from diverse databases onto one another using an underlying ontology describing their biological relationships (Figure 2.16). For instance, the concept Protein is linked to the concept Gene via the relation *is_encoded_by*. Like Cytoscape, Ondex is customisable by the development of plugins for new parsers, filters and statistical analyses.

Datasets are provided for *S. cerevisiae*, *A. thaliana* and *H. sapiens* through the Ondex website¹⁷ [664–666]. A *B. subtilis* dataset is also available through the Newcastle University website¹⁸. The Ondex *S. cerevisiae* dataset was used during functional prediction evaluation in Chapter 7 since it provides a more complete picture of the network interactions than a network with a single node type.

2.4 Network Integration

There are many types of functional data available (see Section 2.1). However, no single data type can completely cover the interactome since each data type provides information about a different aspect of the cellular biology [59, 667]. Therefore, heterogeneous data sources must be integrated to gain a full picture of the cell [51, 420, 497, 668–671].

While LTP data are generally considered very accurate, it is small-scale. In comparison, HTP data are noisy but can provide cell-wide information. Consequently the two data types complement one another [68, 672]. Integration of multiple types of data can also aid in the interpretation of results since combining diverse data sources of different scales provides a fuller picture of the functional interactions occurring in the cell than is possible using a single data sources alone [83, 248]. Integration can reduce experimental noise, and enhance weak interactions present in multiple data sources [48, 49, 51, 66, 67, 109, 342, 673]. Integrated network analyses have been shown to improve accuracy in several applications, for example in inference of gene/protein function [92, 128, 342, 669, 674], in prediction of novel functional interactions [318, 675–677], in detection of protein complexes [90, 429, 534] and in identification of potential disease genes (see Section 2.5.3) [40, 678].

¹⁶<http://www.ondex.org/>

¹⁷<http://www.ondex.org/doc.html>

¹⁸http://research.ncl.ac.uk/synthetic_biology/downloads.html

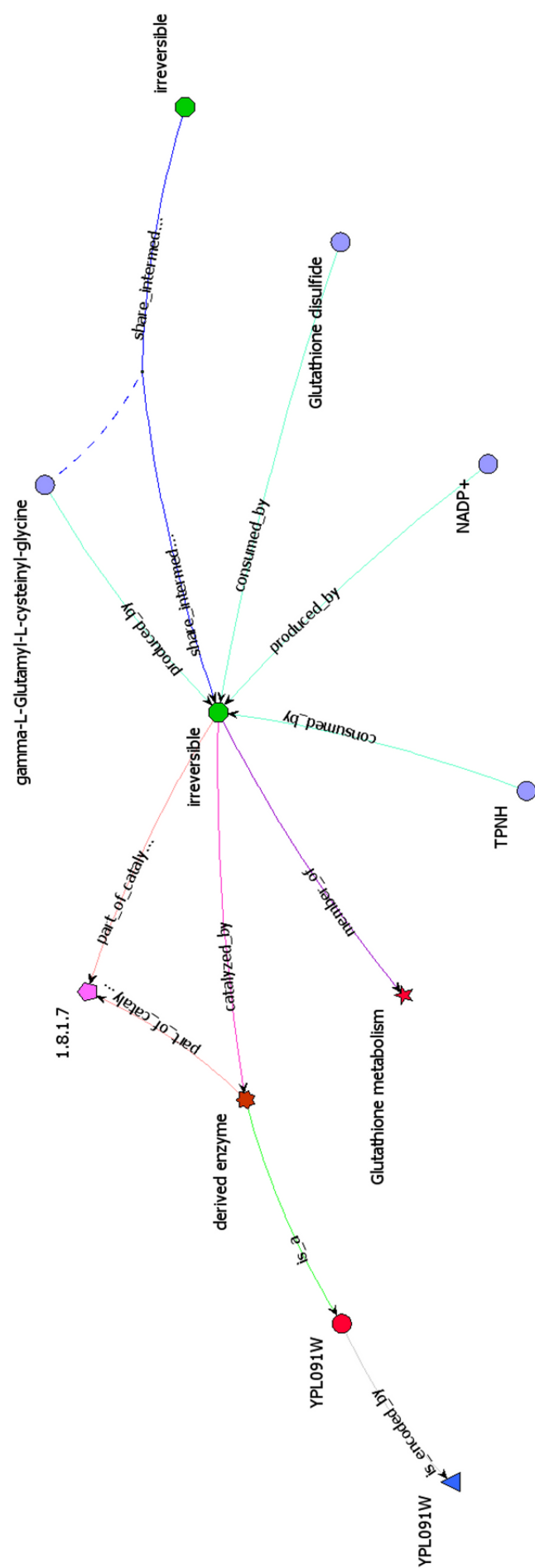


Figure 2.15: Network visualisation in Ondex.

An example Ondex network consisting of seven concept types. Ondex allows dynamic visualisation of the network data. Concepts can be moved, hidden, filtered and tagged with annotations.

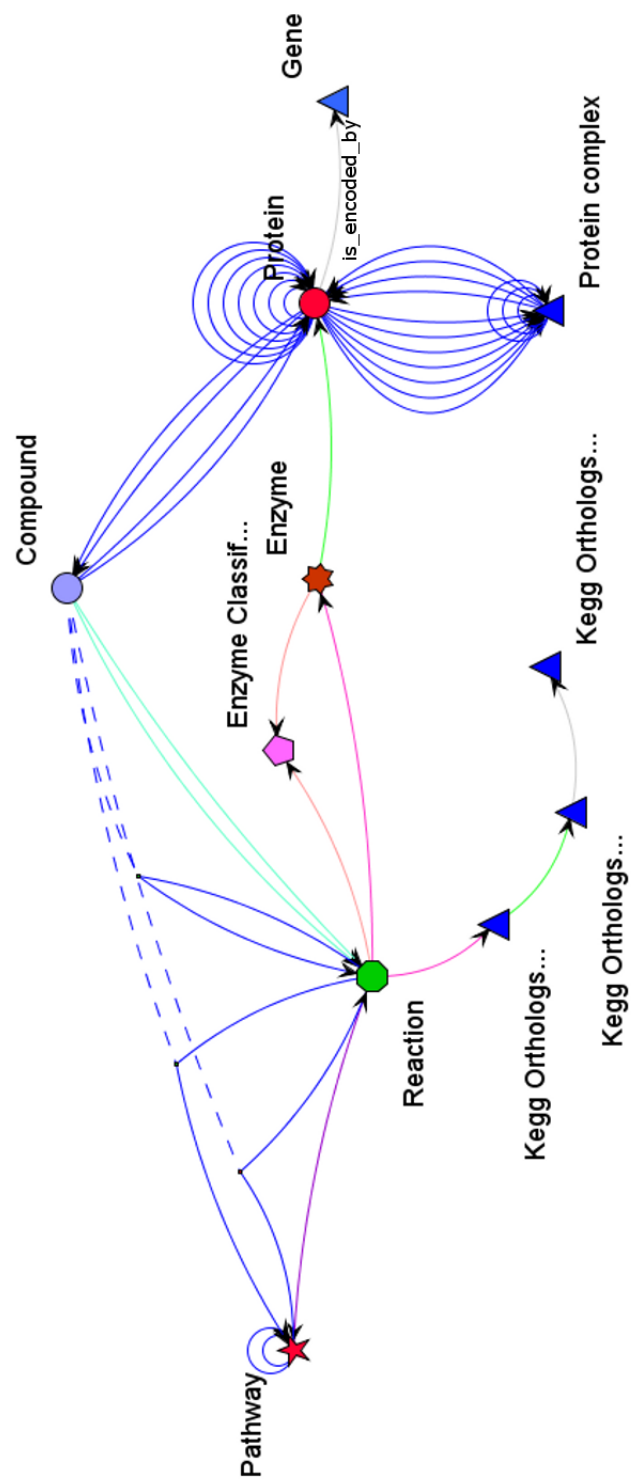


Figure 2.16: Oindex metadata ontology.

The metadata ontology of the network in Figure fig:background:OindexGraph. Oindex networks are based on an underlying ontology such as this which uses a semantically-rich controlled vocabulary to describe the network concepts and the relations between them. For instance, the concept Protein is linked to the concept Gene via the relation is_encoded_by. In this example the other relation labels have been hidden for ease of visualisation.

Network integration can be performed in a variety of ways. At the simplest level datasets can be combined naïvely into a network in which nodes represent genes or gene products, and edges represent any type of interaction between the nodes [79, 508, 679, 680]. In this case no attention is paid to the number of evidence types supporting an edge or the quality of the evidence (Figure 2.17 A). Such networks are useful for the basic visualisation of integrated results.

At a slightly more complex level edge weights can represent the number of lines of evidence for each interaction (Figure 2.17 B). This weighting provides a measure of edge confidence since interactions with several sources of evidence are considered more likely to be true positives than those with only one source of evidence [95, 133, 313, 328, 332, 508, 674, 679, 681–683]. Similarly, interactions detected by two reciprocal bait-prey interactions in an Y2H screen are considered to be of higher confidence than those detected in one bait-prey direction alone [123].

The concept of levels of edge evidence can be extended to consider just the intersection of datasets, thereby discarding interactions with single lines of evidence, and producing a high confidence dataset [184, 489, 508, 684]. However, each functional dataset has its own error rate [45, 313]. Consequently, more sophisticated integration techniques are required to harness these differences in data quality.

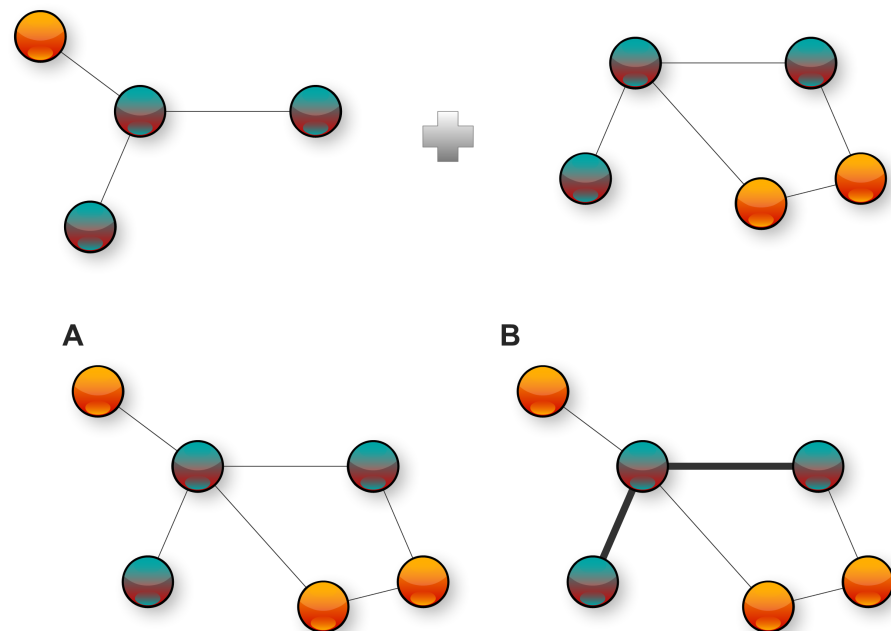


Figure 2.17: Simple network integration.

Datasets may be combined in several different ways. Here two small datasets have three nodes in common (blue circles). **A.** The datasets can be combined naïvely into a union of the available data. Therefore the edges (black lines) have equal weighting. **B.** Edge weights can correspond to the number of lines of evidence for an interaction. Therefore edges in the union of the two datasets are up-weighted (as denoted by thicker lines).

2.4.1 Dataset Noise

Biological data, in particular [HTP](#) data, is incomplete and noisy [[68](#), [69](#), [430](#), [673](#)]. False results can be either *false positives*, interactions that are identified but that do not occur in the cell, or *false negatives*, true interactions that have not been detected. False results arise from two main sources; technical or biological. Technical false results are directly attributable to the experimental method used. If the method has low sensitivity many interactions may be missed leading to false negative results. For instance, in [TAP-MS](#) weak interactions are less likely to be found due to the multiple washing steps required (see Section [2.1.1.2](#)) [[184](#), [185](#)]. Further, some physical interactions are so transient they are difficult to detect.

Experimental methods may also be prone to false positive results. For example, the reporter genes used in the [Y2H](#) technique may be transcribed in the absence of an interaction. The use of multiple testing and multiple techniques to confirm interactions can reduce technical false results [[108](#), [134](#), [184](#)].

False results may also be caused by a number of biological reasons. Many experimental methods are not carried out in natural cellular conditions [[685](#)]. In [TAP-MS](#) experiments proteins, which would not naturally be found in the same time or place within the cell, may interact leading to false positive results.

Protein folding and modification may also affect protein binding. In the [Y2H](#) system binding of the proteins to the Gal4 domains can change some proteins' folding, and alter post-translation changes such as phosphorylations. These changes may lead to false positive interactions between some proteins and to false negative interactions between others.

A surprisingly large amount of functional data are thought to be spurious [[154](#), [235](#)]. Several methods have been used to estimate the level of these false results in *S. cerevisiae* data with widely varying results (Table [2.9](#)) [[686](#)]. Many of these methods utilise high-confidence data to estimate the noise in [HTP](#) datasets. For instance, since [LTP](#) data are considered to be high-quality, comparison with [HTP](#) datasets can reveal noise levels [[121](#)]. Co-expression of genes has been shown to correlate with true positive interactions [[687](#)] and therefore can be used to identify potential false results [[688](#)].

Other data types, such as cellular localisation and other annotation data, can also reveal false results since proteins involved in distinct processes or found in separate cellular locations are not very likely to be functionally related [[313](#)]. Additionally, cross-species comparisons can detect noise in data [[120](#), [689](#)].

Table 2.9: Error rate estimates for *Saccharomyces cerevisiae* experimental data.

A summary of the estimated error rates for HTP technologies in *Saccharomyces cerevisiae* and the analytical methods from which they are derived.

Author	Method	False Positive Rate	False Negative Rate	Ref.
Huang <i>et al.</i> 2007	Extension of capture-recapture theory to estimate parameters prior to calculation using yeast two hybrid data.	25%	15%	[122]
Lin <i>et al.</i> 2008	Bayesian network-based estimation using on multiple-species data.	50%	-	[120]
Mrowka <i>et al.</i> 2001	Comparison of low-throughput and high-throughput datasets.	47-91%		[121]
Hart <i>et al.</i> 2006	Comparison of yeast two hybrid and affinity purification mass spectroscopy data.	46-89%		[154]
Deane <i>et al.</i> 2002	Comparison high throughput data with expression profiles and paralogs.	30-50%	-	[688]
Mering <i>et al.</i> 2002	Comparison of high throughput data with MIPs and YPD benchmark datasets.	50%	-	[184]
Sprinzak <i>et al.</i> 2003	Comparison of yeast two hybrid data with co-localisation and annotation data.	50%	-	[313]
Tong <i>et al.</i> 2004	Comparison of reciprocal bait-prey interactions for false negative detection in genetic data.	-	17-41%	[123]
Edwards <i>et al.</i> 2002	Estimation of false negative rates in yeast two hybrid and affinity capture datasets by comparison. with structural data	-	15-96%	[124]

While estimates of noise in functional data vary, it is clear that a significant number of detected interactions are false and the level of noise in HTP data is non-negligible [184]. Therefore, there is a clear need to distinguish the true interactions from false results. Post-processing of the data can help to reduce false results. For instance, highly promiscuous proteins, termed *sticky* proteins, are thought to participate in many false interactions which do not occur *in vivo*. Promiscuous bait proteins tend to have more reliable interactions than promiscuous prey proteins [157]. Therefore, these sticky proteins can be filtered from the data to improve accuracy [690–692]. However, this method risks the removal of hub proteins and therefore loss of true positive interactions.

The topology of a network can also be used to identify false results. One topology-based method is IRAP (interaction reliability by alternative path) which attempts to remove false positives and restore false negatives (add true positives) based on topological metrics [126, 127]. However this method risks the removal of true positives and the addition of false positives to the network. Alternatively, network models (see Section 2.3.4) can be used to filter the data since true interactions should fit the network model well [693].

Finally, the principles of false result estimation can be used to remove noise from data by comparing data types. Interacting proteins are thought to share similar annotations. Therefore, the similarity of gene annotations, such as GO terms or KEGG Pathways, can be used to identify and filter out potential false positives in the data [694, 695]. Similarly, genome context has been successfully used to calculate the confidence of individual interactions [696].

2.4.2 Gold Standard Data

Methods that attempt to filter out false results risk the loss of true positives and therefore could introduce false negatives to the network. Potentially, a more biologically accurate network can be constructed by taking the quality of each dataset into account, without the loss of data [697]. Since datasets differ in their reliability a significant challenge when integrating diverse datasets is estimating the relative importance and quality of each dataset in a consistent manner [49, 154, 698, 699].

The most commonly used method for calculating dataset quality is scoring against a Gold Standard [43, 59]: a reference network containing a set of interactions believed, with high confidence, to be biologically correct [97, 669]. The Gold Standard represents a benchmark against which to calculate a numerical estimate of dataset confidence and, thereby, allow the consistent integration of interactions from differing experimental types [669, 699]. Additionally, Gold Standards can be used in assessment and evaluation of analysis outcomes [700]. In some cases a second, negative set of interactions that are believed not to occur in the organism is included in a Gold Standard [701].

Gold Standard datasets are also commonly used in evaluation of a final, integrated system. Often the Gold Standard data are split into testing and training sets [112, 683]. Alternatively, a second positive Gold Standard from a different data source can be used in evaluation [49, 128]. In both cases the integrated system is evaluated based on its ability to predict data from the evaluation Gold Standard data, sometimes by *leave-one-out* analysis (see Section 2.5.5).

The quality of a Gold Standard is vital to the accuracy of conclusions drawn from integrated network analysis [699]. Reference data are therefore commonly obtained from expert-curated databases such as the KEGG database [99], MIPS [279], GO [100] or HPRD [111, 389]. The Gold Standard benchmarks typically represent a set of biologically meaningful interactions of a single type, such as:

- Shared metabolic pathway [49, 112, 128, 702].
- Shared biological process [46, 98, 106, 703].
- Shared molecular function [104, 107, 704].
- Shared complex membership [109, 110, 115, 133].

Additionally, small-scale LTP interactions which are considered high-quality can be used as Gold Standards, for instance the High Confidence (HC) dataset available through BioGRID¹⁹ [235].

The creation of a negative Gold Standard can be problematic since it is hard to determine, with high confidence, that an interaction does not occur *in vivo* under any circumstances [705]. A simple method for negative Gold Standard generation is to include any interactions between the genes of the Gold Standard that do not occur in the positive set [49, 60, 98, 106, 107, 112, 114, 115, 128, 706].

Negative Gold Standard datasets can also be based on cellular location, since proteins located in separate areas of the cell are unlikely to interact [103, 109–111, 223, 683]. Alternatively a random set of interactions can be generated to provide a negative Gold Standard [389]. Commonly the randomised set of interactions is generated and then filtered, either to remove interactions of the positive Gold Standard or based on cellular location.

2.4.3 Probabilistic Functional Integrated Networks

One of the most powerful approaches to reduce the extent of dataset noise during network integration is the use of Probabilistic Functional Integrated Networks (PFINs) [49, 115, 128]. In PFINs nodes correspond to genes or gene products and the edges to functional associations between nodes. A

¹⁹<http://thebiogrid.org/downloads/archives/Published%20Datasets/HC-BIOGRID-2.0.31.tab>

PFIN, sometimes referred to as a functional linkage network (FLN), has edge weights which indicate the level of confidence in the combined evidence for that edge. The edges are produced by statistical comparison against a Gold Standard dataset [669, 686]. There are two major approaches to scoring using a Gold Standard. In the first a statistical algorithm may be used to compare each dataset to the Gold Standard prior to integration of the datasets in an effort to estimate the quality of each dataset [49, 102, 342]. The dataset weights can be integrated in several ways. Simple naïve integration involves summing each of the dataset weights to produce a network where an edge weight is the combined sum of all the evidence for that edge (Figure 2.18).

However, this method of integration assumes that there are no dependencies between the available datasets, an assumption which is unlikely given the nature of biological data. Lee and colleagues (2004) attempted to overcome this difficulty by introducing a weighted sum during integration, to successively down-weight evidence scores in order of magnitude:

$$WS = \sum_{i=1}^n \frac{L_i}{D^{(i-1)}} \quad (2.17)$$

where L_1 is the highest weighted line of evidence for the edge and L_n is the lowest weighted line of evidence in a set of n datasets [49]. Division of the score by the D parameter means that, while the highest score is integrated unchanged, subsequent weights are progressively down-weighted. Therefore, a D value of 1.0 would produce a simple summed network as in the example in Figure 2.18, and higher values successively down-weight the confidence scores (Figure 2.19). The Lee method of integration has been used to analyse network data from *S. cerevisiae* [49, 98, 128, 706], *C. elegans* [703, 707], *D. melanogaster* [708] and *A. thaliana* [342].

The second use of Gold Standard data is in machine learning [103, 105, 133, 388]. Machine learning algorithms aim to learn the characteristics of a training dataset in order to predict those of data of the same type. In the case of PFINs Gold Standard data are used to train a machine learning classifier, for instance a support vector machine (SVM) [103], random forest (RF) [388], Markov random field (MRF) [105] or Bayesian network inference [709], to recognise true interactions in the datasets. Once trained the classifier can be used to calculate the probabilities of true interaction between proteins from diverse datasets. A PFIN is produced with edge weights indicating the probability of true interaction.

However, a major drawback to machine learning algorithms is that the training Gold Standard may not cover all characteristics of the real data, leading to uneven training. Therefore, a classifier may be produced which only correctly assigns data that is very similar to the training Gold Standard.

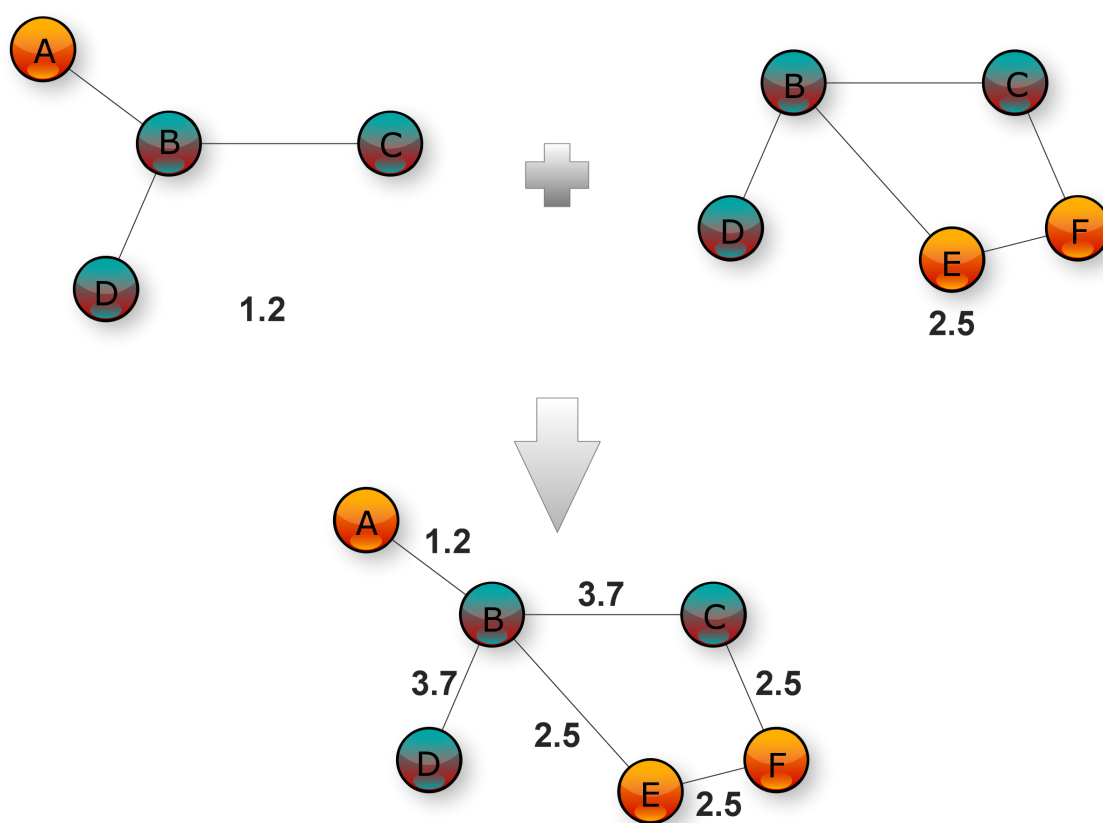


Figure 2.18: Probabilistic functional integrated networks.

PFINs have edge weights representing the confidence in the lines of evidence for each edge. In this example two datasets of confidence 1.2 and 2.5 respectively are integrated by summing the scores over each edge. Therefore, edges with multiple lines of evidence are up-weighted while the datasets' confidence values are also taken into account. Therefore, edges which only occur in the first dataset are weighted 1.2, those only occurring in the second dataset are weighted 2.5, and those edges occurring in both datasets are weighted 3.7. Simple summing of dataset weights in this way is the most naïve method of edge weight calculation.

Additionally, since there is a high level of noise in biological data, overfitting may occur, producing a classifier which models the characteristics of both the true data and the noise. These drawbacks may be addressed by use of the training set expansion method of Yip and colleagues (2009) which uses semi-supervised learning to improve Gold Standard coverage [710] or by the use of multiple Gold Standard sources [101].

In both uses of Gold Standard data discussed above the final network of interactions is annotated with edge weights corresponding to the confidence in that interaction being correct. PFINs have been created for yeast (Table 2.10) and a number of other species (Table 2.11) using a variety of methods and Gold Standards. These networks can then be used to detect protein complexes [90, 534, 592], annotate proteins [92, 674, 711] and predict new interactions (see Sections 2.5.5) [318, 675].

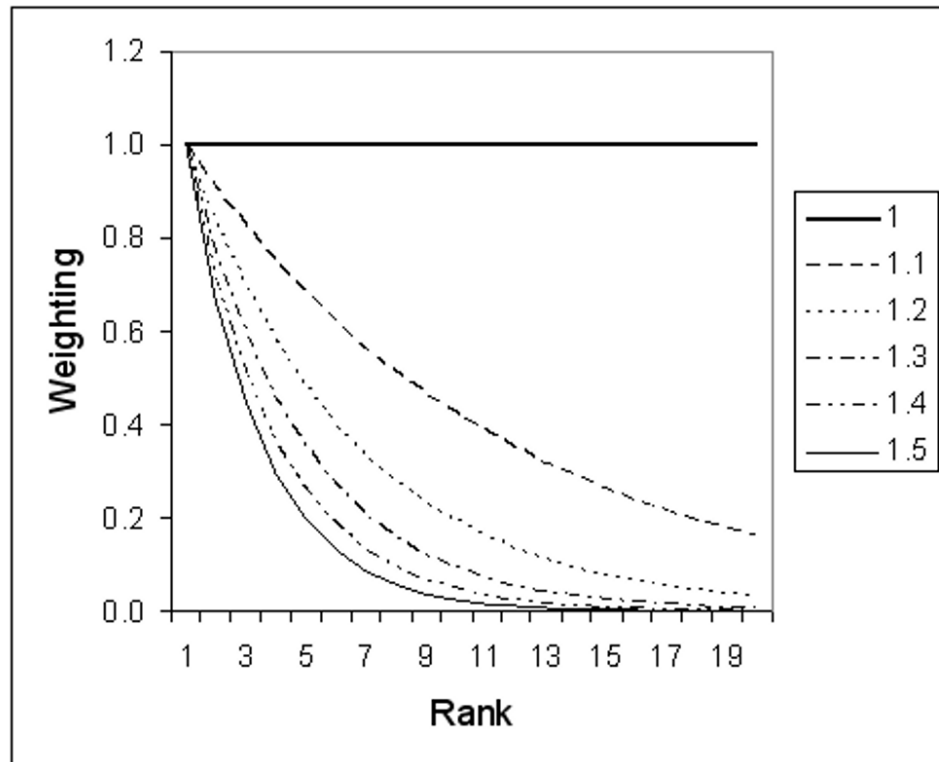


Figure 2.19: The D-value effect.

A D value of one results in a simple sum of the dataset scores. Higher values successively down-weight confidence values.

2.4.4 Dataset Bias

In addition to varying levels of noise, functional datasets have been shown to have significant levels of bias [42, 128]. Bias occurs for a number of reasons:

- Experimental method
- Experimental design
- Publication choice
- Level of interest
- Cellular bias

The experimental method chosen can be a source of bias. Experimental methods each have their own strengths and weaknesses in the type of interaction they can detect [98]. For instance, TAP-MS detects strong stable interactions but is poor at detecting weak transient interactions due to the multiple washing stages of the technique.

Table 2.10: *Saccharomyces cerevisiae* PFINs.

A summary of the PFINs created for the yeast *S. cerevisiae* including the method of integration and Gold Standard choices. Where the positive and negative standards are the same the negative standard is derived from those interactions not present in the positive set. Additionally, where the evaluation standard is the same as the integration standard, cross-fold validation of the integration standard has been used in evaluation.

Key: Co-immunoprecipitation (Co-IP), Gene Ontology (GO), Biological Process (BP), Munich Information Center for Protein Sequences (MIPS), Clusters of Orthologous Groups (COGS), Co-immunoprecipitation (Co-IP), Kyoto Encyclopedia of Genes and Genomes (KEGG), Experimental (EXP), Low-throughput protein-protein interactions (LTP-PPI).

Author	Method	Gold Standard Positive	Gold Standard Negative	Evaluation	Ref.
Bader <i>et al.</i> 2004	Logistic regression	Co-IP	Co-IP	GO	[712]
Lankriet <i>et al.</i> 2004a	Data fusion	MIPS	MIPS	MIPS	[713]
McGary <i>et al.</i> 2007	Probabilistic scoring and integration	GOBP	GOBP	PHENOTYPIC	[706]
Lee <i>et al.</i> 2007	Probabilistic scoring and integration	GOBP	GOBP	MIPS & COG	[98]
Jansen <i>et al.</i> 2002	Data fusion	MIPS	LOCALISATION	MIPS	[683]
Jansen <i>et al.</i> 2003	Bayesian network inference	MIPS	LOCALISATION	MIPS	[109]
James <i>et al.</i> 2009	Probabilistic scoring and integration	KEGG	KEGG	GOBP	[128]
Qi <i>et al.</i> 2005	Random forest	DIP	RANDOM	DIP	[714]
Lee <i>et al.</i> 2004	Probabilistic scoring and fusion	KEGG	KEGG	GO & COG	[49]
Chen <i>et al.</i> 2004	Bayesian network inference	GOBP	GOBP	GOBP	[114]
Troyanskaya <i>et al.</i> 2003	Bayesian network inference	GOBP	GOBP	GOBP	[60]
Antonov <i>et al.</i> 2006	Data fusion	MIPS	MIPS	MIPS	[107]
Kiemer <i>et al.</i> 2007	Probabilistic scoring and integration	LTP PPI & STRUCTURAL	LTP PPI & STRUCTURAL	GO & EXP	[115]
Myers <i>et al.</i> 2005	Bayesian network inference	GOBP	GOBP	GOBP	[106]
Lu <i>et al.</i> 2005	Bayesian classifier	MIPS	LOCALISATION	MIPS	[110]
Yamanishi <i>et al.</i> 2004	Supervised network inference	KEGG	KEGG	KEGG	[112]
Patil & Nakamura 2005	Bayesian network inference	Multiple	Multiple	Multiple	[681]
Jiang & Keating	Probabilistic decision tree	GO	GO	GO	[715]
Ray <i>et al.</i> 2009	Data fusion	GO	GO	GO	[716]
Franzosa <i>et al.</i> 2009	Multiple machine learning	MIPS	CC	MIPS	[133]
Chua <i>et al.</i> 2007	Integrated weighted averaging	GO	GO	GO	[717]
Lanckriet <i>et al.</i> 2004	Support vector machine	MIPS	MIPS	MIPS	[668]

Table 2.11: PFINs for other species.

A summary of the PFINs created for the other species including the method of integration and Gold Standard choices. Where the positive and negative standards are the same the negative standard is derived from those interactions not present in the positive set. Additionally, where the evaluation standard is the same as the integration standard, cross-fold validation of the integration standard has been used in evaluation.

Key: Co-immunoprecipitation (Co-IP), Gene Ontology (GO), Biological Process (BP), Molecular Function(MF), Cellular Component (CC), Munich Information Center for Protein Sequences (MIPS), Co-immunoprecipitation (Co-IP), Kyoto Encyclopedia of Genes and Genomes (KEGG), Experimental (EXP), Human Protein Reference Database (HPRD), Online Mendelian Inheritance in Man (OMIM), Pathway Interaction Database (PID), Eukaryotic Orthologous Groups (KOG), low-throughput (LTP).

Author	Species	Method	Gold Standard Positive	Gold Standard Negative	Evaluation	Ref.
Lee <i>et al.</i> 2008a	<i>Caenorhabditis elegans</i>	Probabilistic scoring and integration	GOBP & KEGG	GOBP	KEGG & GO	[703]
Lee <i>et al.</i> 2010b	<i>Caenorhabditis elegans</i>	Probabilistic scoring and integration	GOBP	GOBP	PHENOTYPIC	[707]
Zhong <i>et al.</i> 2006	<i>Caenorhabditis elegans</i>	Logistic regression	PPI & GI	GI	GI	[718]
Patil & Nakamura 2005	<i>Caenorhabditis elegans</i>	Bayesian network inference	Multiple	Multiple	Multiple	[681]
Li <i>et al.</i> 2006	<i>Arabidopsis thaliana</i>	Bayesian network inference	KOG	KOG	AraCyc	[102]
Lee <i>et al.</i> 2010	<i>Arabidopsis thaliana</i>	Probabilistic scoring and integration	GOBP	GOBP	GOCC & KEGG	[342]
Yellaboina <i>et al.</i> 2007	<i>Escherichia coli</i>	Support vector machine	EcoCyc	CC	EcoCyc	[103]
Date <i>et al.</i> 2006	<i>Plasmodium falciparum</i>	Bayesian Network Inference	GO & KEGG	GO & KEGG	GO & KEGG	[719]
Patil & Nakamura 2005	<i>Drosophila melanogaster</i>	Bayesian network inference	Multiple	Multiple	Multiple	[681]
Costello <i>et al.</i> 2009	<i>Drosophila melanogaster</i>	Probabilistic scoring and integration	GO	GO	GO	[708]
Kim <i>et al.</i> 2008	<i>Mus musculus</i>	Naïve Bayesian classifier	GO	GO	GO	[116]
Guan <i>et al.</i> 2008	<i>Mus musculus</i>	Bayesian network inference	GO	GO	GO	[117]
Franke <i>et al.</i> 2006	<i>Homo sapiens</i>	Bayesian classifier	GOBP & GOMF	GOCC	OMIM	[223]
Huttenhower <i>et al.</i> 2009	<i>Homo sapiens</i>	Bayesian classifier	Multiple	GO, KEGG & PID	Various	[720]
Rhodes <i>et al.</i> 2005	<i>Homo sapiens</i>	Probabilistic scoring and fusion	HPRD	GOCC	HPRD	[111]
Scott <i>et al.</i> 2007	<i>Homo sapiens</i>	Bayesian classifier	HPRD	RANDOM & CC	HPRD	[389]
Patil & Nakamura 2005	<i>Homo sapiens</i>	Bayesian network inference	Multiple	Multiple	Multiple	[681]
Zhong & Sternberg 2006	Multiple eukaryotes	Logistic regression	LTP	GI	PPI	[718]

Conversely, Y2H can detect weak interactions but is poor at detecting interactions involving post-translationally-modified proteins [721]. In addition, some types of proteins are under-represented in functional datasets. For example, membrane proteins are insoluble due to their hydrophobic tail regions and are, therefore, not suited to many experimental methodologies.

Bias also occurs due to experimental design [46]. Individual research groups have their own specific interests. Therefore, each group will naturally design their experiments based on particular areas of cellular biology, for example, by varying the choice of bait proteins in a Y2H screen. These choices bias the final dataset towards the experimental focus.

Additionally, bias can be introduced into functional datasets by the choice of data for publication [121]. Interactions detected in an HTP experiment may not be included in a published dataset if they are not related to the major conclusions of the publication. Alternatively, since many journals issues have a specific focus, only data related to that area may be chosen for publication.

Some areas of cellular biology are highly studied and are, therefore, over-represented in functional data [98, 223]. For example, proteins which are associated with distinct phenotypes or diseases are highly studied in many species [133]. Finally, there are natural cellular biases which lead to dataset bias. In particular, the large number of ribosomal proteins of the cell tend to be over-represented in functional data [125, 722].

In addition to bias in functional data, manually-curated Gold Standard data are thought to be biased, particularly towards well-studied proteins and processes, and many of the databases chosen as Gold Standard sources have a biased focus [133]. For instance KEGG focuses on metabolic pathways, while MIPS focuses on physical protein-protein interactions. Consequently, a KEGG Gold Standard would bias an integrated network towards metabolic pathways, while a MIPS Gold Standard would bias the network towards PPIs [722]. The effect of Gold Standard bias may be addressed using a bespoke hand-curated Gold Standard [125] or using multiple Gold Standard sources [101].

Bias in functional data may affect analyses of the integrated data and possibly mask areas of cellular biology which may be of interest. Therefore, several studies have addressed this issue by attempting to identify and remove bias from the data. For instance, Lee and co-workers discarded the GO term protein biosynthesis from their Gold Standard to minimise training bias since it represented almost a third of the Gold Standard protein pairs [98]. By removing these terms from the Gold Standard, datasets biased to this process would be down-weighted during confidence scoring. The same group used a similar method to discard GO terms from their Gold Standards for *A. thaliana* and *C. elegans* [342, 703, 707]. However, removal of data has the disadvantage of introducing false negative interactions into the network [128].

2.5 Beyond the Interactome

Interactions between proteins form the structural basis of cellular biology. However, the verification of all the interactions comprising the interactome of a species is only part of the task of systems biology. For a complete understanding of cellular biology we must also know which cellular processes the interacting proteins are involved in, what functions the proteins perform, where the proteins act within the cell and under what circumstances [31, 495, 723, 724].

2.5.1 Protein Function

Protein functional prediction and interactome mapping are fundamentally linked and complement one another. In one direction, knowledge of shared role can be used to infer a link between protein pairs and aid in the prediction of interactions. In the other, knowledge of a protein's interactions can be used to assign a role, since proteins involved in the same process tend to interact (Figure 2.20) [725].

Methods to assign function have evolved rapidly in the post-genomic era, and what was once a time consuming task, undertaken on a protein-by-protein basis, can now be carried out on a genome-wide scale using various HTP experimental and *in silico* techniques. The annotation schemas used to record protein function have also changed over time to include new evidence codes reflecting the wide range of prediction methods available to assign function. Nevertheless, a large number of proteins remain uncharacterised [674]. Several computational techniques have been developed to predict protein function (for instance [726–728]), many of which rely on the similar principles to those used in computational PPI prediction (see Section 2.2.2).

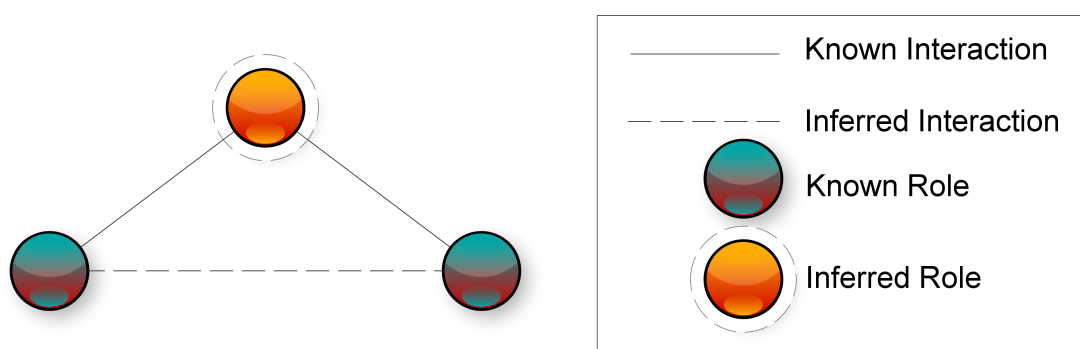


Figure 2.20: Protein Role and PPI Prediction.

A protein's interactions are fundamentally linked to its cellular role. Therefore, interactions can be inferred from cellular role and, conversely, role may be inferred from interactions. In this example the orange node is predicted to be involved in the same role as its two neighbouring blue nodes and, and conversely the blue nodes are predicted to have an interaction.

Early functional prediction relied on sequence similarity and sequence features. Algorithms such as FASTA [729], BLAST [300] and PSI-BLAST [730] were developed to compare the nucleotide sequences of genes, and the amino acid sequences of proteins, with those of genes and proteins of known function. By measuring pairwise alignment between two sequences, similarity scores are produced for a pair of genes or proteins and, where similarity is high, shared function may be inferred [731]. Sequence similarity can indicate divergence from a shared common ancestor, termed homology. However, similar sequences are not always homologous and, while high sequence similarity can indicate similar function, this is not always the case since some homologous proteins have diverged to perform distinct functions [732, 733]. Conversely, other proteins have converged to share similar function, but do not share a common ancestor, and are therefore not homologous. Consequently, while sequence similarity can be indicative of shared function, there are limits to the accuracy of these methods [734–738].

Areas of high sequence similarity between proteins may be restricted to small local sequence features, rather than the global gene or protein sequence [739]. These features usually correspond to the active sites of proteins. Mutations in the active site are far more likely to inactivate a protein and, therefore, these sites are highly conserved between species [740]. Due to this conservation the multiple alignment of protein or gene sequences can reveal areas of sequence that are of importance to protein function [741, 742]. Consequently, multiple sequence alignment of proteins can be utilised for functional prediction [743].

The availability of more extensive sequence data in recent years has allowed the development of several functional prediction techniques which utilise genomic context [113, 744]. For example, in prokaryotes protein function can be predicted based on the conserved chromosomal proximity of genes in multiple species using the gene neighbourhood method [329, 331, 745]. In addition, groups of prokaryotic proteins with highly conserved proximity, termed operons, often interact physically and have related functions [746].

Similarly, the Rosetta Stone method, which identifies PPIs based on patterns of domain fusions, may also be used to infer function [332–334]. This method identifies proteins in which domains are found as a single polypeptide in one species but are found as separate proteins in other species. Additionally, the phylogenetic profiles of proteins can also be utilised to predict function since the correlated inheritance of protein pairs can indicate conservation of functional association due to shared evolutionary pressure [94, 354, 355, 359, 747]. Finally, the conservation of interacting protein pairs, termed interologs, or regulatory interactions, termed regulogs, can be predictive of protein function [346, 351, 748]. The combination of several genome context methods can improve the

Table 2.12: Protein structural alignment.

A summary of several protein structural alignment tools. Structural alignment can be used to infer protein function since proteins with similar 3D structures tend to have similar cellular roles.

Name	Method	Availability	Ref.
MATT	Aligned Fragment Pair Chaining	http://groups.csail.mit.edu/cb/matt	[760]
SANA	Core Alignment and Optimisation	http://zhangroup.aporc.org/bioinfo/SANA	[754]
RSE	Refinement with Seed Extension	http://lmbbi.nci.nih.gov/	[761]
MAMMOTH	Sequence Independent Heuristic	http://physbio.mssm.edu/~ortizg/	[762]
MAPSCI	Coordinate-based alignment	http://www.geom-comp.umn.edu/mapsci	[763]
CE	Combinational Extension	http://cl.sdsc.edu/	[764]
FATCAT	Alignment of Fragment Pairs	http://fatcat.burnham.org	[765]
SAlign	Dynamic Programming	http://modbase.compbio.ucsf.edu/salign-cgi/index.cgi	[766]
TALI	Torsion Angle Alignment	http://redcat.cse.sc.edu/index.php/Project:TALI/	[767]

accuracy of prediction [332, 508]. The Prolinks²⁰ database provides genome context predictions for a number of organisms and includes an interactive navigator for the predicted functional linkages [330].

Protein structure can also be used to predict protein function since the 3D structure of a polypeptide is fundamentally linked its role [749]. Proteins with similar structures are likely to have a similar cellular roles and, therefore, high structural similarity may be indicative of shared function [750–753]. Protein 3D structures can be aligned in a similar fashion to protein sequences. Several structural alignment and functional prediction algorithms have been developed, many of which are freely available either online or as stand-alone packages (Table 2.12). In most cases these algorithms identify and align the core of the protein, often the active site, before aligning the remaining residues [754]. Additionally, a combination of multiple structure-based methods can improve prediction accuracy [755].

Several benchmark databases have been developed to store protein structures and alignments, for instance the Homologous Structure Alignment Database²¹ (HOMSTRAD) [756], S4²² [757] and the Protein Data Bank²³ (PDB) [758]. In addition, a standardised classification system, the Structural Classification of Proteins (SCOP), has been developed to describe protein structures [759].

Following the availability of whole genome sequences it became possible to measure gene expression on a genome-wide scale. Gene expression data are useful in inferring protein function since the proteins involved in the same cellular process are likely to be expressed at the same time. Several methods have been developed to cluster gene expression data and identify coexpressed protein pairs (see Section 2.1.2.1). Conserved coexpression between species may also be used to infer protein

²⁰<http://mysql5.mbi.ucla.edu/cgi-bin/functionator/pronav>

²¹<http://www-cryst.bioc.cam.ac.uk/~homstrad/>

²²<http://compbio.mds.qmw.ac.uk/~james/S4.shtml>

²³<http://www.rcsb.org/pdb/>

function [373, 374]. However, coexpression-based techniques have the same limitations as they do for PPI prediction due to noise and interpretation.

The wealth of available biological literature is also a source of functional data. Several methods have been developed for the automatic annotation of proteins by literature mining [259, 768–775]. Finally, machine learning classifiers may also be used to assign probabilities to potential functional annotations following training based on various properties of genes with known annotations [114, 388, 719, 776–778]. These properties range from sequence-based features to functional interaction data. The use of multiple evidence types produces improved classification over a single data types alone [779, 780].

2.5.2 Cellular Location

While knowledge of a protein's interactions and the processes it is involved in is informative, a large part of our understanding of the interactome also involves knowing a protein's cellular location; where it interacts and performs its function. For instance, a protein may be known to be involved in transmembrane transport, however, there are several membranes in the cell and different molecules are transported across different membranes [781]. Therefore, knowledge of cellular location can enhance the understanding of a protein's cellular role.

As with a protein's function, its cellular location can be indicative of its potential interaction partners, since interacting proteins must be located together. Conversely, protein location data can be used to detect interactions unlikely to occur and, therefore, identify false positives in HTP data. In particular, cellular location data are commonly used to filter positive Gold Standard data and to create negative Gold Standards (see Section 2.4.2). The cellular location of a protein can also indicate its potential function. For instance, a protein located at the telomere is very likely to be involved in telomeric processes such as telomere capping or telomere maintenance.

Cellular location prediction, both experimentally and computationally, is non-trivial. Experimentally, cellular location has been predicted using fluorescence tagged proteins [3], for instance by fusing a target protein to green fluorescent protein (GFP) [782]. A protein's location can then be identified by microscopic visualisation of the cell. Proteins of the major structures of the cell, such as the membranes and ribosomes, can be easily visualised in this way. However, the cell is a moving system of parts. The majority of cellular proteins are in free fluid and can, therefore, be potentially found anywhere in the cell, making determination of a protein's functional location difficult.

Due to the difficulties of experimental prediction, computational cellular compartment prediction has become important [783, 784]. Like the prediction of protein function, many of the computational

methods are sequence based. Protein sequences contain sorting motifs that direct the protein to its correct location [781, 785, 786]. Therefore shared signals can be used to identify location [787–789]. However, there are several drawbacks to sequence-based methods. Firstly, proteins which act in the cytoplasm, the compartment where protein production occurs, do not have sorting signals [727]. Additionally, some proteins are transported as a complex in which only one protein may contain a sorting signal [790]. These proteins may be mistaken for cytoplasmic proteins due to their lack of signal. Further, a sorting signal may involve several areas of a protein’s sequence making its identification difficult [791].

A protein’s composition can also be used to predict its location since it has been observed that the amino acid composition of a protein can correspond to its cellular compartment [792–795]. Additionally a protein’s cellular location may also be inferred using its phylogenetic profile using the same method with which function may be inferred (see Section 2.5.1) [796].

Several classifier-based methods have been applied to the identification of cellular location including neural networks [797, 798], SVMs [799–805], k-nearest neighbour [806, 807], semi-supervised learning [808], Bayesian classifiers [809] and ensemble classifiers [810, 811]. Additionally, mining of the literature for functional keywords may also be used to infer protein location [812, 813]. For instance, a protein associated with the keywords *mitochondrial chromosome segregation* could naturally be annotated to the cellular compartment *mitochondrion*.

2.5.3 Human Disease

In multicellular organisms, particularly those with multiple organs such as humans, genes and proteins may be associated with disease states and distinct phenotypes. Identification of the genes associated with specific phenotypic states, termed candidate genes, is important and challenging [814, 815]. Consequently, disease related genes and their orthologs tend to be highly studied and are over-represented in functional data (see Section 2.4.4) [223].

Candidate genes can be identified using similar methods to those used in function prediction, for instance sequence-based methods such as domain fusion patterns or homology [816–821]. Additionally, conservation of expression across several species or similarity of phylogenetic profiles can be used to prioritise candidate genes [822–826].

2.5.4 Annotation Data

Due to the complexity of cellular biology it is often difficult to consistently interpret the cellular role of a protein. Many cellular processes may be described using several descriptions. For instance, a

Table 2.13: Protein annotation schemas.

Annotation schemas use controlled vocabularies to describe various aspects of genes and their products including function, location and disease association.

Type	Name	Species	Ref
Biological Process	GO Biological Process	Multiple	[100]
	MIPS Functional Catalogue (FunCat)	Multiple	[829]
Molecular Function	GO Molecular Function	Multiple	[100]
	MIPS FunCat	Multiple	[829]
	Enzyme Commission (EC)	Multiple	[830]
Cellular Compartment	GO Cellular Component	Multiple	[100]
	MIPS FunCat	Multiple	[829]
Structural	SCOP	Multiple	[759]
Metabolic Pathway	KEGG PATHWAYS	Multiple	[99]
Orthology	KEGG Orthology	Multiple	[99]
Disease	Online Mendelian Inheritance in Man (OMIM)	<i>Homo sapiens</i>	[831]

process may be referred to as "transcription" by one study but described as "mRNA synthesis" or "RNA biosynthesis" by another. Consequently, it is difficult to integrate data from diverse sources in a consistent manner [827]. Therefore, there is a need for uniform descriptors of cellular biology to produce consistency across biological datasets [273].

To fulfil this need, several annotation schemas have been developed to describe cellular processes including protein functions, structures and locations, and the orthological and disease associations of genes and proteins (Table 2.13). Annotation schemas use controlled vocabularies, often ontologies, which have unique identifiers to consistently describe genes and their products. There are many ontology schemas, some of which are species or area specific²⁴ [828].

The following sections briefly describe the most comprehensive and widely used annotation schemas which are of particular relevance to the work presented in this thesis; the Enzyme Classification (EC), The Kyoto Encyclopedia of Genes and Genomes (KEGG) and the Gene Ontology (GO).

2.5.4.1 Enzyme Classification

The EC system was one of the first protein classification schemas designed to describe enzymatic functions. Each EC annotation is a unique identifier which consists of four numbers in the format n1.n2.n3.n4 [830]. The first number represents a high level of classification, with subsequent numbers representing increasingly precise descriptions of protein function.

²⁴see <http://www.obofoundry.org/>

For example the enzyme glutathione reductase, a cellular antioxidant (see Section 2.6.2.2), has the classification 1.8.1.7²⁵ which refers to:

1 Oxidoreductase

.8 Acting on a sulfur group of donors

.1 With NAD⁺ or NADP⁺ as acceptor

.7 Glutathione-disulphide reductase

This tiered system of classification allows proteins to be classified at higher levels of function when their specific function remains unknown. For instance, a protein may be classified as 1.8.-.- if it is known to be an oxidoreductase acting on a sulfur group of donors but where its specific action is unknown.

2.5.4.2 KEGG

The KEGG database was originally created to store metabolic information based on the EC schema [277]. The KEGG resource now consists of 16 databases describing several areas of biology including metabolism, human disease, compounds and orthology [832–838].

The main KEGG database is the KEGG PATHWAYS resource which stores pathway data using the EC annotation schema. A KEGG pathway consists of the enzymes, substrates and cofactors. Each pathway may be visualized online using graph theoretic representation (Figure 2.21). Pathways may also be downloaded in KEGG Markup Language (KGML) and tab-delimited format²⁶.

KEGG reference pathways are manually curated from the literature by mapping EC numbers to genes. KEGG then uses orthology to create species specific metabolic pathways from the reference pathways. KEGG pathways are often used as Gold Standard data due to their high level of curation [49, 112, 128, 703, 719].

In a KEGG Gold Standard protein pairs annotated to the same pathway form the positive Gold Standard, while those annotated to separate pathways form the negative Gold Standard. In addition to its use as a Gold Standard, several software tools have been developed to utilise KEGG data during computational analysis [839–847].

²⁵http://www.genome.jp/dbget-bin/www_bget?enzyme+1.8.1.7

²⁶<http://www.genome.jp/kegg/download/ftp.html>. It should be noted that the KEGG database will be available only to paid subscribers from 1st July 2011.

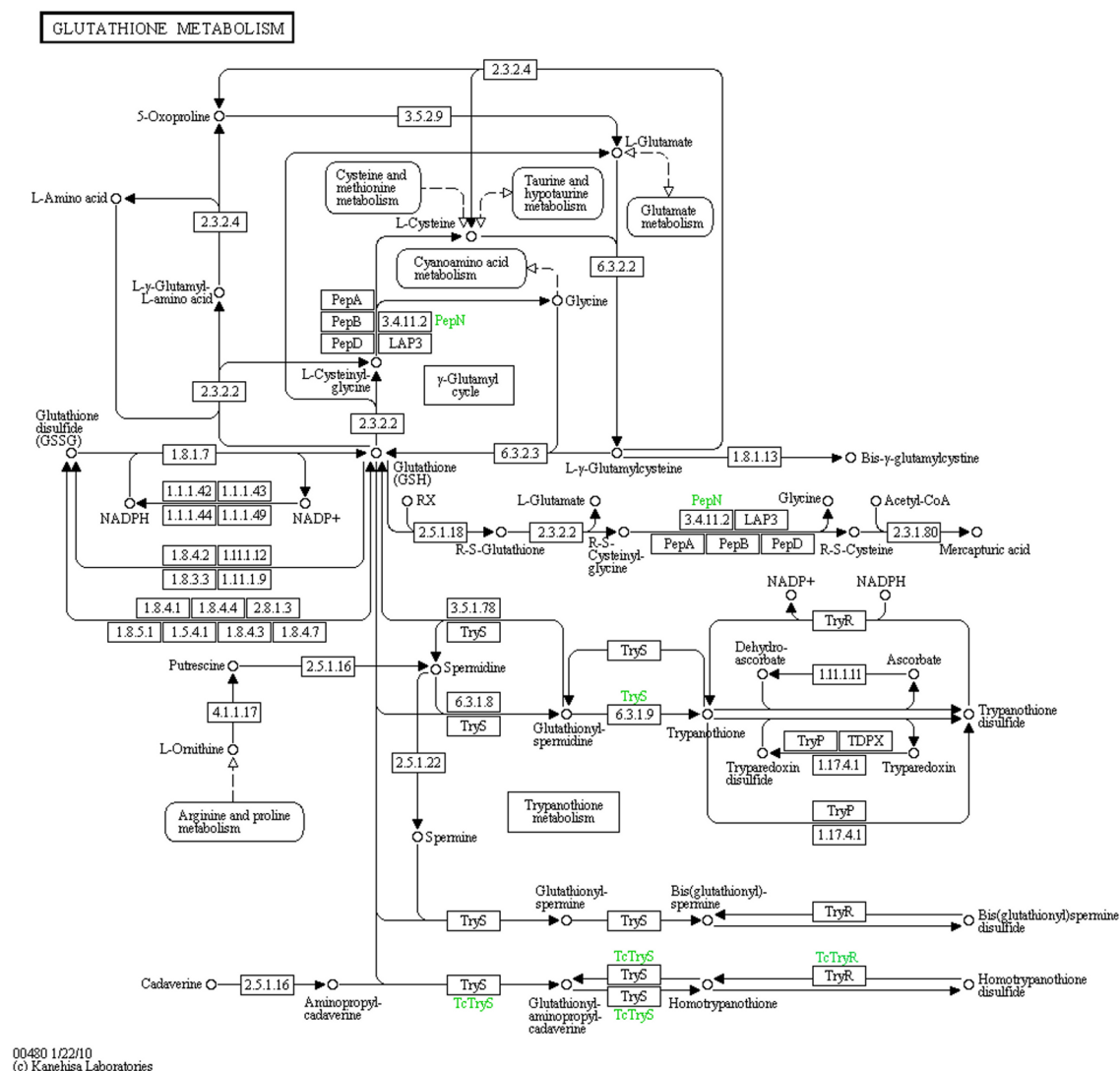


Figure 2.21: The KEGG pathway for glutathione metabolism.

KEGG Pathways are networks of enzymes, substrates and cofactors. Enzymes, such as TryS, are shown as rectangles while substrates and cofactors, such as glutathione and NADPH, are shown as small circles. Arrows representing the direction of reaction flux link the rectangles and circles. Rounded boxes represent links to and from other KEGG Pathways. All elements of the pathway link to detailed information through the KEGG databases. Alternate enzyme names based on orthologous groups are shown in green text.

2.5.4.3 Gene Ontology

The Gene Ontology (GO) is the most comprehensive and widely used hierarchical annotation schema [100, 848]; there are 3142 PubMed²⁷ hits for the phrase "Gene Ontology" in comparison with 722 for "KEGG" and just 20 for the similar hierarchical annotation schema "FunCat" [829].

GO is a controlled vocabulary, which describes the molecular function of proteins, the biological processes they are involved in and the cellular compartments in which they are found. Despite its

²⁷<http://www.ncbi.nlm.nih.gov/pubmed> (accessed 18th March 2011)

name GO is not a true ontology but three controlled vocabularies each in the form of a [DAG](#) with well-defined, structured terms describing three branches of biological knowledge²⁸:

- **molecular_function** (MF) - the protein's activity at the molecular level, for example telomeric DNA binding.
- **biological_process** (BP) - the process in which the protein functions, for example telomere maintenance.
- **cellular_compartment** (CC) - the protein's cellular location, for example telomeric region.

Each branch of GO is a collection of terms all of which have a unique identifier, name and description. For example the term identifier GO:0000723 refers to telomere maintenance and has the description²⁹:

"Any process that contributes to the maintenance of proper telomeric length and structure by affecting and monitoring the activity of telomeric proteins and the length of telomeric DNA. These processes includes those that shorten and lengthen the telomeric DNA sequences."

While the three branches of GO were originally intended to be treated separately, recently efforts have been made to connect the three areas of biology in order to harness more complex cellular relationships [849–854].

Terms in GO are connected to one another in a parent to child hierarchy with specific terms at the bottom and general terms at the top, below the root term. The relationships between the terms have direction. Therefore if **term a** is_a **term b** is_a **term c** then **term c** cannot be a **term a**. The GO [DAG](#) currently has five types of relationship:

- is_a - **term a** is a subtype of **term b**.
- part_of - **term a** is a subpart of **term b**.
- regulates³⁰ - **term a** regulates **term b**.
- positively_regulates³⁰ - **term a** positively regulates **term b**.
- negatively_regulates³⁰ - **term a** negatively regulates **term b**.

²⁸<http://www.geneontology.org/GO.doc.shtml#ontologies>

²⁹http://amigo.geneontology.org/cgi-bin/amigo/term_details?term=GO:0000723&session_id=589amigo1301231935

³⁰The regulates, positively_regulates and negatively_regulates relationships were added to the Gene Ontology structure in April 2008 and were, therefore, not present during in the earlier parts of this project.

Proteins may be annotated to multiple GO terms based on experimental evidence. Due to the hierarchical structure of the GO DAG, annotation has transitivity. Therefore, annotation to a child term automatically implies annotation to all parent terms of that term. For instance, in the example in Figure 2.22, a protein annotated to the lowest term, cellular response to reactive oxygen species (GO:0034614) is automatically annotated to all the terms in the example, since they are parent terms of the annotation. Species-specific annotation files can be downloaded from GO³¹ and from other species-specific databases such as SGD [855].

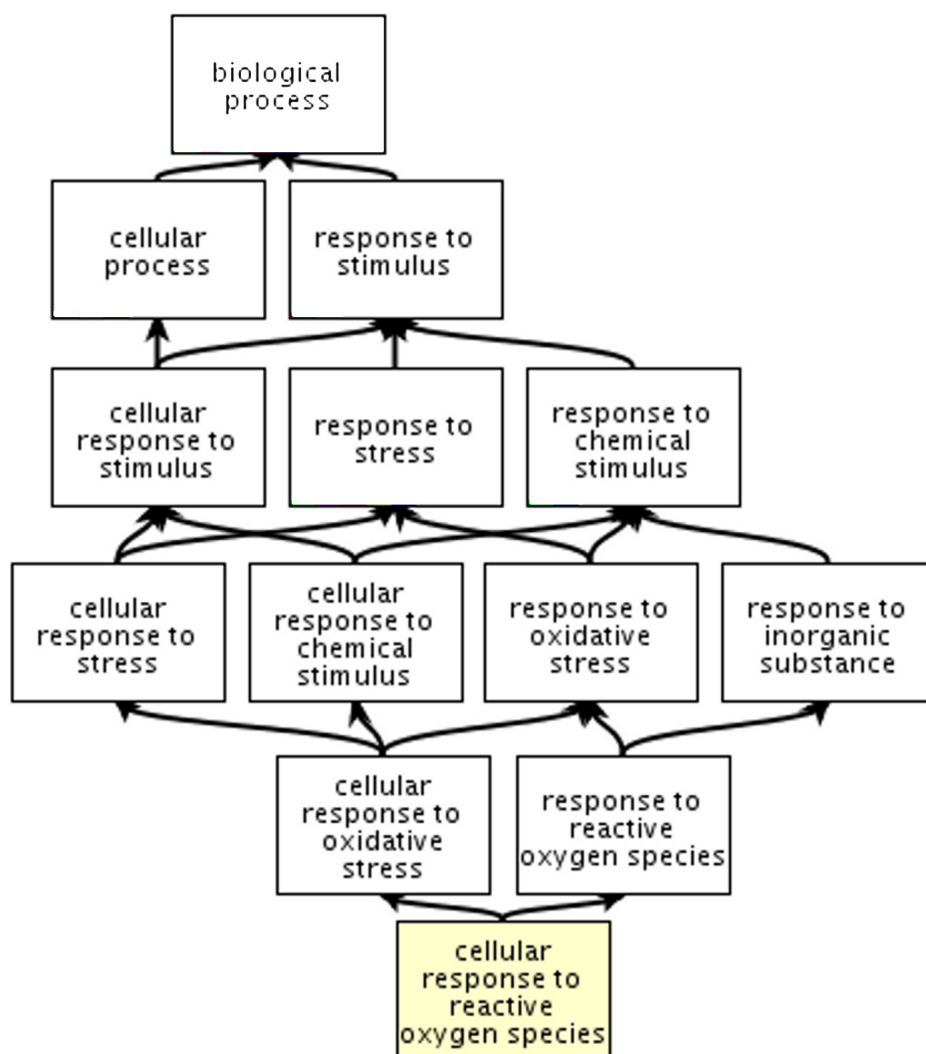


Figure 2.22: The Gene Ontology structure.

GO is structured as a hierarchical DAG with general terms at the top and increasingly specific terms lower down. Terms are connected in directional parent to child relationships. In this example, cellular response to stimulus is_a cellular process which in turn is_a biological process. Therefore, biological process cannot be a cellular response to stimulus or a cellular process. GO terms annotation also has transitivity. Therefore, any protein annotated to the term cellular response to reactive oxygen species (yellow box) is automatically annotated to all the parent terms displayed in this example).

³¹<http://www.geneontology.org/GO.downloads.annotations.shtml>

Extensive metadata describes each GO annotation including the source, date and the type of evidence for assignment of the term. The GO Consortium provides a range of experimental and computational evidence codes for manually-curated annotations (full descriptions of these evidence types are supplied in Appendix B):

Experimental Evidence Codes:

EXP: Inferred from Experiment

IDA: Inferred from Direct Assay

IPI: Inferred from Physical Interaction

IMP: Inferred from Mutant Phenotype

IGI: Inferred from Genetic Interaction

IEP: Inferred from Expression Pattern

Author Statement Evidence Codes:

TAS: Traceable Author Statement

NAS: Non-traceable Author Statement

Computational Analysis Evidence Codes:

ISS: Inferred from Sequence or Structural Similarity

ISO: Inferred from Sequence Orthology

ISA: Inferred from Sequence Alignment

ISM: Inferred from Sequence Model

IGC: Inferred from Genomic Context

RCA: Inferred from Reviewed Computational Analysis

Additionally, one evidence code, IEA: Inferred from Electronic Annotation, is reserved for the annotations produced by the Gene Ontology Annotation (GOA) project which are not curated [856]. The evidence codes are hierarchical with general codes acting as parents codes to more specific evidence types. For instance the experimental EXP code acts as a parent to the other, more specific experimental codes³².

Due to the hierarchical nature of the GO evidence codes, and the diversity of the evidence types, some GO annotations are considered more reliable than others. Annotations with the IEA evidence code are considered the least reliable since they are not manually-curated [857]. While the remaining GO annotations are manually-curated, the different evidence types are also thought to differ in their accuracy. In particular, computational evidence codes are generally considered to be less accurate than the experimental codes [857] and the evidence with the codes inferred from sequence or structural similarity (ISS), inferred from expression pattern (IEP) and non-traceable author statement (NAS) is considered lower reliability than the other codes of their class [722].

Despite its manual curation there are several drawbacks to the computational analysis of GO. Several studies have found that sequence-based annotations are error prone and inconsistent [272, 738, 858]. Further, inconsistencies have been identified between annotations from different evidence types [297]. Despite these caveats GO is the most accurate and comprehensive gene annotation source

³²<http://www.geneontology.org/GO.evidence.tree.shtml>

available and is widely used in computational analysis. In particular, GO is commonly used as an evaluation Gold Standard during functional prediction (see Section 2.5.5) [114, 859–861]. GO has also been used as a Gold Standard for dataset scoring prior to PFIN integration [60, 98, 106, 114, 223, 342, 703, 719]. Finally, the representation of GO terms within groups of genes is commonly assessed. When a term is over-represented in a group in comparison to its annotation in the genome as a whole, functional hypotheses may be inferred for the genes [862].

The use of GO as a Gold Standard presents some problems due to the hierarchical nature of the database. Many high level terms are too general to imply a realistic functional association [125, 863]. For example, the term *metabolic process* (GO:0008152) is a child of the root term *biological_process* (GO:0008150). A total of 4315 yeast genes are annotated to this term³³ and there are many diverse metabolic processes occurring in the cell. Clearly assuming a functional link based on this term would add noise to the Gold Standard by linking many protein pairs which do not participate in the same process.

This problem may be overcome to some extent by ignoring the high-level terms at the top of the DAG [98, 864, 865]. However, despite GO's hierarchical nature, the level of a term in the DAG is not necessarily indicative of a term's specificity. Several methods have been applied to overcome this problem. Some studies have looked at the number of annotations to a term and discarded those terms above, and sometimes also below, a certain threshold [866]. An alternative method is to use a GO term specificity score to choose appropriate terms for the Gold Standard. These scores combine the number of annotations to a term with its position in the DAG, thereby gaining a more accurate measure of the term's specificity (Figure 2.23) [867].

Specificity measures may also be used to compare the semantic similarity of GO terms in the DAG. A large number of semantic similarity measures have been developed which range in accuracy and bias (for example [868–870]). The semantic similarity of GO terms can be used in a number of ways; to compare the functional similarity of proteins [864, 871–874], to identify network clusters [875, 876], to evaluate PPIs [581, 870], to predict function [711, 876–880], to analyse coexpression data [881–884] and, in taxonomic analysis [863, 867].

Another alternative method of generation of a GO-based Gold Standard data involves expert curation of the annotation data [125, 157]. For instance a group of six expert biologists were chosen by Myers and co-workers (2006) to vote on which GO terms to include in their Gold Standard, based on whether the experts considered each term specific enough to infer a functional link [125]. Once voting was complete only those terms with four or more votes were included in the Gold Standard.

³³http://amigo.geneontology.org/cgi-bin/amigo/term_details?term=GO:0008152#lineage (accessed 29th January 2011)

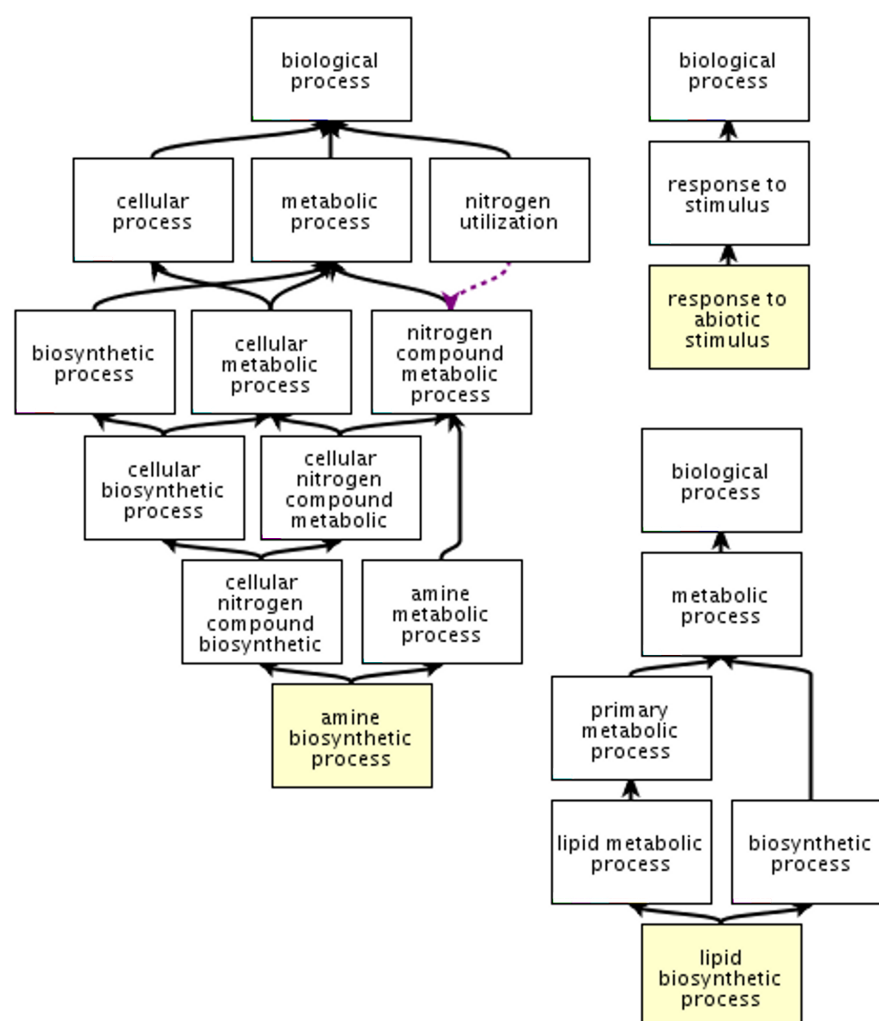


Figure 2.23: GO term specificity.

The level of a term in the DAG is not necessarily indicative of the term's specificity. This example shows three terms (yellow boxes), amine biosynthetic process, response to abiotic stimulus and lipid biosynthetic process, which have the same specificity score but have different depths from the root term in the DAG (4, 3 & 2 respectively). The *is_a* relationships are shown in black while *part_of* relationships are shown as dashed red.

While manual efforts such as this overcome many of the problems associated with GO data, they suffer from a lack of reproducibility. In this case approximately 1/5 of the terms (2031 of 9295) were in the borderline three or four vote bracket. In fact, of the discarded terms over half (716 of 1366) had 3 votes and only 31 terms had no votes. The large percentage of borderline terms makes it likely that a different group of experts could produce a markedly different final list of GO terms and, therefore, a significantly different Gold Standard. In another expert-curated approach by Giot and colleagues (2003), the interactions themselves were analysed based on complex and location data in order to identify pairs of proteins that had a high probability of interaction [157].

Due to GO's popularity for computational analysis a large number of related tools have been developed to perform GO-based analyses. These tools include platforms for GO visualisation [850, 885–887], GO term over-representation [862, 886, 888–892], GO-based gene list comparisons [846, 871, 893] and functional prediction [771, 894–896]. However, due to the hierarchical nature of GO the performance of these tools varies [897].

2.5.5 Network-Based Prediction of Annotation

Since proteins which are involved in the same cellular process or located in the same cellular compartment tend to interact [184, 898], functional interaction networks contain a valuable wealth of data that may be utilised during annotation prediction [49, 60, 106, 508, 668, 708, 777, 780, 859, 899–903]. Network-based annotation prediction is often referred to as guilt-by-association (GBA) prediction since annotations are transferred between pairs of connected nodes within the network [898, 904–906].

GBA functional prediction algorithms each differ in their complexity and accuracy [509]. At the most naïve level, GBA prediction may locally transfer annotations to a node from all nodes with which it has a functional association. However, this level of annotation transfer can be noisy, particularly for hub nodes, and consequently may transfer a high proportion of false annotations. For instance, the central node in Figure 2.24 A has seven neighbouring nodes, which collectively have four different annotations (blue, red, yellow and green). It is unlikely that the central node would be involved in all four processes, given the level of noise in functional data.

Several studies have extended this naïve GBA method to take dataset noise into account. One approach is known as the *Majority Rule* [725]. In this case the annotation which is most highly represented in a node's neighbourhood is transferred to that node. Therefore, the central node in Figure 2.24 A would be annotated to the blue process. A cut-off may also be applied to the number of surrounding annotations whereby only annotations above the cut-off are transferred. A cut-off of 2 annotations in Figure 2.24 A would transfer both the red and blue annotations to the central node, while a cut-off of 4 would not transfer any annotations.

PFINs have edge weights which may be taken into account during GBA annotation prediction. Edge weights are particularly useful when calculated as the probability of functional-relatedness, such as the weights produced by machine learning algorithms [60, 104, 109, 222, 507, 527, 704, 719, 907–909]. Annotations may be transferred between nodes above a particular edge weight cut-off [728, 898, 904]. For example, an edge weight cut-off of 1.5 in Figure 2.24 B would transfer the green and yellow annotations to the central node. Alternatively, the sum of edge weights for each annotation can be calculated to take both frequency and confidence scores into account [910]. Therefore, the

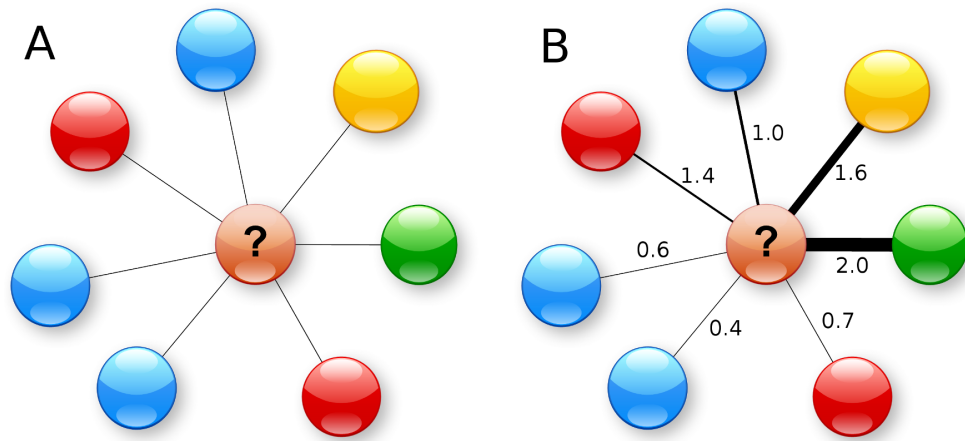


Figure 2.24: Guilt-by-association functional prediction.

Guilt-by-association algorithms differ in their complexity and each may produce a different outcome. In this example the central node has seven neighbours which are collectively annotated to four processes: blue, red, yellow and green. **A.** In an unweighted network naïve GBA would transfer all of the surrounding annotations to the central node. However, this method may be noisy. Using the Majority Rule method the most highly represented annotation (blue) would be transferred to the central node. A cut-off may also be applied to the number of surrounding annotations. Here a cut-off of 2 would transfer both the blue and red annotations to the central node while a cut-off of 4 would not transfer any annotations. **B.** In a weighted network the edge weights can be taken into account during annotation transfer. A cut-off may be applied to the edge weights so that annotations are only transferred along edges above this value. Here a cut-off of 1.5 would annotate the central node to the yellow and green annotations (1.6 and 2.0). The edge weights may also be summed for each annotation and the highest-weighted chosen, in this case the red annotation (2.1). Finally, Maximum Weight rule transfers an annotation along the highest-weighted edge. Therefore the central node would be annotated to the green process (2.0).

highest weighted annotation in Figure 2.24 B would be the red annotation with a sum of 2.1. Finally, annotations may only be transferred from only the neighbouring node attached along the highest-weighted edge. In this case the central node in Figure 2.24 B would transfer the green annotation. This method of GBA prediction is known as the *Maximum Weight* rule and has been shown to have improved accuracy over other local GBA approaches [57].

Local GBA is restricted to a node's immediate neighbourhood in the network, termed its *level-1 neighbours*, and is consequently of limited use in poorly annotated areas of the network. Therefore, several functional prediction methods have been developed which take a larger area of the gene's surrounding network topology into account. Proteins which share annotation partners have a significant chance of sharing function. Consequently, annotations can be transferred between nodes which are connected by a path length of 2, termed *level-2 neighbours* [91, 911, 912]. A combination of level-1 and level-2 annotation transfer produces improved performance over level-1 annotation transfer alone [911]. Alternatively, Chi-squared statistics can be used to extend the Majority Rule to a specified radius around a node of interest [913].

Clustering of the network also allows annotations to be transferred between genes which are not directly connected, since clusters are thought to represent the functional modules of the cell (see Section 2.3.3.2) [725]. Therefore, a node clustering with a large number of other genes annotated to a specific process can be predicted to be involved in that process [90, 564, 860, 914–916]. However, the performance of clustering-based methods is low in comparison to local GBA for some networks [291].

Annotations may also be transferred globally using the full topology of the network. Many of these methods aim to globally maximise the edge weights of functionally associated proteins [92, 580, 894, 917, 918]. The Functional Flow algorithm is a widely used global functional prediction method that simulates the flow of annotations through the network, from annotated to unannotated genes, based on edge weights [902]. The algorithm is particularly useful in poorly annotated areas of the network since the propagation of annotations is not restricted by path length. In other words, an annotation may be propagated through several unannotated nodes if the edge weights are above a specified threshold. Global functional prediction can be combined with local methods using machine learning in order to optimise performance [114].

Evaluation of the quality of an integrated system is difficult due to the level of noise in functional data. However, functional prediction algorithms can provide an objective method of evaluation by utilising known annotation data. In these evaluations the network's ability to predict the known annotations is assessed using one of the functional prediction methods discussed above. A common assessment technique is that of *leave-one-out* cross-validation [104, 510, 676, 676, 907, 911, 919]. In this technique each known annotation is removed from the annotation set and the network assessed by its ability to replace the missing annotation. In some cases, particularly with machine learning-based algorithms, the known annotations are partitioned into groups prior to prediction and cross-validation is carried out using all combinations of the groups as training and testing data.

Functional prediction evaluation allows the sensitivity and specificity of a network's performance to be calculated. Sensitivity is equivalent to the true positive prediction rate and is calculated as $TP / (TP + FN)$, where TP is the number of true positive hits and FN is the number of false negative hits produced by the algorithm. Specificity represents the true negative prediction rate and is calculated $TN / (FP + TN)$, where TN is the number of true negative hits and FP is the number of false positive hits. A receiver operator characteristic (ROC) curve can be plotted as sensitivity against 1-specificity [920]. The area-under-the-curve (AUC) of a ROC curve represents the predictive power of the network, with higher values indicating increased performance [921]. An AUC of 1.0 indicates perfect classification of the known annotations, and an AUC of 0.5 indicates random classification (Figure 2.25).

Network-based analyses, in particular those involving multiple data types, can also be used to predict candidate genes for diseases and phenotypes [68, 118, 220, 678, 922–936]. Where some disease or phenotype genes are known GBA can be used to propagate the annotation to candidate genes along network edges [223, 591, 706, 825, 937–943]. The identification of potential drug targets is also essential to combat disease [944]. Many drugs combat disease by interrupting the cellular processes associated with the disease phenotype. Essential genes, in particular essential network hubs, are potentially useful targets for drugs [598]. Therefore the topology of integrated networks may also be used to identify potential targets [11, 118, 217, 664, 695, 945] and to predict possible side effects of new drugs [946].

Despite the numerous network analysis tools available, the scale of functional networks often makes them difficult to work with. In many cases research groups have their own specific research interests which focus one area of cellular biology. Therefore, only a subset of an integrated functional network may be of interest to any individual group [83]. Consequently many network studies use known annotation data, in particular from GO, to produce area-specific subnetworks from functional interaction data.

There are two main methods with which to build these subnetworks. In the first, a network is integrated using all the available data and the process-specific sub-network extracted from it [41, 59, 129, 708, 947]. In the second, a process-specific subnetwork is built using a subset of the available

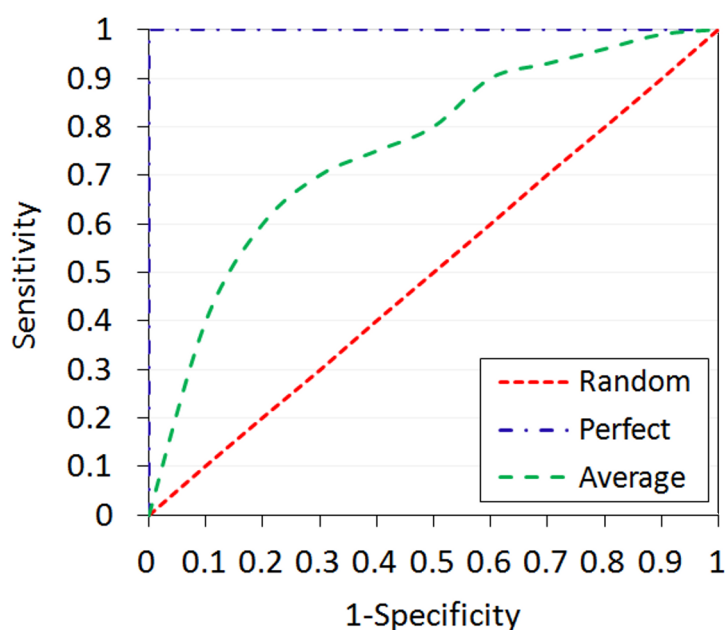


Figure 2.25: Area under curve.

The AUC of a ROC curve is indicative of network performance. Networks which produce an AUC of 0.5 do not perform any better than random (red), while those which produce an AUC of 1.0 perform perfectly (blue). Generally, most networks produce an intermediate AUC (green).

data [46, 130, 307, 720]. However, since these methods rely on existing annotations they have several drawbacks. Like other biological data annotation data are biased towards highly-studied proteins (see Section 2.4.4). Therefore, these methods are of little use for processes with few available annotations or in poorly annotated areas of the integrated network. Further, since some types of annotation data, particularly sequence-based data, are thought to be unreliable (see Section 2.5.4.3), the resulting sub-network may also be unreliable. Finally, these methods potentially discard useful data which would be of interest in relation to the process being studied [128].

2.6 Yeast as a Model Organism to Study Human Ageing

Saccharomyces cerevisiae, commonly known as baker's yeast or budding yeast, is a eukaryotic model organism [948]. *S. cerevisiae* is widely used to study human cellular processes since its genes encode similar proteins to *Homo sapiens* and many of its genes complement human mutations [949–952]. In 1997 31% of yeast genes were known to have human homologs with conserved function [948] and many human disease genes have yeast homologs [953]. The InParanoid database of eukaryotic orthologs [954] listed 2154 orthologous clusters for *S. cerevisiae* and *H. sapiens* in November 2010³⁴.

Yeast is an ideal model organism for a number of reasons. As a eukaryote, yeast has a similar subcellular structure to human cells [955] and its single-celled nature makes it cheap and easy to work with in comparison to higher eukaryotes [949, 956, 957]. Additionally, yeast grows rapidly making it an ideal experimental organism. The full genome sequence of *S. cerevisiae* has been available for several years and the entire complement of ORFs has been identified [320]. A number of powerful genome-wide techniques have been developed using yeast and providing a wealth of genome-wide data to enhance analysis [958, 959]. For instance deletion mutants are available for all non-essential yeast genes through the *Saccharomyces* Gene Deletion Project³⁵ [6, 230] and mutant sets containing genome-wide random insertions have been produced (see Section 2.1.2) [960].

Ageing is a complex, multi-factorial, systems-wide phenomenon [961]. Chronological ageing of *S. cerevisiae* is highly similar to the mammalian ageing process [962] and many associated processes are conserved among eukaryotes [963]. Consequently, *S. cerevisiae* has been widely used to study human ageing and ageing-related diseases [964, 965]. For example, the human premature ageing diseases Werner's Syndrome and Bloom's Syndrome have been extensively studied in *S. cerevisiae* [966–970]. The human genes associated with the two diseases have a single homolog in *S. cerevisiae*, *SGS1*. When this gene is deleted, it results in a similar cellular phenotype and reduced life-span to that seen in humans with the diseases [948, 968].

³⁴<http://inparanoid.sbc.su.se/cgi-bin/e.cgi> (Data accessed 3rd November 2010).

³⁵http://www-sequence.stanford.edu/group/yeast_deletion_project/deletions3.html

There are many theories relating to the ageing process, all of which have some overlap. Two groups of theories concern two major aspects of cellular biology. The first is based on chromosome structure and maintenance, in particular the maintenance of the telomere and the accumulation of DNA damage. The second is associated with reactive oxygen species (ROS), oxidative stress and mitochondria [971]. However, there is significant overlap between the theories in terms of biology and evidence [972–974] and the specific relationships between the processes involved in the ageing process still remains unclear [961].

2.6.1 Telomere Maintenance

The integrity of the genome is essential for a cell's survival. Telomeres are repetitive regions of non-coding DNA (in humans n(TTAGGG) repeats [975, 976]) and associated proteins that protect the end of the chromosome from degradation and telomere-telomere fusion [977–979]. Telomeres are also required for chromosome positioning and replication [976]. The telomere's structure and telomere related processes are highly conserved in eukaryotes [980, 981].

Every time a cell divides its telomeres become progressively shorter. The shortening occurs as the replication machinery, DNA polymerase, is unable to replicate the full length of the DNA due to the need for a RNA template at the end of the telomere to initiate lengthening [982]. This phenomenon is referred to as the *end replication problem* [983]. Without maintenance the telomere eventually becomes too small to protect the chromosome leading to replicative senescence, an irreversible arrest of growth [984]. Due to this effect telomere length regulation and maintenance is central to the telomere's function.

One of the most important components of the eukaryotic maintenance system is the reverse- transcriptase enzyme, telomerase. Telomerase utilises a segment of its RNA as a template to lengthen the telomere by the addition of repeats onto its 3' end, maintaining an average telomere length (Figure 2.26) [980, 981].

In *S. cerevisiae* telomere maintenance involves a large number of genes, covering a wide range of biological processes and several genes of unknown function [985]. *S. cerevisiae* has a telomerase enzyme which is similar to its *H. sapiens* equivalent [985–988]. In addition a number of the gene products are known to bind the telomeric repeats [987]. These proteins cover a wide range of biological processes including DNA repair, DNA damage responses, telomere regulation and chromatid cohesion [988–991]. Deletion or mutation of these genes can lead to shortening or lengthening of the telomere [978, 992]. Baker's yeast has been used as a model to study human telomere maintenance and ageing since many of the genes involved in telomere function in *S. cerevisiae* have been found to have homologs in *H. sapiens*.

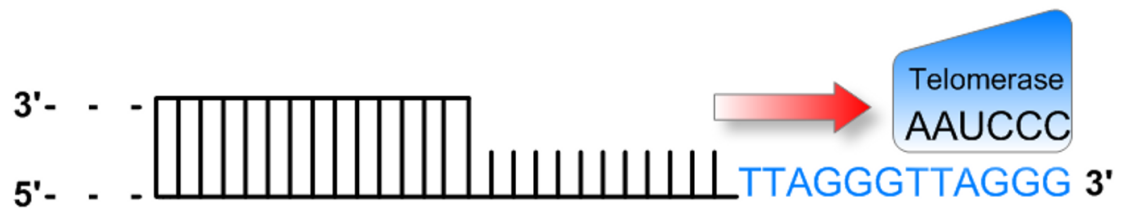


Figure 2.26: Telomere maintenance by telomerase.

The enzyme telomerase maintains the telomeres length by the addition of TTAGGG short repeats using an RNA template.

Of particular note is the telomere capping protein Cdc13, which recruits telomerase to the telomere. *CDC13* is an essential gene and therefore cannot be deleted by conventional means. However, the temperature sensitive mutant *cdc13-1* provides a basis for genetic analysis. In the mutant strain Cdc13 caps the telomere normally at low temperatures. At high temperatures ($> 27^{\circ}\text{C}$) telomeric uncapping occurs. Two studies at the Centre for Integrated Biology of Ageing and Nutrition (CISBAN) have studied this mutant using [HTP](#) technologies. In the first study, microarrays were used to study the transcription of genes in response to telomeric uncapping in order to identify uncapping response genes [993]. In the second study, a genome-wide [SGA](#) was carried out in which the *cdc13-1* mutant was crossed with the yeast deletion mutant collection to identify synthetic [GIs](#) associated to telomere uncapping [16].

2.6.1.1 Telomere Shortening and Disease

In humans telomerase activity in germline cells maintains the telomeres at an average length [994]. However, the somatic cells (non-germline, diploid, body cells) have very low telomerase activity and undergo a natural telomere shortening with age [995, 996]. Somatic cells have a finite lifespan [997] and can divide a limited number of times [998]. The shortening of the telomeres has been linked with the ageing process, since as telomeres become critically short it leads to the non-dividing final state termed *replicative senescence* [972, 999]. Telomeric shortening is thought to act as a "mitotic countdown" limiting the number of cell divisions [1000]. The *telomere theory of ageing*, sometimes referred to as the Hayflick limit theory remains one of the major ageing paradigms [982, 997, 1001–1003].

Telomere length and the breakdown of the telomere maintenance system has been associated with cancer [995, 1000]. If senescence does not occur the telomeres become increasingly shorter, chromosomes undergo damage and telomere–telomere fusions may occur. These circumstances lead to a state known as *cell crisis* [979, 996]. When telomerase is present at high levels in somatic cells both

senescence and crisis are prevented and the cells are known as *immortal* [979, 994, 996]. Telomere shortening has been observed in cancerous tumours [1004] however many tumours also express telomerase [1005]. Therefore reactivation of telomerase in somatic cells may be an important step in tumour progression [995]. Potentially, telomerase is an effective target for cancer treatments [1000]. Telomere maintenance has also been associated with several other human diseases including premature ageing diseases [1006–1008]. Shortening of the telomere has been observed in a number of these diseases, for example in Down's syndrome [1009]. Many of the proteins implicated with premature ageing diseases are associated with the telomere and DNA repair [1006, 1010]. For instance, the *WRN* gene of Werner Syndrome [1011] and *ATM* of Ataxia Telangiectasia [1012].

2.6.2 Oxidative Stress

Oxygen is essential for cell viability, but is deadly in high quantities. Cellular enzymes maintain a redox homeostasis within the cell which balances the production and consumption of reactive oxygen species (ROS). Oxidative stress occurs when this system becomes unbalanced [1013, 1014]. The imbalance can occur in three ways: by an increase in ROS, for instance due to chemical exposure; by a decrease in antioxidants and the other ROS defence systems; or by a combination of the two [1015]. When oxidative stress occurs an excess of ROS builds up in the cell causing macromolecular damage and leading to growth arrest [1016]. The cell has several enzymatic and non-enzymatic defence systems which protect against ROS and repair any subsequent damage. However at high levels of ROS these repair systems are overwhelmed and controlled cell death, known as *apoptosis*, occurs [1017]. In extreme circumstances the apoptosis pathways are bypassed and uncontrolled cell death, termed *necrosis*, can occur [1017].

Oxidative stress has been linked to several human diseases and to the ageing process. *S. cerevisiae* provides an excellent model with which to study the oxidative stress response since baker's yeast generates ROS through the same mechanisms as humans and other mammals, and has many of the same antioxidant features [1018–1022]. ROS are produced during aerobic metabolism and in response to chemical exposure. *S. cerevisiae* provides a convenient system to distinguish between these responses since it can grow either by fermentation or by aerobic respiration. Therefore, growth on a fermentable media allows the study of the chemical responses. Switching between fermentable and non-fermentable growth media induces oxidative stress through respiration [1023].

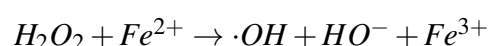
2.6.2.1 Reactive Oxygen Species

ROS are oxygen-derived oxidants and reductants which alter the natural redox homeostasis of the cell. Many ROS are free radicals; that is, they contain an unpaired electron. Other ROS are molecules which are easily converted to free radicals due to their reactivity [1024]. The oxidising potential of ROS range from the relatively low reactivity of hydrogen peroxide (H_2O_2) to the highly reactive hydroxyl radical ($\cdot OH$) [1025].

ROS can be beneficial if the redox homeostasis of the cell is maintained correctly. Some immune cells utilise the superoxide anion ($\cdot O_2^-$) to kill invading pathogens by inducing apoptosis [1026, 1027]. ROS are also messengers in several cellular signalling cascades where they act by oxidising proteins [1028–1031].

ROS are produced naturally in a number of ways. The coenzyme adenosine triphosphate (ATP) is generated in the mitochondrion via oxidative phosphorylation and acts as the source of energy for cellular reactions [1032]. Approximately 80% of cellular oxygen is utilised in this manner [1033]. Electrons produced by the electron transport chain can leak from the mitochondria leading to the production of ROS. It is estimated that around 1% of O_2 in the cell is converted into ROS in this way [1034]. In humans the production of ROS has been linked with specific tissues. For instance hydrogen peroxide production in the thyroid gland during hormone production [1035] and potentially by wounded tissues during leukocyte recruitment [1036].

ROS can also be produced in the presence of metal ions. All transition metals, with the exception of zinc, contain unpaired electrons and can change their oxidative state. For instance iron can cycle between Fe^{2+} and Fe^{3+} . Many of these metals are essential to the cell and play important roles in enzymes where they act by redox cycling [1037]. However, as metal ions have variable oxidation states they can catalyse ROS generation [1038, 1039]. Hydrogen peroxide can produce hydroxyl radicals in the presence of transition metals such as iron by the Fenton Reaction [1040]:

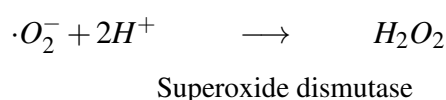


Consequently there is an overlap between metal ion homeostasis and the oxidative stress response. Several proteins involved in the oxidative stress response, such as metal binding proteins, have been linked to iron homeostasis [1017].

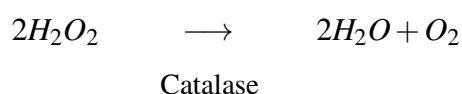
2.6.2.2 Redox Homeostasis

The cell has several systems to maintain redox homeostasis by protecting against ROS, detoxifying oxidants and repairing cellular damage. These systems are made up of factors, termed antioxidants [1041]. The enzymes superoxide dismutase (SOD), catalases and peroxidases are important in this process.

The superoxide anion is relatively stable in comparison with the hydroxyl radical but it can be the precursor for several more reactive species. SOD reduces the superoxide anion to hydrogen peroxide using a bound metal ion [1042]:

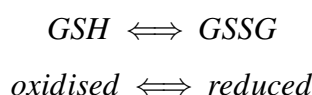


The highly conserved enzyme catalase converts hydrogen peroxide to water and oxygen using a haem group [1043–1045]:



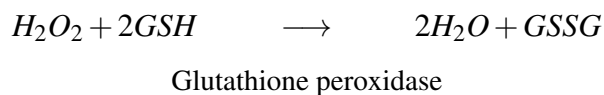
The tripeptide glutathione has several roles in redox homeostasis and is the most abundant antioxidant in the cell [1046–1048]. When oxidised glutathione is a ROS scavenger that utilises its reactive thiol group to detoxify free radicals. Additionally, oxidised glutathione (GSH) also protects the cysteine side chains of proteins and other vulnerable groups from oxidative damage by the formation of disulphide bonds to prevent irreversible protein damage [1040, 1045].

Glutathione can be oxidised or reduced:



High levels of reduced glutathione (GSSG) is indicative of oxidative stress and, therefore, the GSH: GSSG ratio is indicative of the cellular redox state [1049, 1050]. Glutathione is produced by glutathione synthase (Gsh1). Mutants with reduced glutathione, for instance by the deletion of *GSH1*, display ROS hypersensitivity and are vulnerable to chemical oxidants [1051, 1052].

Hydroperoxidases, such as glutathione peroxidase and thioredoxin peroxidase [1053, 1054] are ROS scavengers involved in redox sensing [1055]. Peroxidases regulate cellular peroxide levels by converting hydrogen peroxide to water using an electron donor such as glutathione [1056]:



Two further redox maintenance systems are the glutaredoxins (GRXs) and thioredoxins (TRXs). These systems utilise redox cycling to reduce oxidised proteins, and act to protect proteins against oxidative damage [1057]. The GRXs utilise GSH as an electron donor [1058] and TRXs utilise nicotinamide adenine dinucleotide phosphate (NADPH) [1059]. Both systems have cysteine groups and are involved in the regulation of iron homeostasis [1060, 1061], mainly through negative regulation of the iron-dependent transcription factor, Aft1 [1062, 1063]. In addition, the GRXs have a role in the synthesis and protection of iron/sulphur clusters in enzymes [1064, 1065]. GRX and TRX are therefore essential regulators of the redox state [1066].

There are many other defence systems against ROS including several non-enzymatic molecules (reviewed in [1017]). Vitamin C (ascorbate) and vitamin E in humans [1037, 1040, 1067], and their yeast equivalents [1068, 1069], have antioxidant properties. Some quinones, for instance ubiquinone, are lipid soluble antioxidants that can remove ROS by oxide reduction. Metals binding proteins such as ferritin and metallothionein prevent ROS production by binding free metal ions [1070].

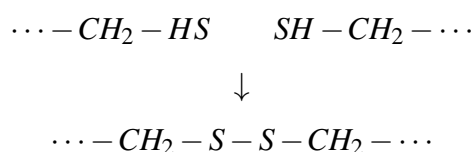
2.6.2.3 ROS Effects

High levels of ROS can overwhelm the natural cellular defences and cause damage to the macromolecular components of the cell by oxidation [1016]. Damage can occur to DNA, proteins and lipids.

DNA damage can occur to individual nucleic acids and to the phosphate backbone. DNA base modifications can cause distortion and mis-pairing leading to GC → AT transversions [1071, 1072]. Damage to the DNA backbone can cause breakage of the DNA strands. Unsurprisingly mitochondrial DNA incurs more damage than nuclear DNA due to the levels of ROS produced during respiration [1073, 1074]. Oxidative damage due to ROS has also been associated with telomere shortening [1075, 1076].

Damage to proteins can occur in several different ways. Protein side chains, in particular cysteine and methionine groups, and the protein backbone can be damaged by ROS. Cross-links often occur

which inactivate the proteins function [1077]. For instance by disulphide bridge formation between thiol side chains:



The introduction of carbonyls to proteins by ROS also commonly occurs. For example the carbonylation of proline, arginine, lysine, and threonine residues [1078]. Protein sulphydryl groups, in particular those on cysteine residues, are very vulnerable to oxidation by ROS into sulphinic, sulphonic and sulphenic acids. Some of these reactions can be beneficial to the cell. For instance, the sulfinic acid switch is used in redox sensing and signalling [1079]. However, many of these reactions are irreversible [1040, 1080]. Damage to a protein can alter its structure, function and hydrophobicity causing incorrect interaction and denaturation. In some cases the damaged proteins aggregate with potentially harmful effects [1081].

Lipids are also damaged under oxidative stress conditions. In particular the hydroxyl radical attacks fatty acids creating a chain reaction of peroxy radical production [1082]. Damage to the lipid bilayers of cellular membranes can cause leakage and change the cell's internal chemistry [1083].

2.6.2.4 The Oxidative Stress Response

Oxidative stress occurs when the redox state of the cell becomes unbalanced. Redox imbalance occurs when the levels of ROS exceed the natural cell defences, when the antioxidant levels of the cell are reduced, or by a combination of the two conditions. Oxidative stress can be induced by chemical exposure, ionising radiation or UV light [1084, 1085]. Several compounds can induce oxidative stress in *S. cerevisiae* by increasing ROS production and are consequently used in the study of oxidative stress such as menadione, diamide, hydrogen peroxide, paraquat and tetra-butyl hydroperoxide (tBOOH).

Many of these chemicals induce oxidative stress by redox cycling. The quinone menadione reacts with the thiol side chains of some molecules to produce superoxide anions and reduces glutathione levels [1086]. Diamides oxidise sulphydryls leading to decreased cellular glutathione and an increase in ROS [1087]. Hydrogen peroxide induces oxidative stress when present in higher quantities than naturally found in the cell [1086]. Paraquat (1,1'-dimethyl-4,4'-bipyridylum) is a herbicide that generates superoxide anions, hydrogen peroxide and hydroxyl radicals through the oxidation of re-

duced glutathione and NADPH [1088]. Finally, the synthetic alkyl hydroperoxide tBOOH can cause free radical chain reactions leading to oxidative stress [1089].

Since heavy metals can induce ROS through the Fenton reaction they can also be used to induce oxidative stress. Arsenic and cadmium are not naturally found in the cell but can induce ROS production following environmental exposure [1090, 1091]. These elements react with sulphur-containing compounds such as thiol-containing glutathione, interrupting the redox balance of the cell [1092, 1093]. Cadmium is also thought to cause mitochondrial dysfunction in humans leading to increased ROS production and oxidative stress [1094]. Interestingly it has been shown that different sources of ROS produce different oxidative stress responses [1017, 1086, 1095–1097].

Several studies have investigated oxidative stress in *S. cerevisiae* using chemical oxidants [1091, 1097–1099]. Under oxidative stress conditions a number of genes are up-regulated by what is termed the *oxidative stress response* [1100]. Many of the proteins involved in redox homeostasis (Section 2.6.2.2) are involved in this response, some of which remove excess ROS and others that repair oxidative damage. For instance glutathione synthase increases the levels of the ROS scavenger GSH [1101] and the TRXs repair oxidised proteins [1041, 1102]. During the response arrest of the cell cycle occurs [1103] and additional systems are up-regulated to repair oxidative damage. These systems include several DNA repair proteins [1104] and proteases which remove damaged proteins [1105]. While the responses to different chemical oxidants have some overlap, each response has unique features [1041, 1095]. Several transcription factors have been implicated in mediating the oxidative stress response including Yap1 [1101, 1106–1108], Skn7 [1108, 1109], Msm2 and Msm4 [1110, 1111], Rox1 [1112, 1113], Met4 [1091], Mga2 [1112], and the metal ion sensing factor Aft1 [1060, 1114]. Many of these factors have roles in other stress responses [1115].

2.6.2.5 ROS, Ageing and Disease

There has long been evidence that ROS and oxidative stress-induced damage are linked with the ageing process [1116]. The *free radical theory of ageing* was first coined in the 1950s. It postulated that the accumulation of ROS damage over time is the cause of cellular ageing [1117]. The author expanded his theory several years later to incorporate mitochondria as the major source of ROS within the cells [1118]. The theory has evolved over the following years and remains one of the major ageing theories, sometimes referred to as the *oxidative stress theory* or *mitochondrial theory* [1119].

A great deal of evidence supports this theory [1120]. Damage produced by ROS has been shown to increase with age. For instance the mitochondrial damage leading to decline in function [1121, 1122] and the oxidative damage to proteins [1077]. Aged *S. cerevisiae* cells contain far higher levels of ROS

[1123] and oxidatively modified proteins than young cells [1124, 1125]. Aged human cells also have more carbonylated proteins [1126].

Redox homeostasis and oxidative stress response components are vital to lifespan in *S. cerevisiae*. For example deletion of the SOD gene decreases lifespan [1127, 1128] while expression of other redox homeostasis genes can overcome this deletion and lengthens lifespan [1129]. Similar evidence has also been seen in several other species [1022].

Oxidative stress and ROS are linked to several diseases including include Huntington's, Ataxia-Telangiectasia, Down Syndrome and Werner Syndrome [1130, 1131]. ROS are also linked to cardiac problems, autoimmune disease, cancers and chronic fatigue syndrome [1132]. Lowered SOD expression is linked to osteoarthritis [1133]. Increased ROS and oxidative damage have been seen in neurological disorders [1134] and cancer [1104, 1135].

Protein carbonylation has been linked with several diseases including include Parkinson's disease, Alzheimer's disease and some cancers [1080]. The brain requires high levels of oxygen in comparison with other organs therefore is naturally more prone to oxidative stress. Protein aggregation due to oxidative stress is associated with neurodegenerative disorders and ageing [1081]. In Alzheimer's and Parkinson's diseases the aggregations are of a single protein type which can compromise cell viability. Notably many of the diseases associated with oxidative stress are also associated with telomere dysfunction (see Section 2.6.1.1) [1131]. Due to these similarities the effects of ROS in yeast and humans, *S. cerevisiae* is an ideal organisms in which to study the potential effects of drugs in humans.

Oxidative stress responses have also been used in the development of disease treatments. Many anti-cancer drugs are designed to induce programmed cell death in cancer cells through ROS production [1028, 1092]. For instance arsenic has also been used to treat cancers in this way.

Although ROS and oxidative stress have long been believed to reduce lifespan it been observed that at low levels ROS lead to an adaptive response resulting increased protection from higher doses [1112]. This adaptation seems to be specific to the type of ROS utilised [1095]. The response can be induced by calorie restriction since this state induces mitochondrial metabolism leading to the gradual accumulation of ROS. When subsequently subjected to higher ROS levels a significant increase in lifespan is observed. This process is termed *mitochondrial hormesis* has been seen in *S. cerevisiae* [1136] and several other mammalian species [1137]. However, mitochondrial hormesis has not been seen in humans to date [1138] and evidence suggests this effect may not be seen in *H. sapiens* [1139].

2.7 Summary

Deciphering the interactome is a massive undertaking involving understanding the interactions of hundreds of proteins in multiple cellular conditions and phenotypic states. Recent developments in high-throughput experimental technologies have produced a wealth of functional data which can aid in achieving this task. Additionally, a large number of databases have been designed to store these datasets, and several annotation schemas have been developed to describe functional data in a consistent manner. There are many diverse types of functional data ranging from direct protein-protein interactions to indirect functional relationships such as genetic interactions.

Each type of functional data provides information about a different aspect of cellular biology. Therefore, integration of the heterogeneous data types can provide a fuller picture of the interactome than any dataset alone and can reveal global properties which are not evident in a single data type. However, analysis of the available data is a non-trivial task due to its scale, levels of noise and biases.

Biological data can be visualised as a network and graph theory used to interpret the data in a manner that is both human-friendly and computationally amenable. Many network tools have been developed for the visualisation and manipulation of functional data, and several types of graph theoretic algorithms can be used to study network data, for instance to compute topological parameters, to cluster the data and to align distinct networks.

PFINs are powerful tools with which to generate new hypotheses from functional data, since their edge weights provide a measure of dataset accuracy. These weights can be incorporated into network analyses to improve the accuracy of results. **PFINs** have been developed for multiple species and used for a number of applications, including protein functional prediction, **PPI** prediction, module detection and evolutionary studies.

However, while probabilistic scoring assesses the quality of individual datasets, it ignores their content. Each functional dataset has its own biases due to experimental design, technical limitations, analysis methods and cellular bias. While several previous studies have attempted to remove these biases, these approaches risk the loss of valid and useful data.

Individual research groups each have their own specific interests. Consequently, while a global network analysis may produce a wealth of data, only a small fraction of the results will be relevant to each specific biological question. Historically, this problem has been approached by using existing annotation data to produce process-specific subnetworks from subsets of the available data. However, this approach may also discard valid and useful data and is of limited use where annotation data are sparse. Therefore, there is need for network integration and analysis methods that overcome this drawback and produce process-relevant hypotheses without loss of data, allowing their use in

unannotated areas of the interactome. The work presented in this thesis addresses this need.

Ageing provides an excellent exemplar process against which to develop and evaluate process-relevant techniques. *S. cerevisiae* ageing has been extensively studied, due to its similarity to the human ageing process and its links to human disease, producing a large amount of data. Ageing's multi-factorial nature provides several related and overlapping processes with which to assess network accuracy. Further, yeast is inexpensive, fast-growing and easy to work with experimentally, allowing straightforward laboratory evaluation of new hypotheses.

Chapter 3

Methods

The aim of this project was the development and systematic evaluation of novel biological network integration and analysis techniques which harness dataset relevance. In order to achieve this aim several existing tools were utilised to build and assess networks, prior to the development of the process-relevant network integration method. Computational and experimental evaluation of the resulting networks was then carried out using several techniques.

This chapter outlines the tools and methods applied. Sections 3.1.1 to 3.1.4.7 deal with the computational analysis of the data and networks: the dataset versions chosen, the initial dataset analysis, and the integration technique developed (see Chapters 4-6). Computational evaluation methods are then discussed in Section 3.1.5. Finally, Section 3.2 describes the experimental techniques applied in the final stage of validation (see Chapter 7).

3.1 Computational Techniques

3.1.1 Data Sources

The BioGRID database¹ [276] for *Saccharomyces cerevisiae* was used as the source of interaction data since it is highly curated and contains interactions of 27 experimental types (see Section 2.1.4.1). The sets of interacting proteins were first split by PubMed ID (PMID)² to identify pairs produced by different experimental studies. Self-interactions, duplicates and interactions with proteins from other species were removed from the dataset.

Datasets containing at least 100 interactions were treated as separate high-throughput (HTP) data sources while the remaining low-throughput (LTP) studies were grouped together in accordance with

¹<http://thebiogrid.org/>

²<http://www.ncbi.nlm.nih.gov/pubmed/>

Table 3.1: Dataset file versions.

The file versions used in this study. Version numbers correspond to the BioGRID version number. KEGG and GO data was taken from the file versions available at the BioGRID release date.

Chapter	Section	Version	Release Date
4	4.1	Version 27	May 2007
4	4.3	Version 38	March 2008
4	4.19	Version 50	March 2009
5	5.4	Versions 17-52	July 2006-May 2009
6	6.2	Version 52	May 2009
7	7.2	Version 65	June 2010

the experimental categories provided by BioGRID (see Section 4.1.3.3) [235]. This cut-off of 100 interactions was used for all datasets unless otherwise stated.

To avoid ambiguity a standard dataset naming format was used. For HTP data:

[Version].Author.PMID.[Type],

for example V27.Pan.15525520.Synthetic_Lethality; and for LTP data,

[Version].Type,

for example V27.Synthetic_Lethality.

KEGG PATHWAY data was chosen as the source for the Gold Standard since KEGG is highly curated (see Section 2.5.4.2). PATHWAY files for *S. cerevisiae* were downloaded from the KEGG database³. The KEGG Gold Standard was constructed by selecting all possible pairs between genes annotated to the same pathway. A negative Gold Standard was also constructed consisting of all possible pairs of genes that were not annotated to the same pathway, excluding those *S. cerevisiae* genes not annotated in KEGG (see Section 2.4.2).

Finally, GO annotations were chosen for use in network evaluation since GO is also highly curated (see Section 2.5.4.3). Two files were downloaded to provide the annotation information: the Gene Ontology from the GO Consortium⁴, and SGD-GO annotation mapping from the SGD⁵ [296]. The Gene Ontology file provides details of all GO terms in use at the time of release, and of the hierarchical relationships between them, while the SGD file provides annotations to these terms in the form of gene-term pairings.

The KEGG and GO annotation files available at the BioGRID release date were used during integration and evaluation unless otherwise stated. The file versions used in this study are summarised in Table 3.1.

³<http://www.genome.jp/kegg/>

⁴<http://www.geneontology.org/>

⁵<http://www.yeastgenome.org/>

3.1.2 Gene Ontology Analysis

The current version of GStats [862], Version 2.2.6, was used to measure over-representation of GO Biological Process terms (see Section 2.5.4.3) using a p-value cut-off of 0.00001. The low cut-off was chosen to limit the results to only those terms with high over-representation to the POI. The package was run in R Version 2.5.0 using the following input:

```
params <- new("GOHyperGParams",
ontology = "BP",
conditional = TRUE,
geneIds = data,
universalGeneIds = gene,
annotation = "YEAST",
pvalueCutoff = 0.00001,
testDirection = "over")
results <- hyperGTest(params)
results
summary(results)
```

where, *data* is a vector of the genes in the dataset and *gene* is a vector of the *S. cerevisiae* genes from the BioConductor⁶ YEAST annotation package⁷.

Specificity of individual GO terms was calculated from the GO and the SGD data files using the Information Content measure of Resnik [867, 1140] using equation 3.1, where, $n(t)$ is the total number of annotations to term t , including all child terms and N is the total number of annotations to all terms. In all cases the term *biological_process* (GO:0008150) was taken as the root term and annotations with the evidence code inferred from electronic annotation (IEA) were excluded (see Section 2.5.4.3). For the Version 65 data, annotations with the new evidence code reviewed computational analysis (RCA) were also excluded (see Section 7.1.1).

$$IC(t) = \ln \left(\frac{n(t)}{N} \right) \quad (3.1)$$

⁶<http://www.bioconductor.org/>

⁷Since completion of this project the YEAST annotation package has been replaced by the GO.db package

3.1.3 Hierarchical Clustering

Hierarchical clustering was carried out using Cluster 3.0⁸ with the default options [11]. Clusters were visualised using Java Treeview 1.1⁹.

3.1.4 Network Integration

3.1.4.1 Confidence Scoring

The confidence score was calculated by scoring the datasets against the KEGG PATHWAYS [277] Gold Standard using the Bayesian statistics approach developed by Lee and colleagues (see Section 2.4.3) [49], which calculates a log-likelihood score for each dataset (3.2).

$$lls^L(E) = \ln \left(\frac{P(L|E)/\neg P(L|E)}{P(L)/\neg P(L)} \right) \quad (3.2)$$

where, $P(L|E)$ and $\neg P(L|E)$ represent the frequencies of linkages L observed in dataset E between genes annotated to the same and differing KEGG pathways, respectively, and, $P(L)$ and $\neg P(L)$ represent the prior expectation of linkages between genes in the same and differing KEGG pathways, respectively.

A score greater than zero indicates that the dataset links genes annotated to the same pathway. Higher scores indicate greater confidence in the data. Datasets that did not have a positive score were discarded (see Section 4.3.1).

3.1.4.2 Process of Interest

Each process of interest (POI) was based on Gene Ontology biological_process (GOBP) annotations. Due to the hierarchical nature of the GO DAG a POI was defined as a term and any descendant of that term within the DAG, with the exclusion of annotations with the evidence code IEA and, for Version 65, annotations with the evidence code RCA (Figure 3.1).

3.1.4.3 Relevance Scoring

A hypergeometric test (Equation 3.3) [862] was applied to each dataset to score over-representation of genes annotated to the POI, producing a Node Relevance score (see Section 4.2).

⁸<http://bonsai.hgc.jp/~mdehoon/software/cluster/software.htm#ctv>

⁹<http://jtreeview.sourceforge.net>

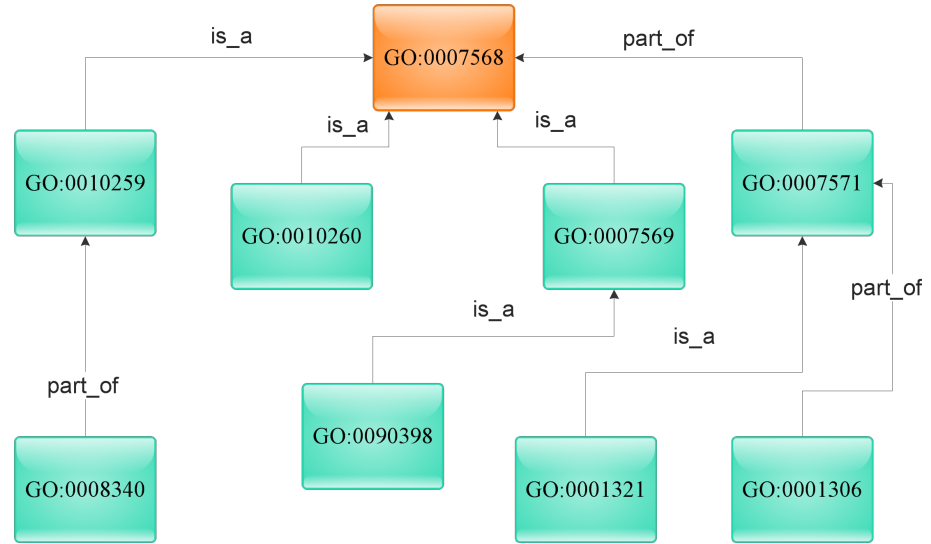


Figure 3.1: Definition of a process of interest.

A process of interest is defined as the POI term and all children of that term in the GO DAG. In this example the POI is the term GO:0007568 (ageing) is the POI (shown in orange). All genes annotated to the term or to the eight child terms (shown in blue) are considered to be annotated to the POI during the hypergeometric test. Note that this figure displays a subset of the child terms of GO:0007568.

$$p(k; N, m, n) = \frac{\binom{m}{k} \binom{N-m}{n-k}}{\binom{N}{n}} \quad (3.3)$$

where m is the number of *S. cerevisiae* genes annotated to the POI, k is the number of genes in the dataset annotated to the POI, n is the dataset size in terms of genes, and, N is the total number of *S. cerevisiae* genes. This test calculates the probability that there would be that many genes annotated to the POI given the total number of genes annotated to the POI in the whole genome. The Node Relevance score was used in Chapters 4 and 5.

In Chapter 6 two further relevance scores, Edge Relevance and Interaction Relevance, are introduced. These scores were also calculated using the hypergeometric test (Equation 3.3).

For Edge Relevance m is the number of possible *S. cerevisiae* annotations containing at least one gene annotated to the POI, k is the number of interactions in the dataset containing at least one gene annotated to the POI, n is the dataset interaction size, and N is the total number of possible *S. cerevisiae* interactions involving genes in the BioGRID database.

For Interaction Relevance m is the number of possible *S. cerevisiae* annotations where both genes are annotated to the POI, k is the number of interactions in the dataset where both genes are annotated to the POI, n is the dataset interaction size, and N is the total number of possible *S. cerevisiae* interactions involving genes in the BioGRID database.

3.1.4.4 Control Integration

Control networks were produced by integrating confidence scores using the weighted sum (Equation 3.4) described by Lee and colleagues [49]. This method integrates the datasets in order of their confidence scores, giving a higher weighting to datasets with higher confidence, while allowing for dependencies between the datasets.

$$WS = \sum_{i=1}^n \frac{L_i}{D^{(i-1)}} \quad (3.4)$$

where L_1 is the highest confidence score and L_n the lowest confidence score of a set of n datasets. Division of the score by the D parameter means that, while the highest score is integrated unchanged, subsequent weights are progressively down-weighted (see Section 2.4.3). With a D value of one the integration is a simple sum of the scores. A D value of 1.1 was chosen for integration since at higher values lower-ranking scores contribute little or nothing to the integration, discarding potentially important information and reducing network performance (Figure 3.2). In the resulting network highly-weighted edges have high confidence.

3.1.4.5 Relevance Integration

To produce the relevance networks the relevance scores were used to re-order the datasets prior to integration of the confidence scores using Equation 3.4, giving a higher weighting to datasets with higher relevance (see Section 4.2). In the resulting network highly-weighted edges have both high confidence and high relevance to the POI. Java code to produce the relevance and control networks is available in Appendix C.

3.1.4.6 Reversed Integration

The reversed networks were produced by integration of the datasets using Equation 3.4 in the reverse order of that used in the control and relevance networks (see Section 4.3.1). Thus, a higher weighting is given to datasets with low confidence and low relevance, respectively.

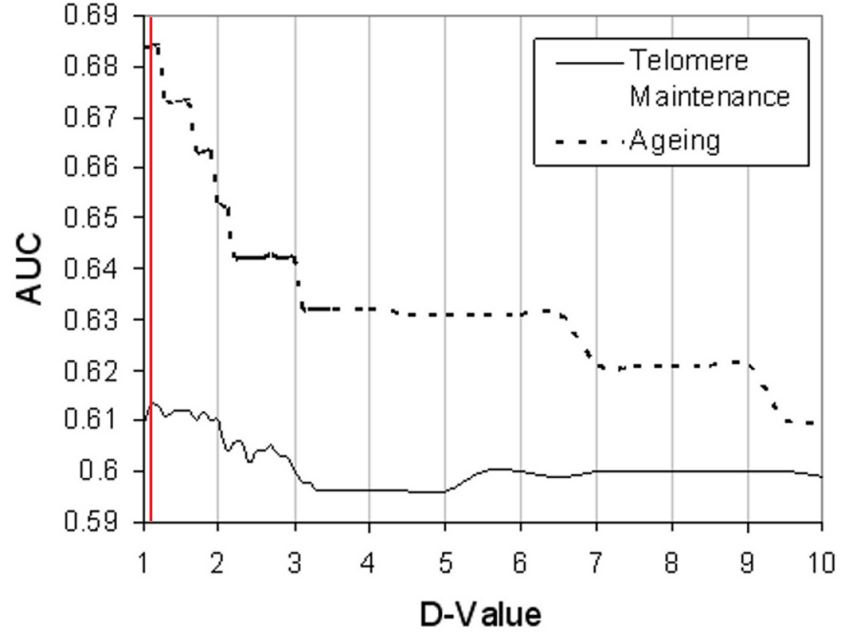


Figure 3.2: D-value choice.

Effect of the D parameter on the AUC for functional prediction of ageing and telomere maintenance. A value of 1.1 was shown to optimise the area under curve value.

3.1.4.7 Composite Integration

Two composite networks were produced. In the first edge weights were calculated for each interaction i using Equation 3.5 developed by Lycett [1141]. In the second the average edge weights were calculated for each interaction using Equation 3.6.

In both cases $WS(i)_{control}$, $WS(i)_{node}$, $WS(i)_{edge}$ and $WS(i)_{interaction}$ are the edge weights for interaction i in the control, Node Relevance, Edge Relevance and Interaction Relevance networks, respectively (see Section 6.4.2).

$$WS(i)_{final} = 1 - ((1 - WS(i)_{control})(1 - WS(i)_{node})(1 - WS(i)_{edge})(1 - WS(i)_{interaction})) \quad (3.5)$$

$$WS(i)_{final} = \frac{WS(i)_{control} + WS(i)_{node} + WS(i)_{edge} + WS(i)_{interaction}}{4} \quad (3.6)$$

3.1.5 Network Visualisation and Evaluation

3.1.5.1 Visualisation and Topological Analysis

Networks were visualised in Cytoscape¹⁰ Version 2.5 [641] and Ondex¹¹ Version 0.3 [80]. The network Analyser¹² plugin version 2.5.1 for Cytoscape was used to calculate topological statistics [400].

The shortest path between all pairs of nodes in the networks was calculated using Dijkstra's algorithm [1142].

3.1.5.2 Network Clustering

The networks were clustered using the Markov clustering algorithm (MCL) algorithm¹³ (see Section 2.3.3.2) [534]. The default inflation value of 1.8 was used in all cases. The clusters containing nodes annotated to the POI were identified in the relevance and control networks for visual comparison.

3.1.5.3 Functional Prediction

Functional prediction was carried out using the Maximum Weight decision rule [57] in which annotations were propagated along the highest weighted edge surrounding a node (see Section 2.5.5). Leave-one-out cross-validation of known annotations to the POI was carried out using this algorithm for both the control, relevance, reversed and composite networks.

Edge weight cut-offs were applied during functional prediction where stated. For combined predictions the maximum weight results for several, topologically identical, networks were calculated and the highest score selected for each network node (see Section 6.4.2).

3.1.5.4 Receiver Operator Characteristic

The performance of the networks was evaluated using ROC curves [920]. The AUC of the ROC curves was used to estimate network functional prediction performance levels. An AUC of 0.5 indicates the network has no predictive power while higher AUC values indicate increasing functional prediction performance, with perfect classification giving an AUC of 1.0.

¹⁰<http://cytoscape.org/>

¹¹<http://www.ondex.org/>

¹²http://chianti.ucsd.edu/cyto_web/plugins/displayplugininfo.php?name=NetworkAnalyzer

¹³<http://www.micans.org/mcl/>

In order to compare the relevance and control curves the error of the AUC was calculated using the standard error of the Wilcoxon statistic $SE(W)$ using Equation 3.7 [920, 921], where θ is the AUC, C_p is the number of positive examples, C_n is the number of negative examples and Q_1 and Q_2 are the probabilities of incorrect annotation assignment as defined by Equations 3.8 and 3.9, respectively.

$$SE(W) = \sqrt{\theta(1 - \theta) + (C_p - 1)(Q_1 - \theta^2) + (C_n - 1)(Q_2 - \theta^2)/C_p C_n} \quad (3.7)$$

$$Q_1 = \frac{\theta}{2 - \theta} \quad (3.8)$$

$$Q_2 = \frac{2\theta^2}{1 + \theta} \quad (3.9)$$

3.2 Laboratory Techniques

3.2.1 Strains and Growth Conditions

Two *Saccharomyces cerevisiae* strains were used in the analysis¹⁴ (see Section 7.5.1); a wild type (wt), BY4741 (*MATa his3Δ1 leu2Δ0 met15Δ0 ura3Δ0*), and, the *AIM1* deletion mutant, BY4741 *aim1Δ* (*MATa his3Δ1 leu2Δ0 met15Δ0 ura3Δ0 aim1::kanMX4*).

S. cerevisiae strains were grown under several different growth conditions. For fermenting growth the strains were grown in YPD medium (1% w/v yeast extract, 2% w/v bacto peptone and 2% w/v glucose). For growth under limited iron conditions (Fe^-) the strains were grown in YPD with 100 μM bathophenanthroline disulfonate (BPS). For growth under increased iron conditions (Fe^+) the strains were grown in YPD with 100 μM iron chloride (Fe^{2+}). All strains were incubated at 30°C in a rotatory shaker at 180 rpm unless stated otherwise.

3.2.2 DNA Extraction

Chromosomal DNA was extracted from the wild type and mutant strains using the "smash and grab" protocol developed by Hoffman [1143]. All reagents were supplied by Sigma unless otherwise stated. To prepare the cells, 10 ml cultures were grown in YPD (1% w/v yeast extract, 2% w/v

¹⁴<http://www.invitrogen.com/site/us/en/home.html>

bacto peptone and 2% w/v glucose) to stationary phase in sterile culture tubes. Cultures were centrifuged for 5 minutes in a table top centrifuge (Sigma) at room temperature, after which supernatant was aspirated. Cells were re-suspended in 0.5 ml water and microcentrifuged for 5 seconds at room temperature. The supernatant was removed and the pellet disrupted by vortexing briefly.

To break open the cells the pellet was re-suspended in 200 ml breaking buffer (2% (v/v) Triton X-100, 1% (v/v) sodium dodecyl sulphate, 100 mM NaCl, 10 mM Tris-Cl, pH 8.0, 1 mM EDTA, pH 8.0), then 0.3 g glass beads and 200 μ l phenol alcohol were added before cell disruption in a Precellys 24 disrupter (Bertin Technologies).

200 ml TE buffer pH 7.5 (10 mM Tris-HCl pH 7.4, 1mM EDTA pH 8) was then added and the solution vortexed briefly. Following microcentrifugation for 5 minutes at high speed the aqueous layer was transferred to a clean tube with 1 ml 100% ethanol and mixed by inversion. Finally, after microcentrifugation for 3 minutes at high speed the supernatant was removed and the pellet re-suspended in 0.4 ml of TE buffer.

To recover the DNA, 30 ml of 1 mg/ml DNase-free RNase A was added to the DNA solution, mixed and incubated for 5 minutes at 37°C. 10 ml of 4 M ammonium acetate and 1 ml of 100% ethanol were then added and mixed by inversion. After microcentrifugation at high speed, at room temperature for 3 minutes the supernatant was discarded and the DNA pellet dried before being re-suspended in 100 ml TE buffer, pH 7.5.

3.2.3 Polymerase Chain Reaction

The polymerase chain reaction (PCR) was carried out at a final volume of 50 μ l:

Kapa 2G Enzyme - Kapa Biosystems ¹⁵ (5 Units/ μ l)	0.5 μ l
Forward Primer (100 μ M)	0.5 μ l
Reverse Primer (100 μ M)	0.5 μ l
Template	0.5 μ l
Buffer	5 μ l
dNTPs	1 μ l
H ₂ O	42 μ l

¹⁵<http://www.kapabiosystems.com/>

Primers as described in Section 7.5.1 were obtained from Integrated DNA Technologies, Glasgow, UK¹⁶:

Forward primer 5' - CGA TGC TAT TCT CTT TTT GAT TCG TC -3'

Reverse primer 5' - GTG AGT AAC CAT GCA TCA TCA GG -3'

PCR was carried out in a T3 Thermocycler (Biometra¹⁷) with the following parameters:

- Step 1: 94°C 2 minutes
- Step 2: 94°C 30 seconds
- Step 3: 50°C 30 seconds
- Step 4: 72°C 1 minute
- Repeat Step 2-4 for 30 cycles
- Step 5: 72°C 10 minutes

3.2.4 Stress Sensitivity Tests

3.2.4.1 Oxidative Stress

For spot tests, the wild type and mutant strains were grown to the middle of logarithmic phase in YPD medium and diluted to an optical density of 0.2 at 600 nm before serial 10-fold dilution (1, 1/10, 1/100, 1/1000). A total of five microliters of each of the dilutions was spotted simultaneously using a 48-prong replicator (Sigma) onto solid YPD medium containing different concentrations of oxidative stress inducing compounds (Table 3.2). Plates were incubated at 30°C for 24 hours, and sensitivity was examined.

Table 3.2: Stress sensitivity testing.

Concentrations of oxidative stress inducing compounds used in the stress sensitivity tests.

Compound	Concentrations
tert-butyl hydroperoxide (tBOOH)	0.4 mM/0.6 mM/0.8 mM/1.0 mM
Diamide (Diazenedicarboxylic acid bis (N,N-dimethylamide))	1.5 mM/2.0 mM/2.5 mM/3.0 mM
Hydrogen Peroxide (H_2O_2)	0.5 mM/1.0 mM/1.5 mM/2.0 mM/2.5 mM/3.0 mM
Menadione (2-methylnaphthalene-1,4-dione)	50 μ M/100 μ M/150 μ M/200 μ M
Cadmium(II) Sulphate ($CdSO_4 \cdot \frac{8}{3}H_2O$)	0.08 mM/0.1 mM/0.12 mM
Sodium Arsenite ($NaAsO_2$)	1.0 mM/1.25 mM/1.5 mM/1.75 mM/2.0 mM

¹⁶<http://eu.idtdna.com/Home/Home.aspx>

¹⁷<http://www.biometra.de/>

3.2.4.2 Iron Response

For spot tests the wild type and mutant strains were grown to the middle of logarithmic phase (OD 0.2 AT 600 nm) in YPD medium at high, average and limited iron concentrations (Section 3.2.1). Serial 10-fold dilutions of each strain were made up to 1/1000.

For fermenting growth 5 ml of each of the dilutions were spotted simultaneously using a 48-prong replicator instrument (Sigma) onto solid YPD medium (1% w/v yeast extract, 2% w/v bacto peptone and 2% w/v glucose plus 2% w/v agar) containing various levels of iron:

- High iron: 100 μ M/500 μ M/1 mM iron sulphate
- Average iron: plain YPD
- Low iron: 100 μ M BPS

For growth under respiring conditions (aerobic) the cells were washed in water to remove glucose before serial 10-fold dilutions of each strain were made up to 1/1000. Each dilution was spotted simultaneously using a 48-prong replicator instrument (Sigma) onto solid YPG medium (1% w/v yeast extract, 2% w/v bacto peptone and 3% w/v glycerol plus 2% w/v agar) .

Plates were incubated at 30°C for 72 hours, and sensitivity was examined as above.

Chapter 4

Harnessing Process-Relevance During Network Integration

The aim of this study was to develop network integration and analysis techniques which harness the inherent biases in functional data in order to improve network performance in relation to specific biological processes. Heterogeneous data sources each contain information about different aspects of the cell. Therefore, integration of these diverse data sources can reveal new aspects of cellular processes that could not be seen in one source alone. Probabilistic Functional Integrated Networks (PFINs) are powerful integration tools, since they take the quality of individual datasets into account by confidence scoring prior to integration (see Section 2.4.3) [49, 115, 128]. However, current techniques ignore the content of the datasets. Individual studies each produce data with its own biases, (see Section 2.4.4). These differences are important and can be harnessed during network integration, in addition to the more conventional score of dataset confidence. Due to the nature of the biases some datasets may be more informative about certain areas of biology than others. In this thesis this property is referred to as *process-relevance*: the more informative a dataset is about a biological process, the higher its relevance is to that process.

In order to identify relevant data and produce process-relevant networks several questions need to be answered:

- What biases occur in functional data?
- How can relevance be quantified?
- How can relevance be incorporated during network integration?
- How can networks be evaluated in relation to process-relevance?

Prior to quantifying relevance, and incorporating it during data integration, it is essential to understand the nature of dataset biases. In the Section 4.1 of this chapter the differences and similarities between experimental types, and between individual studies of the same type, are investigated (Objective 1, Section 1.5). A dataset scoring and integration schema, RelCID, is presented in Section 4.2 which allows a dataset's relevance to a process to be quantified is then developed. Finally, the Section 4.2.1 of this chapter describes a comprehensive evaluation of the RelCID schema as it pertains to yeast ageing.

4.1 Harnessing Process Relevance

4.1.1 Source Data

Datasets form the basis of any integrated resource. However, defining what constitutes a single dataset is far from straightforward, and this definition can affect the performance of an integrated system. In PFINs the individual datasets are confidence-scored prior to integration. Therefore several aspects of source data choice and dataset definition can affect the final network:

- **Redundancy:** different databases can contain duplicate data, which can lead to biases within the network, upweighting edges with duplicate evidence.
- **Type:** different labelling schemas exist to divide data by experiment type and each schema may produce different final datasets.
- **Accuracy:** incorrect data can adversely influence dataset scoring and network performance.
- **Standardisation:** lack of unique identifiers can produce inaccurate mapping between datasets during integration.

Consequently, it is important that the chosen data are up-to-date and contains metadata that accurately describes its sources and methodologies. A large number of databases have been developed to store functional data (Section 2.1.4). In this study the BioGRID was chosen as the data source for a number of reasons. BioGRID is one of the most comprehensive databases available for the yeast *S. cerevisiae*, comprising 22 diverse data types, each with source metadata linking to the original publication [276]. The BioGRID dataset is manually curated to avoid errors and redundancy. The curators use standardised unique gene identifiers and supply synonyms for all genes and proteins. Importantly, the database curators actively encourage community feedback to identify incorrect data. Finally, the BioGRID data are available in a computationally amenable flat-file format (Table 4.1¹).

¹http://wiki.thebiogrid.org/doku.php/biogrid_tab_version_1.0

Table 4.1: The BioGRID flat file format.

BioGRID stores functional interaction data for *S. cerevisiae*, amongst other organisms. The flat-file format includes metadata for each interaction including gene synonyms, methodology and original publication details (http://wiki.thebiogrid.org/doku.php/biogrid_tab_version_1.0).

Column	Description
INTERACTOR_A	Unique ID for Interacting Partner A
INTERACTOR_B	Unique ID for Interacting Partner B
OFFICIAL_SYMBOL_FOR_A	Official name of Interacting Partner A
OFFICIAL_SYMBOL_FOR_B	Official name of Interacting Partner B
ALIASES_FOR_A	List of common names for geneA, separated by 'l'
ALIASES_FOR_B	List of common names for geneB, separated by 'l'
EXPERIMENTAL_SYSTEM	System in which the interaction was shown
SOURCE	Author/s of the interaction
PUBMED_ID	PubMed_ID of the paper, separated by ';'
ORGANISM_A_ID	NCBI ID of Gene A Organism
ORGANISM_B_ID	NCBI ID of Gene B Organism

4.1.2 Evaluation of Dataset Bias

In order to harness inherent dataset bias it is essential to understand how datasets differ, and the nature of these differences. The BioGRID dataset for *S. cerevisiae* can be subdivided at three different levels: by interaction type; by experimental type; or by individual study.

At the most abstract level the database distinguishes between physical and genetic interactions. The two interaction types are then classified according to 22 evidence types (Table 4.2 and Appendix D). Finally, at the lowest level, the data can be subdivided by [PMID](#)² into individual studies. Each individual [PMID](#) indicates a single study designed to analyse a specific biological question. These studies may range in size from small-scale experiments of a few interactions to genome-wide screens comprising hundreds, or even thousands, of interactions.

The datasets were evaluated and compared using three criteria;

- **Genomic coverage:** the coverage of the yeast genome in terms of the numbers of individual genes.
- **Interactome coverage:** the coverage of the interaction space in terms of the numbers of individual interactions.
- **Biological coverage:** the coverage of biological processes in terms of numbers of GO annotations.

²<http://www.ncbi.nlm.nih.gov/pubmed>

Table 4.2: The BioGRID experimental types.

The BioGRID data can be subdivided into 22 experimental types: 14 physical types and 8 genetic types. Full experimental definitions are supplied in Appendix D.

Physical Interactions		Genetic Interaction
Affinity Capture-MS	Far Western	Dosage Growth Defect
Affinity Capture-RNA	FRET	Dosage Lethality
Affinity Capture-Western	Protein-Peptide	Dosage Rescue
Biochemical Activity	Protein-RNA	Phenotypic Enhancement
Co-Crystal Structure	Reconstituted Complex	Phenotypic Suppression
Co-Fractionation	Two-Hybrid	Synthetic Growth Defect
Co-Localization		Synthetic Lethality
Co-Purification		Synthetic Rescue

Genomic and interactome coverage were assessed by gene identifier comparison. For biological coverage over-representation of GOBP annotations was calculated using the GOstats R package (see Section 3.1.2) [862]. The scale of the data produced by GOstats is beyond the scope of a full discussion in this section. Therefore, some specific examples are presented. The full enrichment analysis results are available in Appendix E.

4.1.3 Results

4.1.3.1 Interaction Type

At the highest level of organisation BioGRID distinguishes between two interaction types: genetic and physical. The genomic and interactomic coverage of these types differ. Version 27 of BioGRID contains 73029 interactions involving 5424 genes. Over 65% of the genes are present in both types of interaction. However, the overlap of interactions between the two data types is significantly lower at approximately 3% (Table 4.3).

Table 4.3: Genetic and physical interactions.

The genes and interactions represented by the physical and genetic interaction types of BioGRID.

	Genes	Pairs
Physical interactions only	1729	38259
Genetic interactions only	139	32508
Physical & genetic interactions	3556	2262
Percentage overlap	65.56%	3.10%
Total	5424	73029

The GOstats R package was used to assess the biological coverage of the two interaction types [862]. GOstats applies a hypergeometric test to calculate over-representation of GO terms in the gene annotations of the dataset (see Section 3.1.2). Of the top ten over-represented terms, eight were common to both the genetic and physical datasets (Table 4.4). Notably the majority of the over-represented

terms were general terms with little specificity, such as high-level metabolic and regulatory processes. The GOSTats analysis was repeated excluding those genes present in the overlap of the physical and genetic datasets. When the overlapping genes were excluded fewer terms were over-represented and the majority of terms were lower level terms, with only one term in common (**cellular process**) between the interaction types (Table 4.5). Moreover, there were distinct areas of biology observed in the two datasets. The genes with physical interactions had high representation of metabolic and biosynthetic processes, particularly those involving proteins and amino acids. Conversely, the genes with genetic interactions were over-represented for processes involving nucleic acids.

Table 4.4: GO biological process enrichment for genetic and physical genes.

The GOSTats output for the top ten over-represented GOBP terms for the genetic and physical interaction types. In total there were 138 and 50 terms over-represented for the genetic and physical datasets, respectively. Terms in common are highlighted in bold. Full enrichment results are supplied in are supplied in Appendix E.

Genetic Interaction			
GOBPID	Pvalue	OddsRatio	Term
GO:0065007	5.84E-72	6.090048	Biological regulation
GO:0051641	2.07E-50	5.816283	Cellular localization
GO:0046907	4.73E-46	5.650174	Intracellular transport
GO:0006996	2.98E-44	3.083049	Organelle organization and biogenesis
GO:0044238	1.96E-40	2.370941	Primary metabolic process
GO:0044260	3.16E-40	2.641645	Cellular macromolecule metabolic process
GO:0051234	1.40E-38	2.91872	Establishment of localization
GO:0019538	5.69E-36	2.481291	Protein metabolic process
GO:0031323	4.85E-34	4.518281	Regulation of cellular metabolic process
GO:0050896	2.17E-32	5.51017	Response to stimulus
Physical Interaction			
GOBPID	Pvalue	OddsRatio	Term
GO:0044260	2.95E-30	5.755886	Cellular macromolecule metabolic process
GO:0019538	1.34E-29	5.553653	Protein metabolic process
GO:0044238	4.40E-25	3.453662	Primary metabolic process
GO:0016070	4.55E-21	4.870293	RNA metabolic process
GO:0051234	5.78E-19	4.049084	Establishment of localization
GO:0065007	6.93E-19	5.149184	Biological regulation
GO:0051641	7.24E-19	9.842371	Cellular localization
GO:0044237	3.55E-18	2.77375	Cellular metabolic process
GO:0006996	6.60E-17	4.774281	Organelle organization and biogenesis
GO:0022402	1.38E-13	9.546076	Cell cycle process

Table 4.5: GO biological process term enrichment for the unique genetic and physical genes.
The GOSTats output for the top ten over-represented GOBP terms for the unique genetic and physical interaction types. In total there were 22 and 17 terms over-represented for the genetic and physical datasets, respectively. The term in common is highlighted in bold. Full enrichment results are supplied in Appendix E.

Unique Genetic			
GOBPID	Pvalue	OddsRatio	Term
GO:0009987	2.50E-28	3.390005	Cellular process
GO:0044255	7.27E-13	3.250492	Cellular lipid metabolic process
GO:0009058	3.69E-12	1.96516	Biosynthetic process
GO:0044271	8.71E-12	4.399649	Nitrogen compound biosynthetic process
GO:0009308	1.09E-10	2.919304	Amine metabolic process
GO:0006766	3.18E-10	4.598552	Vitamin metabolic process
GO:0006519	7.25E-10	2.922348	Amino acid and derivative metabolic process
GO:0019752	7.39E-10	2.466605	Carboxylic acid metabolic process
GO:0008652	8.81E-10	4.025441	Amino acid biosynthetic process
GO:0006811	1.89E-09	3.790611	Ion transport
Unique Physical			
GOBPID	Pvalue	OddsRatio	Term
GO:0006412	7.26E-20	3.159723	Translation
GO:0030490	1.72E-15	12.962713	Processing of 20S pre-rRNA
GO:0022613	3.99E-13	2.701273	Ribonucleoprotein complex biogenesis and assembly
GO:0019538	1.21E-12	1.840102	Protein metabolic process
GO:0044249	2.17E-12	1.962804	Cellular biosynthetic process
GO:0008152	2.35E-12	1.704595	Metabolic process
GO:0006365	1.02E-09	4.583565	35S primary transcript processing
GO:0044260	1.03E-09	1.693187	Cellular macromolecule metabolic process
GO:0009987	5.33E-09	1.711193	Cellular process
GO:0006364	7.16E-08	4.699056	rRNA processing

4.1.3.2 Experimental Type

BioGRID is divided into 22 experimental types. Figure 4.1 depicts a partial hierarchical clustering of these evidence types by genomic coverage (see Section 3.1.3). Although the physical and genetic interaction types tended to cluster together, there were distinct subgroups. Due to the number of potential pairs in the yeast interactome a similar clustering of the datasets by interactomic coverage was not computationally feasible. However, visual analysis of the common interactions using Cytoscape³ revealed different patterns of overlap from those observed for genomic coverage. In particular, several physical data types had higher overlap with genetic data types than with other physical types (Figure 4.2). For instance, the two datasets in pair A of Figure 4.1, Affinity Capture-Western and Reconstituted Complex, had low overlap in terms of interactions. In fact, the Affinity Capture-Western dataset had significantly more interactions in common with several other datasets than with the Reconstituted Complex dataset.

³<http://www.cytoscape.org/>

GOstats analysis of the datasets' biological coverage produced a varying number of common terms between closely clustered datasets. Several of the over-represented terms reflected the type of experimental method applied. For instance, the two RNA-based datasets of Group F in Figure 4.2 had high enrichment of RNA-associated terms. Due to the scale of the data produced by GOstats six clustered pairs were chosen for further analysis (A-F in Figure 4.1). The top five over-represented terms for the example dataset pairs are shown in Table 4.6. In many cases the overlap for the clustered datasets was limited to general, high-level processes, while specific low-level terms were unique to the individual data types.

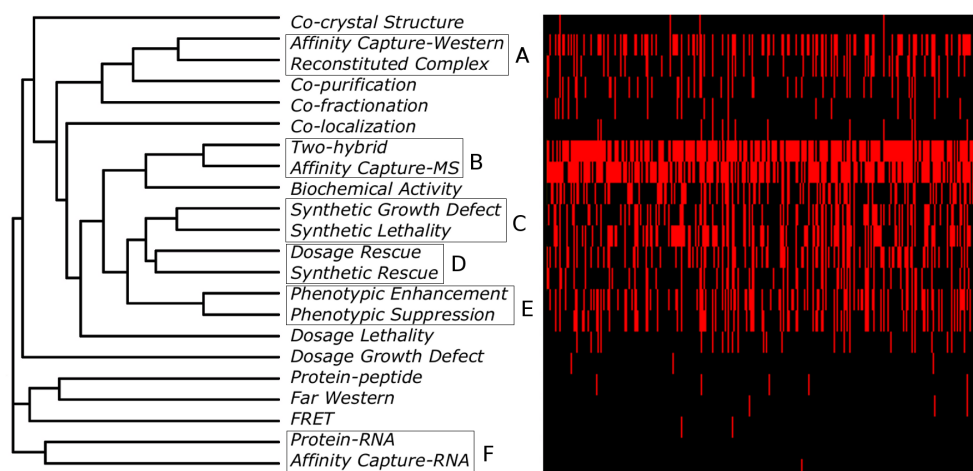


Figure 4.1: Genomic coverage of the experimental datatypes.

Partial heatmap of the data types clustered by genomic coverage. The six dataset pairs (A-F) used as biological coverage examples are highlighted in colour.

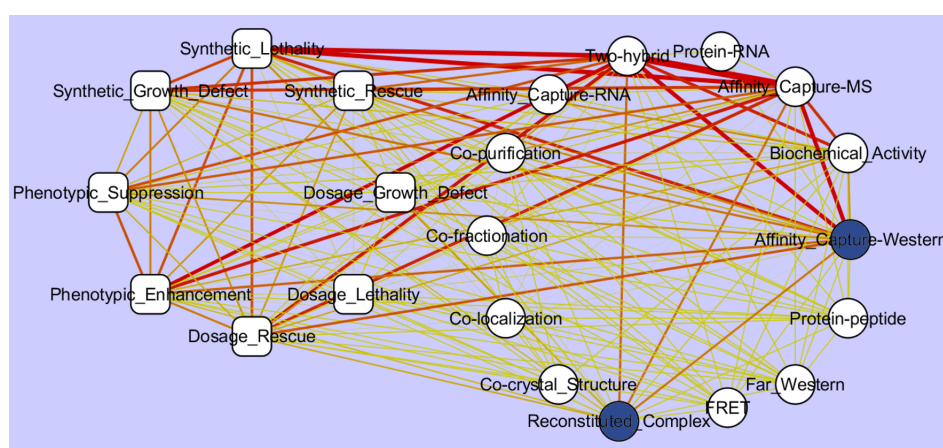


Figure 4.2: Interactome coverage of the experimental datatypes.

Genetic data types are shown as squares and the physical data types as circles. The line widths represent the number of shared interactions. The line colours vary from yellow (few interactions in common) to red (many interactions in common). The datasets of pair A in Figure 4.1 (highlighted in blue) clustered together by genomic coverage. However, their interactomic overlap is low. In particular the Affinity Capture Western dataset has significantly more interactions in common with several other datasets than with the Reconstituted Complex dataset.

Table 4.6: The top five over-represented GO terms in pairs A-F of Figure 4.1.

The terms are displayed in order of over-representation. Terms in common between the pairs of datasets are highlighted in bold. In the case of the Protein-RNA dataset only two GOBP terms were over-represented. The full results including scores are presented in Appendix E.

Pair A	
Affinity Capture Western	Reconstituted Complex
Biopolymer metabolic process	Organelle organization and biogenesis
Cellular localization	Biological regulation
Organelle organization and biogenesis	Cellular localization
Biological regulation	Mitotic cell cycle
Regulation of cellular process	Response to DNA damage stimulus
Pair B	
Two Hybrid	Affinity Capture Mass Spectroscopy
Localization	Translation
Establishment of cellular localization	Biological regulation
Biological regulation	Chromosome organization and biogenesis (sensu Eukaryota)
Biopolymer metabolic process	Cellular process
Response to stimulus	Macromolecule metabolic process
Pair C	
Synthetic Growth Defect	Synthetic Lethality
Organelle organization and biogenesis	Chromosome organization and biogenesis (sensu Eukaryota)
Response to stimulus	Organelle organization and biogenesis
Telomere maintenance	Metabolic process
Chromosome organization and biogenesis (sensu Eukaryota)	Biological regulation
DNA metabolic process	Cellular localization
Pair D	
Dosage Rescue	Synthetic Rescue
Cellular localization	Regulation of cellular process
Secretion	Biological regulation
Biological regulation	Primary metabolic process
Anatomical structure development	Transcription
Regulation of cellular process	RNA biosynthetic process
Pair E	
Phenotypic Enhancement	Phenotypic Suppression
Biological regulation	Biological regulation
Response to stimulus	Regulation of cellular process
Regulation of cellular metabolic process	DNA metabolic process
Organelle organization and biogenesis	Response to stimulus
Secretion	Organelle organization and biogenesis
Pair F	
Affinity Capture RNA	Protein RNA
Nuclear mRNA splicing, via spliceosome	Group I intron splicing
RNA splicing, via transesterification reactions	RNA metabolic process
mRNA catabolic process	
Biopolymer catabolic process	
Cellular macromolecule catabolic process	

4.1.3.3 Experimental Scale

At the lowest level the BioGRID data can be split into individual studies by PMID⁴. However, the majority of these studies are small-scale and contain very few interactions. Datasets of this size are too small to treat as a single data source, since reliable confidence scoring against a Gold Standard would not be possible due to the scale difference [128]. That is, the size of the dataset would be too small in comparison to the Gold Standard's size for accurate confidence assessment since no Gold Standard has complete coverage of the genome. Therefore, a cut-off was chosen to distinguish between large-scale and small-scale data. Figure 4.3 shows the distribution of dataset size within Version 27 of BioGRID. Approximately 1% of the datasets contained greater than 100 interactions. These large-scale studies were treated as single datasets. The remaining datasets were grouped by experimental method using the BioGRID evidence types (Table 4.2) since they were too small for reliable confidence scoring individually. In fact, approximately 90% of the small-scale studies contained less than ten interactions. This cut-off of 100 interactions has been used in previous studies to distinguish between large-scale HTP and small-scale LTP BioGRID data [133, 1141].

When clustering of the 22 experimental types by genomic coverage was repeated in the absence of the HTP data, a markedly different clustering pattern was observed (Figure 4.4). Similarly, the number of over-represented GO terms for the 22 data types was significantly different when the HTP data was excluded (Figure 4.5).

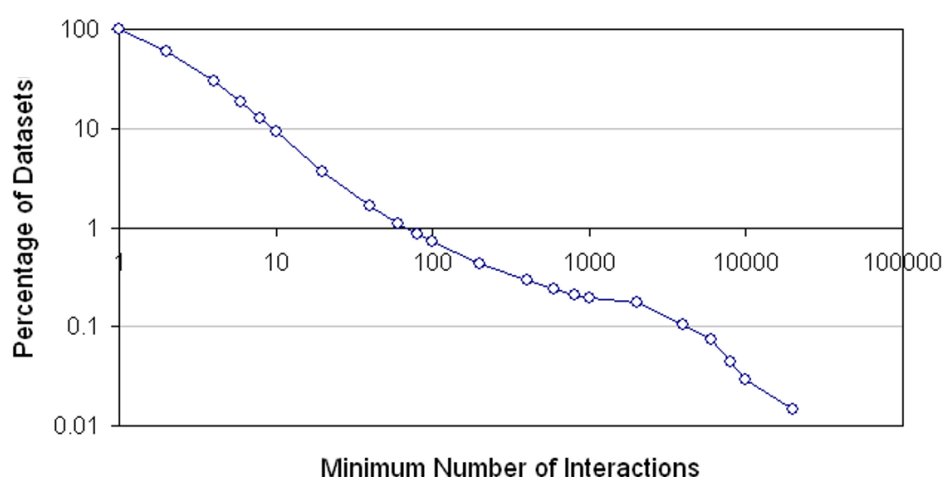


Figure 4.3: BioGRID dataset size.

The distribution of BioGRID dataset size by experimental study as defined by PMID. Approximately 1% of the datasets have >100 interactions. A cut-off was applied at this level to treat studies with 100 or more interactions as separate HTP datasets, and smaller studies as grouped LTP studies based on BioGRID experimental type.

⁴<http://www.ncbi.nlm.nih.gov/pubmed>

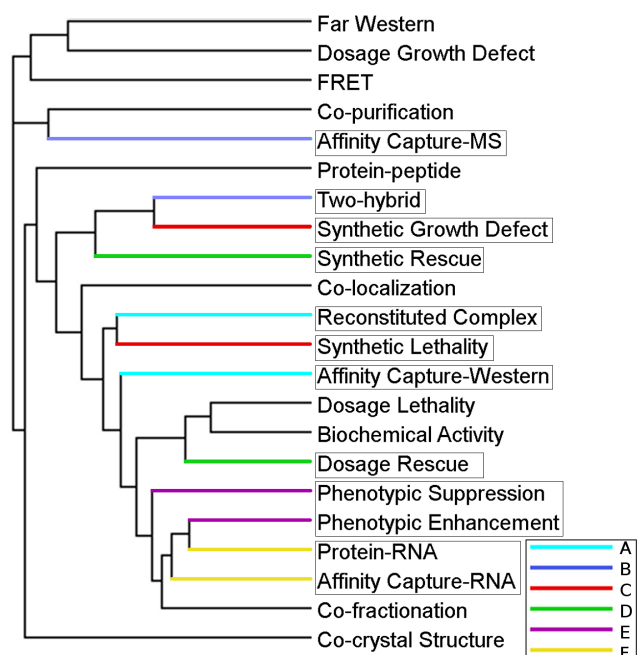


Figure 4.4: The LTP data clustered by genomic coverage.

The dataset clustering was significantly different in the absence of the HTP data. The dataset pairs (A-F) of Figure 4.1 are highlighted.

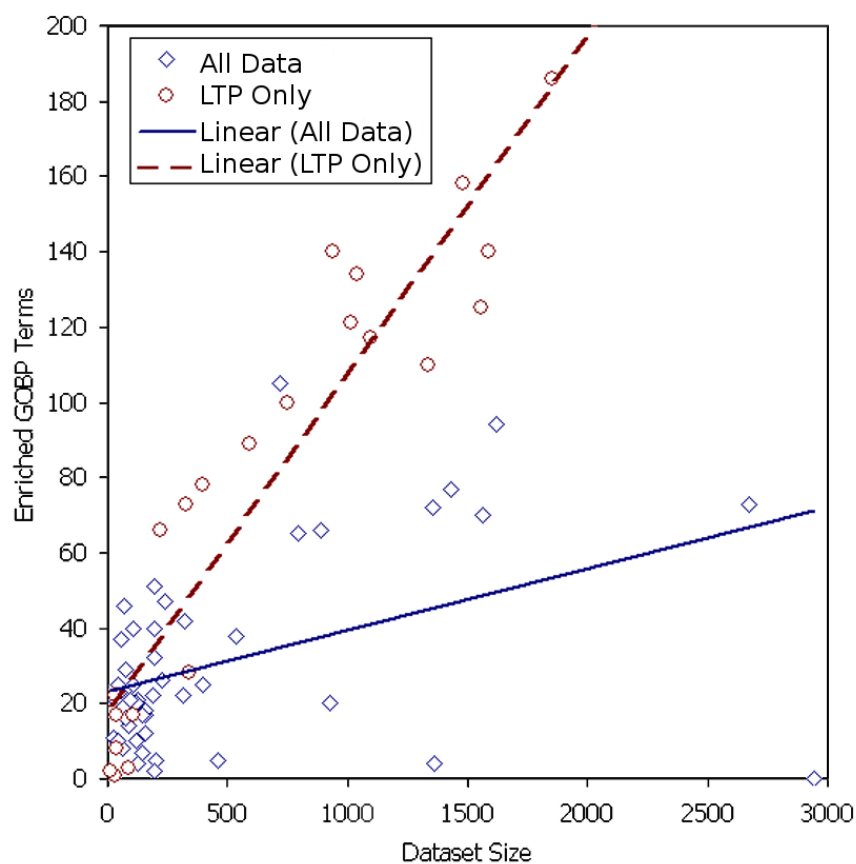


Figure 4.5: Over-represented GO terms.

The number of over-represented terms for the 22 data types. The number of over-represented terms is significantly higher when the HTP data are excluded.

4.1.3.4 Multiple Data Types

Several of the [HTP](#) studies contained multiple evidence types (Table 4.7). Multiple data types complicate the definition of a single dataset since the data may be treated in two ways. On one hand these datasets are a single study, while on the other they may be sub-divided by evidence type. To address this issue it was necessary to refer to the original publications.

In the majority of cases these datasets were primarily of one type (>70% of the interactions). In some cases a second experimental method was used to verify data following a large-scale study. For instance in the study by Sanders and colleagues (Sanders.12052880) [HTP](#) Affinity Capture-[MS](#) was used to detect 480 interactions [1144]. Three of the interactions were then confirmed by a [LTP](#) Affinity Capture Western technique. Mixed datasets of this type were treated as single datasets.

A number of datasets with multiple evidence types were genetic interaction analyses. The multiple data types produced by these studies reflect the nature of genetic interaction detection. For instance, phenotypic experiments, such as that of Collins and co-workers (Collins.17314980) [244], can detect both phenotypic enhancement and phenotypic suppression interactions. These studies were all treated as single datasets.

Two of the studies, Pan.15525520 and Zhao.15766533, did not have a clear majority (>70% interactions) of evidence type. The first study used [dSLAM](#) to detect several types of genetic interactions including synthetic lethals, rescues and growth defects [15]. These data were treated as a single dataset, since all the interactions were detected using the same experimental technique. The second study used a mixture of physical and genetic interaction detection techniques to investigate the chaperone Hsp90 [1145]. Due to the specific focus of this study, the data generated were also treated as a single dataset in order to preserve its unique biases.

4.1.3.5 Individual Studies

The final division of Version 27 of BioGRID at the 100 interaction cut-off produced 70 datasets: 22 [LTP](#) and 48 [HTP](#). Figure 4.6 shows the hierarchical clustering of the datasets by genome coverage. Several of the large-scale studies clustered together by experimental type, but there were distinct clusters of a single type of data. For instance the Affinity Capture-[MS](#) datasets formed a large tightly clustered group of six datasets in the tree (Figure 4.6 cluster 1A), and several separate smaller clusters (Figure 4.6 clusters 1B-1C). In addition, one dataset of this type (Allen.11387327) clustered separately with the FRET dataset (Figure 4.6 cluster 1D).

Similarly, the genetic interaction data types clustered together into three main groups (Figure 4.6 clusters 2A-2C). Finally, there are several clusters of mixed type, such as the cluster containing the

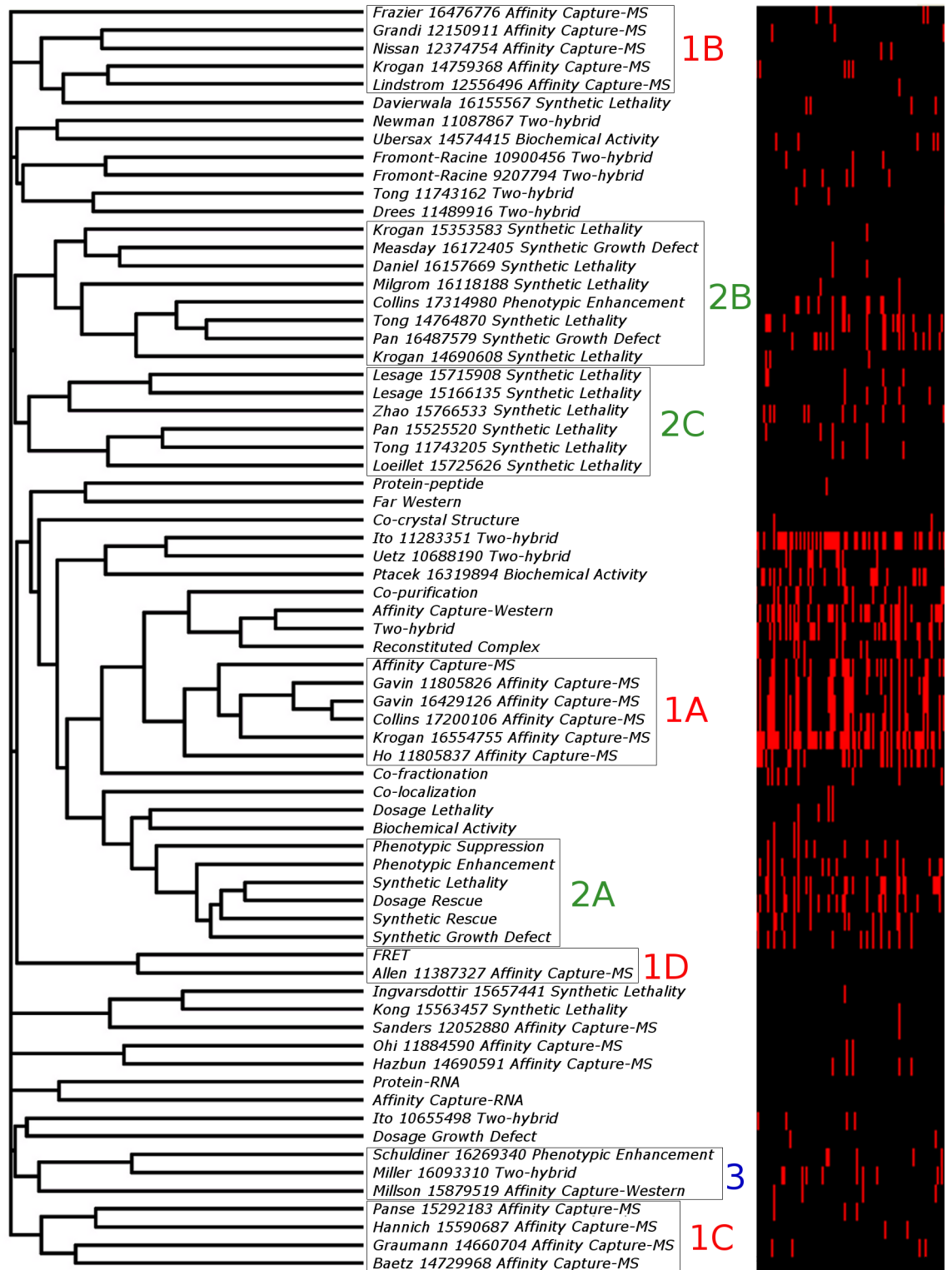


Figure 4.6: Genomic coverage of the individual studies.

A partial heatmap of the 70 datasets clustered by genomic coverage using the same method as that for Figure 4.1. Datasets of the same type tend to cluster together, for instance the Affinity Capture MS datasets (1A-1C) and the genetic interaction types (2A-2C). However, there are clusters that contain multiple data types (1D & 3).

Table 4.7: Multiple data types.

Seventeen of the HTP datasets contain multiple experimental types. The major experimental type and its coverage of the dataset in terms of nodes and interactions are summarised.

Dataset	Type	Major SubType	Nodes (%)	Interactions (%)	Ref
Pan.15525520	Genetic	Synthetic Lethality	57.3	55.7	[15]
Collins.17314980	Genetic	Phenotypic Enhancement	99.3	80.5	[244]
Schuldiner.16269340	Genetic	Phenotypic Enhancement	97.8	79.1	[17]
Pan.16487579	Genetic	Synthetic Growth Defect	92.6	81.2	[243]
Frazier.16476776	Physical	Affinity Capture-MS	88.9	93.2	[1146]
Sanders.12052880	Physical	Affinity Capture-MS	100	100	[1144]
Hazbun.14690591	Physical	Affinity Capture-MS	91.1	90.9	[64]
Measday.16172405	Multi	Synthetic Growth Defect	96.6	81.4	[1147]
Krogan.15353583	Multi	Synthetic Lethality	90.7	90.1	[1148]
Ingvarsdottir.15657441	Multi	Synthetic Lethality	77.2	68.1	[1149]
Zhao.15766533	Multi	Synthetic Lethality	59.6	58.6	[1145]
Tong.11743162	Multi	Two Hybrid	100	99.6	[1150]
Krogan.14690608	Multi	Synthetic Lethality	95.3	94.8	[1151]
Lindstrom.12556496	Multi	Affinity Capture-MS	100	98.1	[1152]
Kong.15563457	Multi	Synthetic Lethality	96.8	99.1	[1153]
Hannich.15590687	Multi	Affinity Capture-MS	94.2	92.9	[1154]
Millson.15879519	Multi	Affinity Capture Western	99.2	99.2	[1155]

Schuldiner.16269340, Miller.16093310 and Milson.15879519 datasets (Figure 4.6 cluster 3).

Due to the number of interactions, hierarchical clustering of the datasets by interactomic coverage was not computationally feasible. However, the overlap between datasets in terms of interactions could be assessed. The datasets varied in size and had large numbers of overlapping interactions. Therefore, the percentage overlap between pairs of datasets was calculated (Figure 4.7). In order to directly compare genomic and interactomic coverage the datasets were ordered as they clustered in Figure 4.6.

The majority of dataset clusters had higher interaction overlap within the cluster than with other clusters in the matrix. In particular, the densely connected Affinity Capture-MS datasets had a high degree of overlap (Figure 4.7 F). However, there were several exceptions. For instance, two Affinity Capture-MS genomic coverage clusters (Figure 4.7 A and B) had very low interaction coverage overlap within the cluster but high overlap with a third Affinity Capture-MS cluster (Figure 4.7 H). Visualisation in Cytoscape revealed that these two groups had very few interactions in common within their clusters in comparison to their overlap with cluster F (Figure 4.8). In addition, three genetic datasets also had higher interactomic overlap with cluster F than within their own cluster (Figure 4.7 I).

GOstats analysis of the datasets' biological coverage produced a varying number of over-represented biological processes for the clustered datasets. In many cases the over-represented terms for the individual studies were more specific than those of the experimental types (see Appendix E). Additionally, the over-represented terms for the datasets tended to reflect the genomic coverage clustering

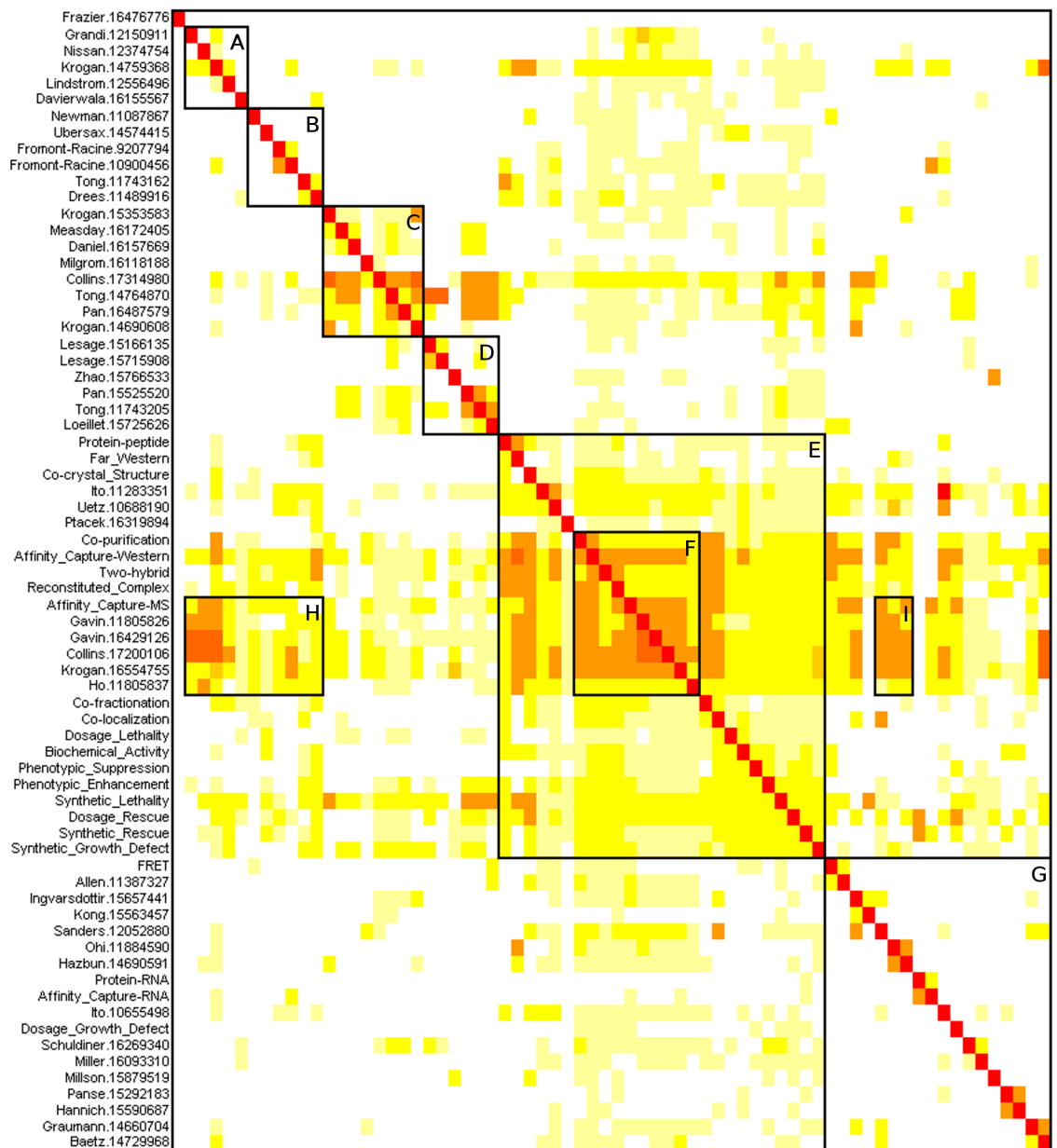


Figure 4.7: Genomic coverage of the individual studies.

The percentage overlap of the datasets in terms of individual interactions coloured from white (0% overlap) to red (100% overlap). The matrix is not symmetrical, as it reflects the differing sizes of the dataset pairs. The datasets are ordered as they clustered by genomic coverage in Figure 4.6 . The clusters genomic coverage clusters tend to have high interaction coverage overlap (C-F). However, several clusters have low overlap (A-B, G). Additionally there are two areas of significant overlap between clusters (H-I).

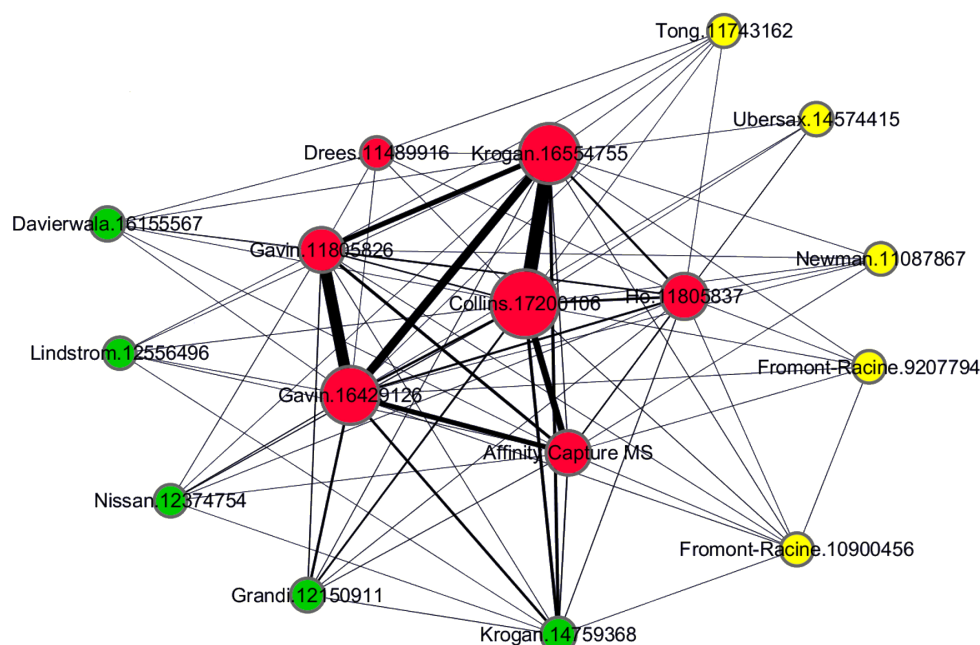


Figure 4.8: Interaction overlap of clusters A, B and F.

Cluster A is coloured green, Cluster B yellow and Cluster F red. Edge weights reflect the number of interactions in common, with thick edges indicating high overlap and thin edges indicating low overlap. Clusters A and B have high genomic overlap within their datasets. However, interaction overlap within these clusters is very low and many of the datasets do not have any interactions in common. However, both clusters have high interaction overlap with cluster F.

to a greater extent than that for the experimental types (Section 4.1.3.2). Of particular note were two clusters which contained mixed data types. The first cluster contained the FRET and Allen.11387327 datasets (Figure 4.6 cluster 1D). The over-represented terms for these datasets were very similar and contained a large number of nucleic acid- and nuclear transport-related terms (Table 4.8). These terms were consistent with the focus of the Allen.11387327 study which analysed the physical interactions involved in the nuclear pore complex [1156].

The second cluster of mixed data types contained three datasets: Schuldiner.16269340, Miller.16093310 and Milson.15879519 (Figure 4.6 cluster 3). The over-represented terms for these datasets were very similar and contained a high proportion of transport related terms (Table 4.9). The terms reflected the focus of the original experimental studies; Schuldiner and colleagues used E-MAPs to find GIs involved in the yeast early secretory pathway [17], Miller and colleagues performed a Y2H analysis of membrane proteins [1157] and Milson and colleagues used TAP to study the cytoplasmic chaperone Hsp90 [1155]. Each of these areas of cellular biology involves an aspect of cellular transportation. Notably, the dataset produced by Zhao and co-workers [1145], which also focused on Hsp90, did not cluster with the Milson dataset, but did share enrichment of the GO term protein folding (GO:0006457) which is consistent with their shared experimental subject (Table 4.10).

Table 4.8: GO biological process term enrichment for the datasets in cluster 1C.

The top five over-represented GOBP terms for the FRET and Allen.11387327 datasets. Full enrichment results are supplied in are supplied in Appendix E.

FRET			
GOBPID	Pvalue	OddsRatio	Term
GO:0006408	4.63E-25	375.05	snRNA export from nucleus
GO:0006610	4.63E-25	375.05	Ribosomal protein import into nucleus
GO:0006607	4.63E-25	375.05	NLS-bearing substrate import into nucleus
GO:0006608	4.63E-25	375.05	snRNP protein import into nucleus
GO:0006609	2.09E-24	312.433333	mRNA-binding (hnRNP) protein import into nucleus
Allen.11387327			
GOBPID	Pvalue	OddsRatio	Term
GO:0006913	2.33E-47	137.038352	Nucleocytoplasmic transport
GO:0051169	3.92E-44	121.659973	Nuclear transport
GO:0050658	1.46E-31	90.059671	RNA transport
GO:0006609	3.76E-31	370.587121	mRNA-binding (hnRNP) protein import into nucleus
GO:0015931	2.72E-30	77.07231	Nucleobase, nucleoside, nucleotide and nucleic acid transport

Table 4.9: GO biological process term enrichment for the datasets in cluster 3.

The top five over-represented GOBP terms for the Schuldiner.16269340, Miller.16093310 and Milson.15879519 datasets. In the case of the Milson dataset only four terms were over-represented. Full enrichment results are supplied in are supplied in Appendix E.

Schuldiner.16269340			
GOBPID	Pvalue	OddsRatio	Term
GO:0044255	9.41E-34	8.25	Cellular lipid metabolic process
GO:0016192	1.11E-31	6.18	Vesicle-mediated transport
GO:0006888	2.39E-30	18.51	ER to Golgi vesicle-mediated transport
GO:0051234	7.04E-24	3.24	Establishment of localization
GO:0043413	3.56E-20	12.39	Biopolymer glycosylation
Miller.16093310			
GOBPID	Pvalue	OddsRatio	Term
GO:0051179	9.74E-89	11.2	Localization
GO:0006810	3.05E-74	10.77	Transport
GO:0009100	4.98E-25	13.09	Glycoprotein metabolic process
GO:0006812	2.70E-24	10.7	Cation transport
GO:0006865	9.83E-17	25.11	Amino acid transport
Milson.15879519			
GOBPID	Pvalue	OddsRatio	Term
GO:0015849	9.63E-08	11.87	Organic acid transport
GO:0006457	1.67E-07	8.33	Protein folding
GO:0051234	3.07E-07	2.79	Establishment of localization
GO:0006865	7.09E-07	13.7	Amino acid transport

4.1.4 Discussion

Section 4.1.3 fulfils the first and second objectives of this project (see Section 1.5). Data of different types show clear differences in genomic, interactomic and biological process coverage. In particular, interactomic coverage is significantly different from genomic and biological coverage at all three levels of dataset division, while genomic coverage and biological coverage are similar. This pattern of similarity suggests that the genes represented by a dataset, rather than the interactions detected, are an indication of the dataset's biases.

At the highest level, the genetic and physical interactions comprise two overlapping groups of genes. Physical interactions outnumber the purely genetic interactions, with the majority of genes being involved in both types of interaction. However, while the genomic coverage is relatively high, overlap in terms of interactomic coverage is small. Unsurprisingly, physical interactions have bias towards protein-related processes while genetic interactions have bias towards nucleic acid-related areas of biology. These distinctions reflect the differing nature of the two interaction types since proteins with GIs are less likely to interact physically (see Section 2.1.2.3) [200].

Splitting the data at this relatively high level reveals differences in data coverage which can be of use during network, and other integrated, analyses. For instance, the 139 genetically interacting proteins represent different areas of biology than the physically interacting group. Therefore, these genes would probably be of less relevance to the analysis of the physical components of the cell, such as the cell membranes. However, as biological data are incomplete and noisy the differences in coverage observed may not represent the true ratio between genetic and physical interactions in the cell. It is therefore necessary to split the data at a higher level of specificity.

Division of the data by experimental type reveals deeper biases, many of which reflect the underlying experimental method. For instance, RNA-related terms are over-represented for RNA-based experimental types. In the absence of the HTP data the experimental types cluster differently. In particular, the two different interaction types-genetic and physical-are not clearly grouped. Additionally, the

Table 4.10: GO biological process term enrichment for the Zhao dataset.

The five over-represented GOBP terms for the Zhao.15766533 dataset. Full enrichment results are supplied in are supplied in Appendix E.

GOBPID	Pvalue	OddsRatio	Term
GO:0000723	1.52E-09	2.95	Telomere maintenance
GO:0009987	5.80E-09	2.07	Cellular process
GO:0006996	4.59E-08	1.8	Organelle organization and biogenesis
GO:0007001	8.78E-08	2.12	Chromosome organization and biogenesis (sensu Eukaryota)
GO:0006457	4.28E-06	3.84	Protein folding

removal of the [HTP](#) data significantly increases the number of over-represented terms. These results suggest that [HTP](#) data heavily influences the biases of the experimental type and may mask the underlying biases of the [LTP](#) data.

These [HTP](#) studies should therefore be treated as individual datasets, revealing the biases of experimental design and focus. Many datasets of the same type cluster together, reflecting the similarities in their experimental method. However, individual studies are tailored to answer specific questions and these questions are reflected in the genomic coverage clustering of some datasets. The biological process coverage at this level has little overlap since many of the over-represented terms are low-level, specific terms. However, similar areas of biology are observed for some dataset groups and this aspect of the datasets is reflected in the genomic clustering (Figure 4.6). For instance the Schuldiner, Miller and Milson datasets cluster together and have enrichment of transport-related terms, reflecting their experimental focuses. Therefore, division of the data into individual studies in this way reveals the specific biases of experimental design and focus.

Splitting the data into [HTP](#) studies (>100 interactions) and grouping the [LTP](#) data by experimental type allows dataset relevance to be assessed, while maintaining an adequate dataset size for confidence scoring. GOSTats uses a hypergeometric test to identify over-representation of biological process annotations by testing all possible GO terms. The test produces a p-value for the over-representation of each GO term. This value provides an ideal method with which to quantify dataset relevance in relation to individual biological questions and therefore fulfils Objective 2 of this project (see Section 1.5). For instance the Schuldiner, Miller and Milson studies have high over-representation of terms involved in cellular transport and, therefore, appear to have high relevance to this process. Consequently a hypergeometric test may be used to measure dataset relevance to specific processes. However, unlike GOSTats, which tests all GO terms at the same time, the test can be used to measure enrichment of a single term, reducing computational time and complexity. This measure of dataset relevance can then be incorporated during network integration, as described below.

4.2 The Integration RelCID Schema

The hypergeometric test produces a score of relevance to a process of interest ([POI](#)) between zero and one, where zero represents high relevance and one represents low relevance to the [POI](#) (Section 3.1.4.3). These relevance scores allow the datasets to be ranked in order of relevance to the [POI](#) (Figure 4.9).

Rank	Dataset	Relevance
1	Pan.16487579	5.69E-43
2	Synthetic_Lethality	1.01E-31
3	Synthetic_Growth_Defect	2.58E-28
4	Phenotypic_Suppression	6.06E-28
5	Tong.14764870	5.88E-25
6	Ye.16729061	1.89E-23
7	Collins.17314980	1.46E-21
8	Affinity_Capture-Western	2.06E-21
9	Synthetic_Rescue	3.93E-19
10	Phenotypic_Enhancement	2.22E-16
11	Loeillet.15725626	3.15E-16
12	Milgrom.16118188	6.07E-16
13	Collins.17200106	7.99E-16
14	Dosage_Rescue	2.28E-13
15	Krogan.14690608	1.21E-12
16	Pan.15525520	6.65E-12
17	Affinity_Capture-MS	1.45E-11
18	Krogan.16554755	2.46E-11
19	Two-hybrid	4.82E-11
20	Zhao.15766533	5.83E-11

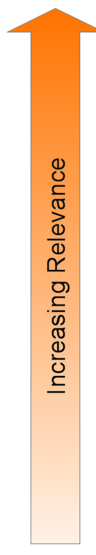


Figure 4.9: Relevance rankings.

The hypergeometric test allows the datasets to be ranked in order of relevance from zero (high relevance) to one (low relevance). Here, datasets are ranked in order of relevance to the process telomere maintenance (GO:0000723). Pan.16487579 has the highest relevance to this process and Zhao.15766533 has the lowest relevance.

The Lee and colleagues integration method [49] applies a weighted sum which integrates the datasets in order of their confidence scores, giving a higher weighting to datasets with higher confidence (see Section 3.1.4.4). In the resulting network highly weighted edges have high confidence. The weighted sum was used to incorporate the relevance rankings during integration by re-ordering the datasets prior to integration of the confidence scores, giving a higher weighting to datasets with higher relevance (Figure 4.10). Therefore, in the resulting relevance network highly weighted edges have both high confidence and high relevance to the POI (Figure 4.11).

4.2.1 Evaluation Strategy

Relevance networks were produced for *S. cerevisiae* using two ageing-related POIs; telomere maintenance (GO:0000723), and ageing (GO:0007568). Functional prediction was used as the basis of network evaluation since it is the most objective evaluation available (see Section 2.5.5).

The networks were evaluated against the control network in four ways:

- **Functional Prediction**

The ability of the networks to predict known GOBP annotations was evaluated by leave-one-out cross-validation of annotations to the POI using the Maximum Weight decision rule [57] (see Section 3.1.5.3).

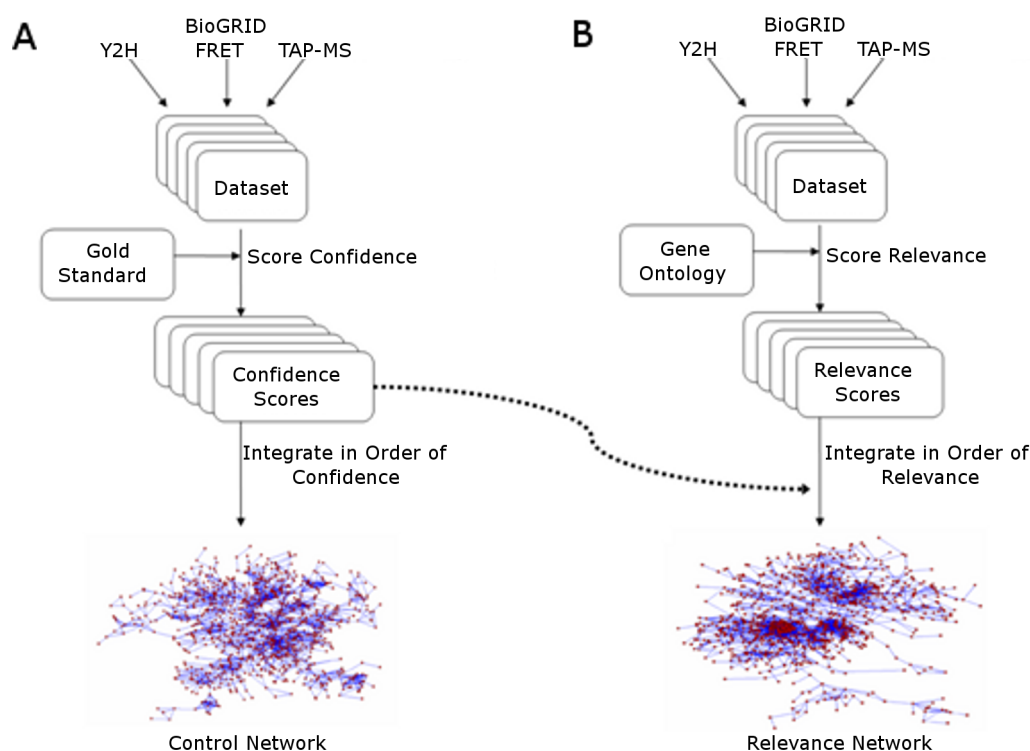


Figure 4.10: Overview of the RelCID integration method.

A. A control network is produced using the method developed by Lee and colleagues [49] of integrating the datasets in order of confidence rank. **B.** Relevance networks are produced by integration of the confidence scores in order of relevance to the process of interest. The two networks have identical topology but differ in the edge weightings between pairs of nodes.

- **Clustering**

The networks were clustered using the MCL algorithm [534] and the co-clustering of genes annotated to the POI and to several other ageing-related terms (such as DNA repair, mitochondrion and telomeric region) was evaluated (see Section 3.1.5.2).

- **Application to Real Data**

The telomere maintenance relevant network performance was evaluated using two telomere-related datasets produced by the Centre for Integrated Systems Biology of Ageing and Nutrition (CISBAN)⁵ (see Section 2.6.1) [16, 993].

- **New Predictions**

New predictions to telomere maintenance (GO:0000723) and ageing (GO:0007568) were produced for previously un-annotated genes using the maximum weight decision rule. These annotations were subsequently compared with new annotations to these terms added to the GO database in the year following network integration (March 2008 to March 2009).

⁵<http://www.ncl.ac.uk/cisban/>

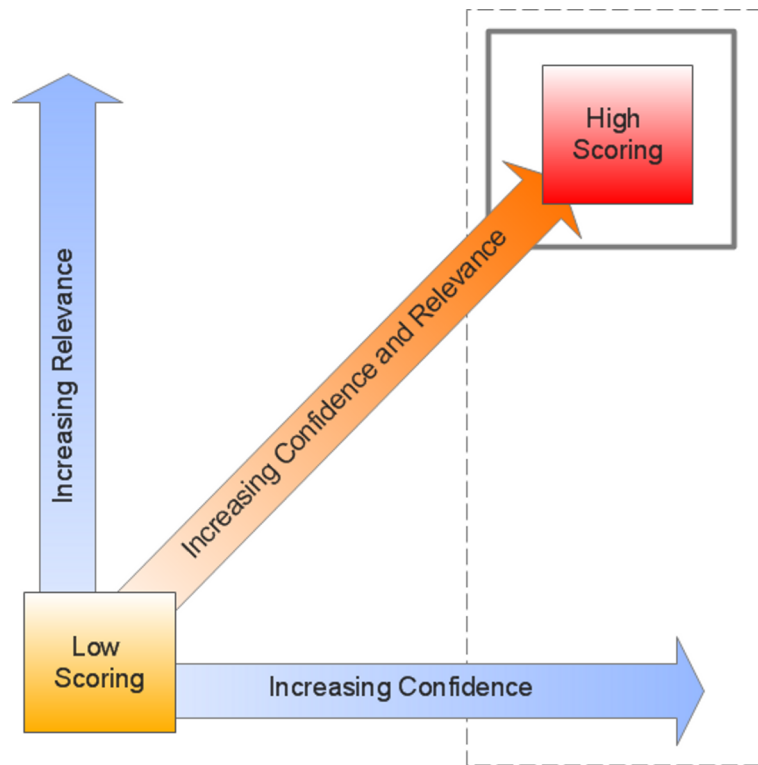


Figure 4.11: The relevance and confidence scores.

The datasets have a score of confidence and a score of relevance. In the control networks the edges have high confidence (dashed area). However, high confidence edges may have low relevance. In the relevance networks highly-weighted edges have both high confidence and high relevance (solid box).

4.3 Results

4.3.1 Network Integration

Division of Version 38 of the BioGRID data for *S. cerevisiae* as described above produced 72 datasets: 50 HTP and 22 LTP. Dataset size ranged from 14421 interactions (Collins.17314980) to as few as 33 interactions (Protein-RNA). Twenty-seven datasets were discarded due to negative scores against the KEGG Gold Standard data. Negative scores occur when the dataset and Gold Standard have little or no overlap. These datasets included many of the smaller HTP datasets and the smallest of the combined LTP datasets (<100 interactions). Therefore, a final set of 45 datasets were integrated.

Three networks were integrated using the RelCID ranked integration method (Section 3.1.4.5): a network with relevance to the GO term ageing (GO:0007568); a network with relevance to telomere maintenance (GO:0000723); and one with relevance to both ageing and telomere maintenance, produced by combining the annotations of both ageing and telomere maintenance.

The control network was integrated using the confidence ranked scores (Section 3.1.4.4). In addition, four networks were produced by reversing the order of relevance and confidence integration respectively, to act as a null hypothesis during functional prediction (Section 3.1.4.6).

Therefore, a total of eight networks were created:

- Control
- Telomere Maintenance
- Ageing
- Combined Telomere-Ageing
- Reversed Control
- Reversed Telomere Maintenance
- Reversed Ageing
- Reversed Combined Telomere-Ageing

All eight networks were topologically identical with 5143 nodes, 69091 edges and the same topological properties (Table 4.11). However the edge weights differed, reflecting the different orders of confidence score integration. Table 4.12 summarises the relevance ranks produced for the top twenty of the 45 datasets ordered by confidence score (the full rankings are presented in Appendix F). The combined ageing-telomere rankings were more similar to those of the telomere relevance rankings than the ageing relevance rankings (Figure 4.12).

Since the control network edges represented the highest possible sum of confidence scores, the distribution of edge weights was significantly lower for the relevance networks (Figure 4.13). However, by altering the order of integration, datasets with high relevance were given a higher weighting in the relevance networks. Therefore, edges with high-relevance evidence were up-weighted and those with low-relevance evidence were down-weighted (Figure 4.14). For instance, the edge between the genes YDR334W and YDL074C, which had evidence with high relevance to telomere maintenance, scored 4.98 and 4.37 in the control and relevance networks, respectively. Conversely, the edge between genes YMR200W and YGL027C, which had low relevance evidence, scored 2.49 in the control network, but was down-weighted to just 0.05 in the relevance network.

Table 4.11: Network topology.

The topological properties of the eight networks. The networks were topologically identical but each had different edge weights reflecting the different order of dataset integration.

Topological Property	Value
Characteristic path length	3.194
Clustering coefficient	0.189
Average degree	26.9
Diameter	10

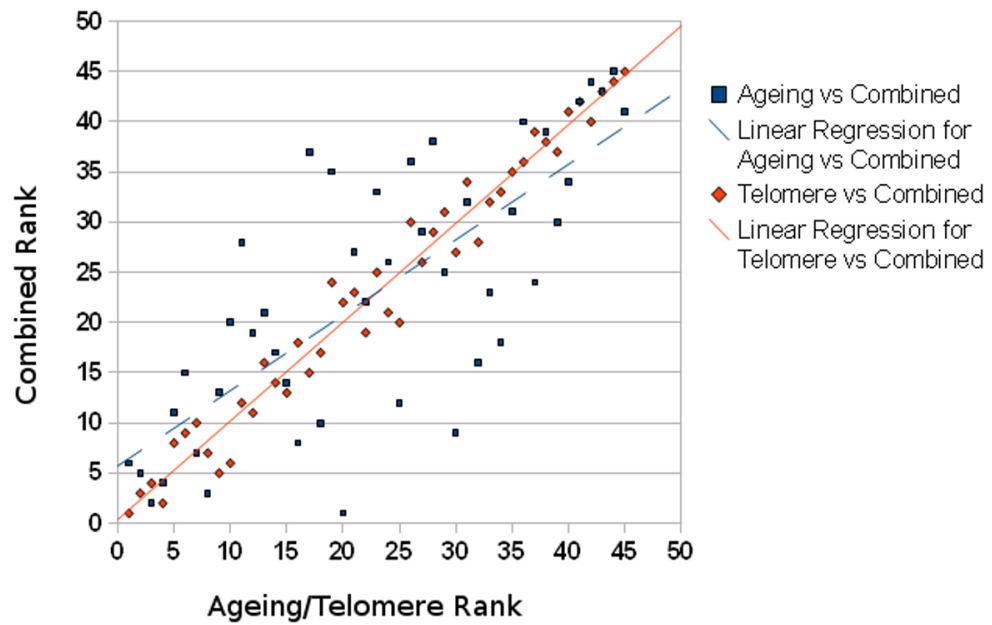


Figure 4.12: Ranking comparison.

The dataset rankings of the combined telomere-ageing network are closer to the telomere maintenance network (orange) than to the ageing relevance network (blue).

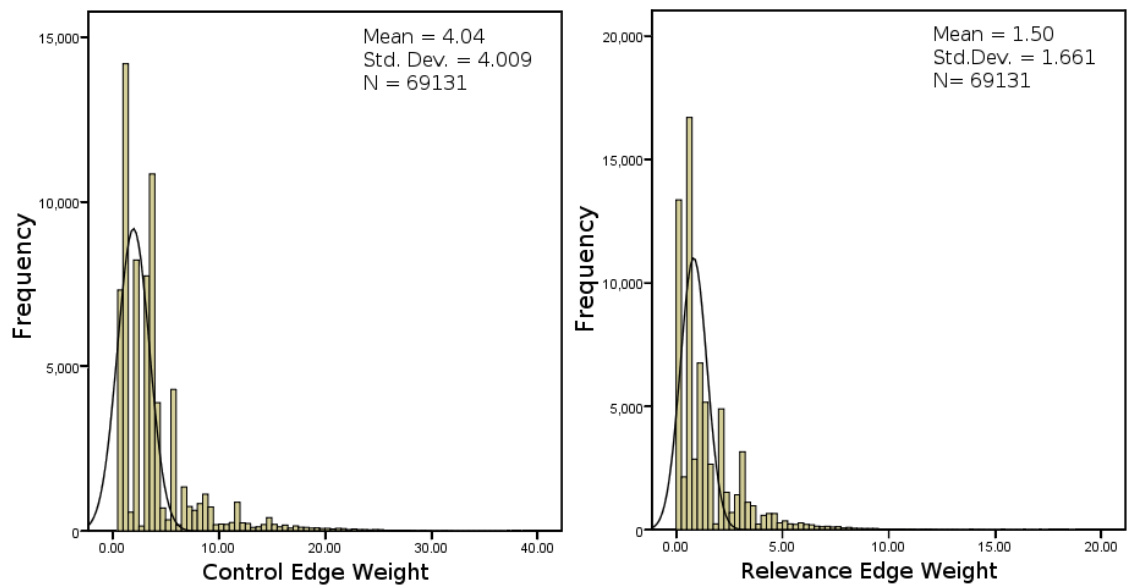


Figure 4.13: Edge weight distribution.

The distribution of edge weights for the control and telomere maintenance networks. The control network has a larger range of edge weights since its edges are integrated in order of magnitude, highest confidence to lowest.0. The ageing and combined ageing-telomere network have similar edge weight distribution and range to that of the telomere network.

Table 4.12: Dataset rankings.

A comparison of integration order for 20 datasets in the control and relevance networks. Datasets were integrated in order from rank 1 to 45. In the reversed ranking networks the datasets were integrated in the opposite order, from 45 to 1.

Dataset	Confidence Rank	Ageing Rank (A)	Telomere Rank (T)	Combined A & T Rank
Protein-peptide	1	39	17	37
Newman.11087867	2	40	45	41
Ingvarsdottir.15657441	3	19	37	24
Tong.11743162	4	42	36	40
FRET	5	45	44	45
Co-crystal Structure	6	30	21	27
Krogan.14759368	7	31	40	34
Collins.17200106	8	11	25	12
Co-localization	9	22	12	19
Co-purification	10	23	29	25
Co-fractionation	11	34	23	33
Affinity Capture-Western	12	8	7	7
Two-hybrid	13	15	9	13
Reconstituted Complex	14	17	6	15
Phenotypic Enhancement	15	10	1	6
Gavin.11805826	16	20	22	22
Far Western	17	37	38	39
Dosage Growth Defect	18	27	24	26
Biochemical Activity	19	24	13	21
Krogan.16554755	20	14	15	14

4.3.2 Network Evaluation

4.3.2.1 Functional Prediction

Leave-one-out functional prediction of annotations to the POIs was carried out for each of the relevance networks, the reversed networks and the control network. The results were analysed using ROC curves (Section 3.1.5.4). The control network produced an AUC of 0.684, 0.613 and 0.640 for ageing, telomere maintenance and the combined terms, respectively (Figure 4.15). The telomere maintenance and combined relevance networks' AUCs were both improved by 0.005, while the ageing network's AUC increased by 0.018. Computation of the standard error of the Wilcoxon statistic, SE(W) showed the improvements were statistically significant in each case (Table 4.13).

Table 4.13: Area under the curve.

Summary of the AUC measurements for the three relevance networks and the control. Standard error of the Wilcoxon statistic, SE(W), measures the statistical significance of the change in AUC.

Process of Interest	Relevant AUC	Relevant SE(W)	Control AUC	Control SE(W)
Telomere Maintenance	0.618	0.00035	0.613	0.00035
Ageing	0.702	0.00177	0.684	0.00180
Combined	0.640	0.00030	0.635	0.00030

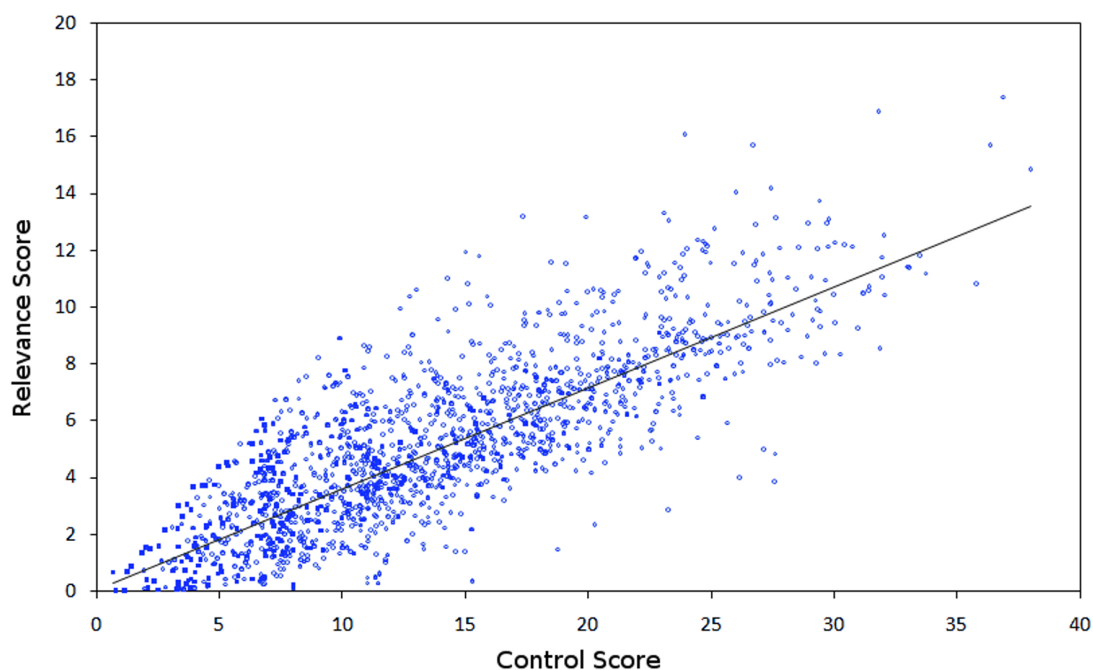


Figure 4.14: Edge weight comparison.

Each point represents a single interaction in the control and telomere maintenance networks. The points above the line are up-weighted in the relevance network and those below the line are down-weighted.

The reversed control network produced an [AUC](#) of 0.684, 0.609 and 0.629 for ageing, telomere maintenance and the combined terms, respectively. The changes in comparison to the control network were not statistically significant in all three cases.

Conversely, the reversed relevance ranked network [AUCs](#) were significantly reduced at 0.632, 0.588 and 0.606, for ageing, telomere maintenance and the combined terms, respectively (Figure 4.15). The drop in [AUC](#) in comparison with the control was larger than the increase in [AUC](#) produced by the relevance networks for all three [POIs](#), and all the changes were statistically significant.

4.3.2.2 Clustering

Clustering of the four networks was carried out using the [MCL](#) algorithm (Section 3.1.5.2) [534]. The clusters were analysed for the presence of nodes annotated to the [POI](#), and to other ageing-associated processes such as DNA repair genes, mitochondrial genes and genes annotated to the Gene Ontology cellular_compartment ([GOCC](#)) telomeric region. Additionally, nodes with unknown annotation were noted as potential telomere maintenance genes.

The control network produced 573 clusters ranging in size from 164 nodes to one node. The relevance networks each produced fewer clusters than the control: the telomere maintenance network had 523 clusters, the ageing network had 508, and combined ageing and telomere maintenance network had

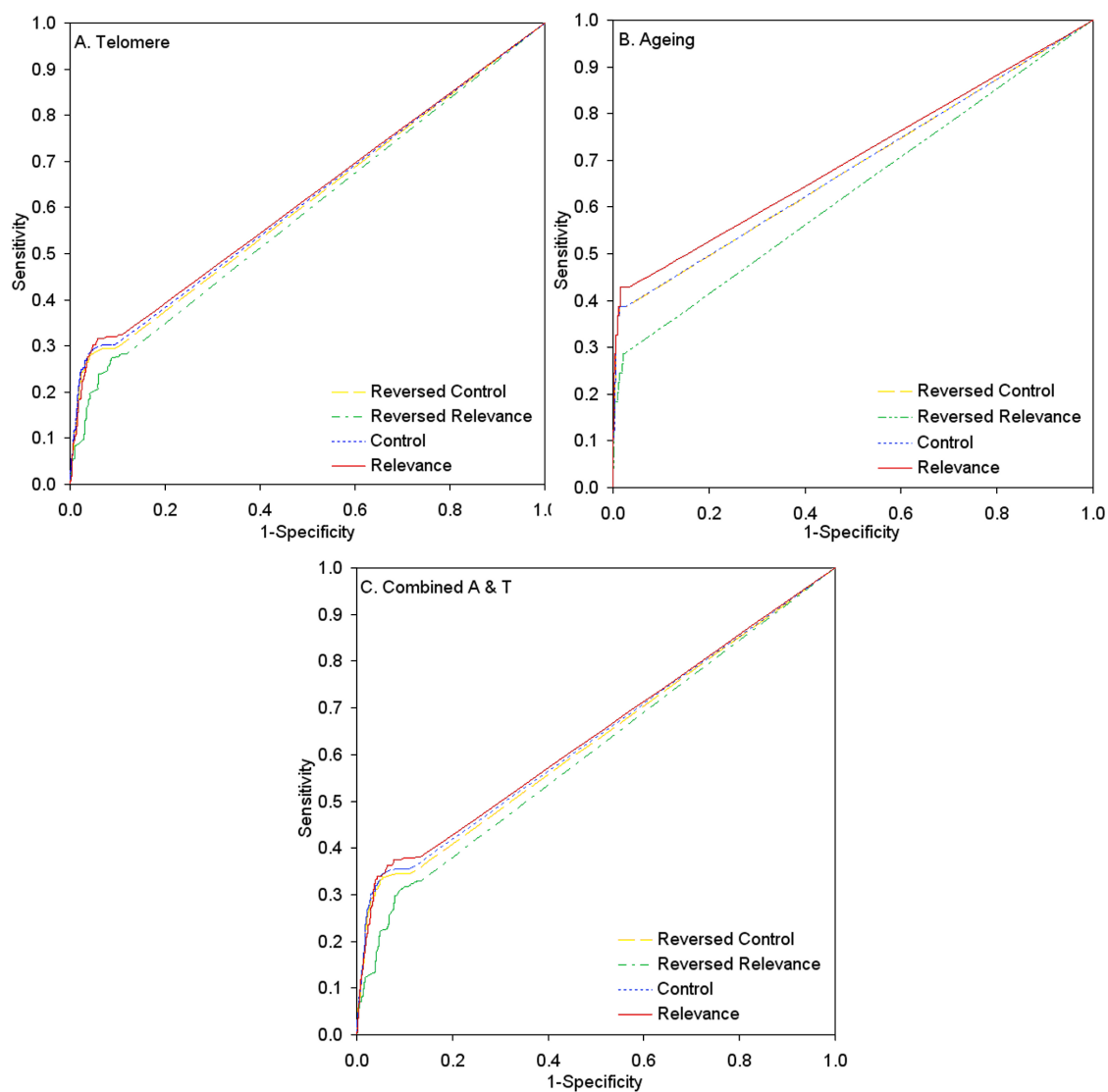


Figure 4.15: Receiver operator characteristic (ROC) curves.

The ROC curves for functional prediction of telomere maintenance, ageing and combined annotations produced by the relevance networks in comparison with the control and reversed ranking networks.

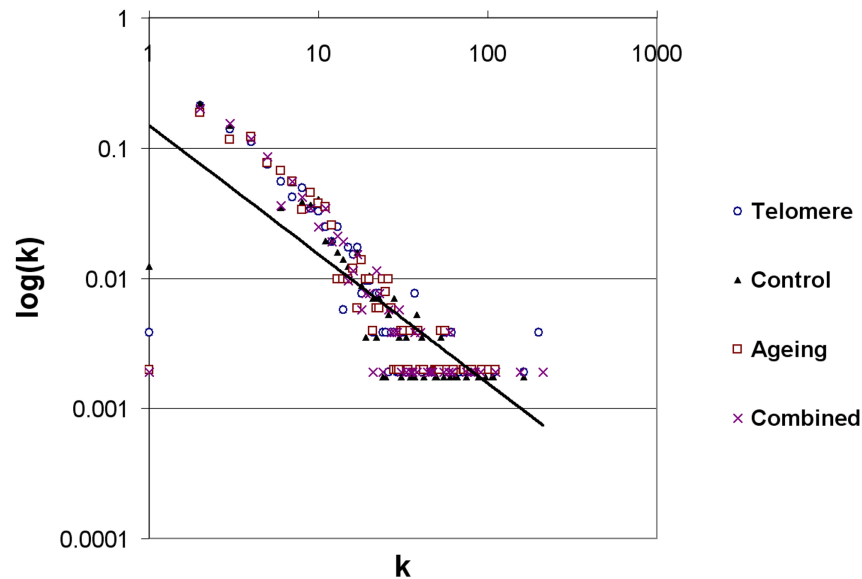


Figure 4.16: Distribution of cluster size for the four networks.

Clustering was carried out using the MCL algorithm with an inflation value of 1.8. Each point represents a single cluster. The linear trend lines, shown in black, are almost identical for each network.

523. The cluster size distribution for all four networks was found to be scale-free (Figure 4.16).

A cluster of interest (COI) was defined as a cluster containing at least one node annotated to the POI. Each of the relevance networks clustered into fewer clusters than the control, but those clusters contained a higher percentage of COIs. Additionally, the average proportion of COIs in the network increased as the minimum cluster size was raised (Table 4.14).

Similar analysis of COIs for a term unrelated to the POI, maintenance of protein location (GO:0045185), showed that the total percentage of COIs was again higher for the relevance network in all three cases. However, as cluster size was increased the percentage COIs slowly dropped below that of the control.

Table 4.14: Clusters of interest (COIs).

A summary of the percentage clusters of interest (COI) for the networks in relation to the process of interest. In all three cases the relevance networks' proportion of COIs increased as minimum cluster size was raised.

	Network	Clusters	Total % COIs	>2 nodes	>3 nodes	>4 nodes
Telomere Maintenance (T)	Control	573	21.29	26.14	28.86	35.19
	Relevant	523	22.37	27.73	31.75	36.92
Ageing (A)	Control	573	5.06	6.14	7.02	6.53
	Relevant	508	6.50	7.73	8.90	8.59
Combined A & T	Control	573	24.26	29.55	33.80	37.98
	Relevant	523	24.67	29.83	33.33	38.35

Telomere Maintenance

In the telomere maintenance network there were 117 COIs. Several of these clusters contained a large number of genes of interest. The largest COI contained 37 genes annotated to telomere maintenance (Figure 4.17 A). Of these genes 29 were clustered together in the control network. However, the remainder were spread between six other smaller clusters (Figure 4.17 B).

When assessed for the ageing-associated genes and unknowns, the relevance network cluster (Figure 4.18 A) contained far more genes of interest, and genes of unknown function, than the equivalent control clusters (Figure 4.18 B). The ageing-related genes in this cluster are summarised in Table 4.15.

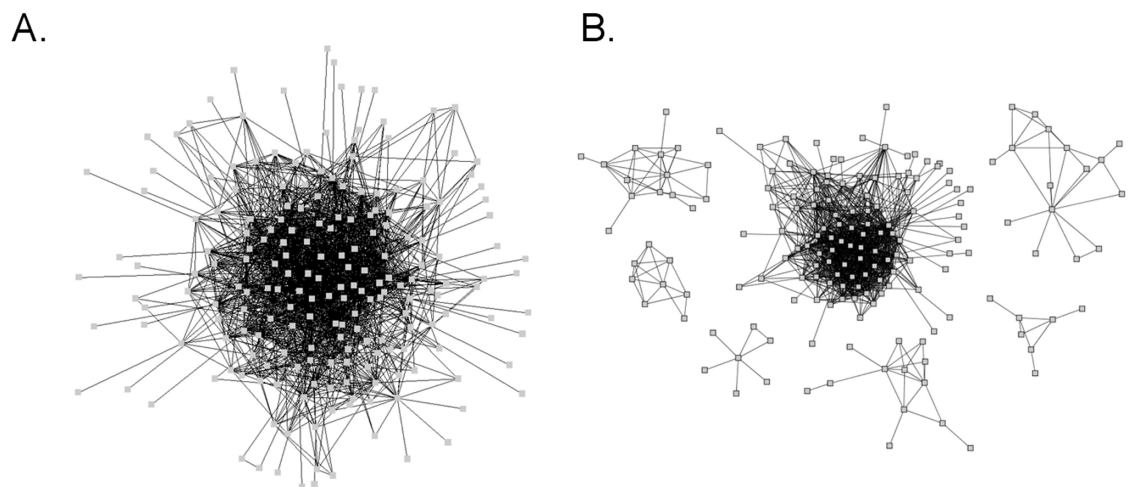


Figure 4.17: Telomere maintenance cluster.

Overview of the largest telomere maintenance cluster and equivalent clusters in the control network. **A.** Whole COI in the telomere maintenance relevance network. **B.** Seven equivalent clusters in the control network.

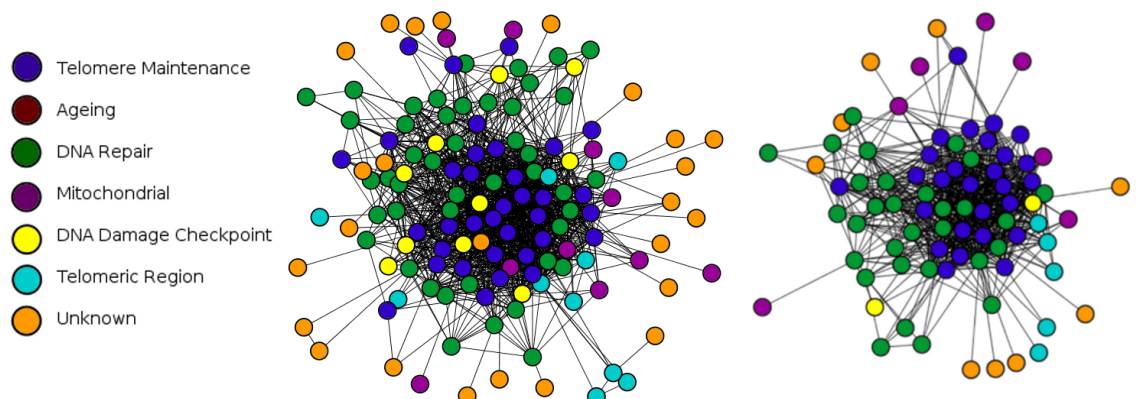


Figure 4.18: Telomere maintenance clustering.

The clusters from Figure 4.17 displaying nodes annotated to ageing-associated processes and unknowns. **A.** Relevant cluster. **B.** Largest control cluster.

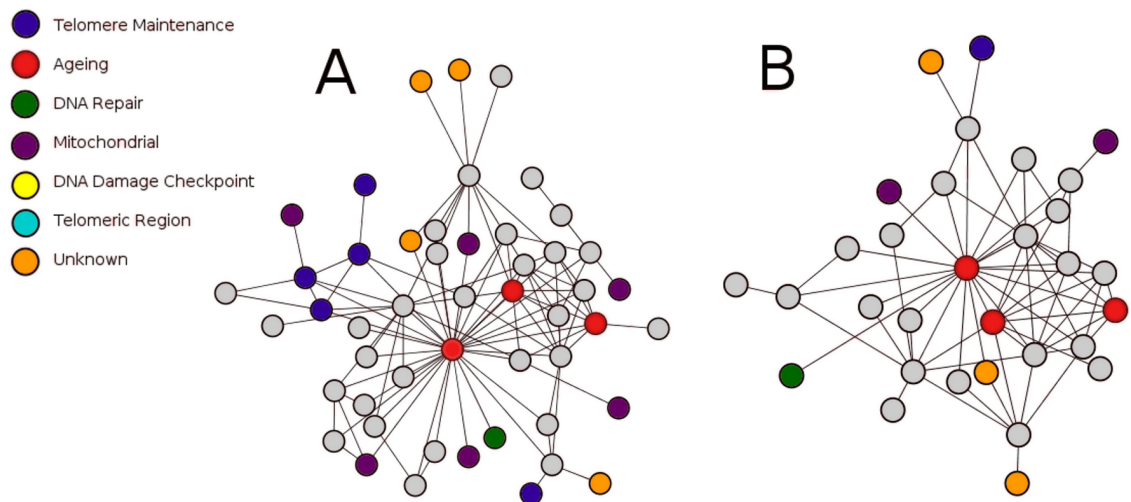
Table 4.15: Cluster annotations.

The number of ageing-related annotations and unknowns in the telomere maintenance and ageing clusters in comparison to the equivalent control network clusters.

GO Term	Telomere Maintenance	Control	Ageing	Control
Telomere maintenance	35	27	5	1
Ageing	0	0	3	3
DNA repair	45	28	1	1
Mitochondrial	10	7	6	2
DNA damage checkpoint	10	2	0	0
Telomeric region	10	5	0	0
Unknown	25	9	4	3

Ageing

There were 33 COIs in the ageing relevance network. The genes annotated to the POI clustered together in the same small groups as they did in the control network. However, the clusters contained a greater number of additional nodes. For instance, a cluster of three ageing genes occurred in the relevance network together with ten genes annotated to ageing-related processes, including five additional telomere maintenance genes and four unknowns (Figure 4.19 A). In the control network the same three ageing genes clustered with only four genes annotated to the ageing-related processes and three genes of unknown function (Figure 4.19 B).

**Figure 4.19: Telomere maintenance clustering.**

Example of the clustering of three ageing genes in the control network (A) and in the relevance network (B).

4.3.2.3 Application to Experimental Data

The functional prediction results and clusters of the telomere maintenance relevance and control networks were compared with two telomere maintenance-related datasets: a microarray dataset of genes that were up- and down-regulated during telomeric uncapping [993]; and a dataset of synthetic interactions involving the temperature-sensitive telomeric capping mutant *cdc13-1* [16] (see Section 2.6.1). In total the telomere relevance network correctly predicted 14 of the up-regulated genes, 13 down-regulated genes and 16 synthetic *CDC13* interactions. The control network predicted 12 up-regulated genes, 9 down-regulated genes and an overlapping but slightly different set of 16 synthetic *CDC13* interactions.

Several of the COIs contained genes from the datasets. In particular, three clusters contained significant numbers of ageing-related genes and several candidates for annotations to telomere maintenance:

Cluster 1

The first cluster contained 34 nodes and 75 edges (Figure 4.20). Of these genes eleven were down-regulated during telomeric uncapping and two were involved in a synthetic interaction with *CDC13*. Additionally, the cluster contained twelve genes annotated to telomere maintenance, some of which overlapped with the telomere uncapping data. Several of the down-regulated genes (YAL021C, YOL145C, YNL273W and YJL168C) were located between genes annotated to telomere maintenance in the cluster, making them potential candidates for annotation to the telomere maintenance GOBP.

A wide variety of GO term annotations were represented by the genes of the cluster, including a large number of histone methylation and transcriptional genes (Table 4.16). In particular the cluster contained the SET1 complex, a protein complex of seven proteins involved in histone methylation. Several of the SET1 genes are annotated to telomere maintenance. The control network clustered the genes of cluster 1 into several smaller clusters and did not cluster all members of the SET1 complex together.

Cluster 2

The second cluster contained 12 nodes and 24 edges. Eight of the genes were down-regulated during telomere uncapping. These genes included two nodes, YLR357W and YCR020W-B, which were already annotated to telomere maintenance (Figure 4.21). The majority of the cluster's genes were annotated to chromatin remodelling and RNA elongation. There were also several DNA repair and cell cycle genes in the cluster. Of particular note were the three down-regulated genes, YML127W, YFR037C and YDR303C, which were connected to the two telomere maintenance genes and, there-

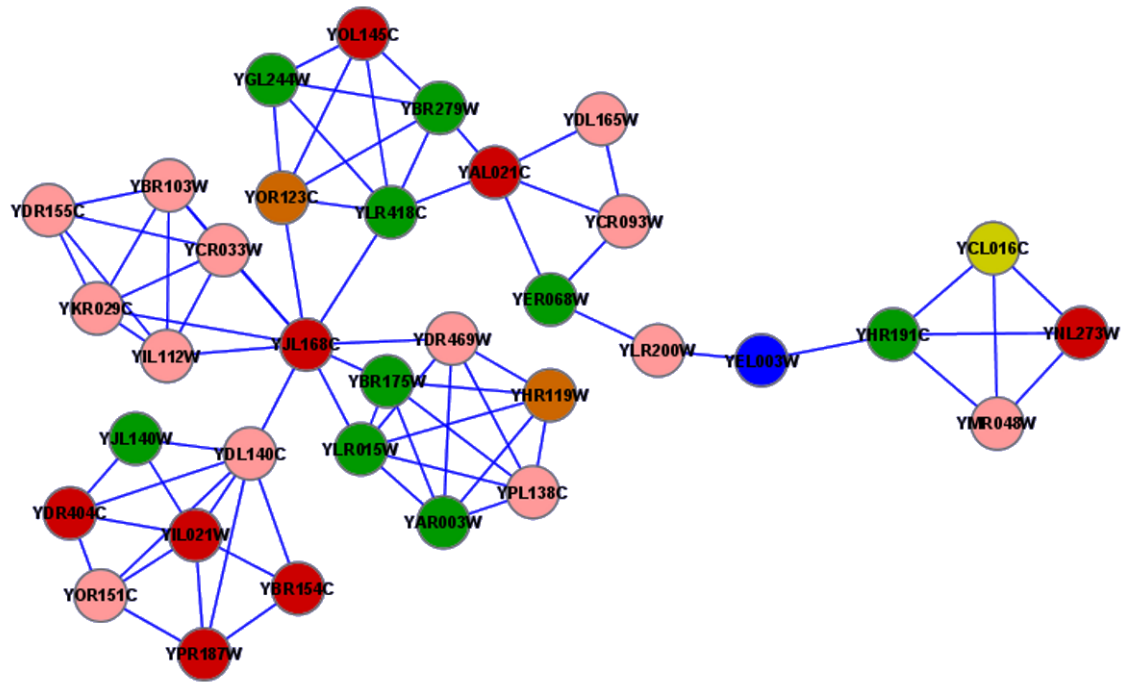


Figure 4.20: Cluster 1.

A cluster of thirty-four nodes from the telomere maintenance network containing eleven down-regulated genes, one synthetic genes and twelve telomere maintenance genes.

Node colouring: Telomere maintenance - green; Down-regulated - red; Synthetic - blue; Down-regulated & telomere maintenance - orange; Down-regulated, synthetic & telomere maintenance - yellow.

fore, could be considered candidates for annotation to the telomere maintenance [GOBP](#). The control network clustered this cluster's genes into two smaller clusters of five genes (top) and seven genes (bottom).

Cluster 3

The third cluster contained 18 nodes and 33 edges (Figure 4.22). Of these genes only one, YHR031C, was down-regulated during telomeric uncapping. However, the cluster also contained five telomere maintenance genes, two of which were directly connected to YHR031C. Additionally, the cluster contained five mitochondrial genes in a fully connected clique (bottom right) and a large proportion of DNA repair genes (Table 4.18). Of particular interest were the three DNA Repair genes, YDR386W, YPL024W and YBR098W, which were located between the down-regulated gene and a telomere maintenance gene in the cluster, making them potential candidates for annotation to the telomere maintenance [GOBP](#). The control network clustered the genes of this cluster into several smaller clusters and did not connect the three repair genes to YHR031C.

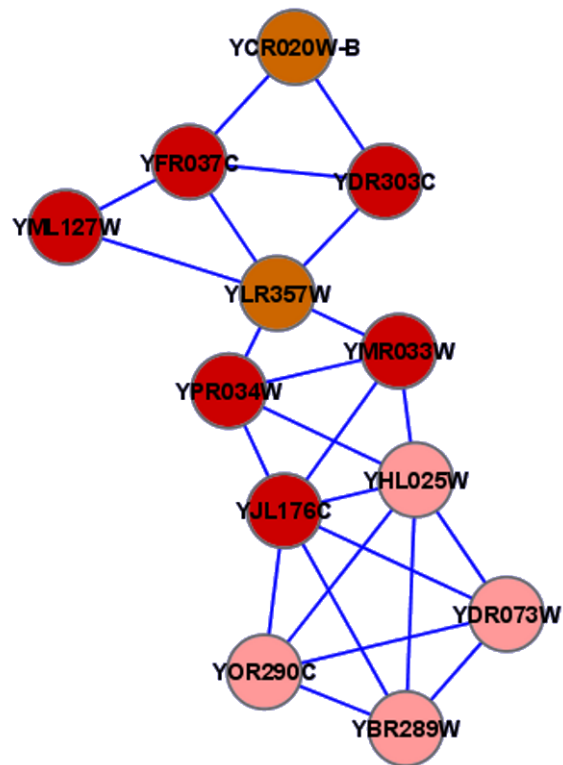


Figure 4.21: Cluster 2.

A cluster of twelve nodes from the telomere maintenance network containing eight down-regulated genes two of which were annotated telomere maintenance genes. The control network clustered these genes into two smaller clusters of five genes (top) and seven genes (bottom).

Node colouring: Down-regulated & telomere maintenance - orange; Down-regulated - red.

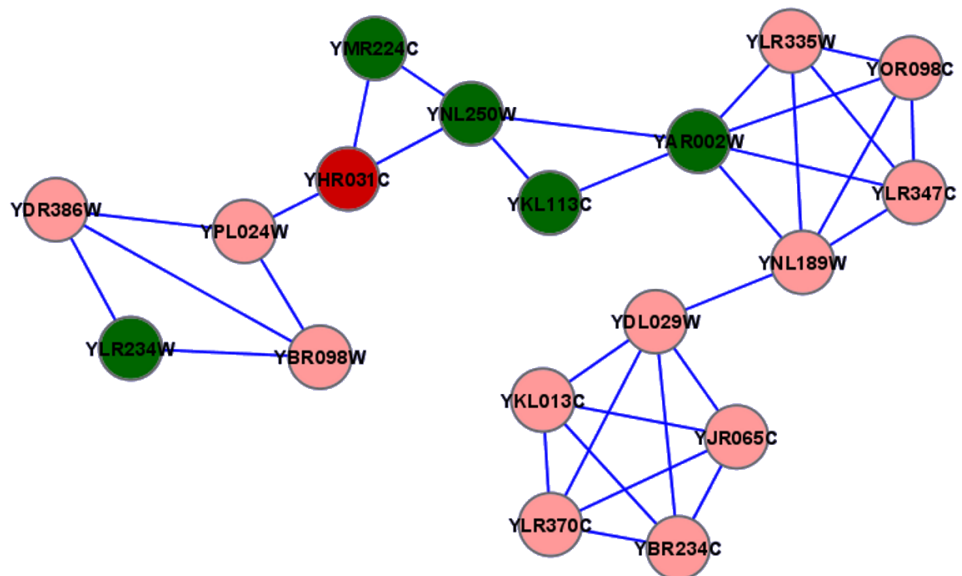


Figure 4.22: Cluster 3.

A cluster of eighteen nodes from the telomere maintenance network containing one down-regulated gene and five telomere maintenance genes.

Node colouring: Telomere maintenance - green; Down-regulated - red.

Table 4.16: Cluster 1 annotations.

The GOBP and GOCC annotations of cluster 1 (Figure 4.20). Only terms with more than one gene annotation are displayed.

Term	Annotations
Histone methylation	12
Telomere maintenance	12
Transcription from RNA polymerase II promoter	11
RNA elongation from RNA polymerase II promoter	9
Chromatin silencing at telomere	7
Histone deacetylation	4
Mitotic sister chromatid cohesion	4
Negative regulation of meiosis	4
Negative regulation of transposition, RNA-mediated	4
Nuclear-transcribed mRNA catabolic process, deadenylation-dependent decay	4
Nuclear-transcribed mRNA poly(a) tail shortening	4
Regulation of transcription from RNA polymerase II promoter	4
Ascospore formation	3
Negative regulation of transcription from RNA polymerase II promoter	3
Regulation of transcription, DNA-dependent	3
DNA recombination	2
DNA replication checkpoint	2
Protein ubiquitination	2
Regulation of cell cycle	2
Response to DNA damage stimulus	2
Response to pheromone during conjugation with cellular fusion	2
Tubulin folding	2

Table 4.17: Cluster 2 annotations.

The GOBP and GOCC annotations of cluster 2 (Figure 4.21). Only terms with more than one annotation are displayed.

Term	Annotations
Chromatin remodeling	7
RNA elongation from RNA polymerase II promoter	7
ATP-dependent chromatin remodeling	6
Double-strand break repair	2
Double-strand break repair via nonhomologous end joining	2
Establishment and/or maintenance of chromatin architecture	2
G1/S transition of mitotic cell cycle	2
G2/M transition of mitotic cell cycle	2
Regulation of cell cycle	2
Telomere maintenance	2

Table 4.18: Cluster 3 annotations.

The GOBP and GOCC annotations of cluster 3 (Figure 4.22). Only terms with more than one gene annotation are displayed.

Term	Annotations
Mitochondrion inheritance	5
Telomere maintenance (htp data)	5
Actin filament organization	4
DNA repair	3
Double-strand break repair via nonhomologous end joining	3
Meiotic recombination	3
Protein targeting to membrane	3
Response to DNA damage stimulus	3
RNA export from nucleus	3
DNA topological change	2
Double-strand break repair via break-induced replication	2
Meiotic DNA double-strand break formation	2
Meiotic DNA double-strand break processing	2
mRNA export from nucleus	2
mRNA-binding (hnRNP) protein import into nucleus	2
NLS-bearing substrate import into nucleus	2
Nuclear pore organization and biogenesis	2
Nucleocytoplasmic transport	2
Protein export from nucleus	2
Protein import into nucleus	2
Ribosomal protein import into nucleus	2
snRNA export from nucleus	2
snRNP protein import into nucleus	2
tRNA export from nucleus	2

4.3.2.4 New Predictions

The Version 50 GO annotation file (March 2009) was compared to the Version 38 file (March 2008) to identify genes that were annotated to ageing (GO:0007568) and telomere maintenance (GO:0000723) after the networks were integrated in March 2008 and which, therefore, had no influence on the edge weights of the relevance networks. Eighteen new annotations to ageing were identified. However, the number of annotations to telomere maintenance had been reduced from 276 to 68, with only five new annotations.

Functional prediction of ageing and telomere maintenance annotations was carried out using the ageing, telomere maintenance and control networks. The predictions were compared with the new annotations to assess the networks' predictive power. Interestingly, prediction performance was extremely poor, with only three annotations predicted by the relevance networks and five predicted by the control (Table 4.19). Only one annotation to telomere maintenance was predicted by both the control and relevance networks.

The Version 38 clusters for the telomere maintenance, ageing and control networks were then compared with the new ageing and telomere maintenance annotations to assess how well the newly-annotated genes clustered with other annotations to the POIs. In the majority of cases, (4/5 for telomere maintenance and 12/18 for ageing) the new annotations were clustered with at least one other annotation to the POI in both the relevance networks and the control networks. However, the control network clusters were smaller and contained fewer ageing-related genes than the relevance clusters.

For example, the telomere maintenance annotations for the genes YJR078W and YPL157W were not predicted by either the relevance or control networks, but were clustered with ageing-related genes in the telomere maintenance relevance network. YJR078W clustered with a large number of mitochondrial genes and a single ageing gene in the relevance network (Table 4.20). In the control network YJR078W clustered with a single mitochondrial gene and the same ageing-related gene (Table 4.21). Additionally, there were a significant number of unannotated genes in the relevance network cluster. YPL157W was clustered with a large number of nuclear, ribosomal and mitochondrial genes, including one gene annotated to telomere maintenance, in the telomere maintenance relevance network (Table 4.22). However, in the control network, this gene clustered with a single gene of unknown function.

Table 4.19: Prediction performance for new annotations.

The ageing relevance network correctly predicted two ageing annotations, while the control also predicted two different ageing annotations. The telomere maintenance relevance network correctly predicted only one telomere maintenance annotation, while the control correctly predicted three.

Term	Gene	Predicted	
		Control	Relevance
Ageing	YBR140C	✓	X
	YCR084C	X	X
	YDR310C	X	✓
	YGL035C	✓	X
	YIL065C	X	X
	YIL155C	X	✓
	YKL085W	X	X
	YKL106W	X	X
	YLL001W	X	X
	YLL026W	X	X
	YLR318W	X	X
	YLR319C	X	X
	YLR368W	X	X
	YOR005C	X	X
	YOR360C	X	X
	YOR384W	X	X
	YPL157W	X	X
	YPR024W	X	X
Telomere Maintenance	YJL092W	✓	✓
	YJR078W	X	X
	YLR071C	✓	X
	YNL139C	✓	X
	YPL157W	X	X

Table 4.20: Relevance network cluster annotations.

The GOBP and GOCC annotations represented in the YJR078W cluster of the telomere maintenance relevance network.

Term	No. Annotations
Biological process unknown	7
Mitochondrion	6
NAD biosynthetic process	6
Cytoplasm	5
Nucleus	4
Cellular component unknown	3
Aerobic respiration	1
Chromatin silencing at rDNA	1
Chromatin silencing at telomere	1
Integral to membrane	1
Mitochondrial inner membrane	1
Mitochondrial outer membrane	1
Plasma membrane	1
Replicative cell aging	1
Ribosome assembly	1
Ribosome biogenesis and assembly	1
Translation	1

Table 4.21: Control network cluster annotations.

The GOBP and GOCC annotations represented in the YJR078W cluster of the control network.

Term	Annotations
Nucleus	3
Cytoplasm	2
NAD biosynthetic process	2
Biological process unknown	1
Chromatin silencing at rDNA	1
Chromatin silencing at telomere	1
Endocytosis	1
Mitochondrial outer membrane	1
Mitochondrion	1
Replicative cell aging	1
Ribosome biogenesis and assembly	1
Threonine metabolic process	1
Cellular component unknown	1

Table 4.22: Relevance network cluster annotations.

The GOBP and GOCC annotations represented in the YPL157W cluster of the telomere maintenance relevance network. In the control network YPL157W clustered with a single gene of unknown function.

Term	Annotations
Nucleolus	8
Ribosome biogenesis and assembly	7
rRNA processing	7
Small nucleolar ribonucleoprotein complex	3
Biological process unknown	2
Box H/ACA snoRNP complex	2
Mitochondrion	2
Nucleus	2
Box H/ACA snoRNA 3'-end processing	1
Box H/ACA snoRNP assembly	1
Cell septum	1
Cytokinesis, completion of separation	1
Extracellular region	1
Fungal-type cell wall	1
Mitochondrial outer membrane	1
Nucleolus organization and biogenesis	1
Nucleoplasm	1
Plasma membrane	1
Proton transport	1
Regulation of pH	1
Ribosomal large subunit assembly and maintenance	1
Ribosome	1
rRNA modification	1
snoRNA metabolic process	1
snRNA capping	1
Telomere maintenance	1

4.3.3 Discussion

The RelCID technique described and evaluated here incorporates a measure of process-relevance into probabilistic network integration, therefore addressing the third objective of this project (see Section 1.5). Three relevance networks have been created using two ageing-related POIs; one with relevance to the GO term ageing, one to the term telomere maintenance, and a combined network with relevance to both terms. The shortening of telomeres over time has been linked to the ageing process [972, 982, 997, 999, 1001–1003] and therefore some overlap between the ageing and the telomere maintenance networks was expected. The overlap between the terms should therefore be enhanced in the combined network. The networks were evaluated by comparison with a control network integrated without a measure of relevance [49].

The three relevance networks naturally had the same topology as the control since they were integrated using the same datasets. However, the edges weights differed, reflecting the order of dataset integration. In the control network the datasets are ranked in order of confidence score, with the highest score ranked first [49]. In the relevance networks, although the scores integrated are still measures of dataset confidence, the datasets are ranked prior to integration in order of relevance to the POI, with the most relevant first. Therefore, in the control network highly-weighted edges have high confidence, but in the relevance network the highly weighted edges have both high confidence and high relevance.

Lower ranked datasets are sequentially down-weighted by the integration step. In the control network the highest-ranking dataset is Protein-Peptide, which is integrated without modification. In the telomere maintenance network this dataset is ranked 17th and is therefore down-weighted at the integrated step. Conversely, the Phenotypic Enhancement data, ranked top for telomere maintenance relevance, will be given a higher weighting in the telomere maintenance network than in the control. The rank order of datasets in the combined network was far closer to that of the telomere maintenance network than the ageing network, since telomere maintenance contributes approximately 78% (276 of 355 genes) of the combined POI.

Importantly, while the order of integration differs between the datasets, the final scores themselves are calculated using the dataset confidence scores. Consequently, datasets with high relevance but low confidence will have a high rank, but their contribution to the final edge weight will be based on their confidence score, and hence still take into account the reliability of the data. Since the control integration represents the highest possible sum of the confidence scores, the weights on the edges of the relevance networks are smaller and have a lower range. However, the relative weighting of edges within the relevance networks reflects the level of relevance of its edges, with low relevance edges down-weighted, and high relevance edges up-weighted.

The RelCID integration schema is extremely flexible and can be used with any existing confidence scoring scheme and any Gold Standard dataset [97, 669]. For instance, the 27 datasets lost due to negative scores against the KEGG Gold Standard may score well against an alternate gold standard. The integration method is also easily applicable to any existing annotation schemes, such as MIPS FunCat [279], GO [100] or KEGG PATHWAYS [99] annotation. Additionally, unlike previous methods for process-relevant network integration, no data are lost, even in unannotated areas of the network. Like all weighted networks, the relevance networks produced by this schema can be subject to an edge weight cut-off to identify up-weighted edges.

In order to assess the networks performance several existing global network evaluation techniques were adapted for use with a single POI (Objective 4, Section 1.5). The performance of the rele-

vance networks was compared against that of the control network in several ways. First, leave-one-out functional prediction of known **POI** annotations was carried out using the Maximum Weight algorithm [57]. This algorithm propagates annotations along the highest weighted edge surrounding a node. While more complex global functional prediction algorithms have been developed [92, 580, 894, 917, 918], the local nature of this algorithm provides a simple method with which to directly compare the networks' performance and the effect of different edge-weighting systems. The algorithm's performance was statistically significantly better on the three relevance networks than on the control network.

As a further control, four networks were also integrated with reversed confidence and relevance integration ranks and used for functional prediction. Reversal of the control rankings had little effect on prediction performance. However, reversal of the relevance ranks significantly reduced performance in relation to the **POIs**. Consequently, the relevance of the datasets appears to be more important in functional prediction of a single process than the reliability of the datasets. In other words, if the datasets have no relevance to the **POI** it appears that the order of integration has little effect on prediction of that **POI**. Therefore, the relevance-ranked integration technique developed here provides a method which increases the prediction accuracy of the networks by taking dataset relevance into account, and as such is a valuable extension to standard data integration techniques for functional prediction.

The networks were also clustered using the **MCL** algorithm [534]. Clustering allows large networks to be broken down into manageable pieces for visual analysis and also identifies groups of associated genes. **MCL** utilises the edge weights of a network allowing direct comparison between performance of the relevance and control networks. The relevance networks produced fewer clusters than the control network, with genes annotated to the **POI** in larger clusters than in the control network.

Genes annotated to the **POI** which appear in separate clusters in the control network co-occur in the relevance networks' clusters, making their relationship to each other, and to other genes, easier to observe and investigate. In addition, a larger range of ageing-related genes and unknowns were found clustered with the **POI** in the relevance networks than in the control clusters. Further, the clusters containing genes annotated to telomere maintenance in the relevance network had higher proportions of genes from two telomere capping-related datasets than the control clusters. The telomere uncapping-related datasets were also predicted with more accuracy by the relevance networks than by the control. Telomere uncapping triggers a range of cellular processes, some of which, such as cell cycle arrest, are not directly associated with telomere biology [16, 993]. Therefore, the telomere maintenance relevance network provides a method to identify the telomere-related areas of the dataset by emphasising the relevant information within the network.

While the performance of the relevance networks was improved over the control in relation to known annotations to the [POIs](#), [PFINs](#) are intended to generate new hypotheses and guide future experiments. Therefore, in the final stage of evaluation the networks were used to produce functional predictions to the [POIs](#) which were then compared with newer, up-to-date annotations. Since the new annotations were not available at the time of integration they have no influence on the edge weights of the relevance networks, allowing unbiased evaluation. Functional prediction performance was poor for all the networks, including the control. However, clustering improved identification of the newly annotated genes. The relevance networks clustered the genes in larger clusters containing a higher proportion of ageing-related genes than the control network.

The gene annotations to `ageing` and `telomere maintenance` had changed significantly between March 2008 and March 2009. The number of annotations to each GO term is naturally expected to increase over time as new data are generated. However, this was not the case for both terms; `ageing` had gained 18 annotations but `telomere maintenance` had lost 213 annotations with only 5 new annotations added. This accounts for 76.45% of the annotations used to generate the relevance network in March 2008. The reduction of `telomere maintenance` annotations is likely to have adversely affected the results of this evaluation since the datasets relevance scores were based on the larger set of annotations.

High-quality biological databases, such as GO, are constantly changing due to the curation process [272–275]. While databases are expected to increase in size over time, the addition of data is only one aspect of the curation process. In addition, curators must also identify and remove incorrect data and change their database schemas to reflect current biological knowledge. These changes may have a significant impact on integrated analyses, particularly if they occur in Gold Standard databases.

While the additional information provided by the relevance networks provides greater scope to draw inferences from the data, changes to the source data can affect network performance. In this case a significant number of the annotations used to generate the relevance network were subsequently removed and, therefore, the network's performance could not accurately be assessed. However, the extent to which the database changes affects performance remains unclear. The next chapter presents an investigation of the effect of database changes on [PFIN](#) performance by systematically comparing network performance between all archived versions of the source databases.

Chapter 5

Evaluation of the Effect of Database Curation on PFIN Performance in *Saccharomyces cerevisiae*

The quality and performance of any probabilistic integrated network is dependent upon the quality of its component data, and of the Gold Standards chosen to evaluate it. Therefore, it is essential that the data sources chosen are accurate and up-to-date, reflecting current biological knowledge [264]. Functional prediction for genes with known annotations can be used to evaluate PFIN performance (see Section 2.5.5). However, since many functional annotations are themselves derived from known functional interaction data, it is possible that the ability of the network to predict known annotations is biased. Therefore, evaluation using data not present at the time of integration can more accurately assess the network's ability to predict unknown annotations.

In the previous chapter (Section 4.19), up-to-date GO annotation data was used to evaluate networks integrated with previous dataset versions. However, the GO database had significantly changed over time and, consequently, direct comparison and evaluation using these data was inconclusive.

Well-curated databases, such as KEGG, BioGRID and GO, change all the time as new data are added. However, data addition is only one aspect of curation. Highly-curated databases are constantly evolving, in both content and structure, to reflect current biological knowledge [297]. It is often necessary to edit or remove data when errors or inconsistencies are identified, often following community feedback, or when data are found to be incorrect in light of new studies [266]. For instance, false positives in high-throughput datasets can later be identified by less error-prone small-scale studies.

Database schemas are also subject to change. In the case of annotation data such as GO, identifiers may be added, removed or modified. These changes may in turn necessitate the addition, removal or

reassignment of annotations to these identifiers, such as the changes observed in Section 4.19.

It is commonly assumed that the quality of interaction and annotation data improves as databases grow and change, particularly for manually curated databases. This assumption is generally unwritten; it is implicit in our understanding of the scientific process that the more data available the more can be learned from it. The validity of this assumption can be questioned, given the exponential growth in raw data, coupled with the high false positive rate estimated for high-throughput data (see Section 2.4.1).

Integrated network theory is based on the premise that the “whole is greater than the sum of its parts” [1158], so it follows that the more high quality the parts, the greater the whole. For instance it would be expected that a PFIN integrated using data from 2008 would perform better than one integrated in 2007. However, this may not be the case since structural changes to the databases, such as those presented in the previous chapter, could affect interpretation of the data.

This chapter describes a systematic evaluation of the changes in four manually-curated databases and evaluates the effect of these changes on PFIN performance in relation to the ageing process. In Section 5.2 the databases’ changes are evaluated. The effect of the changes on PFIN performance is then assessed in Sections 5.3-5.4. Finally, the RelCID integration method described in Section 4.2 is applied to the same data to investigate how the use of process-relevance during integration may improve network performance over time.

5.1 Source Data

All available versions of the BioGRID *S. cerevisiae* data files were downloaded from the BioGRID archive¹ [276]. In total 36 monthly versions were available, ranging from V17 released in July 2006, to V52 on 1st May 2009. V27 and V28 were released on the same date and contained the same data. V27 was therefore discarded, resulting in a final set of 35 BioGRID data files. The data from each file was split using the 100 interaction threshold (Section 3.1.1) into HTP and LTP datasets. The datasets were named using the standardised format described in Section 3.1.1.

The KEGG PATHWAY database was used as the source of Gold Standard data [277]. KEGG data files are archived on a different release schedule from that of BioGRID. For instance, the 29/06/07 KEGG version was current for V17-V19 of BioGRID (Figure 5.1). Therefore, those versions of the *S. cerevisiae* PATHWAY files that were current on the release data of each BioGRID version were selected from the KEGG FTP archive². In total 12 KEGG versions were available at approximately

¹<http://www.thebiogrid.org/downloads.php>

²<http://www.genome.jp/kegg/download/>

three-month intervals from June 2006 to March 2009. Gold Standards were constructed from each file by selecting all possible pairs between genes annotated to the same pathway, as described in Section 3.1.1.

Two files were downloaded to provide the annotation information for relevance scoring and functional prediction: the Gene Ontology from the GO Consortium archive³, and the SGD-GO annotation mapping from the SGD archive⁴. These files are updated daily and are archived at regular intervals (GO monthly and SGD approximately weekly). As the SGD annotations are dependent on the Gene Ontology structure, the two files were treated as a single GO data source. The GO and SGD file versions that were current on the release data of each BioGRID version were downloaded, resulting in a total of 35 GO-SGD file pairs corresponding to the 35 BioGRID versions (Figure 5.1).

For ease of analysis the GO files were named using the corresponding BioGRID version number. Since the KEGG file versions span several BioGRID files they were named using the earliest of the corresponding versions. Three ageing-related POIs were chosen for analysis; ageing (GO:0007568), DNA repair (GO:0006281) and telomere maintenance (GO:0000723) following discussion with colleagues at CISBAN⁵.

5.2 Database Changes

5.2.1 BioGRID

There were 35 versions of BioGRID available for download, ranging from V17 (June 2006) to the most recent V52 (April 2009). The number of datasets increased from 63 to 88 between these versions (Figure 5.2 A). During this interval 26 new HTP datasets were added and one HTP dataset (Ito.10655498) was removed. In addition, one of the HTP datasets, Stevens.11804584, was modified by the removal of 75 interactions. This modification reduced this dataset's size below the 100-interaction HTP threshold, resulting in the remaining interactions being added to the LTP dataset for the following versions (V23-V52). A new LTP experimental type, PCA, was also added producing a new LTP dataset. Addition and removal of datasets is summarised in Figure 5.3.

³<ftp://ftp.geneontology.org/pub/go/>

⁴<http://downloads.yeastgenome.org/>

⁵

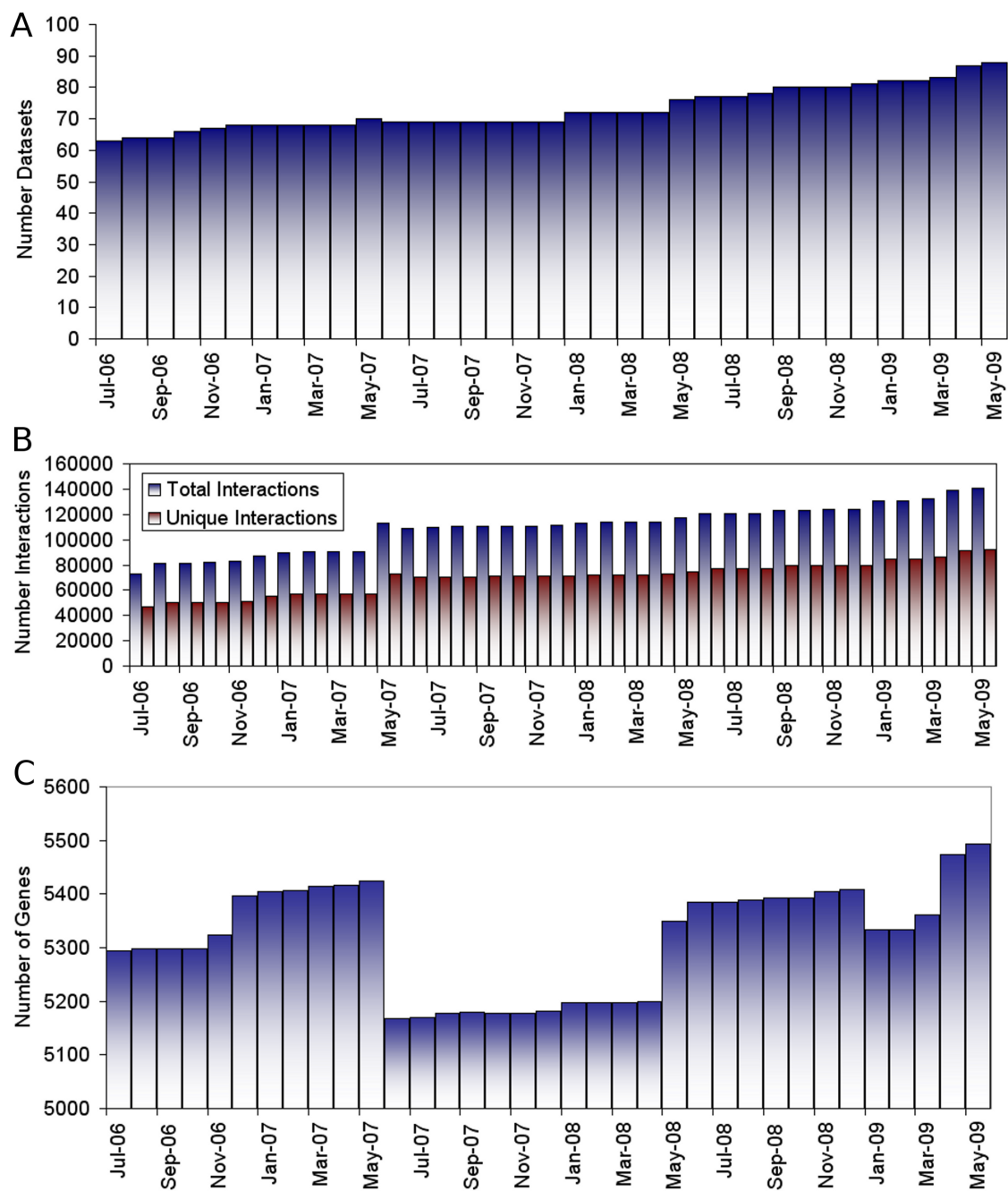


Figure 5.2: Changes to the BioGRID dataset for *Saccharomyces cerevisiae* and its coverage of the genome between July 2006 and May 2009.

A. Number of datasets. **B.** Number of total and unique interactions. **C.** Number of genes.

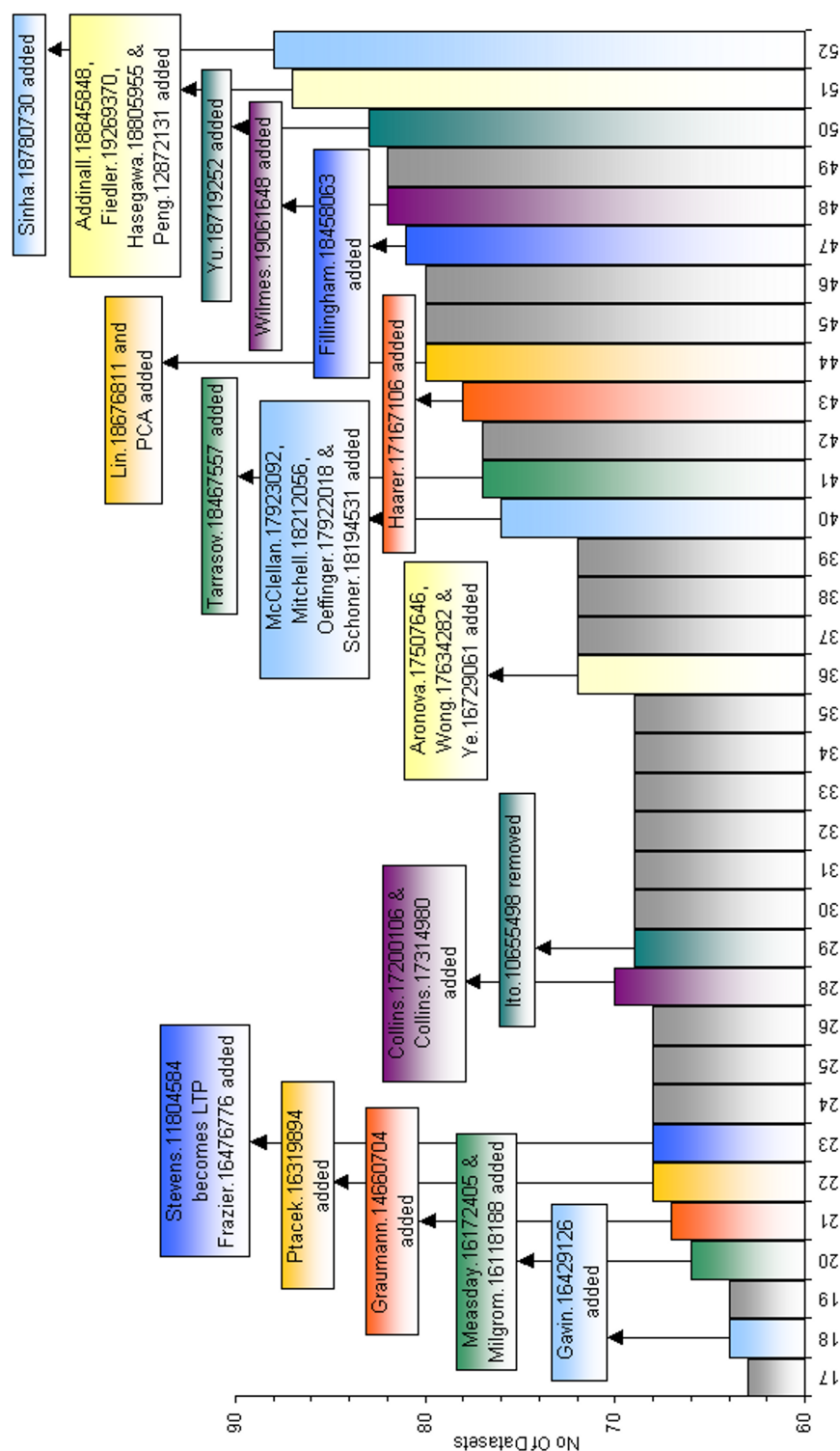


Figure 5.3: Addition and removal of BioGRID datasets between file versions.

In total 26 new HTP datasets were added and one HTP dataset was removed. Additionally, a new LTP experimental type, Protein Fragment Complementation Assay (PCA), was also added at version 44 producing a new LTP dataset.

The total number of interactions in the BioGRID dataset increased steadily with the only decrease occurring in June 2007, corresponding to the removal of the Ito.10655498 dataset. The number of unique interactions in BioGRID increased in a similar pattern (Figure 5.2 B) with the percentage of unique interactions fluctuating between 61% and 66%.

The number of genes covered by the data fluctuated between 5,300 and 5,500 (Figure 5.2 C) with a significant fall in June 2007, corresponding with the removal of the Ito.10655498 dataset, and a significant increase in May 2008, corresponding to the addition of four new HTP datasets (McClellan.17923092, Mitchell.18212056, Oeffinger.17922018 & Schoner.18194531).

The LTP datasets all gradually increased in size as new interactions were added, while the majority of the HTP datasets did not change in size by more than 10 interactions between their earliest and latest versions. There were two exceptions amongst the HTP datasets in addition to the reduction of the Stevens.11804584 dataset from HTP to LTP (Figure 5.4). In the first, the Ito.11283351 dataset was significantly reduced in size in June 2007. This change corresponded with the removal date of another Ito.10655498 dataset. An enquiry to the BioGRID support team revealed that these changes followed feedback about these datasets: "...11283351 contains two sets of interactions, core and non-core. We were informed that only the core set is reliable and removed the rest. The 10655498 publication contained interactions equivalent to the non-core and was also removed" [1159]. The second significant HTP dataset change occurred at January 2009 where 128 interactions of the McClellan.17923092 dataset were removed from the database. This change corresponded with the drop in total coverage of genes in the same BioGRID version (Figure 5.2 C).

5.2.2 KEGG

There were 12 versions of KEGG available for download within the date range of the available BioGRID versions. The number of genes covered by the KEGG datasets increased from 1,189 to 1,274 in this time, with the greatest coverage being 1,294 genes in September 2008. The number of pathways in KEGG increased from 99 to 108, with the greatest number of pathways being 115, also in September 2008. The distribution of genes per pathway and of pathway annotations per gene did not change significantly between KEGG versions (Figure 5.5).

The Gold Standard datasets for this study were generated by selecting all possible pairs between genes annotated to the same pathway in that KEGG version (Section 3.1.1). The number of pairs of genes in the Gold Standard dataset fluctuated between 39,000 and 42,000. The number of connected components in the Gold Standard (groups of genes connected to each other, but not to genes outside the component) varied between 10 and 13 (Table 5.1). The change in the number of genes covered

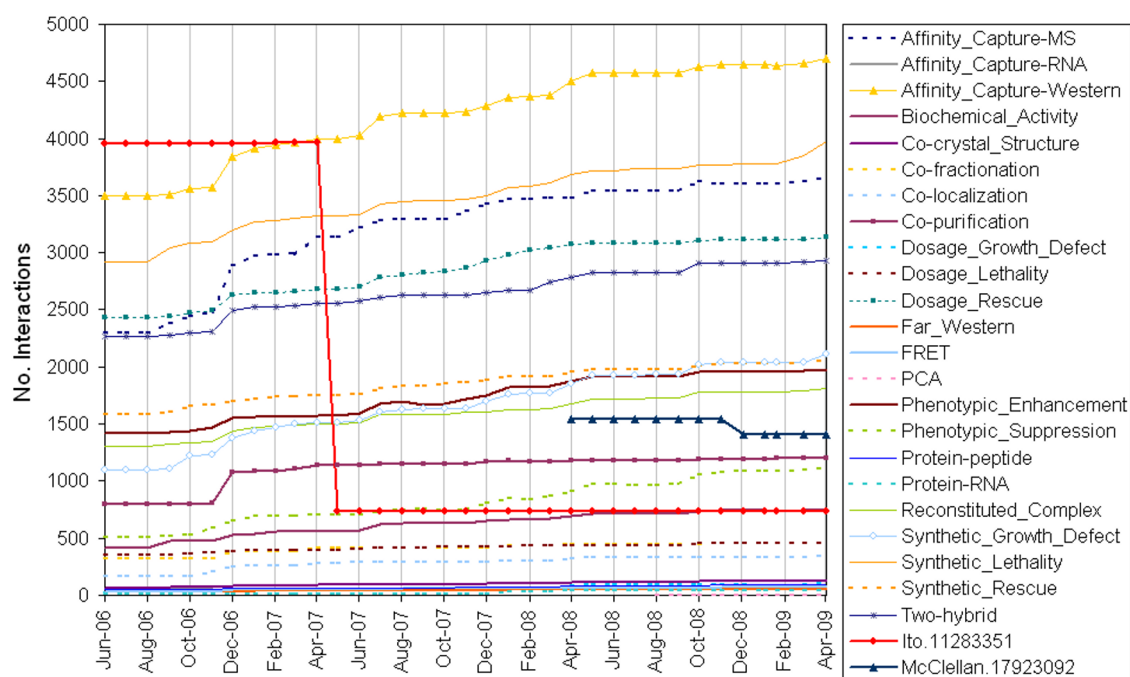


Figure 5.4: Significant changes in BioGRID dataset size from June 2006 to April 2009.

The LTP datasets all gradually increase in size. The majority of HTP datasets do not change by more than ten interactions and are not shown. Two significant changes (>100 interactions) occur in the HTP data: the Ito.11283351 dataset is reduced from around 4,000 interactions to less than 1,000 interactions, and the McClellan.17923092 dataset is reduced by 128 interactions.

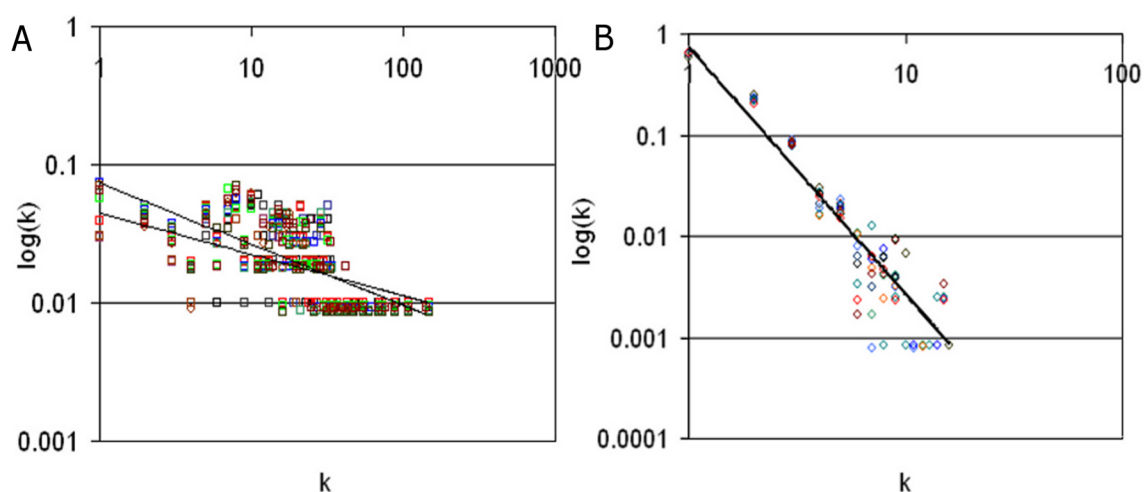


Figure 5.5: KEGG PATHWAY genes and pathways.

The distributions of genes and pathways in the KEGG database. Each point represents a single KEGG version. **A.** The number of genes per pathway. **B.** The number of pathway annotations per gene.

by KEGG did not follow the same pattern as the pairs. For example, between March 2007 and June 2007 the number of genes covered by KEGG increased by one while the number of pairs in the Gold Standard dropped by approximately 200 (Figure 5.6). Similarly, between December 2008 and March 2009, the size of KEGG increased by two genes, while the number of pairs decreased by around 450, and the connected components of the Gold Standard decreased by one. Comparison of the earliest KEGG Gold Standard with the latest version showed that the Gold Standards overlap by 1,127 genes and 35,201 pairs, with 85 genes unique to the earliest version (Figure 5.7) and 235 genes unique to the most recent (Figure 5.8).

Table 5.1: Summary of the Gold Standard produced from each KEGG version.

The changes in Gold Standard for each available version of KEGG. Due to the different release schedule of the KEGG archive each KEGG version corresponds to several BioGRID versions. For the remainder of this chapter each KEGG file is referred to by the earliest version number it covers.

Data	Versions	Total Genes	Pathways	Total Pairs	Connected Components
29/06/2006	V17-V19	1189	99	39770	10
29/09/2006	V20-V22	1189	100	39726	10
27/12/2006	V23-V25	1191	99	39733	10
28/03/2007	V26-V29	1200	99	39732	10
25/06/2007	V30-V32	1201	99	39534	10
24/09/2007	V33-V35	1206	101	39567	11
03/12/2007	V36-V38	1214	105	39715	11
24/03/2008	V39-V41	1238	108	40559	13
30/06/2008	V42-V44	1263	113	41012	12
29/09/2008	V45-V47	1294	115	41378	12
22/12/2008	V48-V50	1272	112	39984	12
30/03/2009	V51-V52	1274	108	39544	11

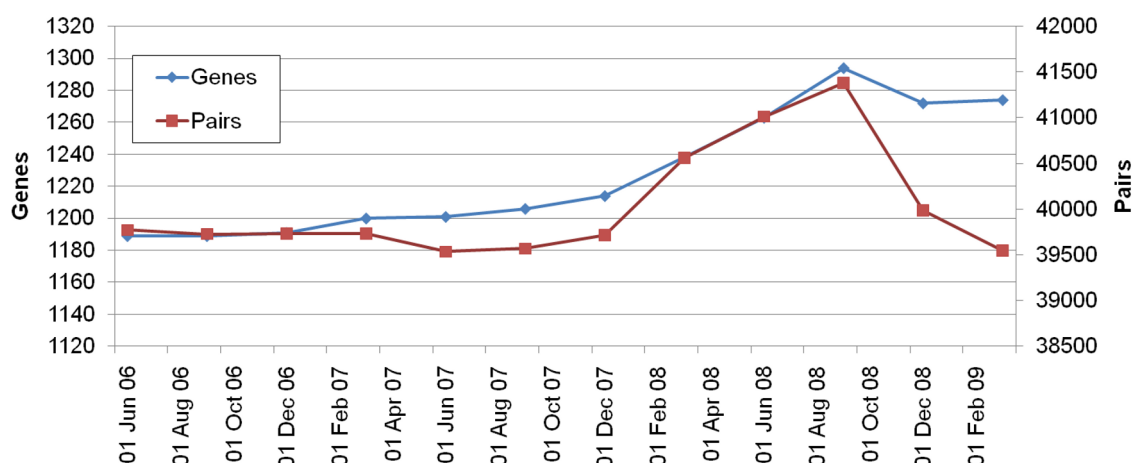


Figure 5.6: Comparison of Gold Standard size between KEGG versions.

Changes in the number of genes and pairs covered by the Gold Standards produced from the available KEGG versions.

5.2.3 Gene Ontology

The numbers of child terms associated with each of the three chosen POIs are presented in Figure 5.9 A. Since GO is structured as a hierarchy, genes annotated to child terms are automatically annotated to the parent term and are, therefore, used in functional prediction (see Section 3.1.4.2). The number of telomere maintenance terms increased from nine terms to 25 terms with two changes occurring in September 2006 and January 2009. DNA repair terms increased from 72 to 81 in a series of changes, including the addition and removal of several child terms between November 2006 and January 2007. The ageing-associated terms changed very little with only two changes: at April 2007 the term GO:0010261 (organ senescence sensu Magnoliophyta) was reassigned as an alternate ID for term GO:0010260 (organ senescence) and, in September 2008 a new term, GO:0034652 (extrachromosomal circular DNA localization during cell ageing) was added as a child term.

The number of genes annotated to each of the test terms is presented in Figure 5.9 B. The telomere maintenance annotations changed very little until August 2008, when 110 annotations were removed. The Gene Ontology Consortium records⁶ indicated that these annotations were based on studies by Gathbonton and colleagues [992] and Askree and colleagues [978]. A decision was made by the curators to represent the data from these studies as phenotypes rather than annotations, resulting in the drop in number of telomere maintenance annotations.

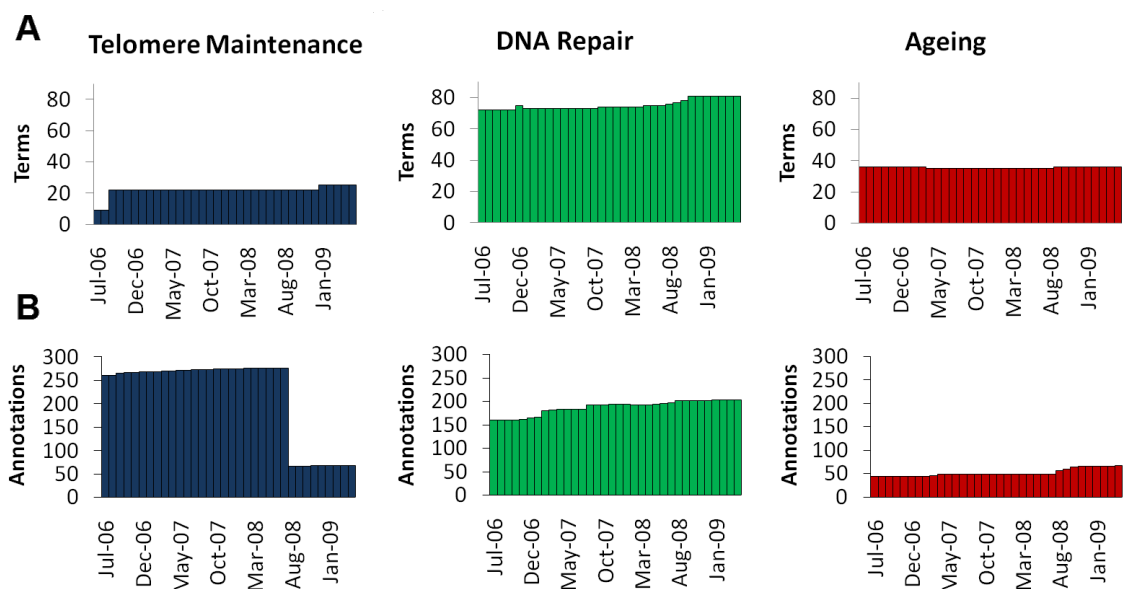


Figure 5.9: Gene Ontology terms and annotations associated with the three test terms.

A. The number of child terms in the GO hierarchy as the Gene Ontology version changes. **B.** The number of annotations to the three test terms in the SGD database for each version.

⁶http://wiki.geneontology.org/index.php/SGD_GO_HTTP_guidelines

The DNA repair annotations rose steadily, with the only decrease occurring in February 2008 when one annotation was removed. The ageing annotations also increased steadily with a single decrease of one annotation at June 2008.

5.3 Evaluation Strategy

Datasets were integrated using the two-step Bayesian statistics approach as described in Chapter 3. The resulting PFINs were evaluated using guilt-by-association for the functional prediction of the three POIs to produce ROC curves (see Section 3.1.5.4) [921].

A triplet of source datasets was required to build and evaluate each PFIN: a BioGRID file as the data source; a KEGG file as a Gold Standard against which to score the datasets; and GO annotation data for functional prediction (Figure 5.10). Each BioGRID-KEGG-GO triplet produces three ROC AUC measurements, representing the PFINs's functional prediction power for the three test GO terms.

In order to analyse the effects of data curation on the PFIN performance a three-stage approach was used as detailed below. In the initial two stages networks were integrated using the control network

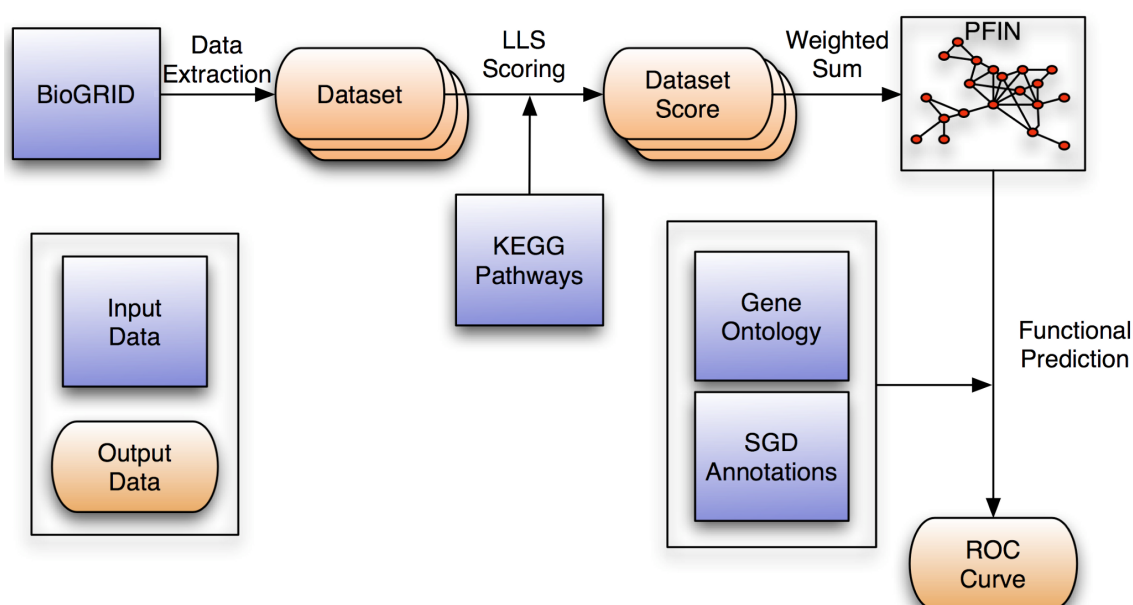


Figure 5.10: Overview of the integration and evaluation method.

Each PFIN is produced using four data sources-BioGRID, KEGG, Gene Ontology and SGD-shown as blue squares. Individual datasets are first extracted from the BioGRID database, and each is scored against the KEGG Gold Standard. Next, a weighted sum of the scores for each edge is calculated to produce the PFIN. The Gene Ontology and SGD annotation files are treated as a single data source, since they are dependent upon one another. Finally, leave-one-out functional prediction of known annotations is used to produce a ROC curve for evaluation.

integration method of Lee and colleagues, with no element of relevance (Section 3.1.4.4).

First, the combined effect of the changes to all three data sources over time was assessed using the BioGRID-KEGG-GO triplets that were current on each BioGRID release date. In total 35 PFINs were produced, one for each BioGRID version (V17-V52) scored using the corresponding KEGG file and evaluated using the corresponding GO file. This procedure produced 35 AUC measurements for each test GO term, each of which represented a monthly time-point in the curation of the three data sources (Figure 5.11 A).

In the second stage of evaluation, the contribution of each individual data source to the overall change in performance was assessed. PFINs were integrated and evaluated in which two data sources in the triplet were kept static and the third was iterated through all its available versions. This process was carried out in two temporal directions (oldest to newest and newest to oldest) to investigate the difference between the oldest and the newest file versions.

First, the oldest available version (V17) was used for the two static files, whilst the third file was varied through V18 to V52. These triplets of data sets were referred to as historic controls (HC). Each triplet in the group represented a monthly time-point of the varying data source (Figure 5.11 B). Fewer KEGG files than GO and BioGRID files were available, so there were fewer triplets required when the KEGG file was varied.

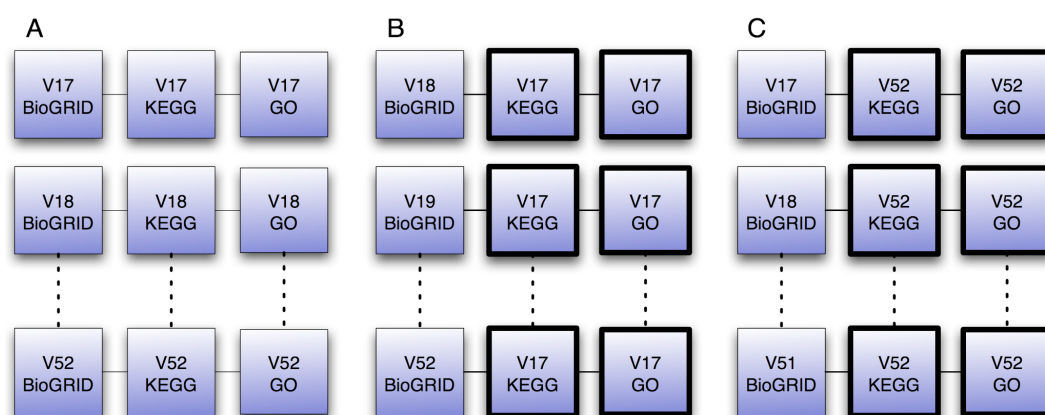


Figure 5.11: Examples of the data file triplets used in this study.

A. In the first stage of evaluation, triplets corresponding to the version dates of each BioGRID file were used to analyse the combined changes to the data sources by integration and evaluation of a PFIN. Each triplet represents a time-point in the curation of all three data sources. In the next stage of evaluation, groups of PFINs were generated in which two data sources were kept static while the other was changed through all its available versions. **B.** In the HC BioGRID study PFINs were produced using each BioGRID version and the V17 KEGG and GO files. **C.** In the RC BioGRID study PFINs were produced using each BioGRID version and the V52 KEGG and GO files. In both cases the static files are outlined in bold. The HC and RC evaluations were then repeated for the KEGG and GO files.

Next, the process was repeated in the reverse direction with the most recent file (V52) used for the two static files and the third file varied backwards through V51-V17. These triplets were referred to as recent controls (RC). Each triplet in the group represented also a monthly time-point of the varying data source (Figure 5.11 C).

Therefore, six groups of networks were produced, three historic controls and three recent controls (Table 5.2). The use of the two control file versions allowed comparison of the oldest and newest file versions (Figure 5.12). The datafile triplets summarised in Figure 5.11 are presented in full in Appendix G.

In the final stage of evaluation networks were generated with relevance to the three POIs using the RelCID integration method described in Section 3.1.4.5 and the file triplets described in Figure 5.11A. This procedure produced 35 additional relevance networks and AUC measurements for each POI, each of which represented a monthly time-point in the curation of the three data sources.

Two relevance score cut-offs were also chosen and PFINs integrated using file triplets of Figure 5.11 A, but excluding those datasets scoring above the cut-offs. This procedure produced a further 70 relevance networks and AUC measurements for each POI, which also represented a monthly time-point in the curation of the three data sources.

Table 5.2: Summary of the BioGRID-KEGG-GO triplets.

The file versions comprising the six groups of triplets used in this study. In total 158 PFINs were integrated and evaluated. The KEGG and GO file versions are numbered by the corresponding BioGRID version as set out in Figure 5.1.

Control	KEGG version	GO version	BioGRID version	Number of PFINs
BioGRID HC	17	17	18 – 52	34
KEGG HC	18 – 52	17	17	11
GO HC	17	18 – 52	17	34
BioGRID RC	52	52	17 – 51	34
KEGG RC	17 – 51	52	52	11
GO RC	52	17 – 51	52	34

5.4 Results

5.4.1 Combined Data Source Changes

Thirty-five versions of BioGRID, spanning a three-year period from July 2006 to May 2009, were available for *S. cerevisiae*. PFINs were integrated by scoring each version using the KEGG PATHWAYS version available at the same date (Appendix G). The PFINs ranged in size from 43,809 interactions to 74,234 interactions, with a gradual increase in size over time corresponding to the increase in size of the BioGRID dataset. The PFINs were evaluated by leave-one-out functional

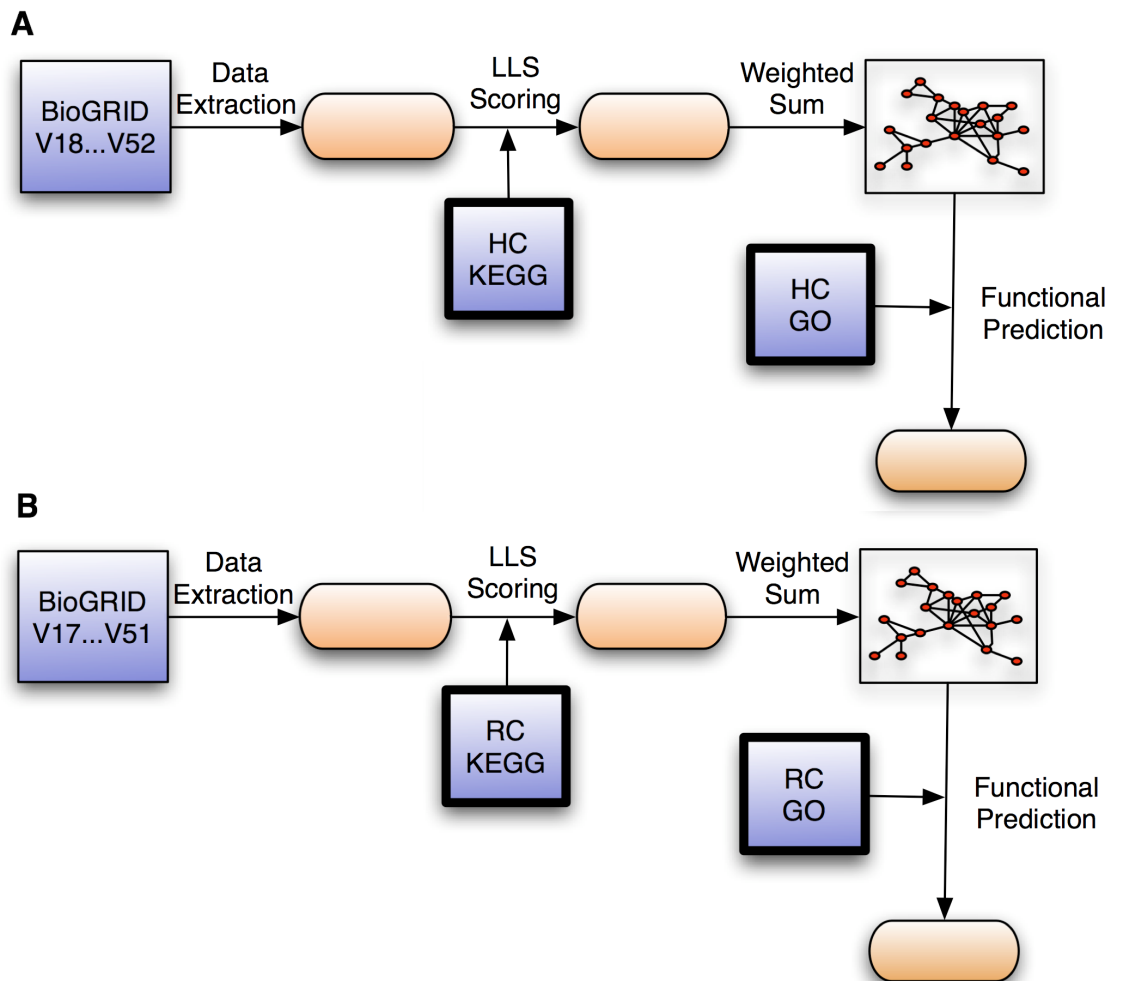


Figure 5.12: Overview of the HC and RC networks.

PFINs were integrated and evaluated in which two data sources in the triplet were kept static (outlined in bold) while the third was iterated through each of its available versions. **A.** In the historic control (HC) V17 is used for two of the triplet files while the other is varied. **B.** In the recent control (RC) V52 is used for two of the triplet files while the third is varied. In these examples, the BioGRID data file is varied, allowing comparison of the RC and HC file versions' performance. This procedure was then repeated for the KEGG and GO files, respectively.

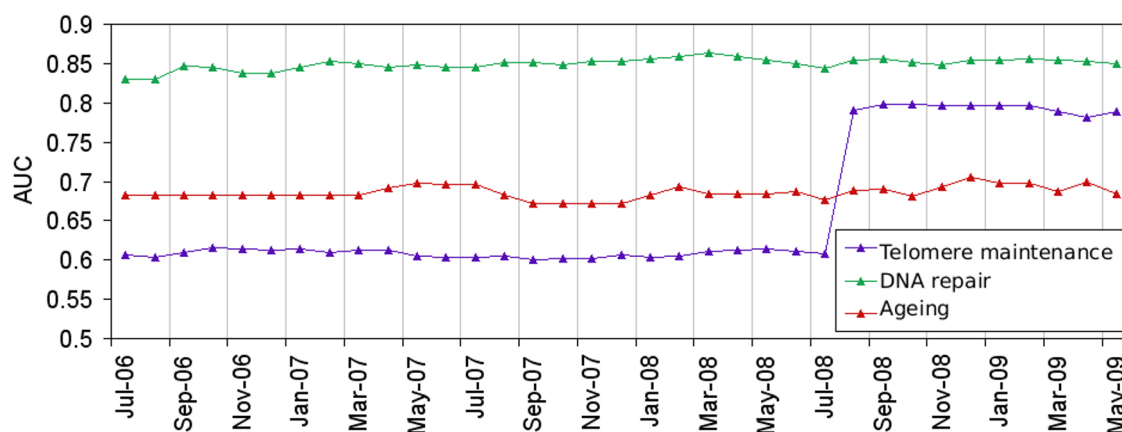


Figure 5.13: The combined effect of changes to the three data sources.

The functional prediction performance of the PFINs using the three test GO biological processes as all three data sources change.

prediction of known GO annotations current at the date of the BioGRID file. Three GO biological process terms were chosen for the evaluation: ageing (GO:0007568), DNA repair (GO:0006281) and telomere maintenance (GO:0000723). The AUC of the ROC curves was calculated for each PFIN in relation to the three test terms (Figure 5.13).

The AUC for telomere maintenance fluctuated between 0.6009 and 0.6159 from July 2006 until August 2008 when it increased to a value between 0.7812 and 0.7983 for the remaining versions. The highest scoring AUC for telomere maintenance was at September 2008, the lowest at September 2007, and the overall increase between first and last versions was 0.1815. The DNA repair AUC fluctuated between 0.8296 and 0.8635 for all versions, with the highest value at March 2008, the lowest at August 2006, and an overall increase of 0.0203. The AUC for ageing fluctuated between 0.6724 and 0.7059, with the highest value at December 2008, the lowest at November 2007, and an overall increase of 0.0010 between V17 and V52.

The standard error of the Wilcoxon statistic showed that all of the changes for telomere maintenance and DNA repair were statistically significant. However, while the range of changes for ageing was significant, the overall change between first and last versions was not (Table 5.3).

Table 5.3: Summary AUC measurements for the combined PFINs.

A summary of the AUC attained for the three test terms including range and Standard Error of the Wilcoxon statistic, SE(W). Statistically significant results are highlighted in bold.

GO Term	V17	V52	V52-V17	High	Low	Range	SE(W) V17	SE(W) V52
Telomere Maintenance	0.6072	0.7887	0.1815	0.7984	0.6009	0.1975	0.0004	0.0011
DNA Repair	0.8302	0.8506	0.0203	0.8635	0.8296	0.0339	0.0004	0.0003
Ageing	0.6824	0.6835	0.0010	0.7059	0.6724	0.0335	0.0020	0.0013

5.4.2 Individual Data Source Changes

In order to assess the individual data sources contribution to the changes in functional prediction performance, multiple [PFINs](#) were then integrated with a different version of one of the source files used for each network. Two control datasets were used: a historic control (HC) corresponding to V17, and a recent control (RC) corresponding to V52 (see Section [5.3](#)).

5.4.2.1 BioGRID

[PFINs](#) were constructed and evaluated for all of the available BioGRID versions using the control KEGG and GO files, resulting in 34 HC [PFINs](#) and 34 RC [PFINs](#) (Appendix [G](#)). The telomere maintenance [AUC](#) for the recent control [PFINs](#) was approximately 0.2 greater than that of the historic control [PFINs](#) for all BioGRID versions. The highest [AUC](#) attained for this term was at March 2007 while the lowest of was at September 2007. The historic control [PFINs](#) showed a slight overall decrease of 0.0075, while the recent control increased by 0.0130 (Figure [5.14](#)).

Using DNA repair as the test term the recent control [PFINs](#) produced a greater [AUC](#) than the historic control at all time points. The highest [AUC](#) for DNA repair was at September 2008 using the recent control and the lowest at July 2007 using the historic control. Overall the historic control [AUC](#) increased by 0.0001 and the recent control by 0.0177 (Figure [5.14](#)).

The ageing [AUC](#) fluctuated between 0.6 and 0.7 in both cases with the historic control [PFINs](#) performing slightly better until March 2008, after which the recent control performed better. The highest [AUC](#) was attained at December 2008 using the recent control; the lowest at August 2007, also using the recent control. The historic control showed a slight overall decrease of 0.0117 while the recent control improved by 0.0285 (Figure [5.14](#)).

In all three cases the range of changes to the [AUC](#) was statistically significant. However, while the overall decrease in the historic control telomere maintenance and ageing [AUCs](#) was also significant, the increase for DNA repair was not. The overall [AUC](#) increase for the recent control [PFINs](#) was significant in all three cases (Table [5.4](#)).

The integration method used to generate the [PFINs](#) ranked the datasets based on their confidence score prior to integration. The datasets were then integrated in order of this score, from highest to lowest, with successively lower weightings given to each dataset. Thus, the addition or removal of BioGRID data affected the order of integration. Datasets scored greater than 0.0 if they tended to link genes in the same pathway. Several of the datasets did not score using either of the Gold Standards and were therefore not used for [PFIN](#) integration (see Section [4.3.1](#)). Three datasets that scored using

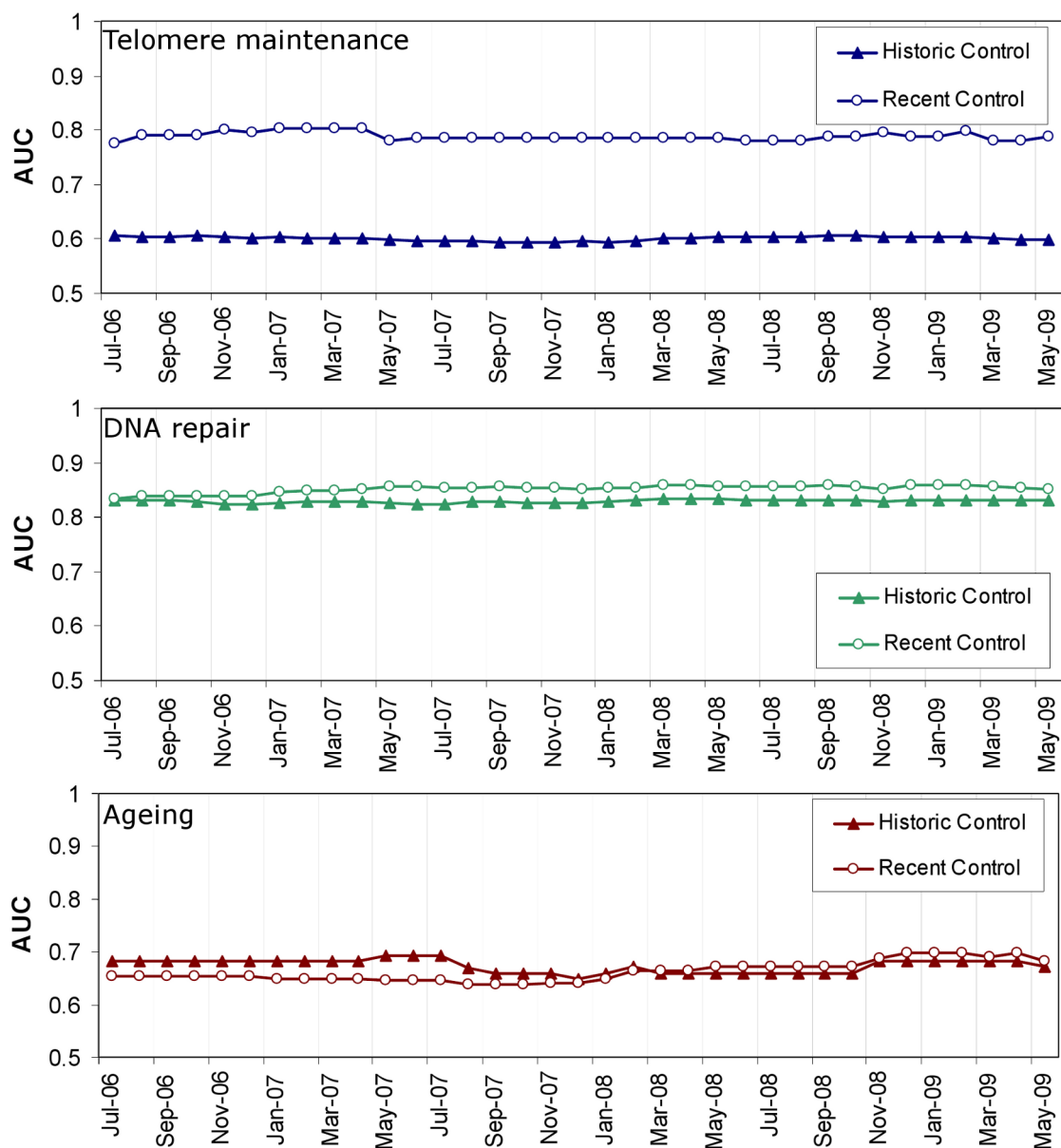


Figure 5.14: PFIN performances for the three test GO terms as the BioGRID file changes. Each graph depicts the changes in functional prediction performance of the PFINs as the BioGRID version is changed against the historic control V17 and recent control V52 files.

Table 5.4: Summary statistics for BioGRID network area-under-curve.

Summary of the AUC gained for the BioGRID data against the historic and recent control datasets, including the overall difference, range of change and Standard Error of the Wilcoxon statistic, SE(W). Statistically significant results are highlighted in bold.

PFINs	V17	V52	V52-V17	High	Low	Range	SE(W) V17	SE(W) V52
HC Telomere Maintenance	0.6072	0.5997	-0.0075	0.6072	0.5931	0.0141	0.0004	0.0004
RC Telomere Maintenance	0.7757	0.7887	0.0130	0.8032	0.7757	0.0274	0.0012	0.0012
HC Repair	0.8302	0.8303	1E-04	0.8329	0.8228	0.0101	0.0004	0.0004
RC Repair	0.8329	0.8506	0.0177	0.8596	0.8329	0.0268	0.0003	0.0003
HC Ageing	0.6824	0.6707	-0.0117	0.6939	0.6475	0.0463	0.0020	0.0020
RC Ageing	0.6549	0.6835	0.0280	0.6989	0.6390	0.0599	0.0014	0.0013

the recent control (Davierwala.16155567, Loeillet.15725626 and Pan.15525520) did not score using the historic control.

Dataset additions, removals and significant log-likelihood score (LLS) changes (>1.0) using the recent control, are presented in Figure 5.15. The majority of LTP datasets increased in confidence score over time with the exception of the Dosage Growth Defect dataset, which decreased. The Ito.11283351 dataset's score increased by 1.64 at the date its non-core data was removed (Figure 5.4). The datasets scored using the historic control showed a similar pattern of changes although the Ito.11283351 score increase was lower, in this case at 1.48.

The changes in LLS altered the final rankings of the datasets and, therefore, the order of their integration. Table 5.5 summarises the top ten ranked datasets for BioGRID versions V17 and V52 when scored using the recent control. Collins.17200106 and the evidence category PCA (Protein Complementation Assay) were new datasets added to the BioGRID database after V17. Both of these datasets scored highly. Consequently, while the LLS scores for Biochemical Activity, Reconstituted Complex and Newman.11087867 changed very little, they dropped down the ranks due to the addition of the higher scoring datasets. Dosage Growth Defect also dropped down the ranks due to a decrease in its score as the dataset grew (Figure 5.15). Inversely, FRET and Protein Peptide were LTP datasets that did not score highly at V17 but increased in score as BioGRID grew. In fact, the FRET dataset scored 0.0 until June 2008. A similar pattern of changes to the top ranked datasets was seen when scored using the historic control.

5.4.2.2 KEGG

PFINs were integrated for each KEGG version in turn using the two controls. In total 22 PFINs were produced: 11 historic controls and 11 recent control PFINs (Appendix G). The telomere maintenance AUC was approximately 0.2 greater for the recent control for all KEGG versions.

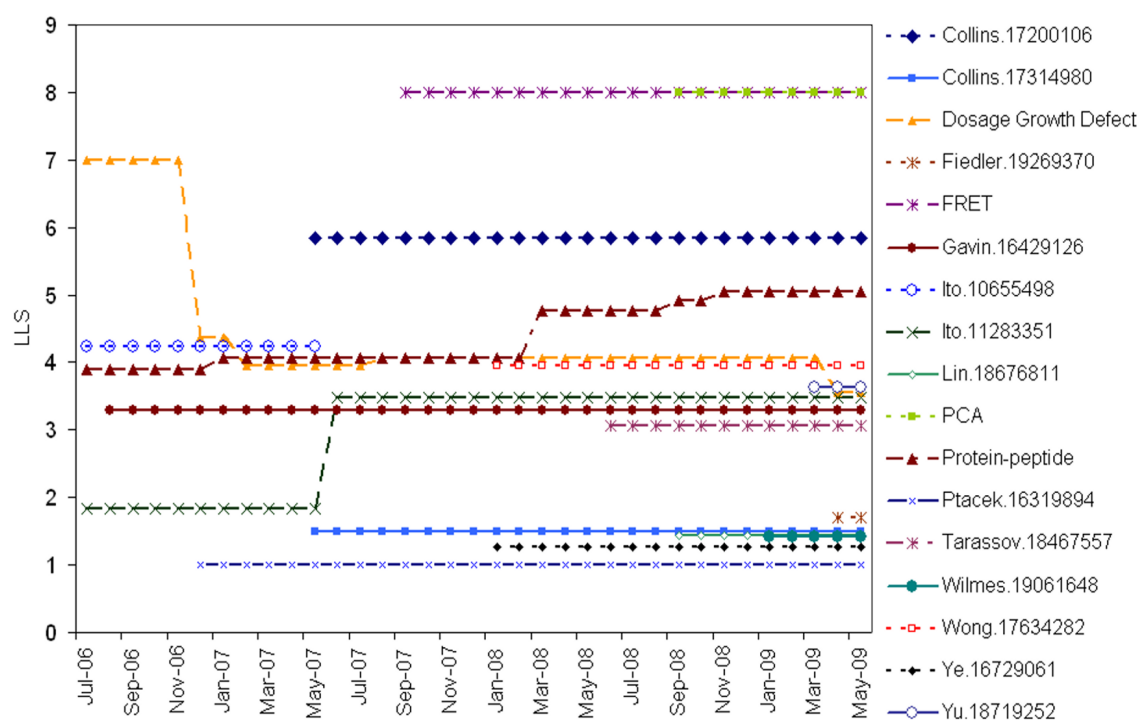


Figure 5.15: Summary of significant changes in LLS for the RC BioGRID PFINs.

Changes in LLS value greater than 1.0 and the addition and removal of datasets. These changes alter the final order of dataset integration and affect the performance of the integrated network.

Table 5.5: The top ten datasets when ranked by confidence score.

The top ten datasets of the RC KEGG scored PFINs for BioGRID version 17 and version 52, ranked in order of confidence score. This ranking is used to determine the order of dataset integration; changes in rank alter the final edge weights of the PFIN and consequently affect the performance of the integrated network.

V17		V52	
Rank	Dataset	Rank	Dataset
1	Ingvarsdottir.15657441	1	Ingvarsdottir.15657441
=	Tong.11743162	=	Tong.11743162
=	Dosage Growth Defect	=	PCA
4	Co-Crystal Structure	=	FRET
5	Krogan.14759368	5	Co-Crystal Structure
6	Co-Fractionation	6	Collins.17200106
7	Co-Purification	7	Krogan.14759368
8	Reconstituted Complex	8	Protein-Peptide
9	Newman.11087867	9	Co-Purification
10	Biochemical Activity	10	Co-Fractionation

The highest and lowest AUC measurements were both attained at December 2008 using the recent control and historic control, respectively. The historic control showed an overall decrease of 0.0046 while the recent control increased by 0.0001 (Figure 5.16).

The DNA repair recent control AUC was approximately 0.02 higher than the historic control AUC for all KEGG versions. The highest AUC for this term was at December 2007 and the lowest at March 2009. The historic control showed an overall decrease of 0.0187 while the recent control also decreased by 0.0125 (Figure 5.16).

The ageing AUC fluctuated at approximately 0.68 in both cases with the recent control performing slightly better at each time point. The highest AUC was attained at September 2006 using the recent control; the lowest was at September 2008, using the historic control. The historic control showed a slight overall decrease of 0.0118, while the recent control decreased by 0.0008 (Figure 5.16).

In all three cases the range of changes to the AUC was statistically significant. The overall decreases in both the historic and recent control measurements for DNA repair and ageing were also significant. However, while the decrease in historic control AUC for telomere maintenance was significant, the change for the recent control, which was the only overall increase, was not (Table 5.6).

Since the integration technique involves ranking datasets, and weighting their contribution according to their ranking, changes to the Gold Standard affect the LLS confidence scores, and thus change the final order of integration. The overall pattern of LLS changes for the KEGG PFINs was far more dynamic than that seen for the BioGRID PFINs in Figure 5.15. Rather than a steady increase or decrease in score, many of the dataset scores fluctuated between versions. In particular, there was a significant increase in three of the LTP datasets, Biochemical Activity, Dosage Lethality and Reconstituted Complex, at January 2006 (Figure 5.17 A), and a fluctuation in the same three datasets between October 2007 and April 2008 (Figure 5.17 B). Additionally three Synthetic Lethality datasets

Table 5.6: Summary statistics for KEGG network area under curve.

Summary of the AUC gained for the KEGG data against the HC and RC datasets including the overall difference, range of change and Standard Error of the Wilcoxon statistic, SE(W). Statistically significant results are highlighted in bold.

PFINs	V17	V52	V52-V17	High	Low	Range	SE(W) V17	SE(W) V52
HC Telomere Maintenance	0.6072	0.6026	-0.0046	0.6114	0.6026	0.0088	0.0004	0.0004
RC Telomere Maintenance	0.7886	0.7887	0.0001	0.7965	0.7811	0.0153	0.001	0.0011
HC Repair	0.8302	0.8115	-0.0187	0.8303	0.8115	0.0188	0.0004	0.0004
RC Repair	0.8631	0.8506	-0.0125	0.8636	0.8506	0.0130	0.0003	0.0003
HC Ageing	0.6824	0.6706	-0.0118	0.6825	0.6704	0.0121	0.0020	0.0020
RC Ageing	0.6910	0.6835	-0.0075	0.6910	0.6756	0.0154	0.0013	0.0013

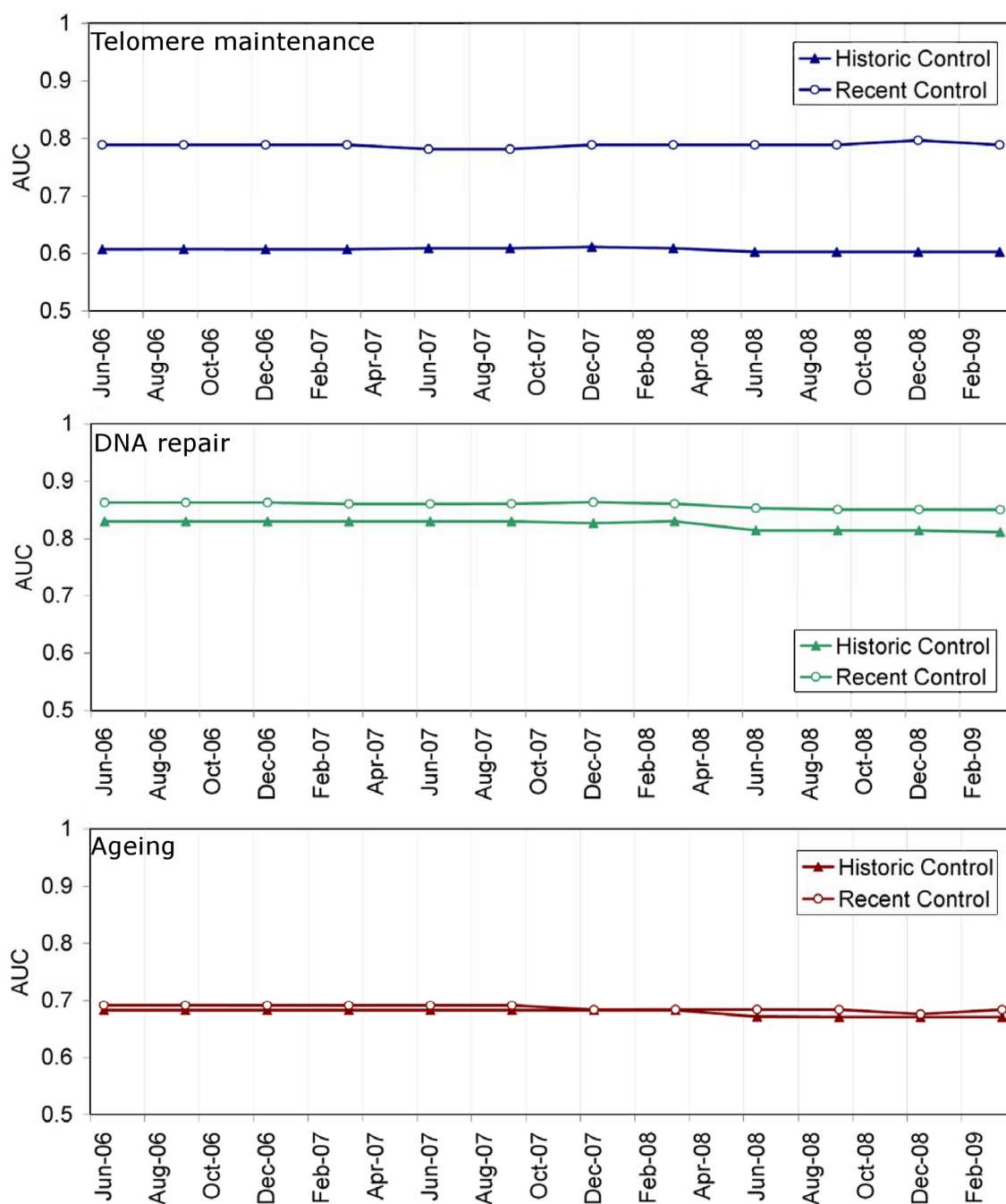


Figure 5.16: PFIN performances for the three test GO terms as the KEGG file changes. Each graph depicts the changes in functional prediction performance of the PFINs as the KEGG version is changed against the V17 historic control and the V52 recent control files.

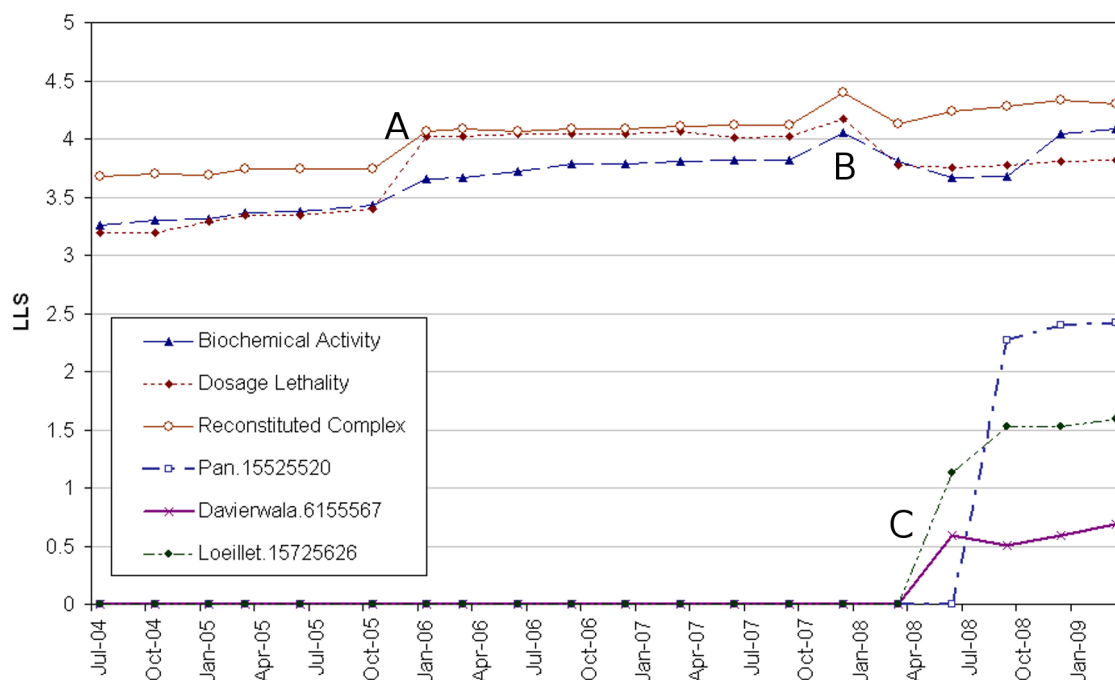


Figure 5.17: Significant changes in LLS score as KEGG changes.

A. An increase in three of the LTP datasets, Biochemical Activity, Dosage Lethality and Reconstituted Complex, at January 2006. **B.** A fluctuation in the Biochemical Activity, Dosage Lethality and Reconstituted Complex datasets between October 2007 and April 2008. **C.** The Loeillet.15725626 and Davierwala.6155567 datasets begin to score at June 2008, while the Pan.15525520 dataset began to score at September 2008.

(Pan.15525520, Loeillet.15725626 and Davierwala.6155567) did not score above 0.0 against the early versions of KEGG. The Loeillet.15725626 and Davierwala.6155567 datasets began to score above 0.0 at June 2008, while the Pan.15525520 dataset began to score above 0.0 at September 2008 (Figure 5.17 C).

The KEGG website provides details of changes to the pathways of the database⁷. Given that the positive Gold Standard is built from genes annotated to the same pathway in KEGG (Section 3.1.1), there are four types of changes to the database which could cause LLS scores to fluctuate:

1. New genes are added to KEGG which were not annotated before, causing an increase in both positive and negative Gold Standard pairs.
2. Genes are lost from KEGG due to their annotations being removed, reducing the number of positive and negative Gold Standard pairs.
3. New annotations are added to a gene, increasing the positive Gold Standard pairs and decreasing the negative pairs.

⁷<http://www.genome.jp/kegg/docs/updnote.html>

4. Annotations are removed from a gene, decreasing the Gold Standard positive pairs and increasing the negative pairs.

At January 2006 several pathways were changed by the addition and removal of genes, including the cell cycle (04110) pathway. A new pathway, snare interactions in vesicular transport (04130), was also added to the database. The additional genes of the cell cycle pathway overlapped with the Biochemical Activity, Dosage Lethality and Reconstituted Complex while the Reconstituted Complex dataset also had overlap with the the new snare interactions in vesicular transport pathway. These pathway changes accounted for the increase in these datasets' scores at this date.

Interestingly, the score changes between October 2007 and April 2008 could be mainly attributed to four genes annotated to the ubiquitin mediated proteolysis pathway (04120). The annotations were removed from KEGG at January 2008 and replaced at April 2008. The Biochemical Activity, Dosage Lethality and Reconstituted Complex datasets each contained at least one of these genes. However, the pairs that contained these genes overlapped the negative Gold Standard. Therefore, removal of the genes reduced the datasets' false positive count and, consequently, accounted for these datasets' score fluctuation between October 2007 and April 2008 (Table 5.7). Unfortunately, although KEGG is a highly curated database, the curators kept no record of the reason for the removal and subsequent re-addition of these annotations [1160].

Three new pathways were added to KEGG at June 2008: mismatch repair (03430), nucleotide excision repair (03420), and base excision repair (03410). Comparison of the KEGG pathway annotations with the Loeillet.15725626 and Davierwala.16155567 datasets indicated that the addition of these three pathways, together with changes to the pathway DNA Replication (03030), accounted for the datasets' positive scores after June 2008. Two pathways were also added to KEGG at September 2008: non-homologous end joining (03450), and homologous recombination (03440). Comparison of these pathways with the Pan.15525520 dataset indicated that the addition of these pathways accounted for this dataset's positive score after September 2008.

Table 5.7: Dataset true positives and false positives.

The true positive (TP) and false positive (FP) counts for the Biochemical Activity, Dosage Lethality and Reconstituted Complex datasets between V44 and V46. The removal of the annotations to four genes of the ubiquitin mediated proteolysis pathway (04120) reduces the datasets' FP count and increases the final LLS scores.

	V44		V45		V46	
Dataset	TP	FP	TP	FP	TP	FP
Biochemical Activity	118	45	118	36	118	47
Dosage Lethality	61	19	59	16	61	25
Reconstituted Complex	234	66	231	50	238	69

Table 5.8: The top ten datasets when ranked by confidence score.

The top ten datasets of the V52 BioGRID datasets scored against KEGG V17 and V52, ranked in order of confidence score. This ranking is used to determine the order of integration; changes in rank alter the final edge weights of the PFIN.

V17		V52	
Rank	Dataset	Rank	Dataset
1	FRET	1	FRET
=	Ingvarsdottir.15657441	=	Ingvarsdottir.15657441
=	Newman.11087867	=	PCA
=	PCA	=	Tong.11743162
=	Protein-peptide	5	Co-crystal Structure
=	Tong.11743162	6	Collins.17200106
7	Co-crystal Structure	7	Krogan.14759368
8	Collins.17200106	8	Protein-peptide
9	Krogan.14759368	9	Co-purification
10	Co-localization	10	Co-fractionation

The changes in [LLS](#) altered the final rankings of the datasets and, therefore the order of their integration. Table 5.8 summarises the top ten ranked datasets of the recent control data when scored using KEGG versions V17 and V52. While Newman.11087867 and Protein-Peptide scored highly against V17, both these datasets drop down the rankings when scored against V52. Inversely, Co-crystal Structure, Krogan.14759368 and Collins.17200106 have a higher ranking at V52 than at V17. A similar pattern of changes to the dataset ranks was seen for the historic control.

5.4.2.3 Gene Ontology

The [PFINs](#) integrated using V17 BioGRID-KEGG and V52 BioGRID-KEGG files were evaluated using each GO file version in turn during functional prediction, resulting in 68 [AUC](#) measurements for each of the three [POIs](#): 34 from the historic controls and 34 using the recent controls. While the [AUC](#) measurements for the three terms fluctuated, the changes were far less frequent than with the KEGG and BioGRID [PFINs](#) and the value of the GO [AUCs](#) remained unchanged for several consecutive versions.

The [AUC](#) for telomere maintenance followed a similar pattern as that of the combined networks and fluctuated around 0.61 until August 2008, when it increased to a value of approximately 0.79 for the remaining versions. The highest [AUC](#) attained was 0.7930 between August and October 2008 using the historic control. The highest [AUC](#) using the recent control was also recorded during the same period. Similarly, the lowest [AUCs](#) for both controls were recorded between July and August 2006, the two earliest versions. The historic control had an overall increase of 0.1793 while the recent control increased by 0.1965 (Figure 5.18).

The historic control DNA repair [AUC](#) was greater than that of the recent control until August 2007

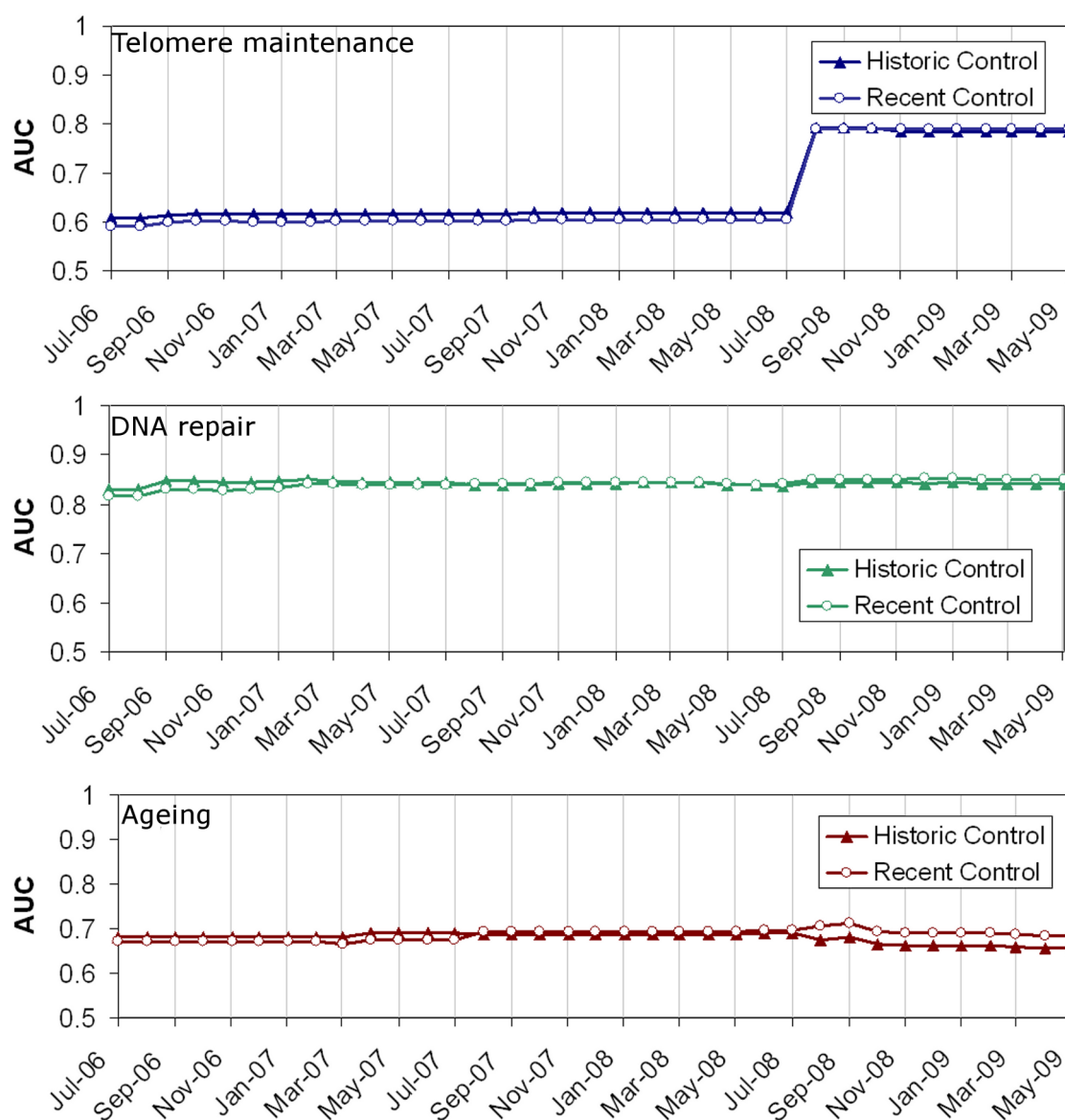


Figure 5.18: PFIN performances for the three test GO terms as the Gene Ontology files change. Each graph depicts the changes in functional prediction performance of the PFINs as the GO version is changed against the historic control and recent control files.

after which the recent control performed better. The highest AUC for this GO term was 0.8530 at January 2009 using the recent control and the lowest of 0.8150 between July and August 2006, also using the recent control. The historic control had an overall increase of 0.0107 while the recent control also increased by 0.0356 (Figure 5.18).

The historic control ageing AUC was greater than that of the recent control until August 2007, after which the recent control attained a higher score for the remaining versions. The highest AUC, 0.7125, was recorded at September 2008 using the recent control; the lowest was 0.6556 between April and May 2009 using the historic control. The ageing historic control showed an overall decrease of 0.0268, while the recent control decreased by 0.0128 (Figure 5.18). The range and overall changes were significant for all three terms (Table 5.9).

These changes in functional prediction of the POIs did not reflect the changes to the structure of the GO DAG (Figure 5.9). However, in several cases the changes in GO annotation to the POIs corresponded to the changes in performance. The 110 telomere maintenance annotations removed from the database in August 2008, caused an increase in functional prediction performance for both controls. The addition of nine DNA repair annotations at August 2007 corresponded with the increase in the recent control functional prediction above that of the historic control. Similarly, the addition of 8 ageing annotations at August 2008 corresponded with an increase in recent control performance and decrease in historic control performance at this version. Additionally, the addition of another ageing annotation in August 2007 corresponded with an increase in recent control performance.

5.4.3 Relevance Networks

In the final stage of evaluation networks were generated with relevance to the three POIs using the RelCID integration method described in Section 3.1.4.5 and the file triplets of Figure 5.11 A. The

Table 5.9: Summary statistics for GO network area under curves.

Summary of the AUC gained for the GO data against the HC and RC datasets including the overall difference, range of change and Standard Error of the Wilcoxon statistic, SE(W). Statistically significant results are highlighted in bold.

PFINs	V17	V52	V52-V17	High	Low	Range	SE(W) V17	SE(W) V52
HC Telomere Maintenance	0.6072	0.7835	0.1763	0.7931	0.6072	0.1859	0.0004	0.0012
RC Telomere maintenance	0.5923	0.7887	0.1965	0.7906	0.5923	0.1983	0.0004	0.0011
HC Repair	0.8302	0.8409	0.0107	0.8497	0.8302	0.0195	0.0004	0.0003
RC Repair	0.8150	0.8506	0.0356	0.8529	0.8150	0.0379	0.0004	0.0003
HC Ageing	0.6824	0.6556	-0.0268	0.6917	0.6556	0.0361	0.0020	0.0014
RC Ageing	0.6705	0.6835	0.0130	0.7125	0.6664	0.0462	0.0020	0.0013

relevance networks were compared with the combined networks' performance as a control (Section 5.4.1 Figure 5.13).

The AUC for telomere maintenance fluctuated between 0.6069 and 0.6198 until August 2008 when it increased to a value between 0.7712 and 0.7867 for the remaining versions. The highest score attained for telomere maintenance was at May 2009 (the most recent file), the lowest at April 2007. The overall increase between first and last versions was 0.1740, 0.0075 lower than the increase for the control network. The relevance network and control performances fluctuated at approximately the same level until September 2008 with the relevance network having a higher AUC at 29 of the 36 time points. After this date the relevance networks performance dropped below the control for the remaining time points (Figure 5.19).

The DNA repair AUC fluctuated between 0.8198 and 0.8732 for all versions, with the highest value at March 2009 and the lowest at July 2006 (the earliest version). The overall increase was 0.0534, 0.0331 greater than the control network's increase. The relevance network's performance was lower than the control network until April 2008, after which it performed higher than that of the control network (Figure 5.19).

The AUC for ageing fluctuated between 0.6770 and 0.7222, with the highest value at August 2008 and the lowest at April 2007. The overall increase between V17 and V52 was 0.0106, 0.0096 greater than the control network's increase. At 23 of the 26 time points the ageing relevance network's performance was higher than the control network (Figure 5.19).

The standard error of the Wilcoxon statistic showed that all of the changes for the relevance networks in comparison with the control were statistically significant. Additionally, the overall change between first and last versions was also significant for all three networks (Table 5.10).

5.4.4 Cut-Off Networks

Two relevance score cut-offs were also chosen (Figure 5.20) and PFINs integrated using only those datasets scoring below the cut-offs. This procedure produced 70 further AUC measurements for each POI; 35 at cut-off 0.1, and 35 at cut-off 0.001 (Figures 5.21 and 5.22).

Table 5.10: Summary AUC measurements for the relevance PFINs.

A summary of the AUC attained for the three test terms including range and Standard Error of the Wilcoxon statistic, SE(W). Statistically significant results highlighted in bold.

GO Term	V17	V52	V52-V17	High	Low	Range	SE(W) V17	SE(W) V52
Telomere Maintenance	0.6127	0.7867	0.1740	0.7867	0.6069	0.1798	0.0004	0.0011
DNA Repair	0.8198	0.8706	0.0534	0.8732	0.8198	0.0534	0.0004	0.0003
Ageing	0.6918	0.7026	0.0452	0.7222	0.6770	0.0452	0.0020	0.0013

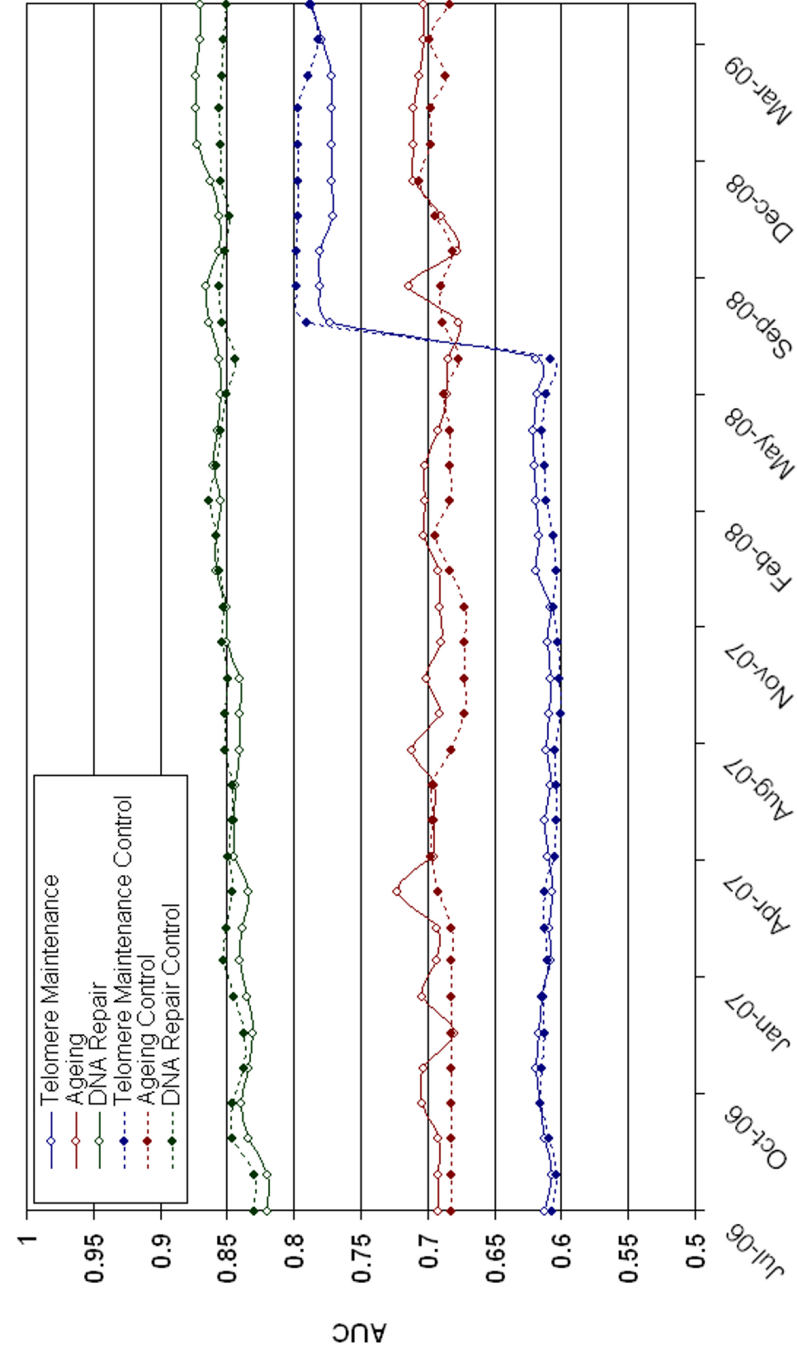


Figure 5.19: The combined effect of changes to the three data sources on relevance network performance.
The functional prediction performance of the relevance PFINs using the three test GO biological processes as all three data sources change.

In general the cut-off networks' performances followed similar patterns to those of the relevance networks with no cut-off. The telomere maintenance network AUCs were all significantly higher before September 2008 for both cut-off networks, with higher values than the control network at all time points, and the highest values attained using the 0.001 cut-off. The ageing cut-off networks fluctuated far more than the relevance networks and had a higher number of AUC values above the control, 32 of 36 at cut-off 0.1, and, 36 of 36 at cut-off 0.001. The DNA repair cut-off networks followed the same fluctuations as those of the relevance network with no significant changes. In all cases the cut-off networks produced a statistically significantly higher final AUC than the control (Table 5.11).

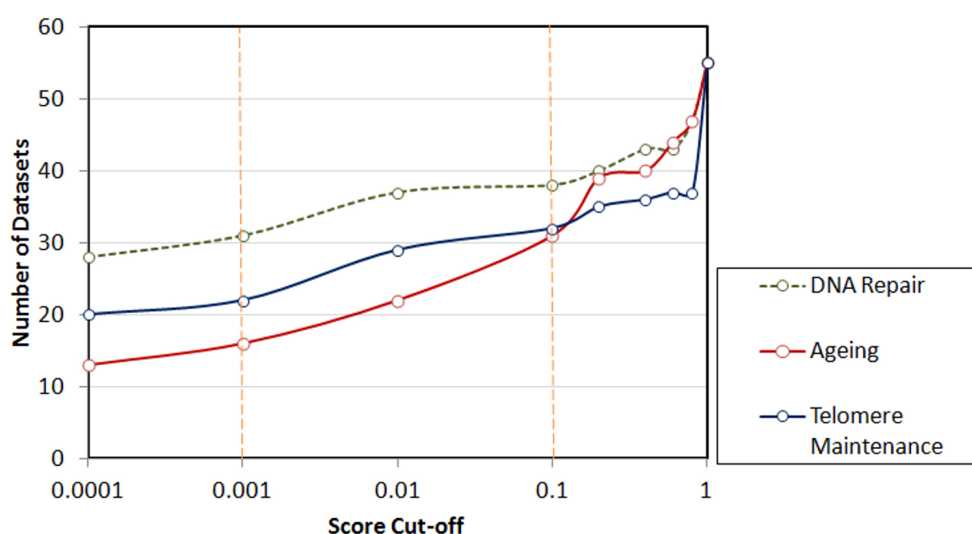


Figure 5.20: The relevance cut-off.

The number of datasets scoring at a range of relevance score cut-offs. Since a score of zero indicates high relevance and a score of one indicates low relevance, the number of datasets increases as the cut-off is raised. Two cut-offs were chosen for networks integration: 0.001 and 0.1 (vertical orange lines). At a cut-off of 0.001 all but the very high relevance datasets were excluded from the integration, while at a cut-off of 0.1 only very low relevance datasets were excluded.

Table 5.11: The relevance cut-off effect.

The final AUC at version 52 in comparison with the control network. With no cut-off the ageing and DNA repair final AUCs were higher than the control, while the telomere maintenance AUC was lower. After the cut-offs were applied, the relevance AUCs were higher than the control network's in all three cases. All the changes were statistically significant.

Network	Telomere Maintenance		DNA Repair		Ageing	
	AUC	Increase	AUC	Increase	AUC	Increase
No cut-off	0.7867	-0.0020	0.8706	+0.0202	0.7026	+0.0200
Cut-off 0.1	0.8008	+0.0121	0.8695	+0.0190	0.6941	+0.0117
Cut-off 0.001	0.7967	+0.0080	0.8697	+0.0191	0.7154	+0.0870

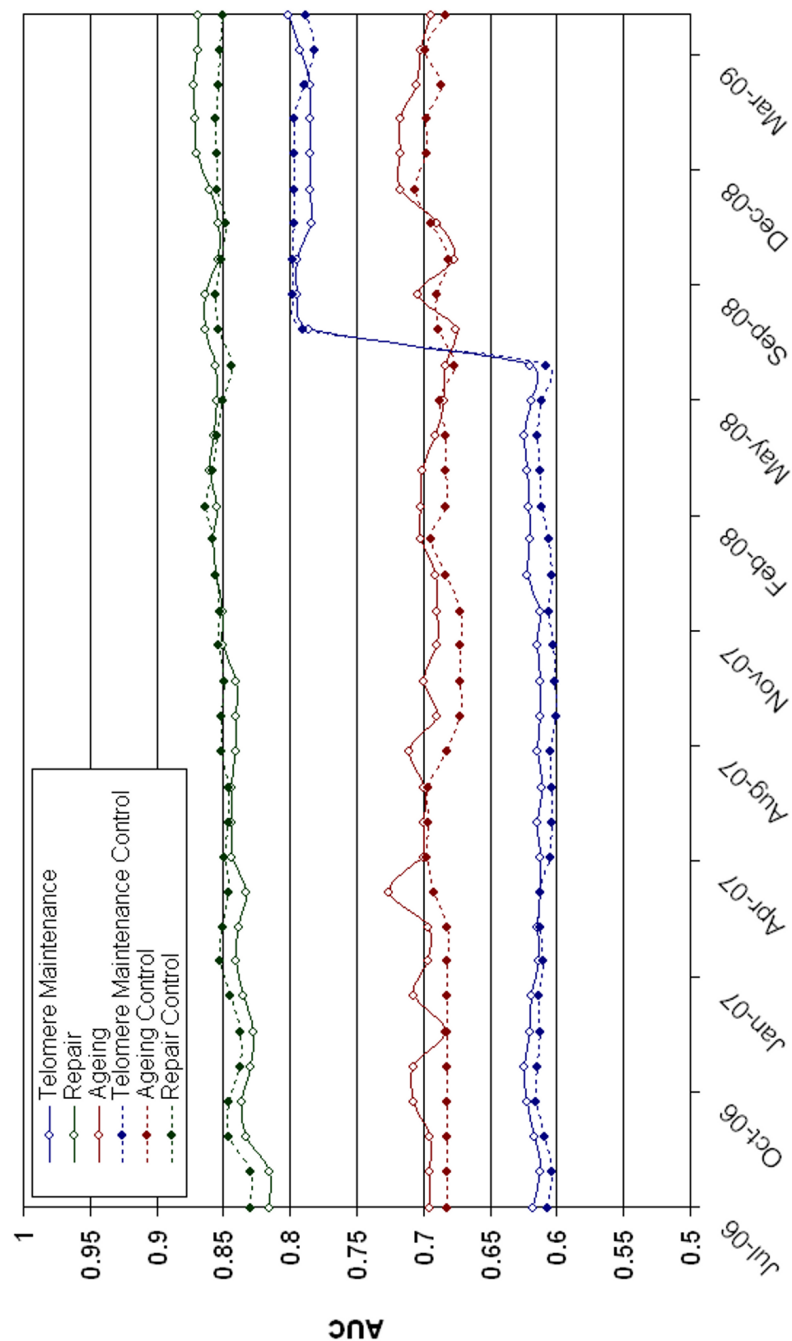


Figure 5.21: The combined effect of changes to the three data sources on relevance network integration at cut-off 0.1.

The functional prediction performance of the relevance PFINs using the three test GO biological processes as all three data sources change. The networks were integrated using only those datasets with a relevance score below 0.1.

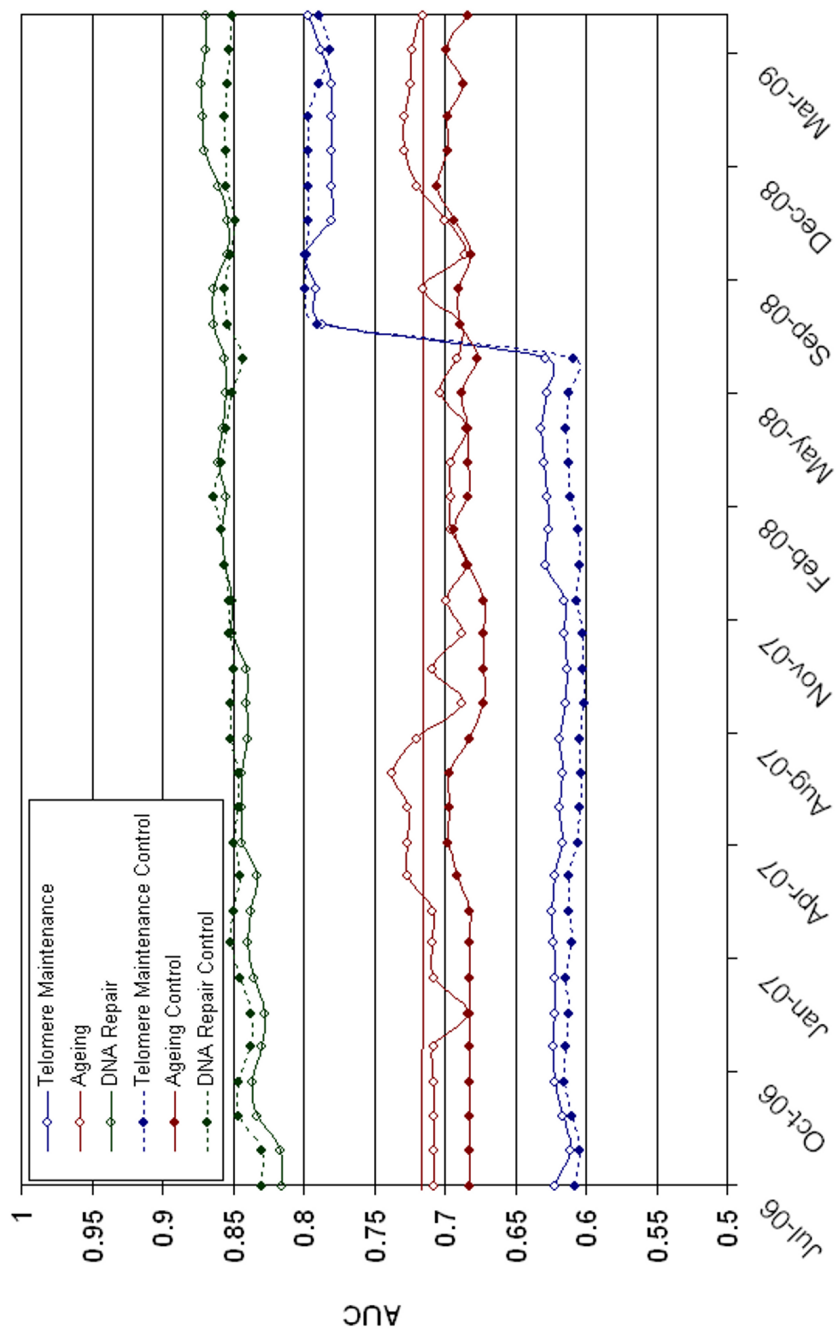


Figure 5.22: The combined effect of changes to the three data sources on relevance network integration at cut-off 0.001.

The functional prediction performance of the relevance PFINs using the three test GO biological processes as all three data sources change. The networks were integrated using only those datasets with a relevance score below 0.001.

5.5 Discussion

The assumption underlying much research into data integration is that the quality, as well as the quantity, of data in the major databases improves more-or-less steadily over time [1161]. For instance, the BioGRID database archive⁸ states:

*"This download directory contains BioGRID releases that have been retired and are no longer representative of the most recent state of our curated interaction set. If you are starting a new project using our data, it is **HIGHLY** recommended that you **NOT** use these data files as they are not the most up to date version of our interaction dataset. The best dataset to use is in the **CURRENT RELEASE** directory."*

While this assumption may be valid globally, it is not necessarily the case when integrated systems are used for a focused task such as the functional prediction of genes involved in specific biological processes [1161].

Here, PFINs were used to study three biological process terms relevant to yeast ageing [957, 1136] in order to further assess network performance in relevance to specific processes (Objective 3, Section 1.5). A chronologically-ordered series of PFINs was created using three highly curated data sources: BioGRID, KEGG and GO. Each PFIN was created using the versions of the datasets current at monthly timepoints. Therefore, the AUC measurements from these PFINs are those that would be seen if PFINs were updated monthly using the latest data. BioGRID was used as the source of the datasets [276] and KEGG as the Gold Standard against which each dataset's confidence was assessed [277]. The PFINs were evaluated by functional prediction of known GO annotations [100]. Three GO ageing-related biological process terms were chosen for the evaluation (telomere maintenance, DNA repair and ageing), and functional prediction performance was measured using the area under the ROC curve [920, 921].

The performance of the integrated networks fluctuated as source data changed through time. The overall pattern of change in functional prediction performance differed for each of the three terms. While the AUC for telomere maintenance was significantly improved between V17 and V52, this increase was not smooth. The ageing and DNA repair AUCs both fluctuated around a single value at all time points, but while two of the terms' overall increase was significant, that of the third was not. In fact, in none of the three cases was the highest AUC attained using the most recent V52 data: telomere maintenance was highest at September 2008, DNA repair at March 2008, and ageing at December 2008. While functional prediction performance increased for all three terms between V17 and V52, it is clear that any assumption of monotonically increasing performance over time is incorrect. However, as the three data sources are all changing simultaneously the extent

⁸<http://thebiogrid.org/download.php>

Table 5.12: Summary of the maximum and minimum AUC achieved.

The maximum and minimum AUC achieved for each of the test GO terms and the file versions used in each case.

Term		AUC	BioGRID	KEGG	GO
Telomere maintenance	Highest	0.80316	V25	V52	V52
	Lowest	0.59226	V52	V52	V17
DNA repair	Highest	0.86358	V52	V23	V52
	Lowest	0.81150	V17	V52	V17
Ageing	Highest	0.71254	V52	V52	V44
	Lowest	0.63901	V31	V52	V52

of their individual contributions is unclear from inspection of the combined results. Therefore, the contribution of individual data sources was systematically evaluated.

PFINs were created in which only one data source version was changed at a time. Each PFIN therefore represents a monthly time point in the curation of an individual data source, and changes in the AUC directly reflect changes in that data source. The analysis was carried out in two temporal directions (oldest to newest and newest to oldest), using V17 as a historic control and V52 as a recent control. Therefore the overall differences between the oldest and newest data could be investigated. Finally, the specific source dataset changes which caused the network performance changes were identified.

The recent control data outperformed the historic control. However, as with the overall changes, the highest performance was not produced by the most recent version of the data (Table 5.12), and performance fluctuated rather than steadily improving. Additionally, changes to the three source data types affected the PFIN performance in different ways.

The BioGRID database is manually curated to ensure the accurate entry of data from the original literature [276]. Additionally, feedback is encouraged from authors and BioGRID users to identify errors and correct any discrepancies⁹. Datasets produced using individual techniques each have their own biases towards different processes (see Section 2.4.4). While the addition, removal, or alteration of a dataset may not affect the functional prediction of one biological process, it may affect the prediction accuracy of another, if the dataset has a bias towards that process. Conversely, the addition of a high-confidence dataset with no relevance to the process being studied may negatively affect performance by masking other more relevant data. Integrated systems such as PFIN are used to generate new hypotheses (see Section 2.5.5), and when a particular biological process is of interest the choice of data prior to integration is vital in order to optimise performance in relation to that process. Importantly, these results indicate that the most recent raw data may not be the optimal data source for any given process of interest.

⁹http://wiki.thebiogrid.org/doku.php/contribute#send_us_your_interaction_data

High-quality Gold Standard data are commonly used to assess experimental dataset quality prior to integration [43, 59, 97, 669]. Here, the Gold Standard dataset was created by selecting all possible pairs of genes annotated to the same KEGG pathway [49, 112, 128, 702]. KEGG is also manually curated, containing species-specific pathways created from reference pathways and orthology data [277]. The scoring metric used during PFIN integration measures a similarity ratio between the Gold Standard and the dataset [49]. Therefore, dataset size and composition is important; a dataset may not score against a Gold Standard containing little similar information, and small changes in the similarity ratios, due to the curation processes of KEGG and BioGRID, can alter the final dataset scores. The overall changes in KEGG caused a statistically significant drop in performance for our three test GO terms between V17 and V52. However, since changes in score are due to changing ratios of similarity, it is likely that the performance of other terms would increase, particularly where the changes to KEGG involve genes annotated to that process. Consequently, the choice of an appropriate Gold Standard is important when scoring dataset confidence prior to data integration in a situation where a specific area of biology is of interest [1161]. If a Gold Standard dataset has little relevance to the process of interest the most useful datasets score poorly and performance may be low.

Annotation data are frequently used to evaluate the performance of an integrated network [104, 510, 676, 676, 907, 911, 919]. Here, the PFINs were evaluated by functional prediction of known GO annotations [100]. The evaluation data was downloaded from two sources: the Gene Ontology structure¹⁰ and the SGD annotation data for *S. cerevisiae*¹¹ [296]. Both of these data sources are manually curated from the literature and are updated regularly as new knowledge is gained. The Gene Ontology Consortium also maintains an extensive website¹² documenting their curation and annotation strategies, as well as any changes made to the database. Changes to the number of terms associated with our POIs did not significantly affect the functional prediction performance of the PFINs. However, the addition and removal of annotations to the terms were directly linked to functional prediction fluctuations. These results indicate that, while annotations are dependent on the GO structure, the observed changes in functional prediction performance are largely due to the curation of annotations, rather than to alteration of the GO structure itself.

As knowledge of cellular biology improves, changes to annotations datasets are inevitable. In many cases these changes are small, such as the removal of single annotations. However, occasionally large changes to annotation data will be required, such as the removal of telomere maintenance annotations observed in August 2008. This change involved the removal of 110 annotations derived from two phenotypic experiments [978, 992]. In these studies disruption of the annotated genes

¹⁰<http://www.geneontology.org/>

¹¹<http://www.yeastgenome.org/>

¹²http://wiki.geneontology.org/index.php/Main_Page

caused either lengthening or shortening of the telomere. While phenotypic data from disruption mutants can give clues to a gene's role, this is not necessarily the case. In fact, many genes in SGD currently have phenotypic data without any corresponding GO annotation, and vice versa. For instance, the gene *VMA1* has a mutant phenotype oxidative stress resistance, decreased¹³ [1041, 1162, 1163] but has no equivalent GO annotation. Conversely, the gene *GRX3* has a GO annotation to cellular response to oxidative stress [1057, 1060] but no associated mutant phenotype¹⁴.

Consequently, in August 2008 the SGD curators made the decision that phenotypic data alone was not enough evidence for annotation to telomere maintenance and the annotations were removed. It is likely that some of these genes will be annotated to telomere maintenance in the future following further experimentation. However, the removal of these annotations appears to have been the correct choice since network performance in functional prediction of telomere maintenance is significantly improved following their removal.

The potential for large changes to annotation datasets, such as the one in August 2008, suggests that in the case of annotation data the most recent data may be the most accurate, particularly when a single GO term is the focus of study. However, while the curation of annotation data can lead to increases in performance, such as that seen for telomere maintenance, there is no clear trend towards improvement and fluctuations do occur. Annotation-based evaluations are only as accurate as the available data and the most current data may not produce the best results. Further, since annotations are derived from experimental data, there is bias towards highly studied processes [133]. This characteristic of annotation data should be taken into account when evaluating a system's performance and interpreting results.

Most research groups using integrated systems such as PFINs are investigating a particular area of biology [128]. Although it is intuitively right to use the most up to data in all analyses, it may not produce the best results when investigating specific processes. Given the increasing volume of data available, it is vital that the correct data are chosen prior to integration, in order to optimise performance in relation to the process of interest. Individual datasets have unique biases towards different biological processes and, therefore, the most recent data may not produce the best results if it does not have relevance to the question being investigated. In fact, in many cases the addition of low-relevance data may mask the information contained in high-relevance data.

In order to minimise the observed effects, a chronologically-ordered series of process-relevant PFINs were created using the same data. Additionally, the integrations were repeated at two relevance cut-offs to produce networks integrated from only high-relevance datasets. As with the control

¹³<http://www.yeastgenome.org/cgi-bin/locus.fpl?locus=VMA1>

¹⁴<http://www.yeastgenome.org/cgi-bin/locus.fpl?locus=GRX3>

networks, the changes over time were not smooth, but a general trend towards improvement was observed. However, the relevance networks produced higher AUCs than the controls and all the changes were statistically significant, particularly when the cut-offs were applied to excluded low relevance datasets. Therefore, while performance still fluctuates, the use of process relevance during network integration can improve prediction performance and overcome some of the effects of dataset curation. These findings should be applicable to any type of integrated system used to study a specific area of biology (see Section 2.4).

The utility of process-relevant networks has been demonstrated in PFINs using yeast ageing as the POI. However, the relationship between dataset relevance and network performance is clearly complex. Ageing is just one of a wide range of cellular processes described by GO. Therefore, to understand and optimise the effects of process-relevant integration, and complete the third objective of this project, network performance must be evaluated in all areas of biology. The next chapter presents an evaluation of process-relevance using all available GOBP terms followed by the optimisation of the RelCID integration schema in light of the results of this evaluation.

Chapter 6

Assessment of GO Biological Processes as POIs and RelCID Performance Optimisation

In Chapter 4 a novel probabilistic network integration technique, RelCID, was presented. RelCID incorporates the datasets' relevance to specific biological processes, termed the **POI**, into the edge weightings of **PFINs**. A dataset's relevance is calculated based on the number of genes in the dataset that are annotated to a specific **GOBP** term. The performance of this technique has been demonstrated using the GO biological process of ageing as an exemplar. Functional prediction was significantly improved over that of the control network, integrated without a measure of relevance. In addition, network clustering produced larger clusters incorporating more nodes annotated to the **POI** and to related processes. The relevance integration technique also produced improved performance to that of the control network using datasets from different points in time (Chapter 5).

These improvements indicate that the bias of a dataset can be captured during integration, in addition to a measure of the dataset's quality, and used to improve network performance in relation to the ageing process. However, ageing is only one aspect of cellular biology described by GO, and functional datasets each have their own biases and are evolving over time. Additionally, some areas of biology, such as the ageing process, are more extensively studied than others. Therefore, it is unlikely that every **GOBP** term will perform equally well when selected as the **POI** during relevance integration.

In the current chapter the RelCID technique is evaluated using all available *S. cerevisiae* **GOBP** terms as **POIs** in turn. The differences in performance of the networks are assessed in light of the GO term, dataset and network properties. Finally, the relevance integration method is extended and optimised to include two further measures of a dataset's relevance to the **POI**.

6.1 Datasets

Version 52 of BioGRID was used as the source of the functional data and the datasets were split using a 100 interaction cut-Off as described in Section 3.1.1. The KEGG PATHWAY and GO datasets for *S. cerevisiae* that were current at Version 52 of BioGRID were used as the sources of the Gold Standard and evaluation data, as described in previous chapters.

6.2 A Full GOBP Sweep

PFINs were integrated using the control and relevance integration technique with a D value of 1.1 (Section 3.1.4) and evaluated by functional prediction of known annotations to each POI term using the maximum weight decision rule (Section 3.1.5.3). Annotations with the evidence code IEA were discarded as in previous evaluations.

All possible GOBP terms were initially chosen as potential POIs. Terms were discarded if they had no annotations in *S. cerevisiae* or if they were not annotated to any genes in the BioGRID dataset. Additionally, the root term biological_process (GO:0008152), was discarded since it is annotated to every yeast gene and would, therefore, always produce perfect classification during functional prediction. Consequently, a total of 2110 GOBP terms were used as POIs.

A set of 2110 relevance networks and a control network were produced. Two AUC measurements were calculated for each of the 2110 terms; one from the control network and one from the relevance network. The AUC for the control networks varied, with 507 terms scoring 0.5 indicating random assignment of the annotations, and 68 scoring 1.0 indicating perfect classification (Figure 6.1).

The extent to which performance was improved using the relevance networks varied with the POI chosen (Figure 6.2). In total, 60.1% of the term AUCs were improved using the relevance integration method and 5.2% of the AUCs were unchanged between the relevance and control networks. The majority of these unchanged terms were at the extremes of the AUC measurements, either having very poor classification in the control network, or close to perfect classification (Figure 6.1).

The overall change in AUC between the relevance and control networks ranged from an increase of 0.38, for the terms:

GO:0000949 - aromatic amino acid family catabolic process to alcohol via Ehrlich pathway

GO:0006559 - L-phenylalanine catabolic process

GO:0006569 - tryptophan catabolic process

GO:0042436 - indole-containing compound catabolic process

GO:0046218 - indolalkylamine catabolic process,

to a decrease of -0.33 for the terms:

GO:0010286 - heat acclimation

GO:0051352 - negative regulation of ligase activity

GO:0051436 - negative regulation of ubiquitin-protein ligase activity involved in mitotic cell cycle

GO:0051444 - negative regulation of ubiquitin-protein ligase activity

GO:0000117 - regulation of transcription involved in G2/M-phase of mitotic cell cycle.

6.3 GO Term Choice

There are four factors which may have affected the relevance network performance in comparison to the control network:

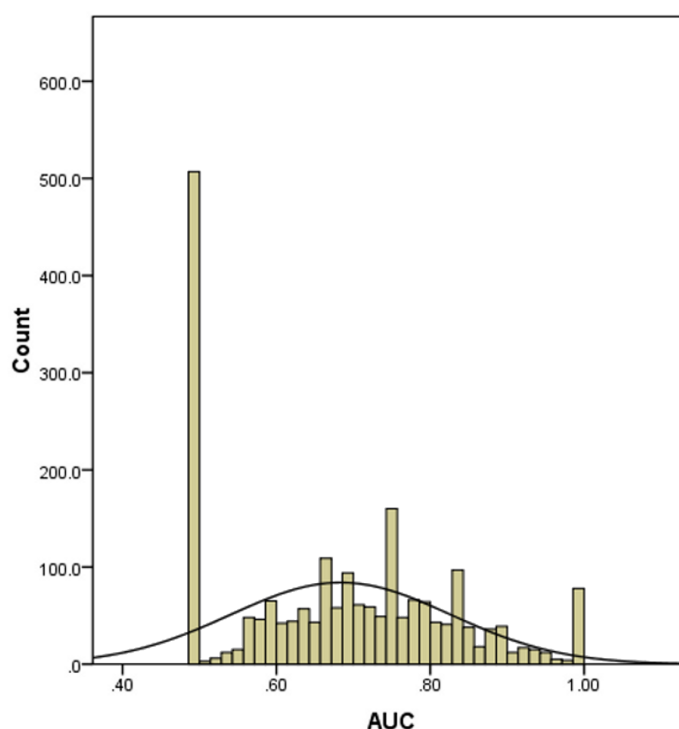


Figure 6.1: Distribution of the control AUCs.

The range of AUC measurements of all 2110 GO terms for the control network. In total 507 terms scored 0.5 indicating random assignment of the annotations during functional prediction, and 68 scored 1.0 indicating perfect classification.

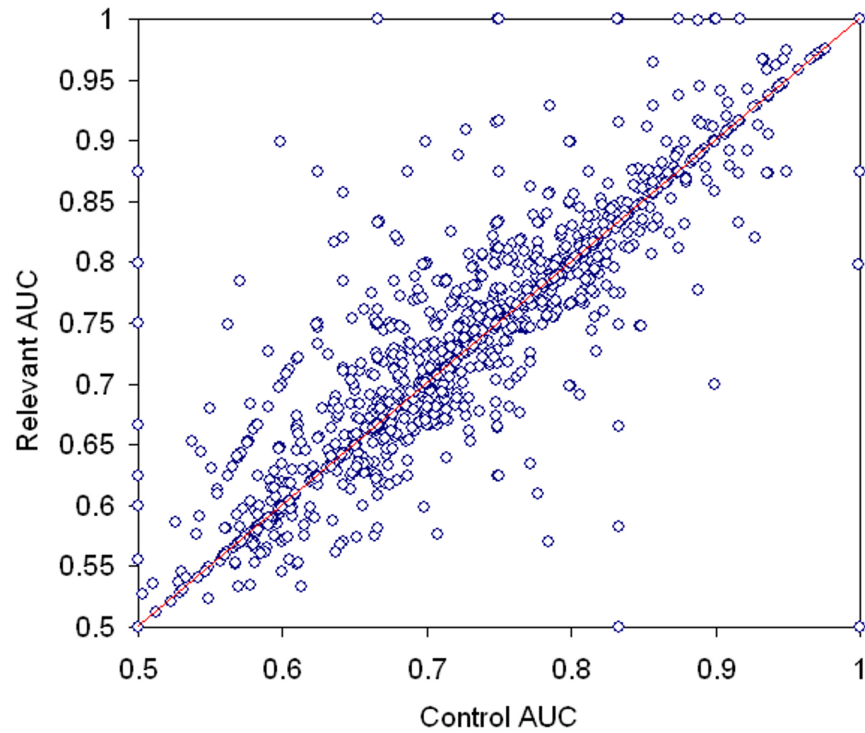


Figure 6.2: The full GOBP sweep.

The AUC of the relevance networks plotted against the control AUC for 2110 GOBP terms. Each point represents a single GO Biological Process term. Those above the diagonal have improved functional prediction when used as the POI, those on the diagonal were unchanged and those below were decreased. In total 60.1% of the AUCs were increased.

1. The GO term properties, such as size, specificity or the specific area of biology the term describes (see Section 2.5.4.3). For instance highly-annotated, highly-specific or highly-studied terms may have improved performance over less annotated, general or under-studied terms.
2. The network topological properties, such as the connectivity of nodes annotated to the [POI](#) (see Section 2.3.2). For instance, since the evaluation utilises local functional prediction, those terms with little connectivity between their annotated genes in the network may not perform well.
3. The scoring properties, such as the relationship between the confidence and relevance scores for highly ranked datasets. For instance, if high relevance datasets have low confidence scores their up-weighting would be limited and performance may not be increased.
4. Individual dataset's topological properties, such as the connectivity of nodes annotated to the [POI](#) term within the high relevance datasets. For instance, if the high relevance datasets have little connectivity between the nodes annotated to the [POI](#) their up-weighting may not increase network performance.

6.3.1 Term Properties

Due to the structure of the GO DAG, each GO term varies in its size and its specificity. A term's size is the number of genes annotated to that term (including genes annotated to its child terms). A term's specificity reflects how specific a biological process the term describes. For example, GOBP terms range from general processes, such as metabolic process (GO:0008152), to highly specific processes, such as protection from non-homologous end joining at telomere (GO:0031848). A term's size may be indicative of its specificity since general terms tend to have more annotations. However, high-specificity terms may have relatively high number of annotations if they describe heavily studied areas of cellular biology. High-specificity terms also tend to be lower in depth in the DAG but, again, this is not always the case (see Section 2.5.4.3). The information content measure of Resnik and colleagues (1999) [1140] measures term specificity by combining a term's size with its location within the GO DAG. This measure provides a more accurate assessment of a term's specificity than size or depth alone (Figure 6.3).

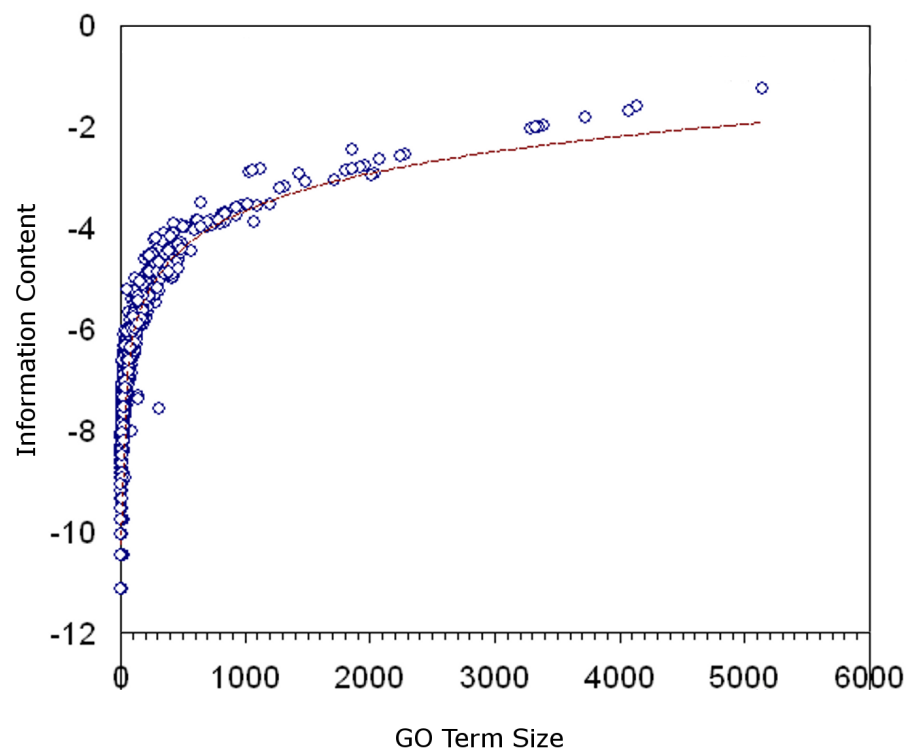


Figure 6.3: GO term size and specificity.

A term's size is the number of genes annotated to it (including to its child terms). A term's specificity reflects how specific a biological process it describes. Here, the term specificity is measured using the information content measure of Resnik and colleagues [1140]. Low specificity terms tend to have more annotations and are generally higher in the hierarchical DAG structure. Conversely, high-specificity terms tend to have fewer annotations and are lower in the DAG structure. However, this is not necessarily the case for highly studied areas of biology, and specificity measures such as this provide a more accurate measurement than size or depth alone.

The information content measure for each term was compared with the change in [AUC](#) values between the control and relevance networks to assess the influence of term specificity on functional prediction performance. Figure 6.4 A depicts the specificity of each term plotted against the change in [AUC](#) between relevance and control networks. The specificity of the terms did not correlate with the [AUC](#) changes to any great extent. In general, the lower the specificity of the terms, the less variable the [AUC](#) was (far right). However, high-specificity terms (far left) had highly variable [AUC](#)s with some terms significantly increased and some significantly decreased.

The specific area of biology represented by a term may also influence its performance as the [POI](#). Therefore, 144 terms which had relatively high specificity but which had low variability in their [AUC](#)s were selected for further analysis (Figure 6.4 B). These terms comprised 89 terms (62%) which had increased performance for the relevance network over the control and 55 terms (41%) which had decreased performance. None of the terms chosen had an unchanged performance between relevance and control networks. The connectivity of the terms in the GO [DAG](#) was visualised in Cytoscape to assess how the [DAG](#) structure related to network performance.

The majority of the chosen terms were connected in one large cluster in the [DAG](#) (Figure 6.5). The cluster consisted of a densely-connected central core with several distinct surrounding groups. A large group of terms (group 3) and three smaller groups of terms were not connected to the main group. The central core contained high-level processes, with a mix of increased (green) and decreased (red) [AUC](#)s. Conversely, the large surrounding groups of terms, and the unconnected groups, contained more specific terms and had distinct areas of increased and decreased [AUC](#)s.

For instance, the terms of group 3 are all transport-related terms 6.6. Although this group had a high level of connectivity in the [DAG](#), a distinct pattern of functional prediction performance was seen within the group (Figure 6.7). Terms directly below the term localisation (GO:0051179) in the GO [DAG](#) had improved performance for the relevance networks in comparison to the control, with the exception of those terms directly below, and including macromolecule localisation (GO:0033036), which all had decreased performance. Each of the other groups 1-10 of Figure 6.5 had similar patterns of distinction within their parent-child term relationships.

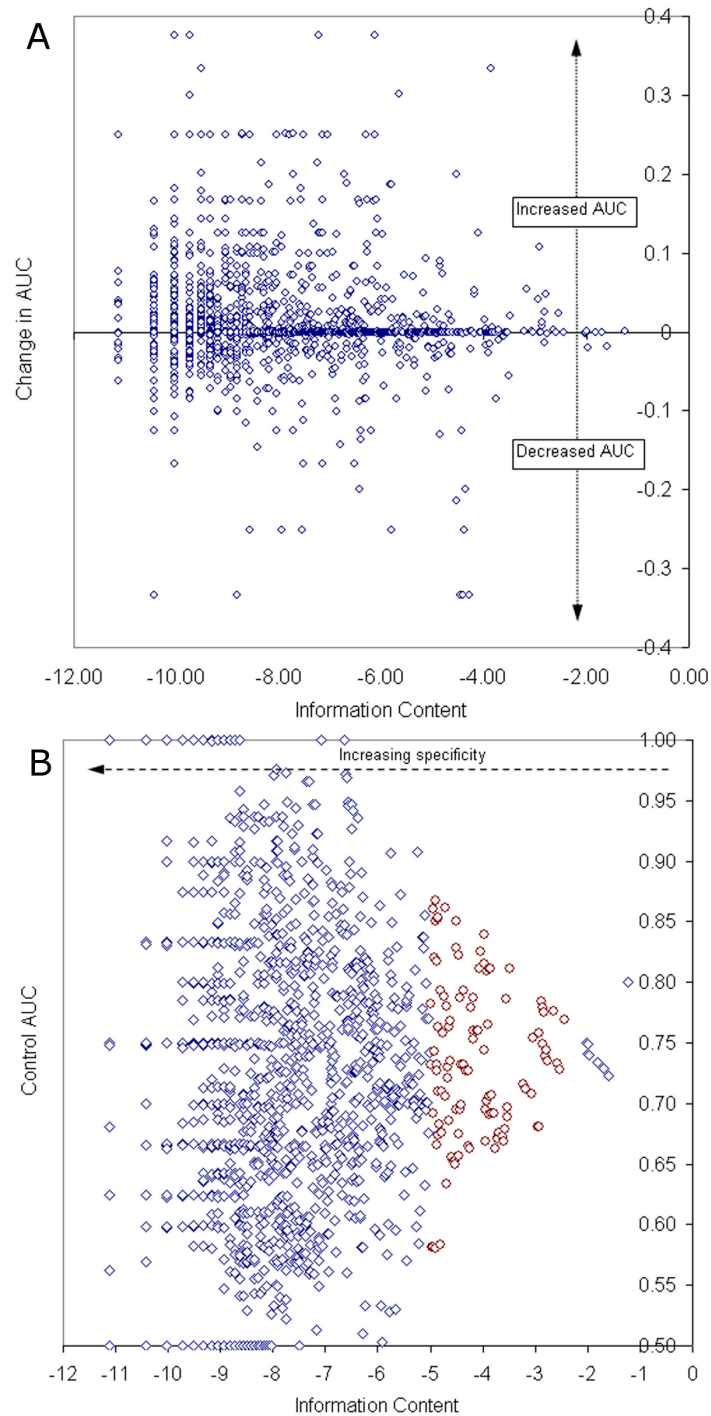


Figure 6.4: GO term specificity.

A. The GO terms' information content measures plotted against the change in AUC of the relevance network in comparison with the control network. Information content is a measure of how specific a biological process the term describes. Each point represents a single GO biological process. Terms above the line have increased AUC for the relevance network, terms below the line have decreased AUC for the relevance networks, and terms on the line have no change between the two networks. Term specificity decreases from left to right with high-level, low-specificity terms at the far right. **B.** The control network AUC plotted against the information content specificity measure for 2110 terms. Each point represents a single GO biological process, with the specificity of the terms increasing right to left as the information content score decreases. Variability of the AUC increases with specificity. A group of 144 terms with relatively low variability (shown in red) were chosen for further analysis.

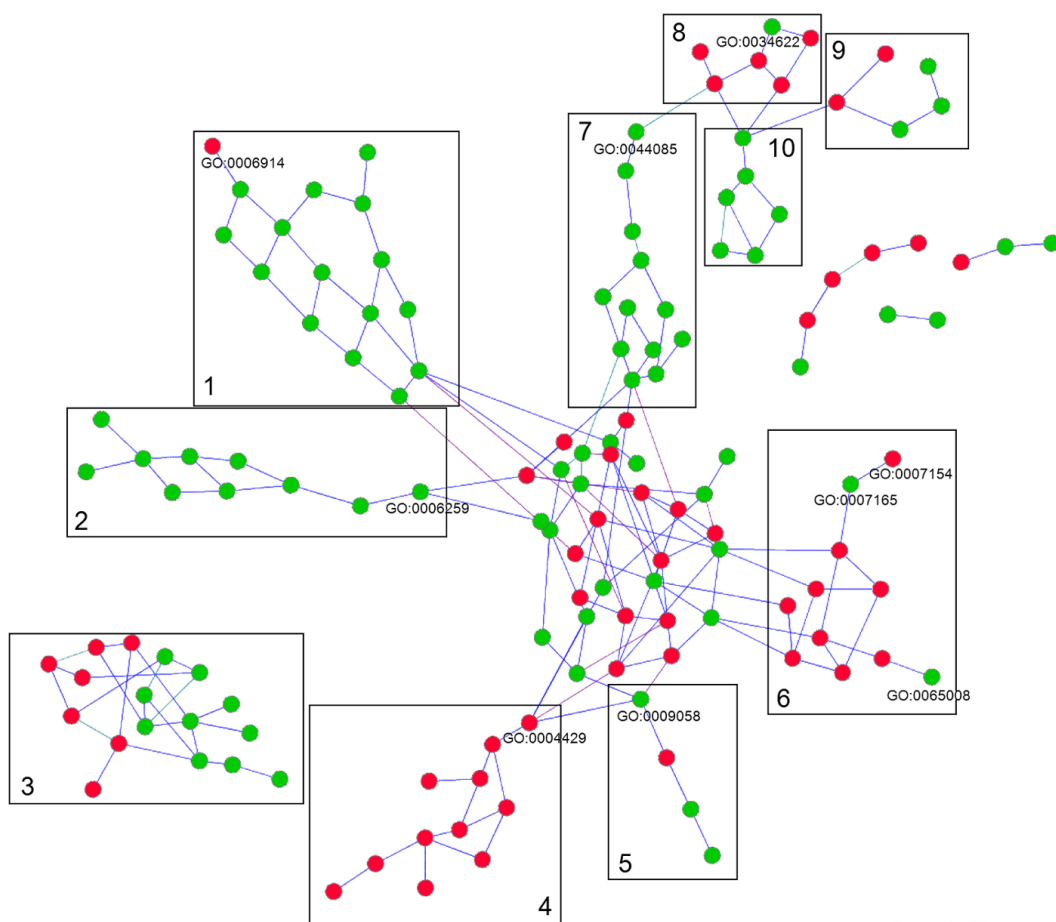


Figure 6.5: The 144 test terms.

The connectivity between the 144 test terms in the GO DAG. Green terms had increased AUC for the relevance networks in comparison with the control and red terms had decreased AUC. The terms form a highly connected core with several distinct surrounding groups (1-10). Group 3 and three smaller groups are not connected to the core. The highly connected core contains high-level terms while the groups each represent distinct lower-level areas of the GO DAG. The details of groups 1-10 are shown overleaf.

<p>1 GO:0006511 ubiquitin-dependent protein catabolic process GO:0043632 modification-dependent macromolecule catabolic process GO:0030163 protein catabolic process GO:0044265 cellular macromolecule catabolic process GO:0044257 cellular protein catabolic process GO:0034962 cellular biopolymer catabolic process GO:0019941 modification-dependent protein catabolic process GO:0019538 protein metabolic process GO:0051603 proteolysis involved in cellular protein catabolic process GO:0043285 biopolymer catabolic process GO:0009056 catabolic process GO:0009057 macromolecule catabolic process GO:0044248 cellular catabolic process GO:0006508 proteolysis GO:0044267 cellular protein metabolic process GO:0006914 autophagy</p>	<p>5 GO:0009058 biosynthetic process GO:0006629 lipid metabolic process GO:0044255 cellular lipid metabolic process GO:0008610 lipid biosynthetic process</p>
<p>2 GO:0009628 response to abiotic stimulus GO:0033554 cellular response to stress GO:0051716 cellular response to stimulus GO:0006950 response to stress GO:0050896 response to stimulus GO:0034984 cellular response to DNA damage GO:0006974 response to DNA damage stimulus GO:0042221 response to chemical stimulus GO:0006281 DNA repair GO:0006259 DNA metabolic process</p>	<p>6 GO:0007165 signal transduction GO:0065008 regulation of biological quality GO:0007154 cell communication GO:0048519 negative regulation of biological process GO:0031324 negative regulation of cellular metabolic process GO:0009892 negative regulation of metabolic process GO:0048523 negative regulation of cellular process GO:0065007 biological regulation GO:0050789 regulation of biological process GO:0050794 regulation of cellular process GO:0010605 negative regulation of macromolecule metabolic process</p>
<p>3 GO:0006810 transport GO:0006811 ion transport GO:0006913 nucleocytoplasmic transport GO:0016192 vesicle-mediated transport GO:0046907 intracellular transport GO:0051169 nuclear transport GO:0051179 localization GO:0051234 establishment of localization GO:0051641 cellular localization GO:0051649 establishment of localization in cell GO:0006605 protein targeting GO:0006886 intracellular protein transport GO:0008104 protein localization GO:0015031 protein transport GO:0033036 macromolecule localization GO:0034613 cellular protein localization GO:0045184 establishment of protein localization</p>	<p>7 GO:0042254 ribosome biogenesis GO:0016072 rRNA metabolic process GO:0006396 RNA processing GO:0022613 ribonucleoprotein complex biogenesis GO:0044085 cellular component biogenesis GO:0006364 rRNA processing GO:0006399 tRNA metabolic process GO:0034470 ncRNA processing GO:0016070 RNA metabolic process GO:0034660 ncRNA metabolic process GO:0006397 mRNA processing GO:0016071 mRNA metabolic process</p>
<p>4 GO:0044271 nitrogen compound biosynthetic process GO:0006082 organic acid metabolic process GO:0019752 carboxylic acid metabolic process GO:0006519 cellular amino acid and derivative metabolic process GO:0009309 amine biosynthetic process GO:0034641 cellular nitrogen compound metabolic process GO:0008652 amino acid biosynthetic process GO:0006520 amino acid metabolic process GO:0009308 cellular amine metabolic process GO:0006807 nitrogen compound metabolic process GO:0044249 cellular biosynthetic process</p>	<p>8 GO:0034622 cellular macromolecular complex assembly GO:0034621 cellular macromolecular complex subunit organization GO:0065003 macromolecular complex assembly GO:0022607 cellular component assembly GO:0010926 anatomical structure formation GO:0043933 macromolecular complex subunit organization</p>
	<p>9 GO:0006325 establishment or maintenance of chromatin architecture GO:0051276 chromosome organization GO:0016568 chromatin modification GO:0007005 mitochondrion organization GO:0006996 organelle organization</p>
	<p>10 GO:0032502 developmental process GO:0032989 cellular component morphogenesis GO:0016043 cellular component organization GO:0009653 anatomical structure morphogenesis GO:0048869 cellular developmental process GO:0048856 anatomical structure development</p>

Figure 6.6: The 144 test term clusters

The term details for groups 1-10 of 6.5. Terms with increased AUC between the relevance and control networks are shown in green and those with decreased AUC in red. Each group contains terms related to a distinct area of cellular biology.

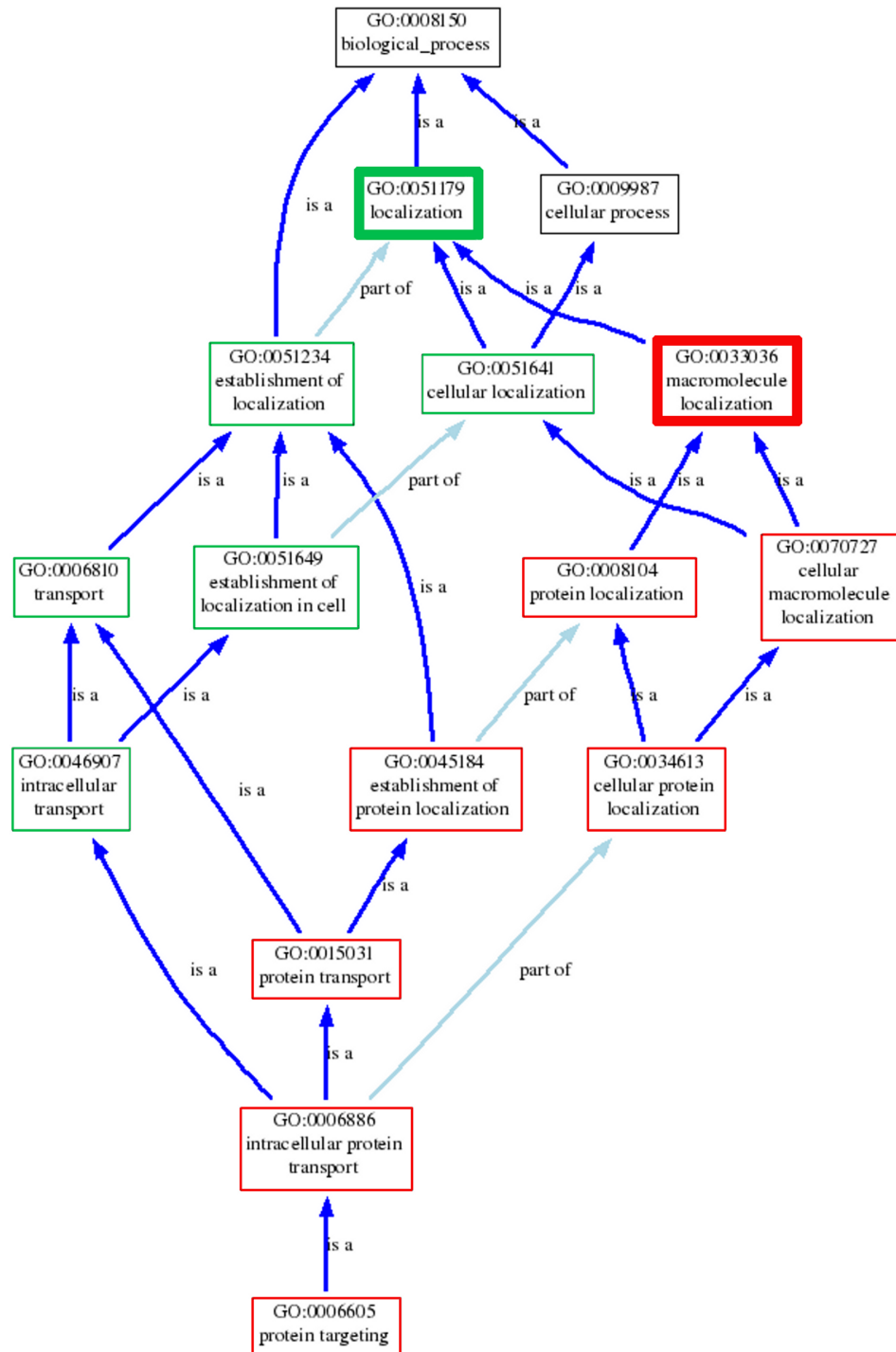


Figure 6.7: Connectivity of the group 3 terms.

The GO DAG structure for terms of group 3 in Figure 6.5. Those in green boxes had increased AUC in the relevance networks while those in red boxes were decreased. A clear topological distinction can be seen between the terms with increased performance (green boxes) and those with decreased performance (red boxes). Terms below the term localisation (GO:0051179) are increased for the relevance networks in comparison to the control, with the exception of those terms directly below, and including macromolecule localisation (GO:0033036), which are all decreased.

6.3.2 Network Topology

Since the functional prediction algorithm chosen for the GO term sweep utilised local rather than global connectivity, the topology of the network in relation to genes annotated to the **POI** may reflect the network's functional prediction performance. For instance, if nodes annotated to the **POI** term have no or low connectivity to one another in the network, performance will be low for the control network. Additionally, if this is the case, the relevance networks' performance will not improve since the relevance measure takes all annotations to the **POI** into account, many of which would not influence functional prediction performance. In other words, while nodes with high relevance have up-weighted edges, the up-weighting would not affect functional prediction due to lack of direct connectivity.

To measure the effect of connectivity on network performance several measures of node connectivity were calculated and compared with the change in **AUC** between the relevance and control networks. First, a baseline **AUC** was calculated for the 2110 available terms by transferring annotations to each **POI** along any connected edge, irrespective of weighting. Therefore, any nodes annotated to the **POI** which had local connections to one another would be correctly classified during functional prediction. This baseline score, therefore, represented a measure of **POI** connectivity within the network's topology.

Terms with a low baseline **AUC** tended to have little or no change in **AUC** between the relevance and control networks (Figure 6.8). Conversely, terms with a high baseline **AUC** have high variability in **AUC** changes, with some terms significantly increased and some significantly decreased.

Two further measures of **POI** connectivity were calculated; the percentage of nodes annotated to the **POI** which were directly connected to each other was calculated, and the average shortest path between all pairs of nodes annotated to the **POIs** was calculated using Dijkstra's algorithm (see Section 3.1.5.1) [1142].

The **AUC** for terms with no or very low connectivity tended not to change irrespective of the term's average shortest path (Figure 6.9). Additionally, terms with a large average shortest path (>3) tended to be less variable in **AUC**. However, neither the average shortest path nor percentage connectivity correlated with the change in **AUC** to any great extent.

6.3.3 Dataset Ranking

The relevance scores of the datasets in relation to the **POI** may influence network performance. For instance, a **POI** with very few high relevance datasets may not perform well during relevance integration, irrespective of the dataset's confidence scores, since there would be too little relevant

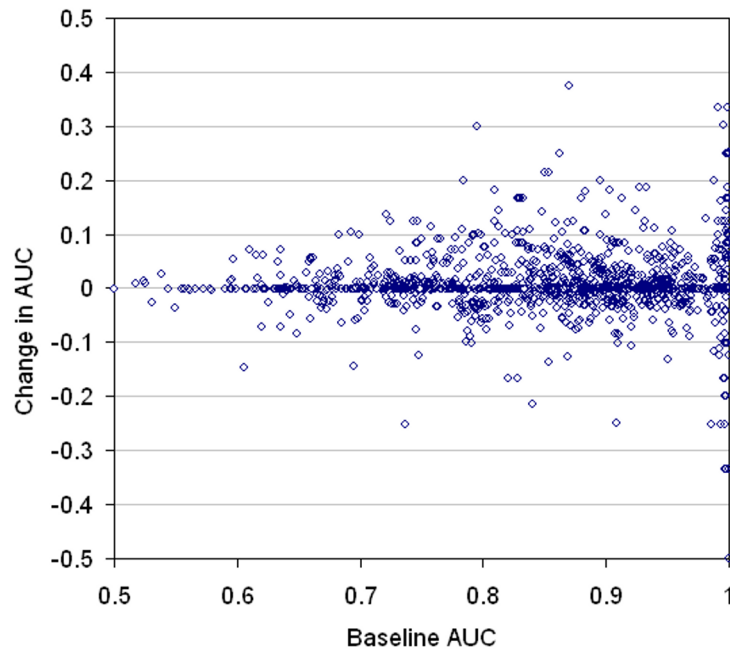


Figure 6.8: The baseline AUC.

The baseline AUC plotted against the change in AUC between the relevance and control networks for each term. The baseline AUC was calculated by transferring annotations to the POI along any connected edge irrespective of weighting and, therefore, provides a basic measure of connectivity between terms annotated to the POI. Each point represents a single GO term with the connectivity of genes annotated to the terms increasing from left to right.

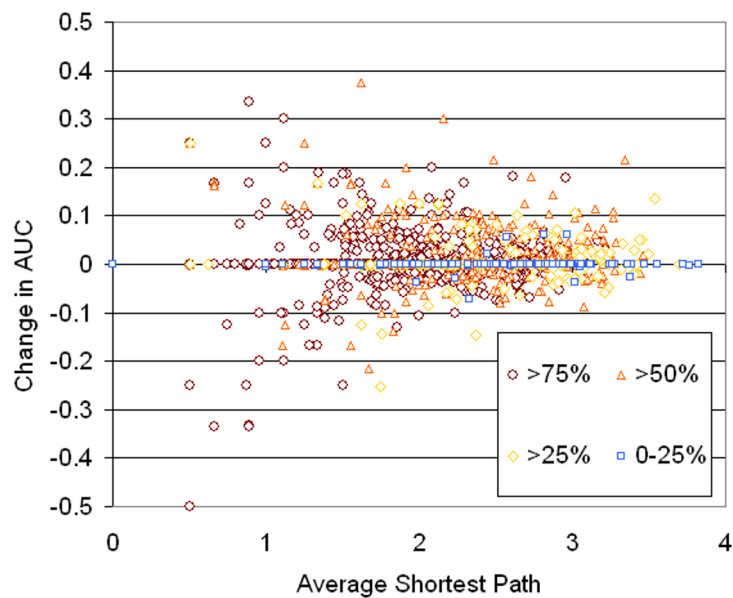


Figure 6.9: Average shortest path and percentage connectivity.

The average shortest path plotted against the change in AUC between the control and relevance networks. Each point represents a single GO term and is coloured by percentage connectivity of the genes annotated to this term.

data to influence the edge weights to any great extent. The dataset relevance ranks and scores for the 144 test terms of Figure 6.5 are depicted in Figure 6.10. Each line represents the ranked order of datasets in relation to a single GO term. The further right the vertical part of the line, the more high-relevance datasets are related to that term. At the left, several terms with few low relevance datasets performed poorly in comparison to the control (shown in red). However, there are also terms with low relevance datasets which performed well (green). Conversely, while the datasets the far right had a large number of high-relevance datasets, a significant number performed poorly.

The relationships between the datasets' confidence and relevance scores may also influence network performance. For instance, if the high relevance datasets have very low confidence the effect of the relevance ranked integration may not up-weight the most relevant edges significantly in comparison to the high-confidence control. Therefore, the relevance networks' performance may be reduced in comparison with the control's performance.

Due to the weighted sum used for integration, the two highest ranked datasets had the greatest contribution to the edge weights (Section 3.1.4.4). Therefore, these datasets had the most influence on the relevant networks performance. Several of the datasets were ranked highest in relation to multiple POIs. However, POIs with the same highest-ranked dataset did not necessarily have similar perfor-

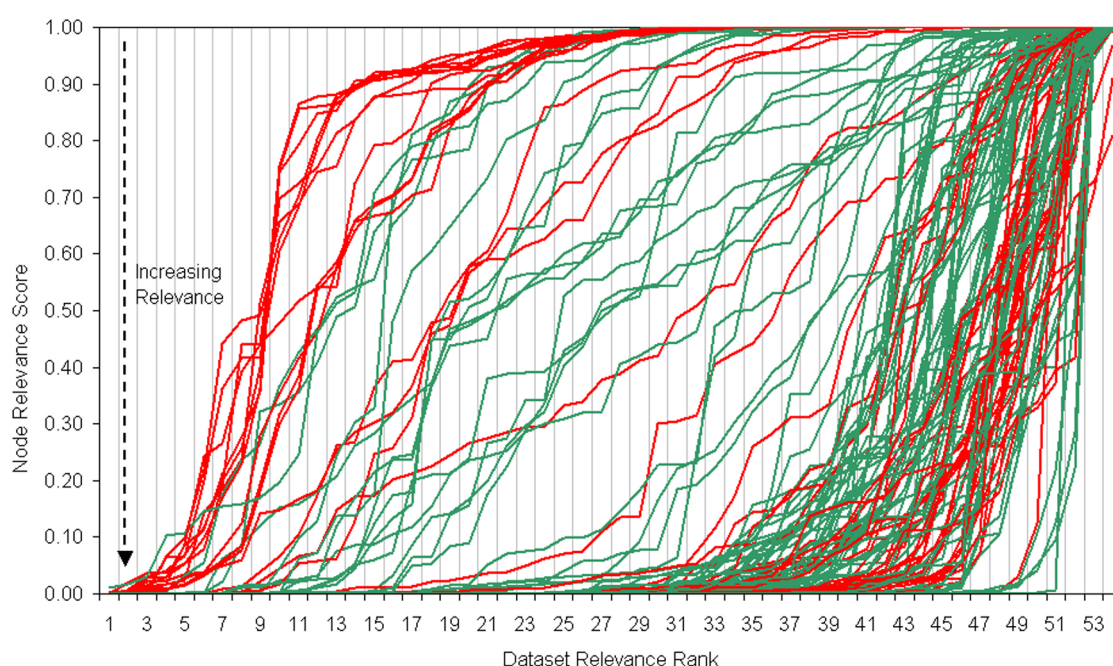


Figure 6.10: The relevance scores and ranks.

The relevance scores plotted against the datasets' relevance ranks. Each line represents the ranked order of datasets in relation to a single GO term. Those terms that had increased AUC between relevance and control network are shown in green and those decreased shown in red. The further to the right the vertical portion of a line is, the greater the number of high-relevance datasets for the GO term.

Table 6.1: The highest relevance ranked datasets for the 144 terms.

Several datasets had the highest relevance score for multiple POIs. The highest-ranked dataset has the greatest contribution to the network edges weights. However, the performance of the networks for POIs sharing the same highest-ranked dataset varied, with some AUCs increased and some decreased.

Dataset	Increased AUC	Decreased AUC
Affinity Capture-MS	3	0
Affinity Capture-Western	24	17
Biochemical Activity	3	0
Collins.17200106.Affinity Capture-MS	6	1
Collins.17314980.Phenotypic Enhancement	9	7
Dosage Rescue	2	0
Drees.11489916.Two-hybrid	3	0
Fiedler.19269370.Phenotypic Enhancement	4	2
Gavin.16429126.Affinity Capture-MS	4	4
Ho.11805837.Affinity Capture-MS	0	5
Krogan.16554755.Affinity Capture-MS	1	2
Miller.16093310.Two-hybrid	4	0
Pan.16487579.Synthetic Growth Defect	0	1
PCA	0	2
Phenotypic Enhancement	5	2
Phenotypic Suppression	0	1
Ptacek.16319894.Biochemical Activity	0	2
Schuldiner.16269340.Phenotypic Enhancement	2	1
Synthetic Lethality	1	0
Synthetic Rescue	6	6
Two-hybrid	1	0
Wilmes.19061648.Phenotypic Enhancement	11	2

mance. In fact, in many cases two POIs for which the same dataset was ranked highest in relevance performed oppositely, with one AUC increased in comparison to the control network, and one AUC reduced (Table 6.1). The proportion of AUC increases for each highest relevance datasets did not correlate with the LLS of the datasets (Figure 6.11). In one case, where the highest relevance dataset was also the highest confidence dataset, there was no improvement between the relevance and control networks (top left). Further, there was no correlation between the LLS of the two highest relevance datasets and the AUC change (Figure 6.12).

6.3.4 Dataset Topology

Although the highest relevance ranked dataset may have high confidence and high relevance, its topology may also affect network performance. For instance, a dataset may contain several nodes of interest that are not directly connected to one another in the dataset, and therefore will not be correctly predicted by the local GBA functional prediction algorithm applied, irrespective of the edge weights. The functional prediction of the POIs was therefore repeated using only the interactions of the highest ranked dataset. Since all edges in a single dataset are equal, edge weights were not

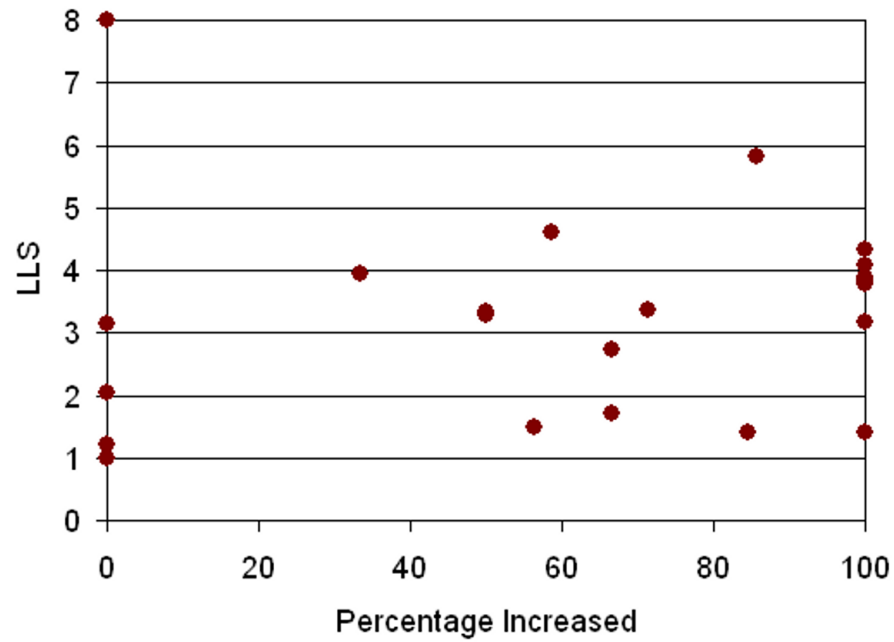


Figure 6.11: The highest ranked datasets.

The LLS score of the highest relevance ranked datasets in Table 6.1 plotted against the percentage POIs increased for the relevance networks in comparison to the control. Each point represents a single dataset.

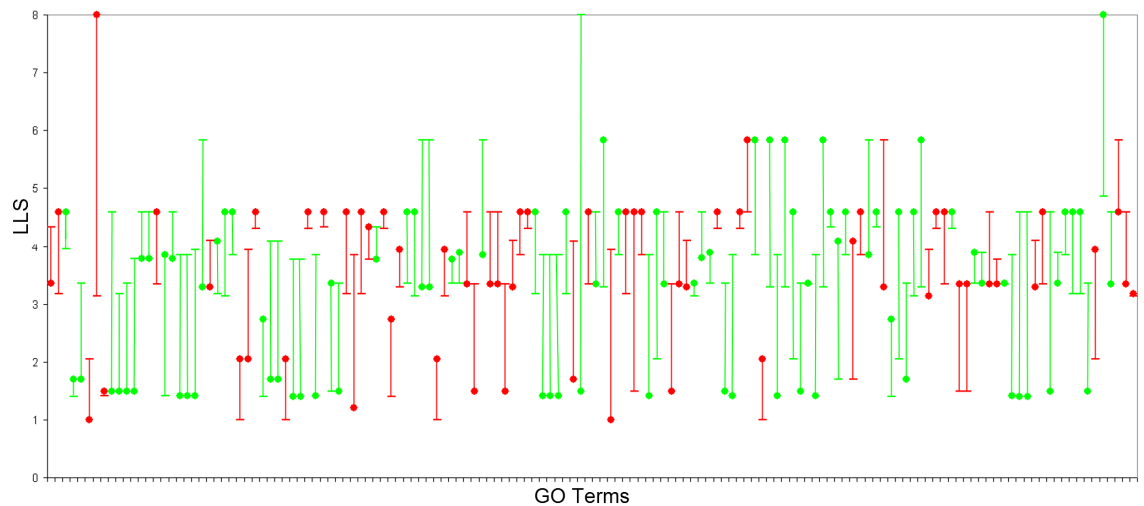


Figure 6.12: The two highest ranked datasets.

The relationship between the LLS scores of the two highest relevance ranked datasets of the 144 terms and the change in AUC. Each line represents a single POI with the highest ranked dataset marked as a circle and the length of the line the difference between the two LLS scores. In some cases the highest ranked dataset has a higher LLS score than the second ranked, while in others the second ranked dataset has the higher LLS score. Terms with increased performance in the relevance network in comparison to the control are marked green and those with decreased performance are marked red.

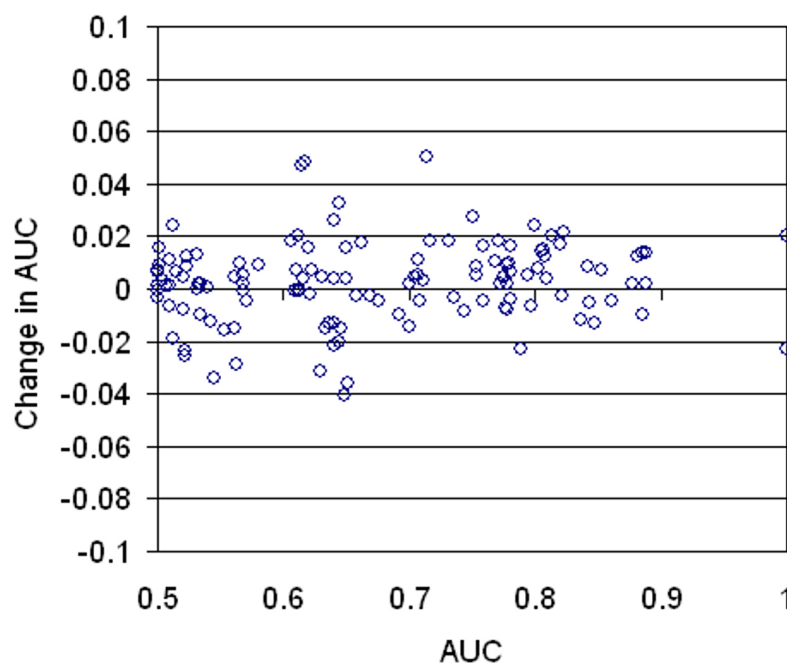


Figure 6.13: The highest ranked datasets AUC.

The AUC for functional prediction of the 144 POIs using the highest relevance dataset only plotted against the change in AUC for the POI between the relevance and control networks. Each point represents a single POI.

required. Figure 6.13 shows the AUC of the highest weighted dataset plotted against the change in AUC between the relevance network and control network for each of the 144 terms in Figure 6.5. There was no correlation between highest dataset's AUC and the change. In fact, even when the highest-ranked dataset produced perfect classification ($AUC = 1.0$), the change in AUC between relevance and control networks varied, with some terms improved and some decreased.

The datasets were then individually visualised in Cytoscape to assess the topology of the nodes annotated to the POI and those annotated to other processes. In many cases the datasets were not fully connected and contained few interactions between POI-annotated genes. Further, in several cases, datasets which visually appeared to be highly-relevant to a specific POI were not scoring highly in the relevance ranking. For example, the Sanders.12052880.Affinity Capture-MS [1144] has a high proportion (45%) of nodes annotated to the POI transcription from RNA polymerase II promoter (GO:0006366) and contains several interactions between these nodes, and a large number of interactions between these nodes and nodes annotated to other processes (Figure 6.14). However, since the relevance score takes the ratio of annotated and un-annotated nodes into account, this dataset does not score highly for relevance to this process. Consequently, datasets with the same number of nodes annotated to the POI may have vastly different numbers of interactions between these nodes and other nodes in the network (Figure 6.15).

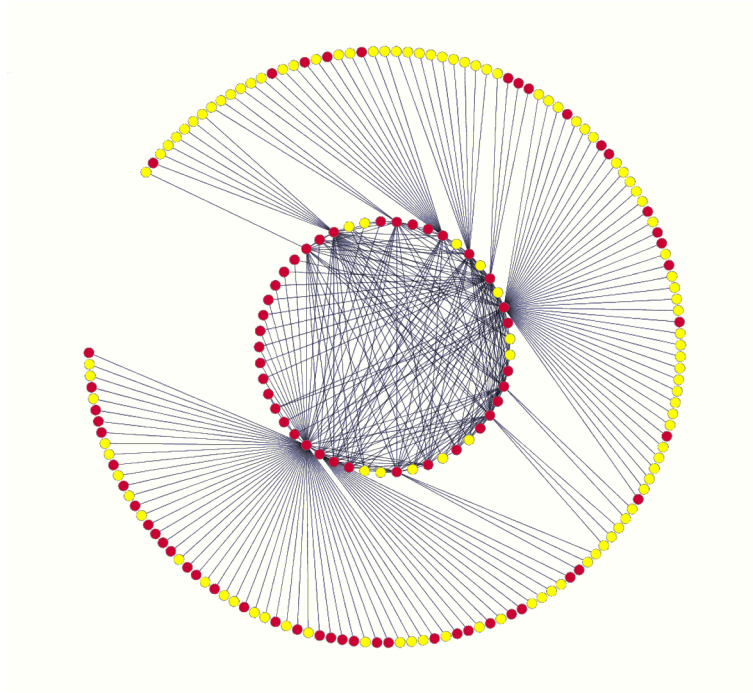


Figure 6.14: The Sanders.12052880.Affinity Capture-MS dataset.

Nodes annotated to the POI transcription from RNA polymerase II promoter (GO:0006366) are coloured red while those annotated to other processes are coloured yellow. The dataset contains 194 genes, 87 of which are annotated to the POI. A large number of interactions involve these genes. However, while the dataset visually appears highly relevant to this process, it does not score well due to the large number of genes which are not annotated to the GO:0006366.

6.4 Extending the Relevance Integration Schema

Given the results presented in Section 6.3 two further relevance scores were introduced to measure additional aspects of dataset relevance:

- Interaction Relevance: the over-representation of edges involving at least one node annotated to the [POI](#).
- Edge Relevance: the over-representation of edges between two nodes annotated to the [POI](#).

The original, node-based score will be referred to as Node Relevance for the remainder of this thesis. Both of the new scores were calculated using the hypergeometric test (see Section 3.1.4.3). The performances of networks integrated using the new scores were compared with the performance of the control network and Node Relevance network. The three aspects of network relevance were then combined to optimise functional prediction performance.

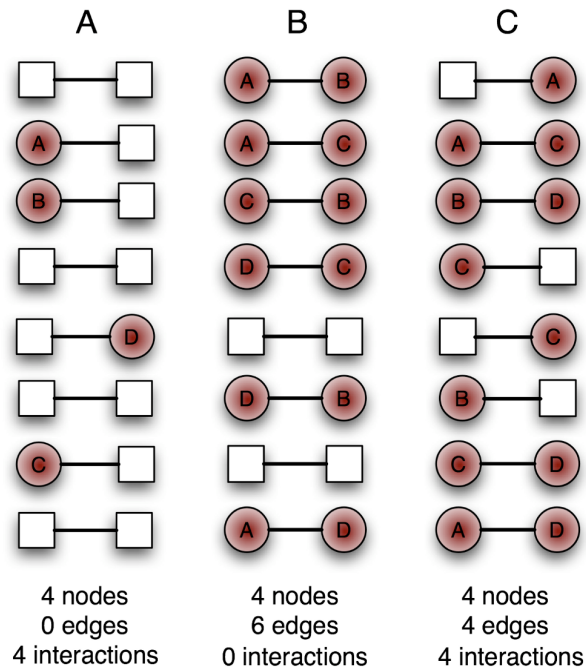


Figure 6.15: Dataset POI connectivity.

Three example datasets of the same size, each containing four nodes annotated to the POI (shown in red). Although the datasets each have the same number of genes annotated to the POI, the number of interactions involving these genes differs. Dataset A has four interactions involving the nodes but no edges connecting them. Dataset B has six edges connecting the nodes but no interactions between the POI nodes and other nodes. Dataset C has a mixture of interactions involving the nodes and edges between them.

6.4.1 Network Performance

Only 20% of the Interaction Relevance networks were improved in comparison with the control, compared to 71% for the Edge Relevance and the 60% for the Node Relevance. Table 6.2 summarises the improvements for individual POIs. Interestingly, no terms were improved using only the Interaction Relevance network alone, while several were only improved by either the Node Relevance or Edge Relevance networks, respectively. The largest group of terms (33) were improved by both the Node and Edge Relevance networks. Only 13 terms were not improved by any of the networks.

The relationships between the three relevance scores were highly variable. Figure 6.16 displays 3D plots of the scores for five terms chosen as examples to illustrate this variability. In the plots the three relevance scores are plotted on a single axis and each point represents a single dataset. Datasets scoring in the lower left corner scored highly in all relevance aspects (red circle), while those in the top right (blue circle) scored poorly in all three aspects of relevance (Figure 6.16 A).

The relationship between the dataset scores for the groups of terms in Table 6.2 were highly variable.

Table 6.2: Relevance score improvements.

The number and percentage of the 144 terms that were improved by the different relevance networks. A total of 20 terms was improved by all three relevance networks, while 13 terms were improved by none of the networks. The majority of the terms (109 of 144) were improved using the Edge Relevance followed by 82 of 144 by the Node Relevance score and just 42 of 144 by the Interaction Relevance network.

Relevance Improvement	Number Terms	Percentage Improved
Node	16	11.11
Edge	33	23.23
Interaction	0	0.00
Node + Edge	40	27.27
Node + Interaction	6	4.04
Edge + Interaction	16	11.11
All three	20	14.14
None	13	9.09

In some cases, terms that did not have improved performance for the relevance networks had several high-relevance datasets. For instance, reproduction (GO:0000003) had a group of high-scoring datasets in the lower left corner of the plot and several other datasets with high Node Relevance and high Edge Relevance (Figure 6.16 B). The term nucleocytoplasmic transport (GO:0006913) also did not have improved performance for any relevance network but also had a large proportion of datasets with high Node Relevance (Figure 6.16 C). Conversely, some terms which did not have any improvement for the relevance networks, such as term autophagy (GO:0006914), had few high-scoring datasets (Figure 6.16 D).

The relationship between the relevance scores for POIs which had improved performance for all three relevance networks were also highly variable. For example the network performance for two terms, modification-dependent protein catabolic process (GO:0019941) and regulation of macromolecule biosynthetic process (GO:0010556), was improved in all three relevance networks. GO:0010556 had an extremely high proportion of high relevance datasets, with the majority of the datasets scoring in the lower left corner of the plot (Figure 6.16 E). Inversely, while GO:0019941 had a high proportion of high Node Relevance datasets, the Interaction and Edge Relevance scores were highly variable with many low scores (Figure 6.16 F).

6.4.2 Combining Relevance

The three aspects of relevance, and the high-confidence control, can be combined to give a single AUC. Combination can occur at two different stages of integration and evaluation. In the first, the individual network scores can be combined into a single network prior to functional prediction. Alternatively, the four separate networks may be used for functional prediction and the results com-

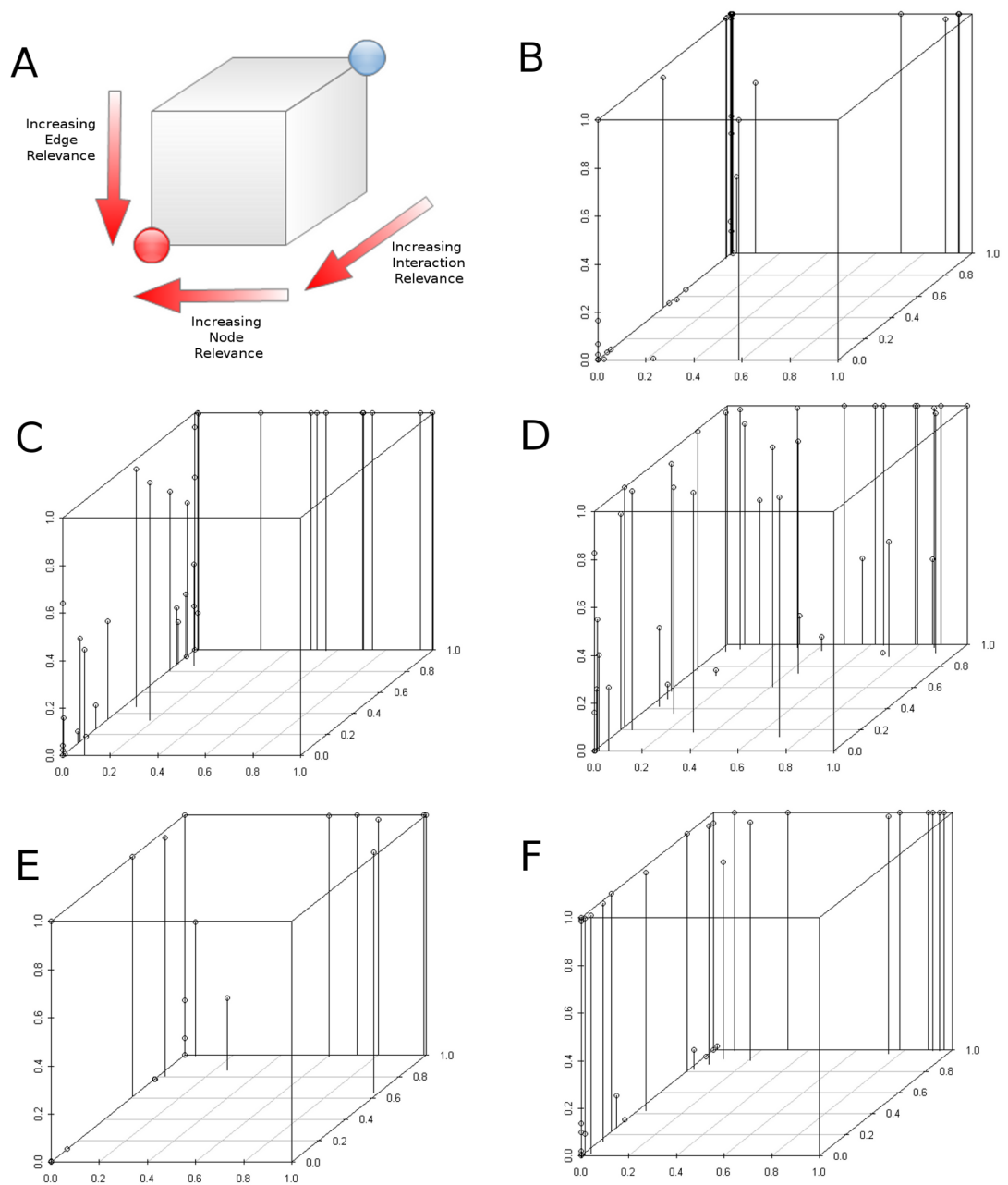


Figure 6.16: The three relevance scores.

3D plots of the three relevance scores for five POIs, GO:0000003, GO:0006913, GO:0006914, GO:0010556 and GO:0019941 (B-F respectively). Datasets scoring in the lower left corner scored highly in all relevance aspects (red circle), while those in the top right (blue circle) scored poorly in all three aspects (A). Datasets B, C and D were not improved using any relevance scored integration, while E and F were improved by all three relevance scores.

bined prior to calculation of the [AUC](#). Since the functional prediction algorithm is computationally intensive the first method of combining the scores has the advantage of reducing computational time, while the second method has the advantage of preserving the distinct areas of dataset relevance during functional prediction.

As the networks were topologically identical there were multiple methods by which their relevance scores could be combined. Here the scores for each edge were combined into a single composite network using two simple approaches:

1. A weighted sum of the four scores [[1141](#)]:

$$WS(i) = 1 - ((1 - WS(i)_{control})(1 - WS(i)_{node})(1 - WS(i)_{edge})(1 - WS(i)_{interaction}))$$

where, $WS(i)_{control}$, $WS(i)_{node}$, $WS(i)_{edge}$ and $WS(i)_{interaction}$ were the edge weights for interaction i in the control, node relevance, edge relevance and interaction relevance networks, respectively. Therefore, while some of the individual edge weights were small, integration of the weights could produce a high final weight.

2. An average of the edge weights over the four networks :

$$WS(i) = \frac{WS(i)_{control} + WS(i)_{node} + WS(i)_{edge} + WS(i)_{interaction}}{4}$$

where, $WS(i)_{control}$, $WS(i)_{node}$, $WS(i)_{edge}$ and $WS(i)_{interaction}$ were the edge weights for interaction i in the control, node relevance, edge relevance and interaction relevance networks, respectively. Therefore, the final score was the average of the four network scores.

Composite networks were integrated for the 144 terms of Figure [6.5](#) using the two methods of score combination. In both cases the majority of the networks had improved performance in functional prediction over the control network. Using the weighted sum 72.2% of the [AUCs](#) were improved (Figure [6.17](#) A), while using the average score 73.6 % of the [AUCs](#) were improved (Figure [6.17](#) B). However, in the majority of cases the improvements were smaller than those produced by the individual relevance networks.

The functional prediction algorithm produces a score for each gene in relation to a [POI](#). Since the algorithm is used to predict a single [POI](#) at a time, any genes that are not predicted to be involved in this term are weighted 0.0. Therefore, while one network will score a particular gene 0.0, another may assign a higher score. The results of functional prediction from the three relevance networks were combined with those of the control by selection of the highest weighted score of the four networks for each individual gene, therefore maximising functional prediction of the [POI](#) (Figure [6.18](#)).

In order to assess the contribution of each network to the final [AUC](#) all possible combinations of network results were computed for the 144 terms. In all cases the combined functional prediction results

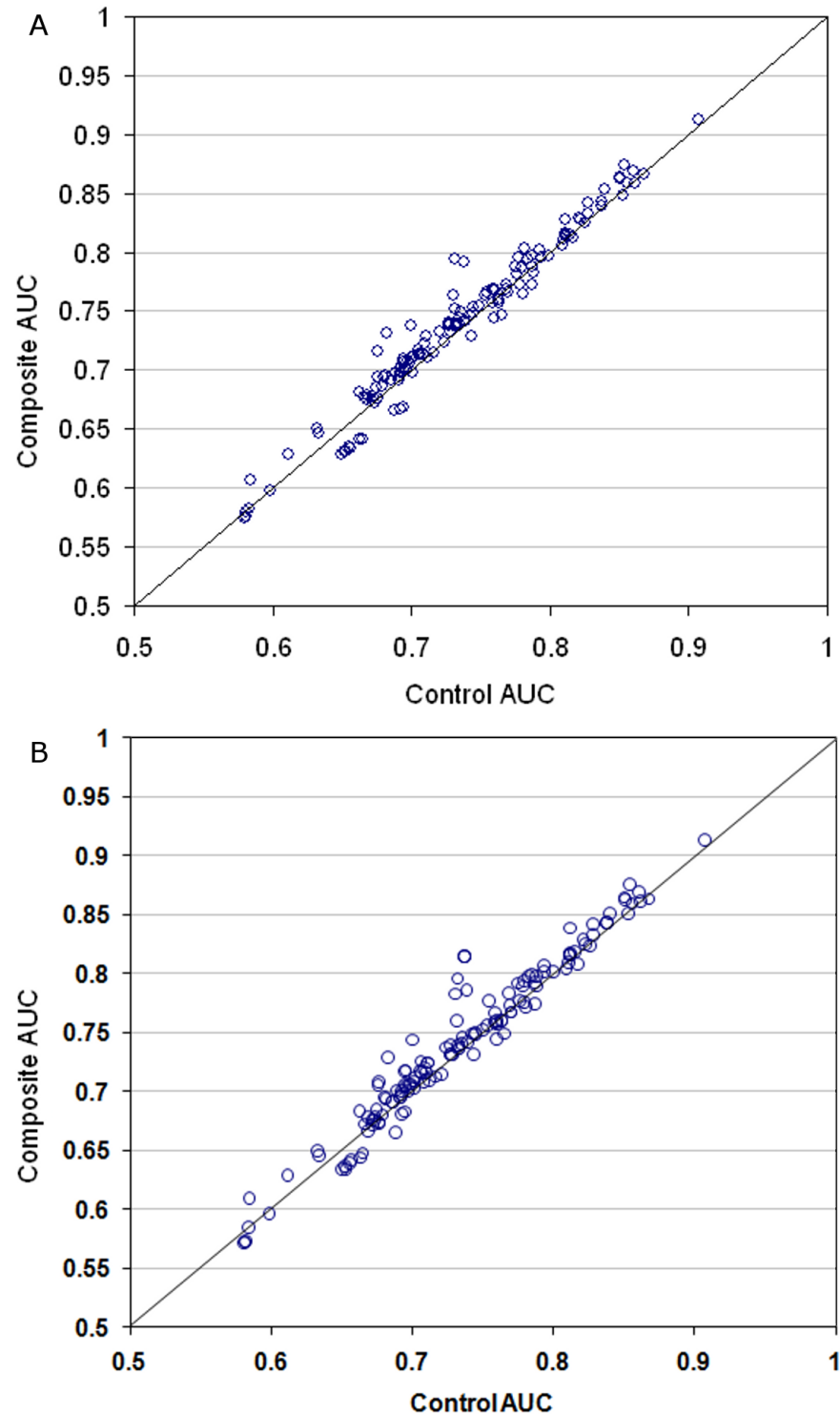


Figure 6.17: Composite network performance.

The AUC of the weighted sum composite networks in comparison with the AUC of the control network. Each point represents a single biological process. Those processes above the line were improved by the composite relevance integration and those below were decreased. In total 72.2% of the terms were improved.

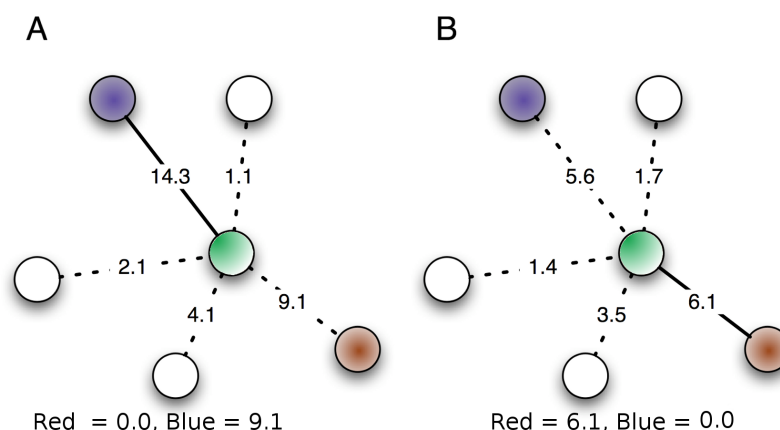


Figure 6.18: Combination of the functional prediction results for two networks.

A simple example of the effect of combining the functional prediction results for an unknown green node which should correctly be assigned to the annotations of both the red node and blue nodes. The first network (A) assigns 0.0 to the node for the red annotation since it is not on the highest edge. In network B the red annotation would be correctly assigned along the highest-weighted edge (solid line). Conversely, network A would assign the blue annotation correctly but network B would assign 0.0. When the predictions from the two networks are combined by the selection of the highest weighted prediction for the each annotation, both the red and blue annotations are correctly assigned. It should be noted that using real data the edge weights would not be identical for the two POIs, however, this effect would still be seen when combining functional prediction results of real data.

produced better performance than the control networks, the relevance networks and the composite networks. Those combinations that included the control and either the Edge or Node Relevance had high performance with 100% of the terms improved in comparison with the control network alone (Table 6.3). However, the combination of all four networks performed best, with 100% of the AUCs increased over those of the control networks and the majority (59%) of the highest AUCs for the terms (Figure 6.19).

Table 6.3: Combined prediction results.

The percentage of terms improved and the proportion of highest AUCs for each of the combined functional predictions.

Network	Percentage Terms Improved	Number Highest AUCs	Percentage Highest AUCs
Node + Interaction	88	3	2.1
Node + Edge	94	3	2.1
Interaction+Edge	97	1	0.7
Control+Node	100	0	0
Control+Interaction	99	0	0
Control+Edge	100	1	0.7
Node+Interaction+Edge	99	6	4.2
Node+Interaction+Control	100	17	11.8
Node+Edge+Control	100	19	13.2
Interaction+Edge+Control	100	9	6.3
All Four	100	85	59

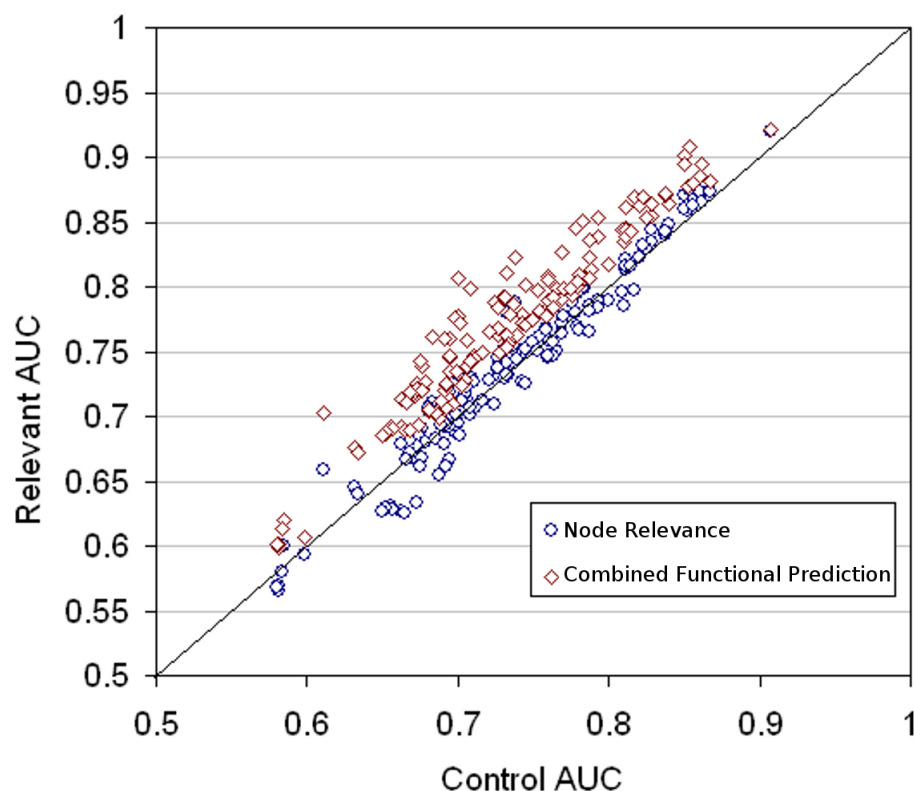


Figure 6.19: Combined functional prediction.

The AUC produced by the combined functional prediction results plotted against the control network AUC. Each point represents one of the 144 GO Biological Process term with the red points representing the combined functional prediction score and the blue points representing the original node relevance scores of Figure 6.2. Those above the diagonal had improved functional prediction and those below were decreased. The combined results were improved for 100% of the terms in, comparison to the 62% increase of the Node Relevance network alone.

6.5 Discussion

While the RelCID integration method improved protein function prediction performance in relation to ageing, the ageing process is only one area of cellular biology described by GO [100, 848]. Therefore, PFINs were produced for all GO terms using the RelCID method and compared with a control network, integrated without a measure of relevance [49]. Network performance was assessed using several novel process-relevant techniques developed during this project to fulfil Objective 4 (see Section 1.5). Performance of the networks varied across the terms with performance increasing for some terms and decreasing for others. Several factors may influence network performance in relation to specific POIs and each has been investigated in order to allow selection of appropriate POIs prior to integration.

The properties of the GO terms themselves seem to have little effect on network performance. In particular, GO term specificity [1140] has no significant correlation with functional prediction per-

formance, although low specificity terms tend to have a smaller range in performance variation. The small range is likely to reflect the number of annotations to these terms. In other words, when terms are annotated to the majority of the genes there is a lower likelihood of incorrect classification during GBA functional prediction [57] and therefore the AUC changes little.

The differences in term performance are reflected in the parent-child relationships of the GO DAG [100]. If a term's performance is decreased by the relevance integration method, its child terms' performances are also decreased. This effect is a reflection of the structure of the DAG and the transitivity of child-parent annotations. Since an annotation to a child term automatically annotates a gene to its parent terms, there are areas of the DAG where groups of connected terms are annotated to the same or very similar groups of genes. Therefore it follows that the relevance scores, and relevance networks, will be highly similar and perform in the same way. The relationships of the DAG could therefore potentially be used to select appropriate POIs prior to integration and to infer the performance of closely-related terms.

The connectivity of genes annotated to the POI within the network would be expected to influence network performance due to the local nature of GBA functional prediction [57]; if the genes are not connected performance will be zero irrespective of edge weights. However, while it would intuitively be expected that increasing connectivity would have a linear relationship with network performance, this was not the case. Where connectivity was very low the performance of the networks did not vary to a great extent. However, at relatively low levels of connectivity (>25%) the performance change became very variable, with some terms increasing and some decreasing. Similarly, the average shortest path between genes annotated to the POI did not correlate well with network performance. Therefore, these factors cannot be used to select appropriate POIs. However, while the connectivity of genes annotated to the POI did not correlate with local GBA functional prediction performance, it may potentially influence global functional prediction. Therefore, the adaptation of a global algorithm, such as Functional Flow [902], for use with the relevance edge weights could potentially improve performance in unannotated areas of the genome, in particular where the POI-annotated genes have a small average shortest path [1142].

The relationship between confidence and relevance scores would also be expected to influence network performance and be potential factors in the selection of appropriate POIs. For example, POIs with few high-relevance datasets or those where high-relevance datasets score poorly for confidence were expected to perform poorly. Interestingly, this was not the case. While most POIs with few high-relevance datasets performed poorly, a significant number performed well. Conversely, many POIs with a large number of high-relevance datasets performed poorly. In addition the confidence scores of the high-relevance datasets appear to have little effect on relevance network performance.

While the GO term chosen, the network topology and the dataset scores were expected to influence networks' performance, it appears that the topology of the individual datasets in relation to the POI is the most important factor in functional prediction. A dataset may contain several nodes annotated to the POI but may have few interactions involving these nodes, and conversely, a dataset may contain only one node annotated to the POI but a large number of interactions involving it. Further, a dataset that appears visually to be highly relevant to a POI may not score well using the original Node Relevance score. Therefore, the connectivity of nodes annotated to the POI must be taken into account to accurately assess dataset relevance.

The original relevance score, termed Node Relevance, does not capture every aspect of dataset relevance since it is simply based on node counting and ignores the number of edges in which nodes are involved. Consequently, two further relevance scores were developed; Edge Relevance, based on edges between nodes annotated to the POI; and Interaction Relevance, based on interactions between nodes annotated to the POI and other nodes. The performance of the relevance networks varied. The Edge Relevance performed best while the Interaction Relevance performed worst. This effect is directly due to the limitations of the evaluation metric applied. Local GBA functional prediction transfers annotations between directly connected nodes [57]. Therefore a dataset with a high level of Edge Relevance-nodes annotated to the POI which are directly connected-will perform well. However, given that a small percentage of protein functions are accurately known, many new predictions are treated as false positives when they may in fact be biologically correct. Given this aspect of the evaluation, datasets with high Interaction Relevance, such as the Sanders.12052880.Affinity Capture-MS dataset [1144], are adversely affected in their performance despite containing relevant data. As the number of known protein functions increases this effect will be reduced and network performance should improve.

Since the three relevance networks capture different aspects of the dataset relevance, integration of these aspects potentially optimises the relevance effect. The results were, therefore, integrated in two ways. First, the four networks were combined into a composite network prior to functional prediction. Second, the functional prediction results of the four networks were integrated into a single combined functional prediction result.

The composite networks were produced in two ways; by the average of the edge weights, and by a weighted sum of the edge weights. Both the composite networks performed well, with the majority of POIs improved. Combination of the networks in this way is computationally less intensive than combining the functional prediction results, since only a single functional prediction calculation is required. However, many of the true positives predictions made by the individual networks are lost, and improvement is not 100%. Conversely, combination of the individual functional prediction

results into a combined functional prediction preserves the distinctions between the networks and produces optimised performance for all GO terms. Therefore, while computationally intensive the combined functional prediction method produces the optimal results and was used in the remainder of this project.

As with the single relevance networks, Interaction Relevance has the poorest performance in the combined networks. However, [PFINs](#) are ultimately used to produce new hypotheses and guide future experiments. Therefore, while the Edge Relevance networks perform well for known annotations, due to up-weighting of the datasets with high connectivity between these known genes, the Interaction Relevance networks are equally valid for inclusion since they capture the interactions between known genes and genes potentially involved in the [POI](#).

While the inclusion of the control network in the combined network may initially appear counter-productive to the relevance integration technique, this is not the case. Some datasets may score poorly in in all three relevance aspects but contain a small number of high-confidence interactions involving the [POI](#) that will not be captured by any of the relevance networks. Even a single interaction involving a node annotated to the [POI](#) is valuable information. Inclusion of the high-confidence control in the combined functional prediction allows these interactions to be correctly classified during functional prediction.

The combination of the three relevance networks with the control captures all aspects of dataset relevance and dataset confidence and produces the highest functional prediction performance when evaluated using known annotations, successfully completing Objective 3 of this project (see Section [1.5](#)). However, since [PFINs](#) are intended to produce new hypotheses (see Section [2.5.5](#)), their performance must be evaluated based on new predictions (Objectives 5 and 6). In the next chapter the combined functional prediction technique is used to produce novel predictions for over 500 [GOBPs](#). The predictions are then evaluated using computational analysis before a single prediction is chosen for experimental validation.

Chapter 7

Computational and Laboratory Analysis of Network-Generated Predictions

In the previous three chapters of this thesis, the development and evaluation of a novel network integration algorithm, RelCID, which incorporates dataset relevance to a specific process during network integration, has been described. The algorithm has been evaluated using a number of network analysis techniques and has been shown to have improved performance over a control network, integrated without a measure of relevance, for over 500 [GOBP](#) terms.

[PFIN](#)s are ultimately intended to generate novel hypotheses based on functional interaction data. In this chapter, the RelCID integration algorithm is applied to *S. cerevisiae* functional data in order to generate new functional predictions (Objective 5, Section [1.5](#)). The new predictions are first evaluated by comparison with known curated GO annotations, and with GO annotations generated by other computational methods. Then, a single prediction to the ageing-related GO term response to oxidative stress (GO:0006979) is chosen and experimentally evaluated (Objective 6).

7.1 Functional Prediction

7.1.1 Datasets

The BioGRID version 65 (June 2010) data for *S. cerevisiae* was used as the source of functional interaction data. Datasets were split at a cut-off of 100 interactions into [HTP](#) and [LTP](#) datasets, following the protocol outlined in Section [3.1.1](#). The equivalent June 2010 KEGG PATHWAYS data was used to generate a Gold Standard, and confidence scoring was carried out as in Section [3.1.4.1](#). GO and SGD annotation data from June 2010 was used for dataset relevance scoring and functional

prediction (see Sections 3.1.4.3 and 3.1.5.3). Annotations with the evidence codes IEA and RCA were excluded during the network integration and functional prediction stages and used in the initial evaluation (see Section 7.3.2).

7.1.2 GO Term Selection

The information content specificity score was calculated for all available GOBP terms (see Section 3.1.2). A total of 505 terms was chosen with a specificity score between -2.0 and -5.0, in order to optimise the number of known annotations, while discarding those general terms at the top of the GO DAG.

Networks were generated and functional predictions produced using the optimised RelCID integration schema described in Section 6.4.2 (Figure 7.1). A D value of 1.1 was chosen as before (Section 3.1.4.4). A total of 1516 networks was produced; 505 Node Relevance, 505 Edge Relevance and 505 Interaction Relevance, and, a control network integrated without a measure of relevance. Functional prediction was carried out for each of the terms using the maximum weight rule and the most highly weighted prediction was selected for each gene (see Section 6.4.2).

7.2 Prediction Results

In total 319766 predictions were made for the 505 terms, covering 5423 *S. cerevisiae* genes (Objective 5, Section 1.5). The highest scoring prediction (26.11711) was produced by the control network to the term `gene-specific transcription from RNA polymerase II promoter` (GO:0032569). The lowest scoring predictions (both 0.00118) were produced by the Edge and Interaction Relevance networks to the terms `regulation of primary metabolic process` (GO:0080090) and `regulation of macromolecule metabolic process` (GO:0060255). These two low-scoring terms are both child terms of the general process `regulation of metabolic process` (GO:0019222). The number of predictions per gene ranged from 26 to 2429 with an average number of predictions of 633.2 (Figure 7.2 A). The gene with the largest number of predictions was YJR120W, a protein of unknown function which had 128 interactions in the network. The number of predictions per term ranged from 1 to 418 with an average of 59 (Figure 7.2 B). The annotation predicted the most often was to the general term `biological regulation` (GO:0065007).

Approximately half of the predictions (161130) were based on control network scores (Figure 7.3). Of the relevance network predictions, the Edge Relevance networks produced the fewest predictions (24310), while the Interaction Relevance networks produced the most predictions (93192). In order

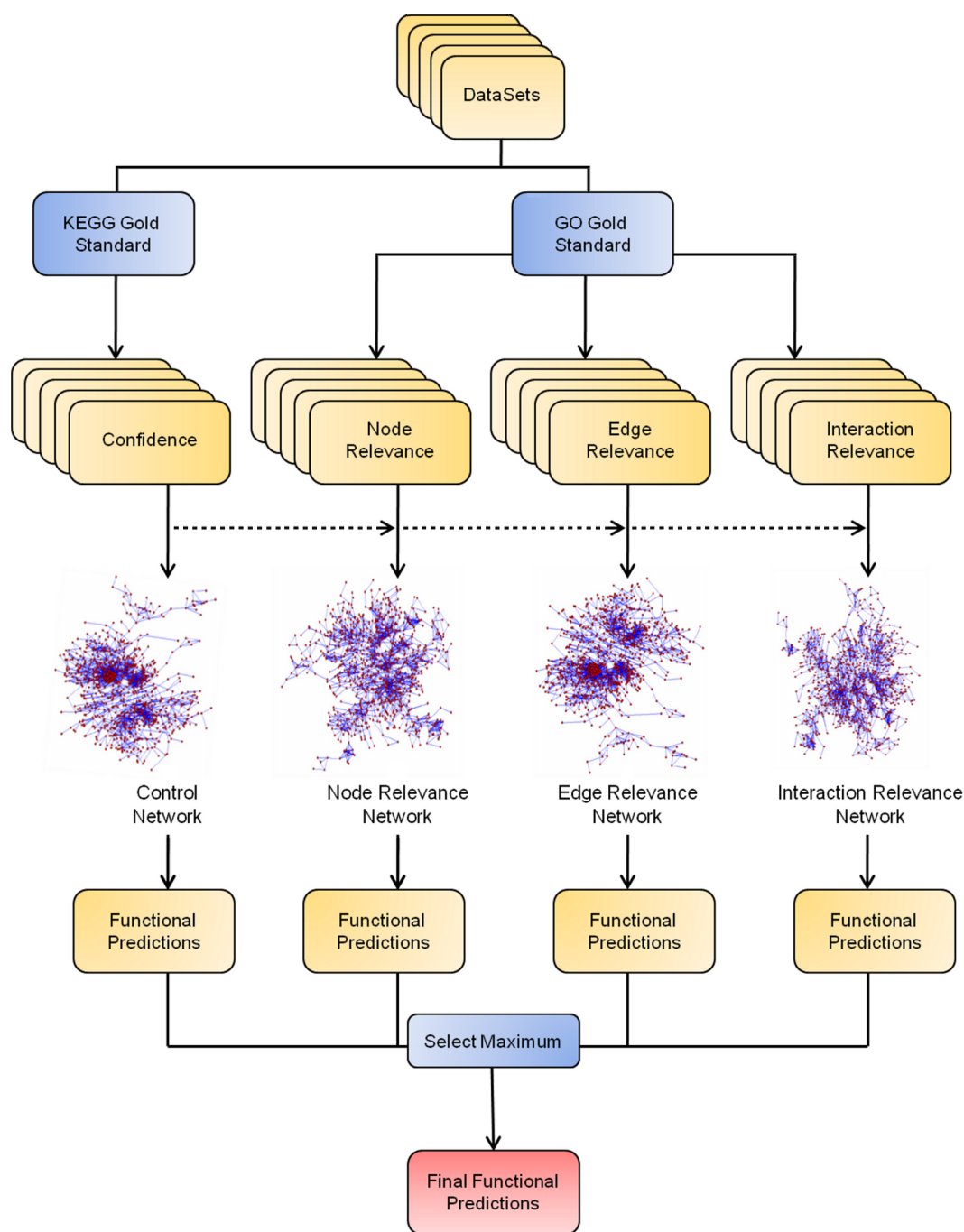


Figure 7.1: The extended RelCID functional prediction schema.

The functional datasets are confidence scored using a KEGG Gold Standard and relevance scored using Gene Ontology data. Four networks are integrated for each GO term using the confidence- and relevance-ranked datasets. In all four cases, the confidence scores are integrated into the edge weights in ranked order. Functional prediction is then carried out using the maximum weight decision rule. The maximum scoring predictions for each gene are chosen from the four prediction sets to produce the final functional predictions.

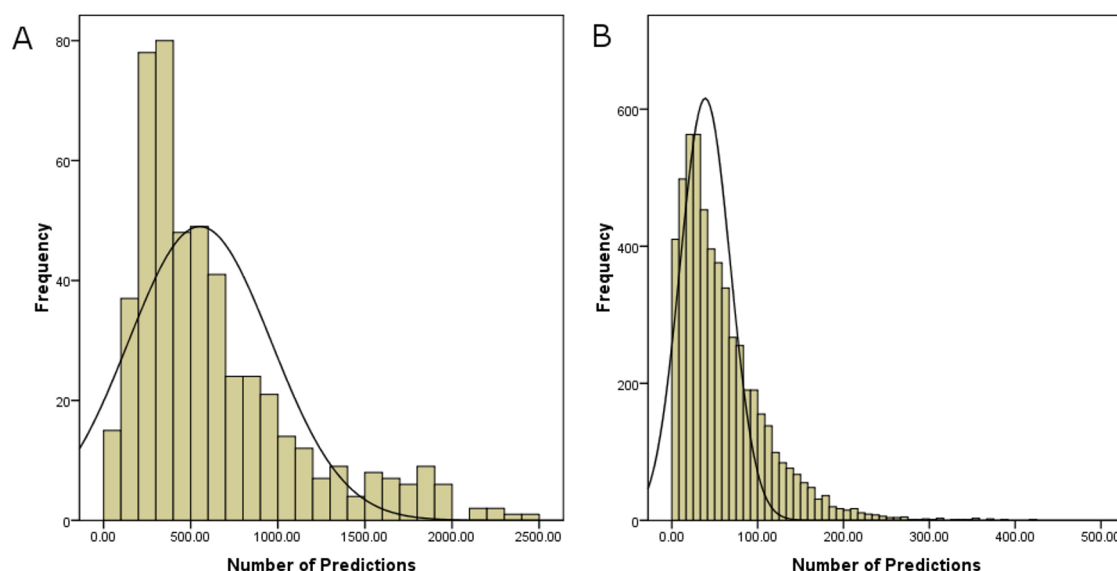


Figure 7.2: Prediction distributions.

A. The number of predictions per gene. **B.** The number of predictions per GOBP term.

to further evaluate the relevance integration algorithms, the predictions produced by the relevance networks were selected for further analysis.

7.3 Computational Evaluation

In total, 158636 functional predictions were produced. The majority of the predictions, 120421, (75.9%), scored below 2.0, with just under half the predictions, 73017 (46.0%), scoring below 1.0 (Figure 7.4). The predictions were evaluated by comparison with known GO annotation data to fulfil the first stage of the final objective of this project (see Section 1.5).

7.3.1 Consistency with Existing Annotations

Predictions were first evaluated for consistency with known curated annotations, based on the parent-child relationships of the GO DAG. A prediction was considered consistent with the known annotations if it was to a child term of an existing annotation to the same gene. In other words, the predicted function for a gene was a sub-process of a process in which the gene is known to be involved. In total 0.57% (906) of the relevance predictions were consistent with known annotations. For example, the gene YBR149W is known to be involved in cellular carbohydrate metabolic process (GO:0044262) and was predicted by the Node Relevance network to be involved in hexose catabolic process (GO:0019320), a child term of GO:0044262 (Figure 7.5)¹.

¹<http://www.ebi.ac.uk/QuickGO/GTerm?id=GO:0019320#term=info>

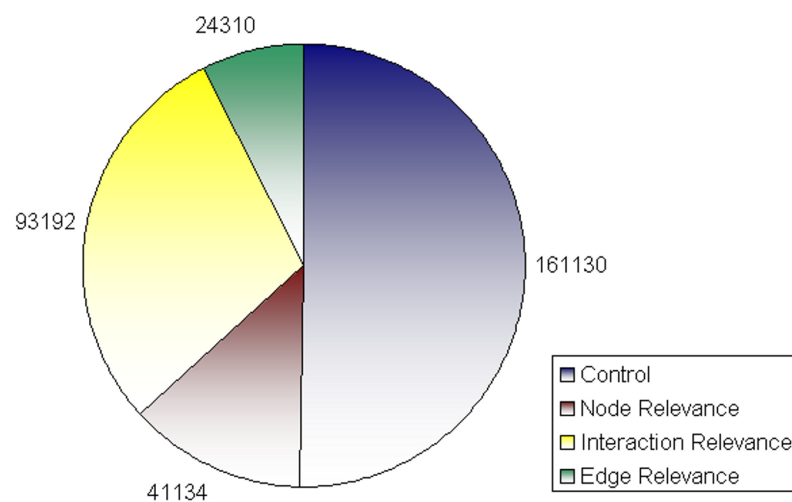


Figure 7.3: Source of the functional predictions.

The maximum scoring prediction was chosen for each gene from the functional predictions produced by the four networks. The majority of the maximum scoring predictions were produced by the control network. Of the three relevance networks the Interaction Relevance networks produced the most predictions and the Edge Relevance the fewest.

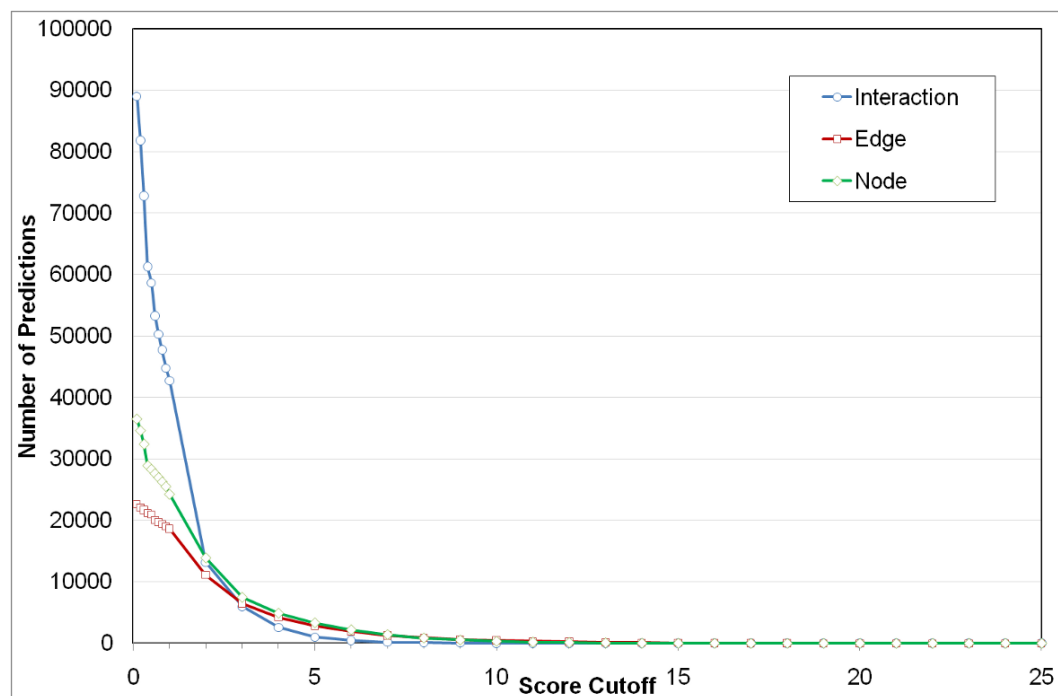


Figure 7.4: Score range of the functional predictions.

The majority of the predictions (75.9%) scored below 2.0, with just below half (46.0%) scoring below 1.0.

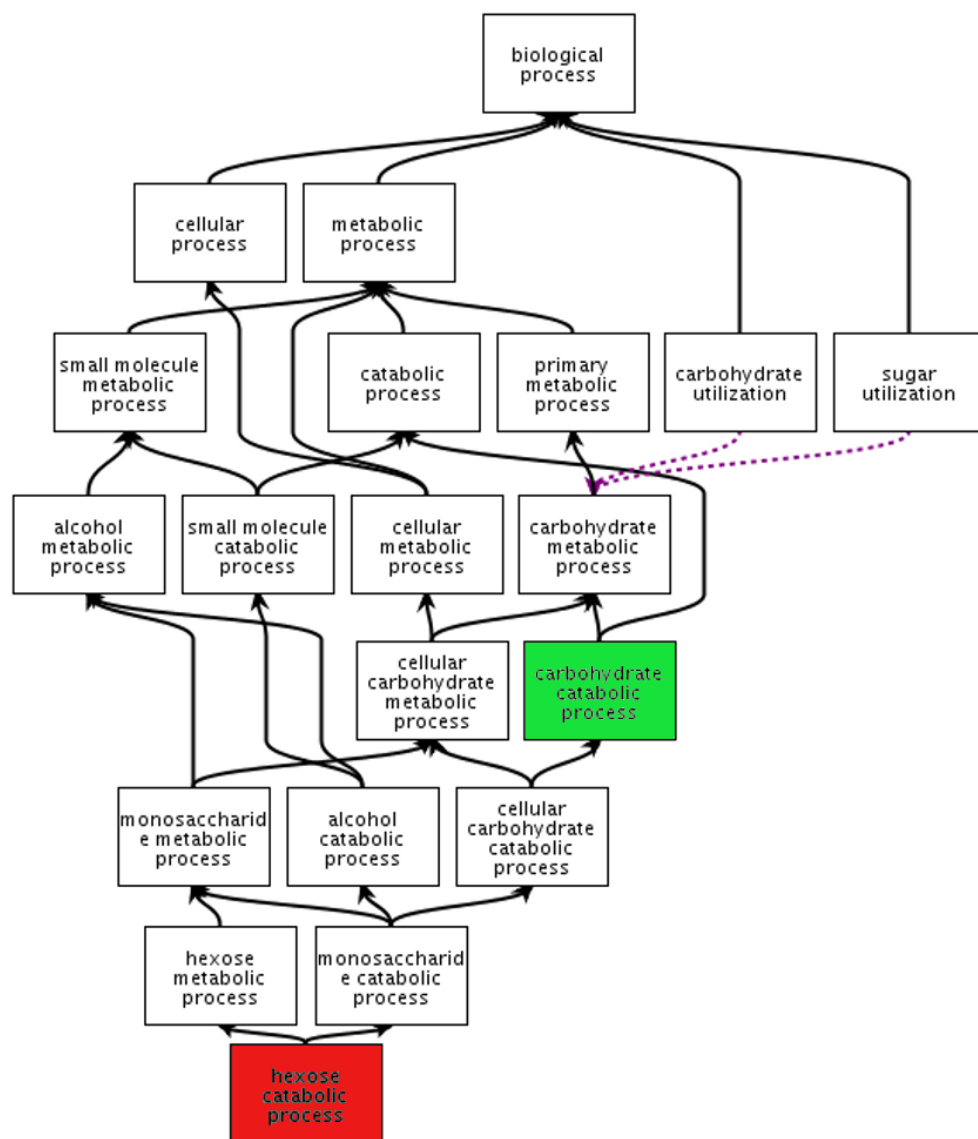


Figure 7.5: Example of a child term prediction.

A section of the GO DAG with *is_a* relationships depicted as black arrows and *part_of* relationships as dashed red arrows. The gene YBR149W is annotated to cellular carbohydrate metabolic process (green). YBR149W is predicted by the Node Relevance network to be involved in hexose catabolic process (red), which is a child term of cellular carbohydrate metabolic process. Therefore, the prediction to hexose catabolic process is consistent with the existing annotation data for this gene.

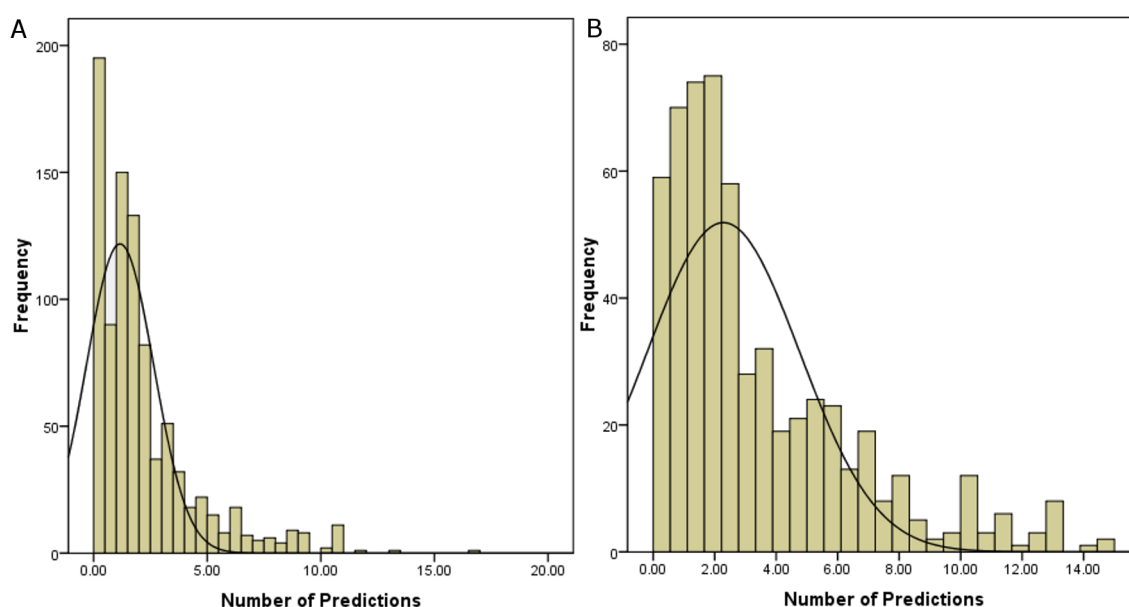


Figure 7.6: Prediction score frequencies.

A. The distribution of the scores for predictions that were consistent with known curated annotations.

B. The distribution of the scores for predictions that were matches to IEA and RCA annotations.

While 40.0% of the new predictions scored below 1.0, the majority of the consistent predictions, 621 (68.5%), scored over 1.0 (Figure 7.6 A). The highest scoring of the consistent annotations was produced by the Edge Relevance network for the gene YIL021W to the process transcription initiation from RNA polymerase II promoter (GO:0006367). YIL021W codes for the protein RNA Polymerase B, which is well known to be involved in transcription [1164]. Consistent predictions of this type are far more likely to be correct than those to terms unrelated to a known annotation.

7.3.2 Consistency with Previous Computational Predictions

Predictions were also evaluated by comparison with those annotations with the evidence codes IEA and RCA (see Section 2.5.4.3). These annotations were excluded during relevance scoring and functional prediction and, therefore, did not have any influence on the network edge weights or prediction results. While annotations with the evidence codes RCA and IEA are considered to be less reliable than other annotations, predictions that are consistent with these annotations may be considered more likely to be correct than those which are inconsistent with known annotations. In total, 0.37% (581) of the relevance predictions were exact matches to these annotations. Again, the majority of the matching predictions, 467 (80.3%), scored over 1.0 (Figure 7.6 B).

The highest-scoring matching prediction was for the gene YGR274C to the term chromatin modification (GO:0016568). YGR274C has an IEA annotation to GO:0016568 based on Swiss-Prot keyword mapping [1165]. The gene codes for a TATA binding protein-associated factor that is known to bind chromatin and is, therefore, annotated with the Gene Ontology molecular_function (GOMF) term chromatin binding (GO:0003682). While the prediction to GO:0016568 has no consistency with YGR274C's curated GOBP annotations, it can be considered consistent with its GOMF annotations.

The evaluation was then repeated to consider consistency with child terms of the RCA and IEA, as above. When the child terms of these annotations were included, 26.23% (41610) of the predicted annotations were consistent with the computational annotations.

7.3.3 Multiple Functional Predictions

Of the 5423 genes in the network, 4515 had multiple functional predictions. In particular 24 genes had multiple predictions scoring above 14.0 (Figure 7.7). Analysis of the prediction scores and dataset ranks indicated that these multiple predictions were to associated terms, and, in many cases the relevance ranks for the datasets were unchanged for the predicted groups of terms. In other words, the same datasets were of high relevance for all the terms. For example, the gene YJL140W had three high scoring predictions, two of which scored the same:

1. GO:0022411 cellular component disassembly - 17.71
2. GO:0032984 macromolecular complex disassembly - 17.83
3. GO:0034623 cellular macromolecular complex disassembly - 17.83

These terms are directly linked in the GO DAG and the dataset rankings for all three terms were the same. Additionally, the dataset scores differed only slightly and were, in fact, identical for GO:0034623 and GO:0032984 (Table 7.1).

7.3.4 Discussion

Sections 7.2 and 7.3 address Objectives 5 and 6 of this project (see Section 1.5) using 505 GOBP terms as POIs. Ultimately, PFINs are intended to produce novel functional predictions prior to laboratory analysis. Therefore, in this section the relevance network integration and functional prediction schema was used to produce new predictions for a variety of POIs. GOBP terms were chosen as POIs based on their information content specificity measure. Therefore, the chosen terms were relatively

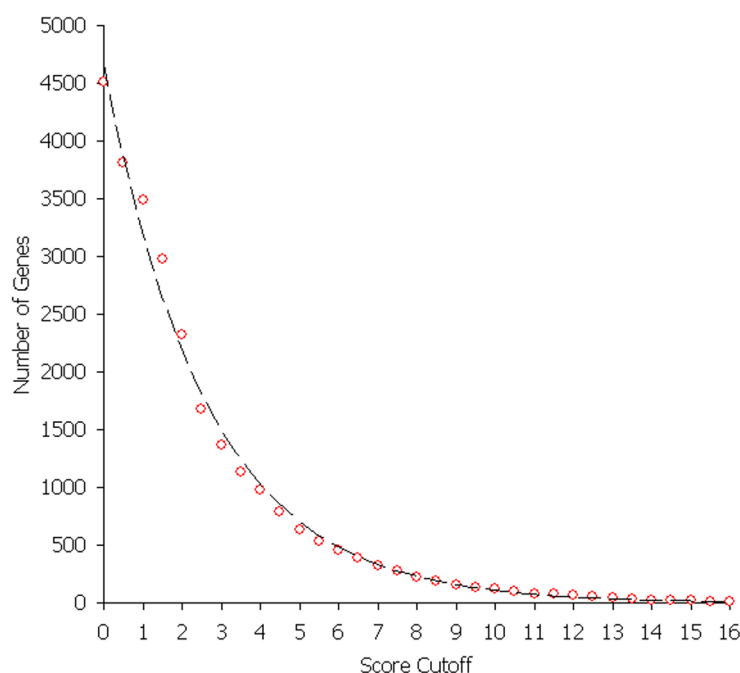


Figure 7.7: Multiple predictions.

The number of genes with multiple GOBP predictions above a specific score cut-off. The majority of the multiple predictions scored below 3.0, with only 24 genes having high-scoring (>14.0) multiple predictions.

Table 7.1: Dataset relevance ranks.

The Node Relevance ranks for the three GOBP terms predicted for the gene YJL140W. The ranks are identical for the three genes and the dataset scores for GO:0032984 and GO:0034623 are identical. The Edge and Interaction Relevance ranks and scores were also highly similar.

GO:0022411		GO:0032984 and GO:0034623	
Dataset	Score	Dataset	Score
Baetz.14729968	3.62E-26	Baetz.14729968	1.58E-26
Affinity_Capture-MS	1.19E-22	Affinity_Capture-MS	1.15E-22
Gavin.16429126	3.80E-20	Gavin.16429126	9.87E-22
Sanders.12052880	1.23E-17	Sanders.12052880	3.74E-18
Krogan.14759368	8.56E-17	Krogan.14759368	2.17E-17
Gavin.11805826	5.62E-15	Gavin.11805826	3.70E-16
Affinity_Capture-Western	9.99E-15	Affinity_Capture-Western	5.10E-15
Collins.17200106	1.58E-13	Collins.17200106	9.16E-15
Dosage_Rescue	3.11E-12	Dosage_Rescue	3.46E-12

specific, while having a sufficient number of annotations to form the basis for relevance scoring and functional prediction. While the relevance integration method would work for more specific terms, the low number of annotations to some of these terms would limit the scope of local [GBA](#) functional prediction.

Prediction was carried out using the optimised RelCID integration and functional prediction method, in which predictions from the three relevance networks and the control are combined by selection of the highest scoring prediction for each gene. The majority of these high-scoring predictions were

produced by the control network. This result is unsurprising since the control network edge weights represent the highest possible sum of the confidence scores (see Section 4.3.1).

While the control network produced more predictions, the relevance network predictions are of more interest since they are based on up-weighting of relevant data. Therefore, these predictions are only produced when a measure of dataset relevance is incorporated during network integration. Of the relevance networks, the Interaction Relevance produced the most predictions and the Edge Relevance the least. The difference in the number of predictions is due to the nature of these relevance scores. Interaction Relevance measures the number of interactions between genes annotated to the POI and other genes in the dataset. Consequently, many edges of this type are up-weighted in the network during integration. Given that the local GBA algorithm transfers annotations along edges from annotated to unannotated genes, the Interaction Relevance network has greater scope for annotation transfer in its up-weighted edges.

Conversely, the Edge Relevance score measures the number of interactions in the dataset involving two genes annotated to the POI. Many of these edges are up-weighted in the resulting network. While these edges are highly relevant to the POI, they offer less scope for annotation transfer. Consequently, while the Edge Relevance networks perform best during leave-one-out evaluation of known annotations, they do not produce as many new predictions as the other relevance scored networks. However, the novel predictions of the Edge Relevance network are as valid as those of the other networks.

Evaluation of new predictions is non-trivial. While known annotation data can be used to assess network performance by the production of ROC curves, as seen in Chapters 4-6, there is very little data with which to assess new functional predictions. Ultimately, the only true evaluation of new predictions is by small-scale experimental analysis. Clearly, the experimental analysis of 158636 functional predictions is beyond the scope of this project, both in time-scale and cost.

However the predictions may be evaluated, to some extent, by their consistency with known annotation data. This type of evaluation is possible due to the hierarchical nature of the GO DAG. In total, 0.57% of the predictions were consistent with known annotation data. That is, the predictions were to child terms of a known annotation for the same gene. While this percentage is low, the other predictions cannot be considered incorrect, since a lack of supporting data does not in itself refute a novel hypothesis. However, the consistent predictions may be considered more likely to be correct given current data.

Interestingly, while 46% of the predictions scored below 1.0, the majority of the consistent predictions, 68.5%, scored over 1.0. The higher scores of the consistent predictions may indicate that higher scoring predictions are of higher quality. However, the difference in scores may also be due

to the levels of evidence involving a particular **GOBP** term. Therefore, the terms with a high level of associated experimental data, such as those describing highly studied processes, will have higher scores and also be more likely to have consistent predictions.

A second computational evaluation was possible due the exclusion of **RCA** and **IEA** annotations during integration and prediction. **RCA** annotations are produced by integrated analyses similar to the method developed in this thesis. In fact, the functional prediction results produced in this chapter can be considered putative **RCA** GO annotations (see Appendix B). Logically, using **RCA** annotations to produce **RCA** annotations is counter-intuitive and, consequently, the **RCA** annotations were excluded.

Additionally, the lower-quality **IEA** annotations were excluded from integration and functional prediction, as in previous chapters. **IEA** are computational annotations produced by automated transfer of annotations from other databases, such as SwissProt², and are not curated. The **RCA** and **IEA** annotations, therefore, provide a further evaluation of the new functional predictions. In total, 0.37% of the predictions matched the **RCA** and **IEA** annotations and 26.23% were consistent with them. Again, while the numbers are relatively low, the other predictions cannot be considered incorrect.

While these evaluation methods are far from ideal and limited in scope, they remain the only useful methods of computational evaluation available, given the lack of data. Potentially, evaluation may be extended to include consistency with **GOMF** and **GOCC** annotations. However, this type of evaluation is difficult and time-consuming, requiring a high level of human curation to map the separate branches of GO to one another. Several projects are currently ongoing to provide these mappings (see Section 2.5.4.3). Once these projects are complete, evaluations using **GOMF** and **GOCC** data would be more feasible.

In addition, a wealth of data which may be used to assess individual predictions, such as domain and phenotypic data, is stored in diverse biological databases. These data may (or may not) be consistent with a novel prediction. For instance, a prediction to a membrane-associated **GOBP** may be made for a protein containing a transmembrane domain, or, a prediction to a stress response **GOBP** may be made for a gene with a known stress-related phenotype. However, database searching is laborious and difficult to carry out in a systematic fashion given the number of databases available and the heterogeneous nature of the data. Therefore, database searching is not feasible for large numbers of predictions, and may only be carried out efficiently on a gene-by-gene basis.

Many of the genes in this study had multiple functional predictions to several GO terms. In many cases, these groups of terms were directly linked in the GO **DAG**. Multiple predictions of this type are a product of the **DAG** structure and the overlap between gene annotations. The same genes are often

²<http://expasy.org/sprot/>

annotated to the same group of annotations. When this is the case, the datasets' relevance scores are highly similar (if not identical) and, therefore, produce similarly weighted networks, leading to similar functional predictions. This DAG-based effect was also seen in Section 6.3.1 where the performance of the POIs was linked to the DAG structure.

Despite the difficulties of computational evaluation where there is available evaluation data for the genes, the relevance integration algorithm appears to perform well. Although the number of consistent and matching annotations was low, it is not possible to say with any certainty that any of the predictions are false. It is impossible to confirm that a novel hypothesis is correct using existing data. New data are required to computationally evaluate each new hypothesis. Therefore, a single functional prediction was chosen for detailed experimental analysis.

7.4 Laboratory Evaluation of a Functional Prediction

Many research groups have a specific focus and would apply integrated network analyses, such as the methods described in this thesis, to specific biological questions, rather than to the broad collection of GO terms used in the previous section. The focus of this project is on the ageing process. One important aspect of ageing is the response to oxidative stress since ROS have been associated with the ageing process (see Section 2.6.2.5). The GO term `response to oxidative stress` (GO:0006979) is defined as: "a change in state or activity of a cell or an organism (in terms of movement, secretion, enzyme production, gene expression, etc.) as a result of oxidative stress, a state often resulting from exposure to high levels of reactive oxygen species, e.g. superoxide anions, hydrogen peroxide (H_2O_2), and hydroxyl radicals"³.

The term `response to oxidative stress` provides an ideal POI for the laboratory evaluation of predictions for a number of reasons. The term is relatively specific, but 88 genes are annotated to it in *S. cerevisiae* which provides a good basis for relevance scoring and functional prediction. In addition, oxidative stress is easy to study in *S. cerevisiae*, since it can be induced by a number of readily available chemicals. Further, a response to oxidative stress can be quantified by simple spot testing of colony growth.

7.4.1 Choice of Prediction

In total, there were 368 predictions to `response to oxidative stress` (GO:0006979) produced by the networks, 181 of which scored greater than 1.0. Predictions to unannotated genes potentially

³http://amigo.geneontology.org/cgi-bin/amigo/term_details?term=GO:0006979&session_id=2617amigo1305289785

provide more interesting and novel hypotheses than those to previously annotated genes. Therefore the 33 predictions to unannotated genes were selected. Predictions produced by the control networks were also discarded, since they were produced in the absence of a relevance measure.

This selection process produced a final short list of two functional predictions to GO:0006979, YGL015C and YAL046C, both with a score of 1.6. The control network predicted YAL046C to be involved in transcriptional termination (GO:0030847 and GO:0030846) and YGL015C to be involved in the osmotic stress response (GO:0006970). Analysis of the genes' neighbourhoods was used to assess the level of evidence supporting the predictions. The first prediction, to the gene YGL015C, was supported by one line of evidence, while the second, to the gene YAL046C, was supported by two lines of evidence (Figure 7.8). Due to the time-scale of this project it was only possible to investigate one of these predictions, therefore the prediction to YAL046C was chosen for laboratory evaluation.

Prior to experimentation, the available data regarding YAL046C was investigated to assess the plausibility of this prediction. YAL046C codes for the protein Aim1, an unannotated 118 amino acid protein which has been linked to mitochondrial genome maintenance [1166]. The prediction to GO:0006979 was transferred to Aim1 along two equally-weighted edges from the genes YDR098C and YER174C. The experimental evidence for both of these edges was produced by a [HTP Y2H](#) screen by Yu and co-workers (2008) [134]. YDR098C and YER174C code for the Grx3 and Grx4 glutaredoxins. Grx3 and Grx4 are involved in the glutathione-glutaredoxin system and iron ion homeostasis [1060], two processes which are important during the oxidative stress (see Section 2.6.2).

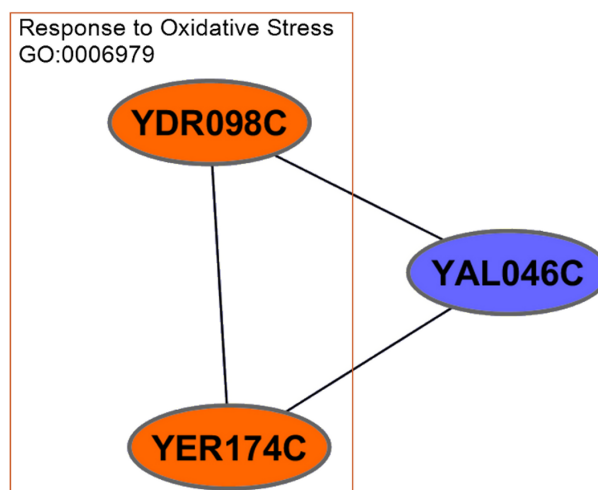


Figure 7.8: The evidence for the Aim1 prediction.

The prediction for the Aim1 gene (YAL046C) was transferred from the genes YDR098C and YER174C during functional prediction. YDR098C and YER174C code for the Grx3 and Grx4 glutaredoxins which also interact with each other in the network.

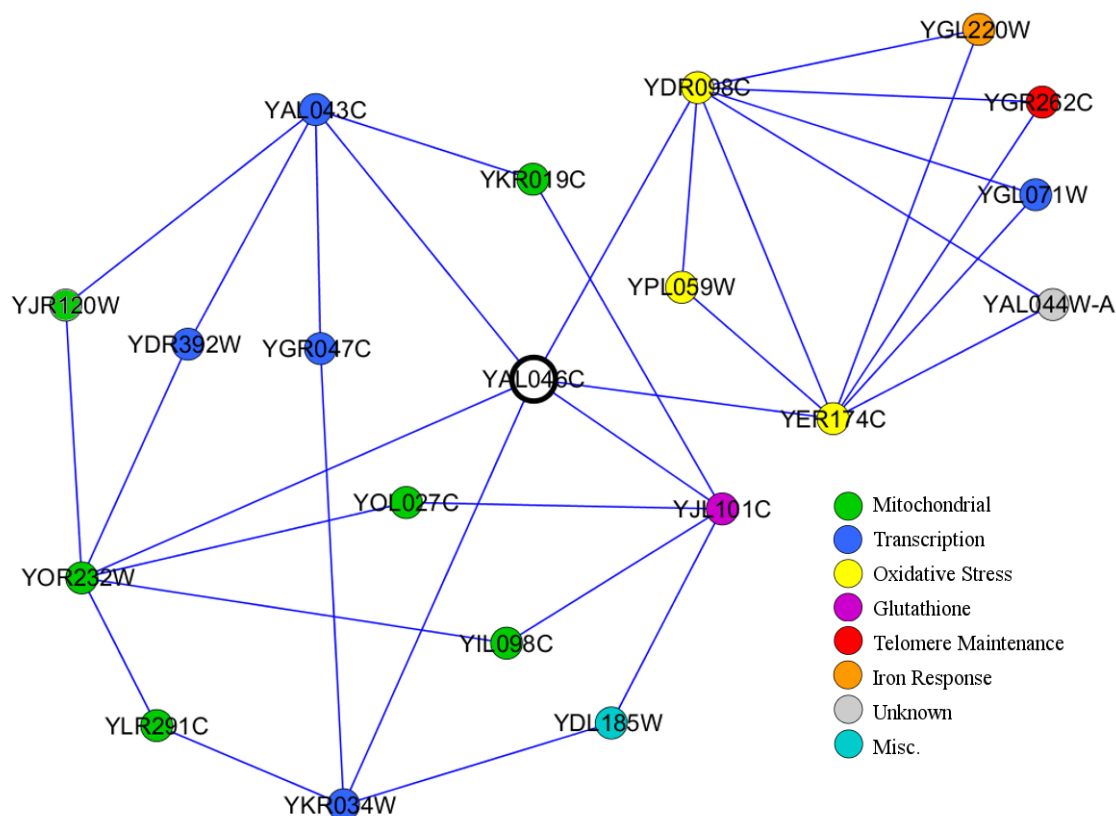


Figure 7.9: The extended neighbourhood of Aim1.

Nodes are coloured by general Gene Ontology biological process groups. It should be noted that there is overlap between the groups displayed. Most notably, the three oxidative stress response genes, YDR098C, YPL059W and YER174C, are also involved in the iron response, and, the iron response gene YGL220W is also involved in transcription.

The neighbourhood of Aim1 was visualised in Cytoscape and Ondex. The extended neighbourhood of the Aim1 in the [PFIN](#) contained several other mitochondrial genes, in addition to genes involved in oxidative stress, glutathione biosynthesis, telomere maintenance and the iron response (Figure 7.9). Notably, the neighbourhood also contained several genes related to transcription, the process in which the control network predicted Aim1 to be involved. The genes of the neighbourhood were annotated to a wide variety of [GOBP](#) terms including several other stress responses (Table 7.2). Additionally, several of the neighbourhood genes had mutant phenotypes associated with ageing-related processes (Table 7.3). Surprisingly, while many of the genes have oxidative stress-related phenotypes recorded in SGD, only four have been annotated to the term `response to oxidative stress`.

The Aim1 protein contains a BolA-like domain. The BolA superfamily⁴ contains homologues of the *E. coli* protein BolA. In *E. coli* BolA is a stress-induced protein which is thought to have a role in the general stress response [1167, 1168]. Interestingly, the protein has also been linked to transcription [1169]. Most notably, two other genes in the Aim1 neighbourhood of the network, YGL220W

⁴<http://supfam.org/SUPERFAMILY/cgi-bin/scop.cgi?sunid=82657>

Table 7.2: Gene Ontology biological process neighbourhood of Aim1.

The GOBP annotations of the genes in the Aim1 neighbourhood. GOBPs associated with ageing and oxidative stress are highlighted in bold, including, glutathione biosynthesis processes, mitochondrial terms and the responses to ROS producing compounds.

Orf	Gene	GOBP	Type
YDR098C	GRX3	Cellular response to oxidative stress	LTP
		Actin cytoskeleton organization	LTP
		Cellular iron ion homeostasis	LTP
YER174C	GRX4	Cellular response to oxidative stress	LTP
		Actin cytoskeleton organization	LTP
		Cellular iron ion homeostasis	LTP
YDR392W	SPT3	Ascospore formation	LTP
		Chromatin modification	LTP
		Conjugation with cellular fusion	LTP
		Gene-specific transcription from RNA polymerase II promoter	LTP
		Histone acetylation	LTP
		Invasive growth in response to glucose limitation	LTP
		Pseudohyphal growth	LTP
YOR232W	MGE1	Protein import mitochondrial matrix	LTP
		Protein refolding	LTP
YLR291C	GCD7	Regulation of translational initiation	LTP
YJL101C	GSH1	Glutathione biosynthetic process	LTP
		Response to cadmium ion	LTP
		Response to hydrogen peroxide	LTP
YKR034W	DAL80	Negative regulation of gene-specific transcription from RNA polymerase II promoter	LTP
		Nitrogen catabolite repression of transcription	LTP
YGL220W	FRA2	Negative regulation of transcription from RNA polymerase II promoter in response to iron	LTP
YAL043C	PTA1	mRNA cleavage	LTP
		mRNA polyadenylation	LTP
		Termination of RNA polymerase II transcription, exosome-dependent	LTP
		Termination of RNA polymerase II transcription, poly(A)-coupled	LTP
		tRNA processing	LTP
YKR019C	IRS4	Autophagy	LTP
		Cellular response to starvation	LTP
		Chromatin silencing at rDNA	LTP
		Fungal-type cell wall organisation	LTP
		Inositol lipid-mediated signalling	LTP
YOL027C	MDM38	Cellular potassium ion homeostasis	LTP
		Mitochondrial respiratory chain complex III biogenesis	LTP
		Mitochondrial respiratory chain complex IV biogenesis	LTP
		Positive regulation of mitochondrial translation	LTP
		Potassium ion transport	LTP
		Protein insertion into mitochondrial membrane	LTP
		Proton transport	LTP
YGR262C	BUD32	Positive regulation of transcription from RNA polymerase II promoter	LTP
		Protein phosphorylation	LTP
		Telomere maintenance	LTP
		Threonylcarbamoyladenine metabolic process	LTP
		Cellular bud site selection	HTP
YPL059W	GRX5	Cellular response to oxidative stress	LTP
		Iron-sulfur cluster assembly	LTP
		Response to osmotic stress	LTP

Table 7.2: Continued

The GOBP annotations of the genes in the Aim1 neighbourhood. GOBPs associated with ageing and oxidative stress are highlighted in bold, including, glutathione biosynthesis processes, mitochondrial terms and the responses to ROS producing compounds.

Orf	Gene	GOBP	Type
YDL185W	VMA1	Cellular protein metabolic process	LTP
		Intron homing	LTP
		Vacuolar acidification	LTP
YIL098C	FMC1	Mitochondrial proton-transporting ATP synthase complex	LTP
YGL071W	AFT1	High-affinity iron ion transport	LTP
		Positive regulation of transcription from RNA polymerase II promoter	LTP
YJR120W	-	Cellular respiration	LTP
		Mitochondrial organization	LTP
		Sterol transport	LTP

Table 7.3: Phenotypic neighbourhood of Aim1.

Several genes in the Aim1 neighbourhood have disruption phenotypes associated with ageing and oxidative stress responses. Oxidative stress response phenotypes are highlighted in bold. Many of these genes are not annotated to the corresponding GOBP terms.

Orf	Gene	Mutant	Scale	Phenotype
YAL046C	AIM1	Null	HTP	Mitochondrial genome maintenance: abnormal
YDR098C	GRX3	Null	HTP	Resistance to BPS: decreased
YER174C	GRX4	Null	HTP	Metal resistance: decreased
YDR392W	SPT3	Null	LTP	Chronological lifespan: increased
		Null	HTP	Metal resistance: decreased
YOR232W	MGE1	Conditional	LTP	Mitochondrial transport: decreased
		Repressible	HTP	Mitochondrial morphology: abnormal
YJL101C	GSH1	Null	LTP	Glutathione accumulation: decreased
		Null	LTP	Oxidative stress resistance: decreased
		Null	HTP	Resistance to cadmium chloride: decreased
YGL220W	FRA2	Null	LTP	Mitochondrial genome maintenance: abnormal
YKR019C	IRS4	Null	HTP	Resistance to BPS: decreased
YOL027C	MDM38	Null	LTP	Mitochondrial morphology: abnormal
		Null	HTP	Glutathione excretion: increased
		Null	HTP	Mitochondrial morphology: abnormal
YPL059W	GRX5	Null	LTP	Oxidative stress resistance: decreased
		Null	LTP	Mn-superoxide dismutase (Sod2p) activity: decreased
		Null	LTP	Replicative lifespan: decreased
		Null	HTP	Oxidative stress resistance: decreased
		Null	HTP	Resistance to arsenite(3-): decreased
YDL185W	VMA1	Null	LTP	Metal resistance: decreased
		Null	LTP	Oxidative stress resistance: decreased
		Null	HTP	Mitochondrial morphology: abnormal
		Null	HTP	Oxidative stress resistance: decreased
YIL098C	FMC1	Null	HTP	Glutathione excretion: increased
		Null	HTP	Mitochondrial genome maintenance: abnormal
		Null	HTP	Oxidative stress resistance: decreased
YGL071W	AFT1	Activation	LTP	Metal resistance: increased
		Null	LTP	Metal resistance: decreased
		Null	LTP	Resistance to BPS: decreased
		Null	HTP	Metal resistance: decreased
		Null	HTP	Oxidative stress resistance: decreased

and YAL044W-A, contain BolA-like domains, and both of the genes also interact with Grx3 and Grx4 (Figure 7.10). YGL220W codes for the protein Fra2, a 120 amino acid protein involved in the regulation of the iron regulon [1065, 1114]. Additionally, like Aim1, Fra2 has also been linked to mitochondrial genome maintenance [1166]. The interactions of Fra2 with Grx3 and Grx4 both have multiple lines of evidence, from both HTP and LTP studies [21, 24, 190, 1065, 1114, 1170]. YAL044W-A is a putative 110 amino acid unannotated protein. Notably, the protein is linked to Grx3 and Grx4 by the Y2H screen by Yu and co-workers which also links Aim1 to these proteins [134]. Interestingly, these are the only three *S. cerevisiae* proteins that possess BolA-like domains⁵.

Since the shared interaction partners of Aim1, Grx3 and Grx4, are both involved in both oxidative stress and iron regulation, a plausible hypothesis is that Aim1's predicted involvement in the oxidative stress response may be linked to iron homeostasis, since excess iron ions can produce ROS via the Fenton Reaction (see Section 2.6.2.1). To test this hypothesis, the network integration and functional prediction was repeated using the GO term iron ion homeostasis (GO:0055072) as the POI. YAL046C was predicted to be involved in this process and, therefore, the involvement of Aim1 in the response to oxidative stress and iron homeostasis can be considered a plausible functional prediction for laboratory analysis. Consequently, several simple stress tests were designed to evaluate the Aim1 deletion mutant's response to oxidative stress and varying iron levels (see Section 3.2).

7.4.2 Comparison with Traditional Database Searching

The RelCID method produces a list of candidate genes for annotation to a POI. Each candidate prediction has an associated score based on the level of evidence for the prediction. Predictions to unannotated genes were chosen here since they present more interesting hypotheses than predictions to genes with known function. Of the predictions, two had the highest score of 1.6. Since the prediction for *AIM1* had more lines of evidence supporting it, this prediction was considered to be the highest confidence prediction.

The total time taken to produce and assess this prediction is shown in Table 7.4. The first two stages of the process were automated, requiring six inputs; the BioGRID, KEGG, SGD and GO input files, a D-value for integration, and, a GO term of interest as POI (see Section 4.2). Stage 2 produced a list of predictions ranked in order of score, from which the highest-confidence prediction was selected. In stages 3-6, manual database searching was carried out to evaluate the plausibility of the prediction. In total the 6 stages took 21 hours.

⁵<http://www.yeastgenome.org/cgi-bin/protein/domainPage.pl?dbid=S000003188#domains>



Figure 7.10: Protein domains of Aim1, Fra2 and YAL044W-A.

The Aim1 (A), Fra2 (B) and YAL044W-A (C) proteins contain BolaA-like domains. In *E. coli* BolaA is a stress-induced protein that is involved in stress responses and transcription.

The RelCID method produces a list of candidate genes for annotation to a **POI**. Each candidate prediction has an associated score based on the level of evidence for the prediction. Predictions to unannotated genes were chosen here since they present more interesting hypotheses than predictions to genes of known function. Of the predictions, two had the highest score of 1.6. Since the prediction for Aim1 had more lines of evidence supporting it, this prediction was considered to be the highest confidence prediction.

The total time taken to produce and assess this prediction is shown in Table 7.4. The first two stages of the process were automated, requiring six inputs; the BioGRID, KEGG, SGD and GO input files, a D-value for integration, and, a GO term of interest (see Section 4.2). Stage 2 produced a list of predictions ranked in order of score, from which the highest-confidence prediction was selected. In stages 3-6, manual database searching was carried out to evaluate the plausibility of the prediction. In total the 6 stages took 21 hours.

In the absence of a network-based, or other statistical, analysis any gene with an interaction involving the **POI** may be considered a candidate for annotation to that term. Given the scale of interaction data, identifying candidate genes is non-trivial, even when the data are represented as a network (Figure 7.11). Traditionally these candidates would have been identified by database searching. For instance, the Gene Ontology database provides the genes annotated to the **POI** and the BioGRID database

Table 7.4: Analysis time for the AIM1 prediction.

Computational stages are shown in italics. Stages 1 and 2 are fully automated. Six inputs are required; the BioGRID, KEGG, SGD and GO input files, a D-value for integration, and, the GO term of interest. Four networks are produced as input for stage 2. Stage 2 then produces a list of functional predictions with associated score from which the highest confidence prediction may be selected for analysis. This stage may involve a small amount of human input if there is more than one highest scoring prediction. Stages 4-6 involve manual database searching for evidence supporting (or disproving) the prediction.

Analysis Stage	Description	Time (hours)
1. <i>Network Build</i>	Dataset scoring and integration into three relevance networks and a control network.	0.5
2. <i>Prediction and Selection</i>	Functional prediction using the maximum weight decision rule and selection of the highest confidence predictions.	6.5
3. <i>Visualisation</i>	Visualisation of Aim1's interactions in Cytoscape and Ondex.	3
4. <i>SGD</i>	Survey of the available data for Aim1 and its surrounding neighbourhood in the network.	3
5. <i>Other databases</i>	Survey of other database evidence regarding Aim1 and its surrounding neighbours in the network.	4
6. <i>Literature</i>	Literature survey of Aim1 and its interaction partners.	4
	Total	21

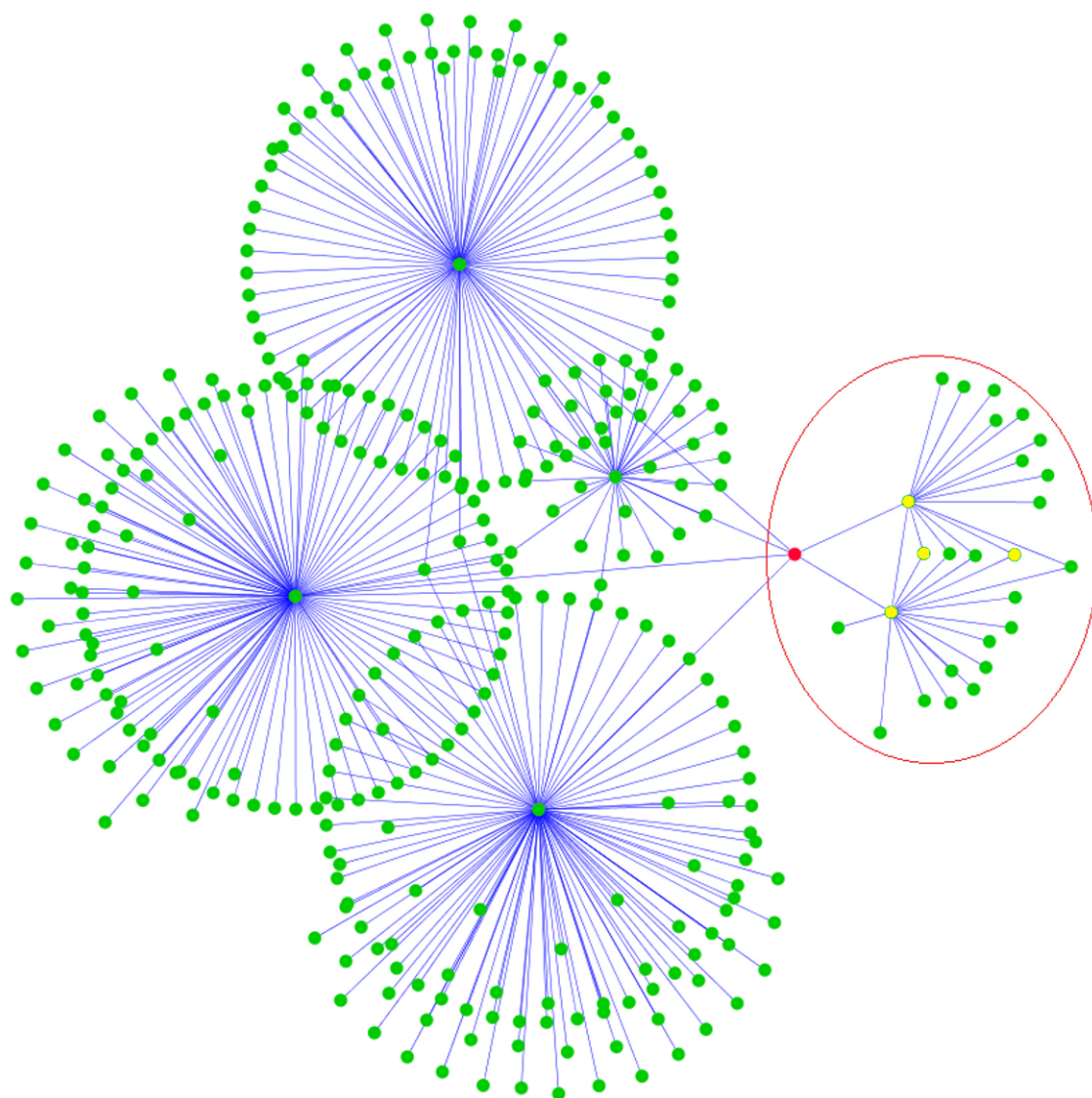


Figure 7.11: The level-2 Aim1 neighbourhood.

The extended neighbourhood of Aim1 (red) includes any node within a shortest path of 2 from the gene. The neighbourhood consists of 339 nodes and 352 edges. In the circled area alone there are 4 genes annotated to the response to oxidative stress (yellow) and, consequently, 23 potential candidate genes including Aim1. In the absence of a statistical method to score candidate genes, such as the one presented in this thesis, there is no way to narrow down the hypothesis field and, therefore, each candidate must be manually assessed. As the size of the network increases to level-3 neighbours and beyond, the network becomes too large for easy visual analysis and the number of genes interacting with the POI increases.

provides the interactions involving the genes. In this case, all candidates are equally weighted and must be individually assessed using the available literature and evidence.

A total of 87 genes were annotated to response to oxidative stress in June 2010. There were 90 interactions involving pairs of these genes and 1984 interactions between these genes and other genes. Of the 1984 interactions with other genes, 51 involved unannotated genes. The identification of the 51 candidate genes took 2 hours.

Since analysis of the Aim1 prediction took 14 hours (stages 3-6 of Table 7.4), it would potentially take 716 hours (2 hours + 14 hours*51) to identify and evaluate all 51 candidates. Therefore, traditional database searching could require up to 7.9 weeks (based on a 40 hour working week) to identify a high confidence and plausible prediction for laboratory analysis, an increase of 97% in analysis time.

7.5 Experimental Results

7.5.1 Strain Confirmation

The wild type (wt) strain, BY4741 (*MATa his3Δ1 leu2Δ0 met15Δ0 ura3Δ0*), and, the *AIM1* deletion mutant, BY4741 *aim1Δ* (*MATa his3Δ1 leu2Δ0 met15Δ0 ura3Δ0 aim1::kanMX4*) were acquired from Invitrogen⁶. The deletion was confirmed by PCR. Two primers, approximately 800 bp apart, were designed to span between the upstream region of *AIM1* gene and the middle of the KanMX module (Figure 7.12). PCR produced a clear 800 bp band for the mutant strain, confirming deletion of *AIM1* (Figure 7.13).

7.5.2 Stress Responses

The wild type and *aim1Δ* mutant strains were subjected to oxidative stress tests by spotting them onto plates containing a variety of oxidative stress-inducing compounds-hydrogen peroxide, menadione, diamide, tBOOH, cadmium and arsenic- at various concentrations (see Section 3.2.4.1). The

⁶<http://www.invitrogen.com/site/us/en/home.html>

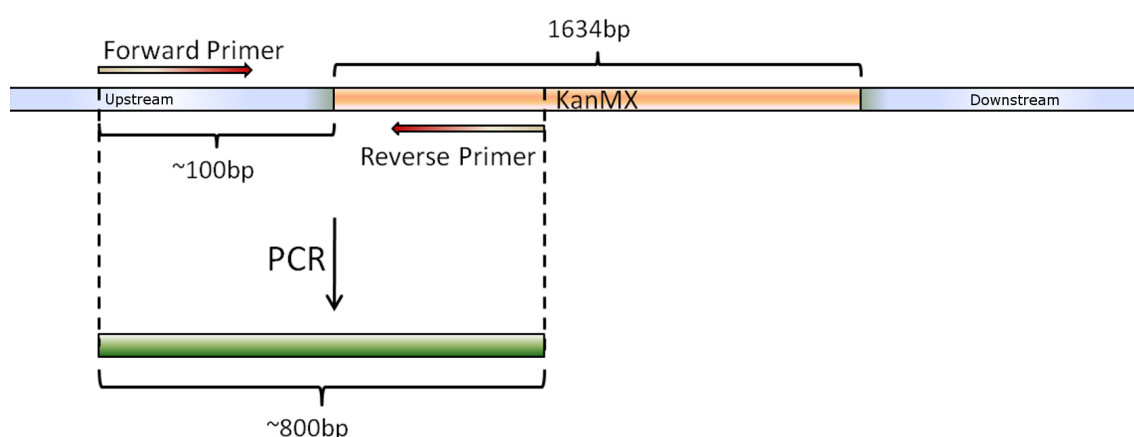


Figure 7.12: Strain confirmation.

In the mutant strain the *AIM1* gene has been deleted by replacement with the KanMX resistance module. This replacement was confirmed by PCR. Primers were designed to amplify a region of approximately 800 bp spanning the *AIM1* upstream region and the KanMX module.

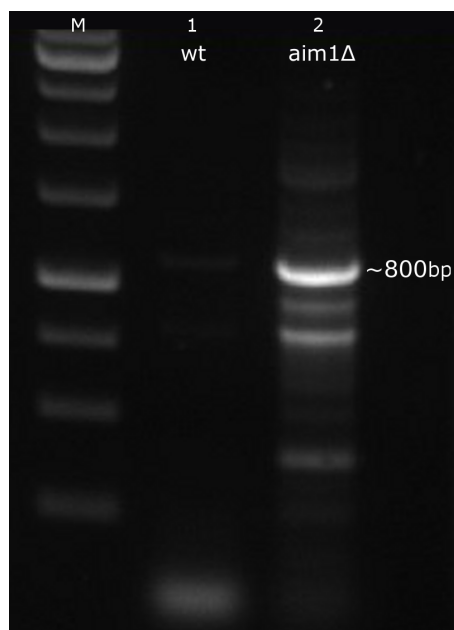


Figure 7.13: PCR result.

Separation of the PCR products following amplification of the target region. Lane 1 contains the PCR products of the wt strain. Lane 2 contains the PCR products of the *aim1Δ* mutant strain. The expected band at approximately 800 bp is labelled.

mutant and wild type strains did not show any difference in growth under oxidative stress-inducing conditions (Figure 7.14).

To test the mutant's response to iron, the wild type and *aim1Δ* mutant strains were grown in liquid culture to stationary phase under three different conditions: average iron, low iron, and, high iron. For average iron the strains were grown in YPD medium. For growth under limited iron conditions (Fe^-) the strains were grown in YPD with 100 μM bathophenanthroline disulfonate (BPS). For growth under increased iron conditions (Fe^+) the strains were grown in YPD with 100 μM iron chloride (see Section 3.2.1).

Therefore, 6 different cultures were produced:

- | | | |
|-----------------|----------------------------|-----------------------------|
| 1. wild type | 3. wild type + low iron | 5. wild type + high iron |
| 2. <i>aim1Δ</i> | 4. <i>aim1Δ</i> + low iron | 6. <i>aim1Δ</i> + high iron |

The six cultures were spotted onto plates containing average, low and high iron. No difference in growth between mutant and wild type strains was observed under any condition (Figure 7.15).

The six cultures were then subjected to oxidative stress testing by spotting onto plates containing several different oxidative stress-inducing reagents, as above. Again, no difference in growth between the mutant and wild type strains was observed under any condition (Figure 7.16).

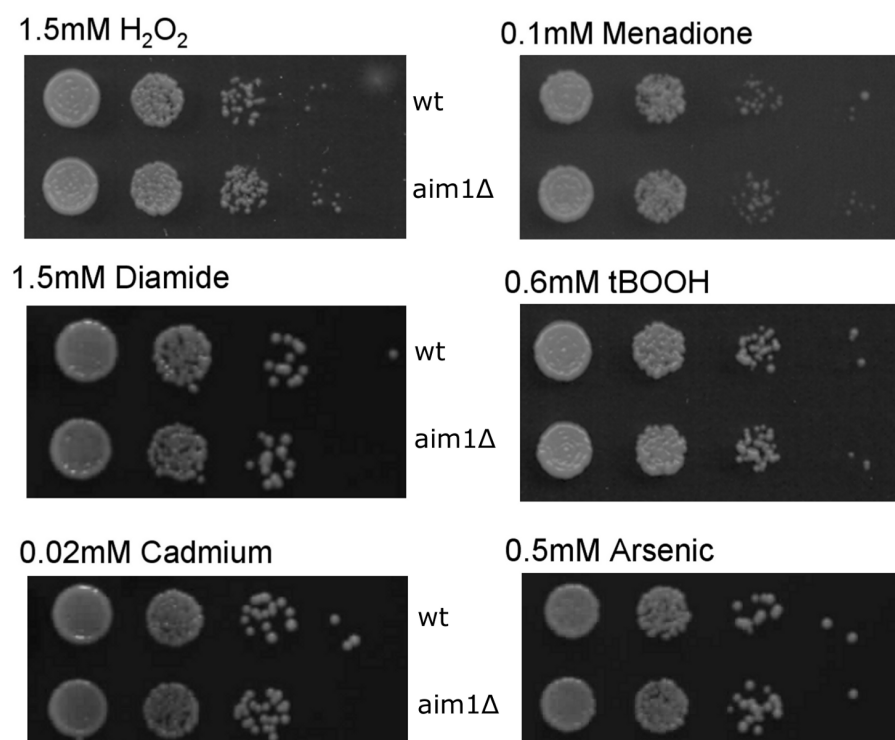


Figure 7.14: Oxidative stress response.

Serial dilutions of the wild type and *aim1Δ* mutant strains were spotted onto plates containing various oxidative stress-inducing reagents. The first row contains the wild type strain and the second row the *aim1Δ* mutant.

Finally, the six cultures were plated on YPG plates in order to provide the growth conditions for respiration (see Section 3.2.4.2). No difference in growth between the wild type and mutant strains was observed under this growth condition (Figure 7.17).

7.6 Discussion

Section 7.4 fulfils the final objective of this project (Objective 6, Section 1.5) using oxidative stress as an exemplar ageing-related POI for laboratory evaluation of a novel prediction. Ageing is a complex phenomenon that has recently been the focus of considerable research due to its association with many diseases [964, 965]. Many biological processes have been linked to the ageing process and several theories of ageing have been postulated, all of which have some overlap (see Section 2.6). Of the two major theories, one is based on chromosome structure and maintenance, in particular the maintenance of the telomere and the accumulation of DNA damage [982, 997, 1001–1003], and the other is associated with ROS, oxidative stress and the mitochondria [1116–1119]. There is thought to be significant overlap between these theories in terms of biology and evidence, but many aspects of the ageing process remain unclear [972–974].

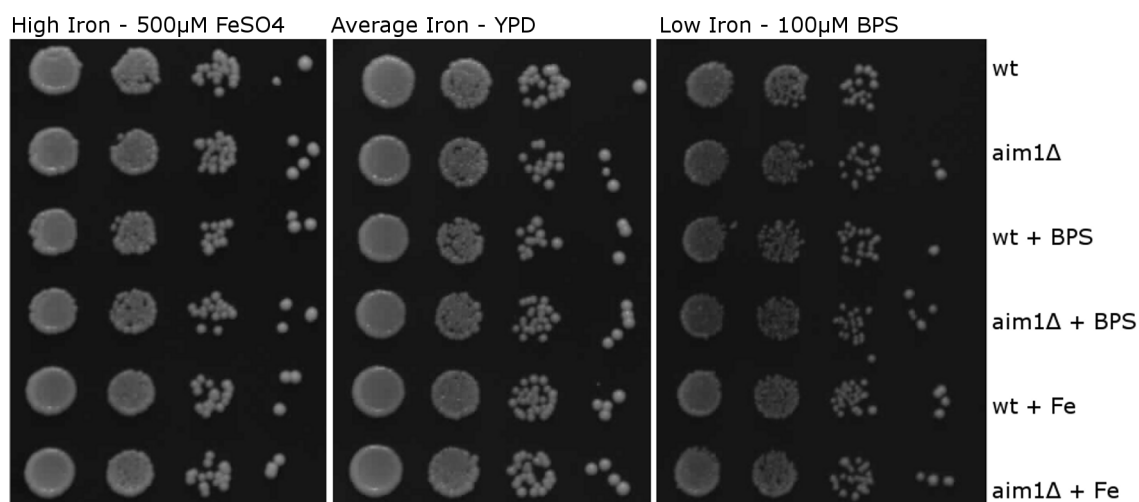


Figure 7.15: Iron response.

The wild type and *aim1Δ* mutant strains were grown under average, low and high iron conditions. Iron chloride was used to produce high iron conditions and the iron chelator bathophenanthroline disulfonate (BPS) to produce low iron conditions. Serial dilutions of the cultures were spotted onto plates containing average, low and high iron. The top two row are average iron growth, the middle two rows are low iron growth and the bottom two rows are high iron growth. The wild type strain is the top row in all three cases.

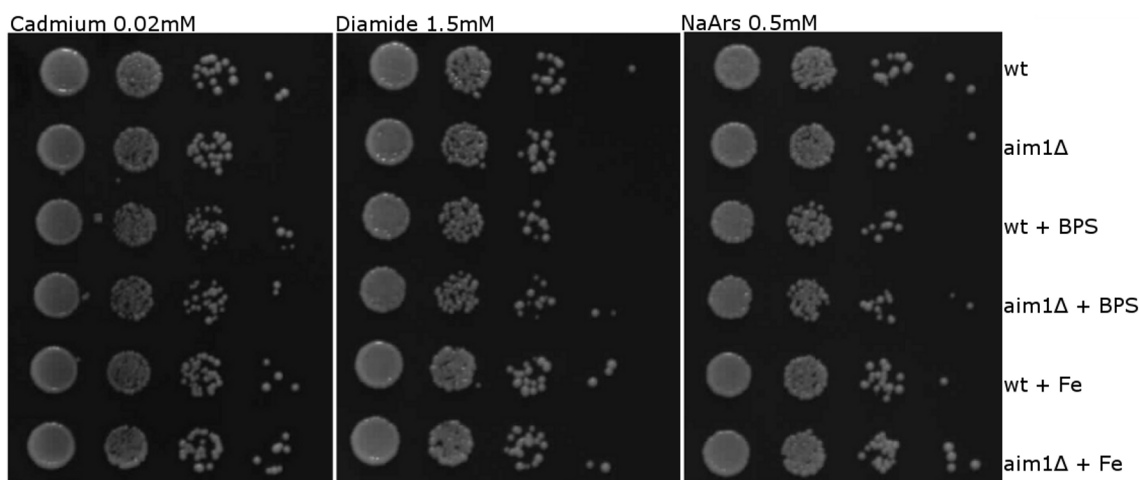


Figure 7.16: Oxidative stress response.

The wild type and *aim1Δ* mutant strains were grown under average, low and high iron conditions and serial dilutions of the cultures were spotted onto plates containing various oxidative stress-inducing reagents. The top two row are average iron growth, the middle two rows are low iron growth and the bottom two rows are high iron growth. The wild type strain is the top row in all three cases.

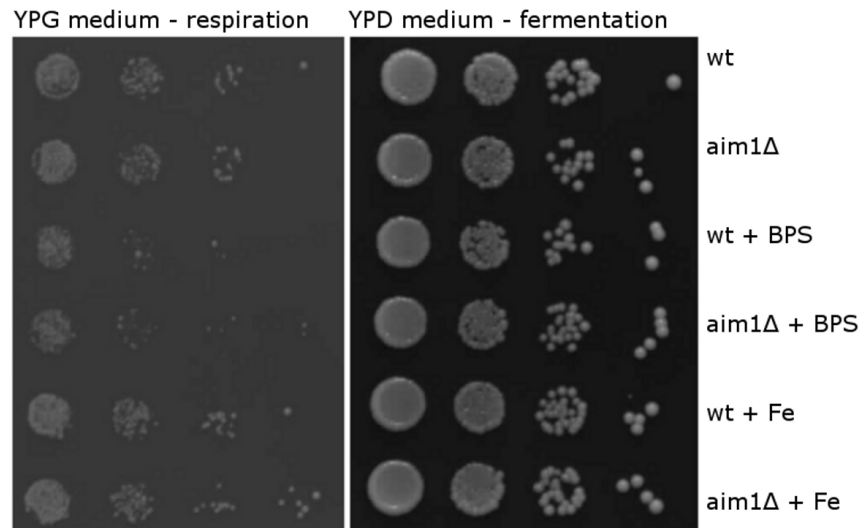


Figure 7.17: Response to differing carbon source.

The wild type and *aim1Δ* mutant strains were grown under average, low and high iron conditions. Serial dilutions of the cultures were spotted onto YPD plates for fermenting growth and YPG plates for respiring growth. The top two row are average iron growth, the middle two rows are low iron growth and the bottom two rows are high iron growth. The wild type strain is the top row in all three cases.

The oxidative stress response is thought to be important to lifespan in yeast and humans [1120]. Oxidative stress is caused by the presence of ROS. ROS are thought to accumulate over time, causing damage to cellular components and, therefore, contributing to the cellular ageing [1016]. Since the mitochondria are the major source of ROS within cells, the mitochondria are also linked to ageing [1118].

The oxidative stress response in *S. cerevisiae* is an ideal POI for laboratory evaluation of the network integration and functional prediction schema developed in this thesis. *S. cerevisiae* is cheap, fast-growing and easy to work with [949, 956, 957]. Further, mutant disruption strains are available for the majority of the *S. cerevisiae* genome (see Section 2.6). Oxidative stress may be induced in *S. cerevisiae* using several relatively cheap reagents such as hydrogen peroxide and diamide [1087]. Additionally, oxidative stress occurs during respiratory growth [1023]. Finally, phenotypic changes in *S. cerevisiae* can be easily identified by spot testing of colony growth.

In this section, 368 functional predictions for the GOBP term response to oxidative stress (GO:0006979) were produced using the RelCID integration and functional prediction method. Since the predictions each have scores, based on the network edge weights, they may be ranked in order of confidence and the highest-scoring prediction selected. Here, predictions to unannotated genes were selected as the focus of analysis since they have more scope for novel discovery. However, it could be argued that predictions for genes which have known annotations to a parent term of the POI, may be of greater interest since they are consistent with known data.

The highest scoring prediction produced by the networks was for the gene *AIM1*. This gene has no annotations, but has been linked to mitochondrial genome maintenance by phenotypic data [1166]. However, despite being associated with the mitochondrion, and having several mitochondrial genes in its level-2 neighbourhood, Aim1 only has one direct interaction with another mitochondrial gene, a negative genetic association from a study by Costanzo and colleagues (2010) [1171]. Nevertheless, given that the mitochondrion is associated with oxidative stress and ageing, the association of Aim1 and the mitochondrion is consistent with the prediction.

Several other aspects of Aim1 and its neighbouring network also support the prediction, including phenotypic (Table 7.3) and domain data (Figure 7.10). Further, the involvement of its neighbours Grx3 and Grx4 in iron ion homeostasis [1060] suggests a possible link between Aim1 and the iron response. This potential link was subsequently supported by further network analysis. Therefore, the highest-scoring prediction produced by the network is plausible and is supported by additional data. Consequently, the final hypothesis for the laboratory evaluation was that Aim1 is involved in the response to oxidative stress, possibly through an association with iron homeostasis.

The control network predicted Aim1 to be involved in several general transcription-related terms. This prediction is also plausible given the link between BolA and transcription [1169]. However, transcription is a very general process. Every gene which is transcribed has a role in a specific process, and many genes are transcribed in response to specific conditions. Therefore, it is reasonable to hypothesise that Aim1 may have a role in transcription as part of the oxidative stress response. If this were the case, both the relevance network and control network predictions would be correct. However, while the control network only predicted an involvement in transcription, the relevance networks predicted a more specific role. Cases such as this highlight some of the drawbacks of the GO structure. When annotations are taken out of context important observations may be missed. However, currently there is no systematic method to link GO annotations to one another beyond the relationships of the GO DAG. A connection between the distinct processes can only be made by referring to the original literature supporting the annotations or, in some cases, by referring to the official SGD description of the gene. Neither method is amenable to large-scale computational analysis. However, it could be argued that developing linkages of this type may over-complicate what is already an intricate annotation schema.

The RelCID approach identified a high-confidence prediction in just seven hours. Evaluation of the prediction then took fourteen hours. In comparison to traditional database searching, the network-based approach reduces analysis time by approximately 97%, since it would take up over 17 weeks to individually evaluate each candidate gene by hand. In reality, this extrapolation in time is likely to be an over-estimate, since it is unlikely that every gene would require fourteen hours of analysis.

Further, in some cases an interesting prediction may be found for experimental validation before all the candidates have been assessed. If this were the case, it is possible that a better and more interesting hypothesis may be present in the group of genes that was not evaluated. Therefore, traditional database searching is more subjective than statistically-based network analyses since it has no numerical aspect. Further, other GOBPs may have more or fewer candidate genes than GO:0006979. Consequently, the RelCID technique reduces analysis time, potentially to a large extent by providing a quantitative measure of prediction confidence.

The initial experimental results did not show any link between Aim1 and the oxidative stress or iron responses. However spot tests, such as the ones carried out, are the simplest and least sophisticated method to identify phenotypic changes in mutant strains. Many genes associated with a particular biological process do not have an equivalent phenotype when disrupted. For instance Grx3 and Grx4 are annotated to the response to oxidative stress but neither gene has a related mutant phenotype for single deletions [1057]. The two genes must be disrupted as a double mutant for a phenotypic effect to be seen.

Consequently, while the prediction for *AIM1* may be correct, more sophisticated experimental techniques are required to validate it. One possible approach is to examine transcript levels in the mutant strain by either real-time PCR or sequencing. Differences in transcript levels for genes associated with the oxidative stress and iron responses between wild type and *aim1*Δ mutant would support the prediction. A second approach would be to examine double mutants in which *AIM1* and a second POI-related gene has been deleted. If Aim1 acts in a parallel role to another protein a phenotypic effect may be seen in the double mutant. Since two other genes in the Aim1 neighbourhood contain BolA-like domains and share interaction partners with Aim1 [21, 24, 134, 190, 1065, 1114, 1170], they are both potential candidates for parallel roles.

Chapter 8

Discussion and Future Work

The aim of this project was to investigate and develop techniques to exploit data bias in order to optimise [PFIN](#) performance in relation to specific biological questions. In order to achieve this the six objectives set out in [Section 1.5](#) have been met. In [Chapter 4](#) the available functional data was evaluated revealing the differences and similarities of the datasets, and the individual dataset biases (Objective 1). These biases allowed dataset relevance to be quantified in relation to specific biological processes (Objective 2). A relevance measure was incorporated into a novel [PFIN](#) integration technique, RelCID, which harnesses dataset relevance in order to direct network analyses to answer specific biological questions (Objective 3).

During the initial stages of the project significant changes to the source databases were observed owing to the curation process. Investigation of the effect of these changes on network performance in [Chapter 5](#) revealed that the assumption of increasing network performance over time is incorrect. In fact network performance fluctuates. However, the RelCID technique was demonstrated to improve network performance and overcome some of the effects of the curation process (Objective 3).

Existing network evaluation techniques were global in that they assessed network performance across all areas of biology. Therefore evaluation of RelCID required the development of several process-specific evaluation techniques throughout this project (Objective 4). Following thorough evaluation and optimisation in [Chapter 6](#), RelCID was applied in [Chapter 7](#) to produce several novel hypotheses for the yeast *Saccharomyces cerevisiae* (Objective 5). Computational assessment of the hypotheses demonstrated RelCID's utility in tailoring analyses to specific biological processes (Objective 6). In particular RelCID significantly reduces analysis time. Finally an initial laboratory evaluation of a single prediction was carried out (Objective 6). The techniques developed and conclusions drawn during the fulfilment of these objectives are discussed in detail in the following sections.

8.1 Introduction

Computational analysis has become essential to biology due to the recent explosion in new high-throughput experimental technologies [1]. There is simply too much data to analyse manually within a reasonable time-scale, and the amount of biological data available continues to grow. The field of Systems Biology aims to analyse this wealth of data by drawing on the principles of computing science, mathematics and statistics [26, 32–36]. At the heart of Systems Biology is the elucidation of the interactome: the entire complement of molecular interactions that may occur within an organism under all circumstances and for all cell types [37]. However, this is a substantial task.

Available biological data are spread over hundreds of databases and are heterogeneous in nature [27, 67]. The datasets have differing levels of noise, bias and genomic coverage [42, 68, 69, 128, 430, 673]. In order to fully characterise every aspect of highly complex cellular systems, and infer new knowledge from the data, diverse data sources must be systematically integrated [48, 49, 51, 420, 497, 668–671].

PFINs are one of the most powerful graph-theoretic approaches to data integration; they reduce the impact of dataset noise by taking a measure of dataset confidence into account during network integration [49, 115, 128]. In **PFINs** nodes correspond to genes or gene products and the edges to functional associations between nodes. The edge weights of **PFINs** indicate a level of confidence in the combined evidence for that edge, usually calculated by statistical comparison against a Gold Standard dataset [669, 699]. Therefore, edges with multiple lines of evidence may be up-weighted, although where the corresponding evidence is of low confidence, the weighting is reduced.

Graph-theoretic algorithms may be adapted to take **PFIN** edge weights into account [57, 88, 431, 517, 533, 534, 568, 569]. For instance, rather than propagating gene annotations along any edge in a network during **GBA** functional prediction, annotations may be propagated only along the highest confidence edges using the Maximum Weight rule [57]. By including this measure of dataset confidence into network analysis, **PFINs** have produced improved accuracy in a number of graph theoretic applications; for instance, in the detection of protein complexes [90, 534, 592], the annotation of proteins [49, 92, 103, 105–112, 674, 711] and the prediction of new interactions [318, 675].

The majority of research groups have specific research interests [128]. While available functional data contains a wealth of valuable information, only a subset of the data has relevance to each specific biological question. Often this relevant data may be difficult to distinguish from the large amount of other data in a network. In particular, high confidence data that is not relevant to the biological question being addressed may obscure more relevant data when edge-weighted analyses are carried out. In other words the data that is of no relevance to the process under investigation can act as

additional noise in the network.

Given the growing amount of the data available it will become increasingly necessary for many purposes to focus on a subset of an integrated network. Previously, methods to produce area- or process-specific subnetworks have relied on known annotation data as the basis for data filtering [46, 59, 129, 130]. However, many areas of the interactome contain very few (or sometime no) annotated genes [902]. Further, many annotations are high-level, general terms which lack the specificity needed for process-specific analysis.

8.2 The RelCID Algorithm

RelCID, the novel algorithm described in this thesis, allows process-relevant PFINs to be integrated without loss of data, by harnessing dataset biases. Process-relevant edge weightings are produced by ordering the datasets according to their relevance to a specific biological process of interest (POI), prior to integration of their confidence scores. Therefore, a higher weighting is given to the most relevant datasets. However, since the edge weights are based on dataset confidence, dataset noise is still reduced as in classical PFINs. Since relevance and confidence are calculated independently, the RelCID method can be applied to any confidence scoring schema using any Gold Standard dataset and any POI.

Importantly, while known annotations are used for the relevance scoring, an entire dataset is scored as a whole. Therefore, all a dataset's edges are treated equally, inclusive of those involving unannotated genes. Therefore, unlike previous process-relevant methods, the up-weighting of relevant data is global and is not limited to well-annotated areas of the interactome. Consequently, the PFINs produced by RelCID have far less bias in weighting towards known and highly-studied genes, giving more scope for the generation of novel process-specific hypotheses.

The relationship between dataset relevance and PFIN performance is complex and involves a number of factors. Several conclusions regarding this relationship and the use of process-relevant networks were drawn during this research:

- As previously demonstrated [46, 98, 121] datasets have their own unique biases in relation to specific biological questions.
- The incorporation of a measure of dataset relevance by the RelCID algorithm improves network performance with respect to the prediction of protein function.
- The relationship between dataset relevance and network performance is highly complex.

- The manual curation of source databases alters network performance over time.
- The RelCID algorithm produces plausible novel hypotheses that cannot be produced in the absence of a measure of relevance.

These conclusions are discussed in the following sections.

8.2.1 Dataset Relevance

Datasets generated by different experimental approaches show clear differences in the areas of biology that they cover (see Section 4.1.2). In particular, genetic data have a distinct focus in comparison with physical data, and different experimental types have different biases in relation to biological processes. These distinctions are not unsurprising, given the nature of the experimental methodologies. For instance, experimental methods that identify physical interactions, such as Y2H [144] and TAP-MS [181], are likely to be biased towards processes involving the physical binding of proteins, such as complex formation. Conversely, experimental methods for the detection of genetic interactions, such as SGAs [14], dSLAM [15] and E-MAPs [17], are biased towards processes that include indirect functional relationships, such as regulatory relationships. However, the differences between physical and genetic data types are generally too high-level to infer relevance to specific biological questions.

It appears that individual studies also have more specific low-level biases due to their experimental design and focus (see Section 4.1.3.5). These biases can reflect similarities between datasets of different experiment types. Naturally, some studies are more relevant to specific areas of biology than others and therefore by measuring the levels of bias a dataset's process-relevance may be quantified.

Given that individual research groups have their own specific interests, these biases are not unsurprising. In fact, a research group conducting both experimental and network-based analyses is likely to find its own experimental datasets have high, if not the highest, relevance to their POI. In this case the use of a process-relevant integration technique allows iterative refinement of a process-relevant network. Experimental data can be up-weighted in the network, analysis of which can, in turn, guide further experiments. This type of iterative analysis is the basis of the field of Systems Biology [26, 32–36].

8.2.2 Harnessing Relevance

The hypergeometric test provides a measure of dataset relevance to specific biological processes. The test produces a p-value for the representation of a GO term in a dataset's gene annotations in

comparison to its representation in the *S. cerevisiae* genome as a whole. A group of datasets may therefore be ranked in order of relevance to a [POI](#). RelCID uses these ranks to extend the integration method developed by Lee and co-workers in 2004 [49].

Lee and colleagues observed that there are likely to be dependencies between available biological datasets (see Section 2.4.3) and consequently attempted to overcome this difficulty by introducing a weighted sum during network integration. The sum successively down-weights dataset confidence scores in order of magnitude. Therefore, the method gives a higher weighting to datasets with higher confidence and produces a network in which highly-weighted edges have high confidence (see Section 3.1.4.4).

In the RelCID schema the weighted sum is adapted to incorporate the relevance rankings by re-ordering the datasets prior to integration of the confidence scores, giving a higher weighting to datasets with higher relevance. Therefore, in the resulting relevance network highly-weighted edges have both high confidence and high relevance to the [POI](#). The relevance network is topologically identical to a network integrated without a measure of relevance and differs only in its edge weights. The weighting differences are directly attributable to the up-weighting of the more relevant datasets in the network.

In this project, four manually-curated data sources are used to integrate the process-relevant networks: BioGRID as source data, KEGG PATHWAYS as the Gold Standard for confidence scoring, and GO and SGD to provide annotations for relevance scoring. While two distinct databases were chosen for confidence and relevance scoring respectively, the relevance integration schema could be applied using GO for both of the scores. The use of GO in this way would be possible since the two scores measure distinct aspects of a dataset's GO annotation. A dataset may have high confidence (many genes sharing the same annotation) but have low relevance (few or no genes annotated to the [POI](#)), and vice versa. A dataset may also have both high confidence and high relevance, or both low confidence and low relevance. Similarly, KEGG PATHWAY annotations could be used as the basis of relevance scoring as could other Gold Standard datasets (see Section 2.4.2). Therefore, the RelCID algorithm is highly versatile and is applicable using any combination of Gold Standard datasets.

8.2.2.1 Evaluation Strategy

Four evaluation methods were chosen to assess the relevance networks' performance in relation to a control network that was integrated without a measure of relevance. A common use for [PFINs](#) is the inference of protein function [104, 510, 676, 676, 907, 911, 919]. Functional inference provides an excellent approach to network evaluation since it produces a numerical measure of network performance as the [AUC](#) of a [ROC](#) curve [920, 921]. Many functional prediction algorithms have

been developed (see Section 2.5.5). Since relevance and control networks are topologically identical, the functional prediction algorithm chosen should utilise the networks' edge weights rather than just their topology.

In this project, the local GBA algorithm Maximum Weight was chosen for network evaluation [57]. This algorithm propagates annotations locally along the highest weighted edge, and has been shown to have the most accurate performance of local GBA algorithms. Since the relevance and control networks differ in edge weighting alone, Maximum Weight provides a simple and direct comparison of network performance.

Both the relevance and the control networks' edge weights are produced by integration of the datasets in a ranked order [128]. Therefore, as a control for the evaluation, functional prediction was repeated for networks integrated with these ranks reversed. These networks acted as a null hypothesis, since if the ranked integration produces optimal performance as expected, reversal of the ranks should decrease network performance in relation to the POIs.

Clustering is a common method by which large complex networks may be divided into smaller, visualisable sub-parts [527]. The MCL clustering algorithm was chosen since it also uses edge weights [534]. Unlike functional prediction algorithms, clustering algorithms do not produce a numerical estimate of network performance. Evaluation of clustering is therefore more subjective, relying on visual analysis and interpretation of the data. Since the networks in this project were focussed on the ageing process, a selection of ageing-related GO terms was chosen to evaluate the clusters, including several DNA damage and repair, mitochondrial and telomeric region terms. The co-clustering of genes annotated to these ageing related terms was then used to assess clustering performance.

Integrated networks are often intended to generate new hypotheses for laboratory validation. Therefore, it was also essential to evaluate the relevance networks' performance in light of new data. This evaluation was achieved in two ways. First, the functional prediction results and network clusters were compared with two ageing-related datasets [16, 993] which had been produced after the networks were integrated. Second, the functional prediction results were compared with new annotations to the POI, which had been added to the GO database after network integration. Many GO annotations are derived from existing functional interaction data, and therefore some dependencies exist between the integrated datasets and the GO annotations used to assess them. By using data that was not available at the time of integration these dependencies were eliminated, giving an objective evaluation of network performance.

8.2.2.2 Network Performance in Relation to Ageing

Initially, a relevance measure was developed based on the number of genes within a dataset that are annotated to a specific GOBP, termed the POI. Due to the transitive nature of the GO DAG, any genes annotated to a child term of the chosen term were included in the POI [100]. This relevance measure was termed Node Relevance and was incorporated into network integration, in addition to a measure of dataset confidence [49]. The performances of the process-relevant networks were compared to those of a control network, integrated based on dataset confidence alone. The networks were topologically identical, since they were integrated using the same source data, but differed in their edge weights. *S. cerevisiae* ageing was used as a test POI.

Several results were apparent regarding the ageing networks' performance in relation to the control:

- The relevance networks' functional prediction performance was significantly improved over that of the control network.
- Reversal of the relevance integration order significantly reduced functional prediction performance, while reversal of the control network integration order had little effect on functional prediction performance.
- The relevance networks had fewer clusters than the controls and a higher percentage of clusters containing genes annotated to ageing-related processes, and unknown genes.
- Genes annotated to the POI in the relevance networks appeared in separate, smaller clusters in the control network.
- The clustering improvements apparent in the relevance networks were not observed when the networks were analysed with regard to processes unrelated to ageing.
- When compared to ageing-related data which was not available at the time of integration, relevance network clustering of the new data was improved over that of the control network.
- Relevance network performance in relation to the assignment of new annotations could not be conclusively assessed due to significant changes in the GO database.

It appears that the content of datasets is of greater importance than the datasets' confidence scores when network analysis is applied in relation to a specific process. Reversal of the order of the confidence-ranked integration of the control networks has little effect on network performance in relation to ageing. This observation is initially surprising, given that the power of PFINs is attributed to their confidence-based edge weights [49, 115, 128]. However, in this study performance is only

assessed in relation to a single biological process. It is highly likely that, if assessed globally for all GOBP terms, reversal of the control network order of integration would cause a larger drop in functional prediction performance. However, due to the scale of GO a full evaluation of the effect of score rank reversal was beyond the time-frame of this project.

Significantly, in relation to ageing reversal of the order of relevance-ranked integration causes a significant performance decrease. Given that this reversal gives lower edge weightings to the most relevant datasets this effect is expected. When relevant data are down-weighted in this way it is masked by less relevant data during functional prediction.

Conversely, when the datasets are integrated in order of relevance a significant performance increase is observed compared to the control network. This observation suggests that high relevance data may be masked by high confidence but low relevance data in the control network. Dataset content is clearly of high importance when network analyses are focused on a specific biological question. Therefore, incorporating a measure of relevance during network integration improves functional prediction performance.

Many PFINs are too large to easily analyse visually. While computational algorithms can identify patterns in data, many novel observations can only be made by visual examination of a network based on expert knowledge, a task which is beyond the current capabilities of computers. Clustering a PFIN allows it to be broken down into several, visually comprehensible parts [527].

Unlike functional prediction, network clustering is generally a non-quantitative technique. Comparison of different clustering patterns is non-trivial in many cases, due to the complexity of biological data. However, if process-relevant networks are integrated to study a specific question, this complexity is reduced by restricting cluster assessment to one area of biology. Therefore, quantitative assessment of the clusters is possible using GO annotations associated with the POI. In the analyses carried out during this project, it is clear both quantitatively and visually that the clustering of process-relevance networks is improved over that of the control since the ageing-related nodes were clustered together in large clusters in the relevance network.

Potentially the most interesting genes in an integrated network are the unknowns: genes with no annotations. In the control network the majority of the unknowns cluster together in small, poorly-annotated groups, a scenario which gives little scope for hypothesis generation. This clustering is likely to be due to bias of the control network edge weightings towards highly-studied processes. Therefore, highly-weighted edges are likely to be between annotated nodes rather than between annotated nodes and unknowns.

In the relevance networks this bias is overcome by the up-weighting of edges that are relevant to the POI. Nodes annotated to the POI cluster together, allowing easy visualisation of the most relevant

data. While many unknown genes still cluster in small, poorly annotated clusters in the relevance networks, several of the clusters containing genes annotated to the **POI** also contain a large proportion of unknowns. Additionally, the clusters contain a large number of nodes from ageing-related datasets [16, 993] which were produced after the networks were integrated. Since the **MCL** algorithm utilises edge weights [534], the difference in clustering of unknown genes is directly attributable to the up-weighting of the ageing-relevant data. The unknowns clustering with genes annotated to the **POI** in the relevance networks are, therefore, potential candidates for involvement in the ageing process.

These functional prediction and clustering differences demonstrate the power of the RelCID algorithm. Incorporation of a measure of relevance during network integration improves the functional predictive power of the networks. The up-weighting of relevant data also allows relevant nodes to cluster together and overcomes the biases of the control network towards high-confidence but potentially low relevance data. Further, the large clusters produced by the relevance networks have greater scope for hypothesis generation than those of the control when focussing on a specific biological process.

Ideally, functional predictions should be compared with new annotations to evaluate the networks' performance in light of new data. However, this is not always possible. High-quality biological databases, such as GO, are constantly changing [297]. As discussed in depth in Section 8.2.4, curators often identify and remove incorrect data and change their database schemas to reflect current biological knowledge. These changes may have a significant impact on integrated analyses. For instance, the removal of 76.45% of telomere maintenance annotations from the data used to generate the relevance networks in March 2008 made the comparison of these networks with those derived from March 2009 data impossible (see Section 4.19). Changes such as these are common in biological data. Therefore, careful selection of source data and analysis of results is crucial to accurate computational hypothesis generation in Systems Biology.

8.2.3 Choice of POI

Ageing is only a small aspect of cellular biology. Therefore, the integration and functional prediction evaluations carried out for the ageing terms were repeated using each **GOBP** in turn as **POI**. Clustering evaluations were not carried out, since clustering does not produce a numerical result that may be compared across multiple networks.

The performance changes of the networks were not consistent. Some relevance networks exhibited increased performance over the control, while network others had decreased performance. Several aspects of the **GOBP** terms, datasets and networks were assessed in relation to functional prediction performance and several results were apparent:

- The performance of **PFIN**s differs for different **GOBP** terms.
- The size and specificity of a **GOBP** term does not significantly influence **PFIN** performance in relation to that term.
- The performance of **GOBP** terms as **POI** was related to GO's **DAG** structure in that parent-child groups of terms exhibited similar performance.
- Network functional prediction performance is not correlated with any of the global network topological properties assessed, such as the average shortest path between the nodes annotated to the **POI**.
- Network performance was correlated neither with dataset relevance or confidence scores, nor with the dataset relevance or confidence ranks.
- The connectivity of genes annotated to a **POI** within individual datasets is related to network performance for that term.
- The Node Relevance score does not capture all aspects of dataset relevance to the **POI**.

The Node Relevance score is not sufficient to capture every aspect of a dataset's relevance to a **POI**. To have relevance to a **POI** a dataset must contain some nodes annotated to the that process, but the connectivity of these nodes can differ greatly. The Sanders.12052880 Affinity Capture-MS dataset [1144] is an excellent example of this effect. This dataset contains 193 nodes, 87 of which are annotated to the **POI**. Due to the high proportion of unannotated nodes, the dataset's Node Relevance score is not very high. However, >95% of this dataset's interactions involve a node annotated to the **POI**. Consequently, this dataset's relevance to this process is far higher than the numerical value produced by the Node Relevance calculation.

Two further relevance scores were therefore introduced, based on interactions involving the **POI**: Edge Relevance, to measure the level of interaction between two nodes each annotated to the **POI**; and Interaction Relevance, to measure the level of interaction between nodes annotated to the **POI** and their non-**POI** neighbours. The performance of the three network scores (Node Relevance, Edge Relevance and Interaction Relevance) differed for each **POI** in relation to known annotations.

In the majority of cases the Edge Relevance networks' performances were the highest. However, this result is largely influenced by the nature of the Edge Relevance score and the functional prediction algorithm chosen. The Edge Relevance algorithm up-weights datasets with a large number of edges between nodes annotated to the **POI**. During leave-one-out functional prediction evaluation, annotations are transferred between nodes annotated to the **POI**. Due to the local nature of the algorithm,

the annotation that is "left out" during evaluation must be attached along its highest-weighted edge to a node which is also annotated to the **POI** in order to be correctly assigned [57]. Therefore, it follows that networks integrated using Edge Relevance perform well. However, despite the bias towards edges between nodes annotated to the **POI**, datasets which score highly for Edge Relevance are also likely to contain other, unannotated, nodes which are relevant to the **POI**.

The Interaction Relevance networks performed the worst. Again, this result is largely due to the nature of the Interaction Relevance score and the functional prediction algorithm applied to the networks. The Interaction Relevance algorithm up-weights edges involving a single node annotated to the **POI**. Therefore, the majority of up-weighted edges have no influence on the functional prediction evaluation. However, this score is important since it includes the unannotated nodes. In fact, due to this aspect of the relevance scores, the Interaction Relevance algorithm produces far more novel hypotheses than the other scores when used to infer functional predictions. For instance, a large proportion of the edges of the Sanders.12052880 Affinity Capture-MS dataset [1144] are up-weighted in the Interaction Relevance network. Interaction Relevance is therefore, potentially very useful in the discovery of novel hypotheses.

The relationship between network functional prediction performance and the three dataset scores was highly variable and complex. In some cases, terms with many high-relevance datasets performed poorly. Conversely, some terms with few high relevance datasets performed well. Further, the relationship between the three dataset scores was also complex. While it was observed that a **POI** required some high relevance datasets to have improved performance, little correlation was observed between relevance scores and the corresponding network's performance. For instance, a **GOBP** term may have many high Interaction Relevance datasets and few high Node or Edge Relevance datasets but may only have improved performance when used as **POI** in the Edge Relevance network. Many of the aspects of dataset relevance to a **POI** are unclear. The most likely explanation for this ambiguity may be noise in the data. Therefore, as biological data slowly improves through curation, false interactions will be identified and removed (see Section 8.2.4), and these relationships should become clearer.

8.2.3.1 Combining Measures of Relevance

The three relevance scores (Node Relevance, Edge Relevance and Interaction Relevance) measure different aspects of a dataset's relevance to the **POI**. In combination, therefore, they could potentially combine all of these aspects into a single network. This combination is possible at different stages of integration and evaluation of the data. The individual network scores can be combined into a single network prior to functional prediction, or the separate networks may be used individually for

functional prediction and the results combined prior to calculation of the [AUC](#). Ideally, the construction of a single network combining all aspects of relevance and confidence is preferable, since the functional prediction stage is the most computationally intensive step of the RelCID algorithm.

Since the networks are topologically identical, differing in edge weights alone, integration of the different relevance measures into a composite network is straightforward and can be accomplished in a number of ways. Here, two simple methods of integration were chosen: averaging over the edge weights from the different networks and calculating a weighted sum of the edge weights from the different networks (see Section [6.4.2](#)). Both types of composite networks perform well in comparison to the control network. However, neither method of score integration produces improved performance for all GO terms. In fact, the performance of the composite networks for a large number of terms is reduced, and many of the true positive predictions of the single relevance networks are actually missed. Additionally, the improvement in performance for the composite networks is often smaller than the improvements produced by some of the networks integrated with a single measure of relevance. However, there may be more complex strategies of edge weight combination which may produce improved performance over the two chosen here.

It appears that the different aspects of network relevance cannot be combined into a single network in these ways without loss of information. To preserve the unique strengths of the different relevance scores, the networks must be created separately, and the functional prediction results combined (see Section [6.4.2](#)). Combination of the results in this way is relatively straightforward, as the networks are used to predict annotations to a single process. Therefore, a node is predicted either to be involved in a process or not by each network. Since the predictions each have scores based on the edge weights of their evidence, the highest score for a gene can be considered the highest confidence prediction.

Combination of the functional prediction results produces improved performance for all of the [POIs](#). As with the single relevance networks, Interaction Relevance performed the poorest and Edge Relevance the best when in combination with the other networks. Again, as discussed above, this reduced performance is due to the nature of the Interaction Relevance score. The information in these networks can be considered more valuable than that of Edge Relevance networks since they take unknowns into account.

The difference in performance between composite networks and combined functional prediction is not surprising. While it is computationally preferable to have a single network combining all aspects of relevance, such a network is not optimal for performance. The relevance scores measure distinct aspects of the datasets' contents, with little dependency between the scores. Therefore an edge may have a very different weighting in each network. If edge weights are combined into a single network, for instance by averaging, the high weighted edges of some networks are potentially down-weighted,

leading to incorrect functional assignment. Conversely, if functional prediction is carried out on the networks individually, the high weighted edges of each network may be utilised, preserving the information harnessed by the individual scores and improving functional assignment.

The best performance produced by the relevance networks is seen when the functional prediction results of all three networks are combined. Network performance is even further improved by the inclusion of the results of the control network. Inclusion of the control network in the combined prediction results may initially appear counter-productive, since relevance network integration is designed specifically to out-perform the control in relation to specific processes. However, due to the nature of functional interaction data, the control networks inclusion is essential to optimise performance. For instance, a dataset may contain 1000 nodes and 3000 edges, with two nodes, annotated to the [POI](#), that interact only with one another. This dataset would score poorly in all three aspects of relevance since it only has two nodes and one interaction involving the [POI](#). Therefore, the edge of relevance since it only has two nodes and one interaction involving the [POI](#). Therefore, the edge between these nodes would be down-weighted in the relevance networks. However, if the dataset is of very high confidence, this edge would be up-weighted in the control network. Further, if the edge was only found in this dataset then it would follow that the control network is likely to be the only network to correctly assign these annotations. Inclusion of the control network in the combined results is therefore essential to account for high confidence data which is not present in high relevance datasets.

8.2.4 Source data

The performance and accuracy of [PFINs](#) are dependent on the quality of the Gold Standards chosen [699]. Currently, Gold Standard datasets are derived from high-quality, manually-curated databases such as KEGG [49, 112, 128, 702] and GO [46, 98, 106, 703]. However, there are drawbacks to the use of Gold Standard data. Manually-curated databases tend to be biased towards intensively studied proteins and processes [98, 133, 223]. Therefore, Gold Standard datasets are non-saturating, in that they do not cover all areas of cellular biology. This situation which may lead to an incorrect assessment of dataset confidence. In this study several of the medium sized datasets were discarded since they did not score against the Gold Standard (see Section 4.3.1). It was, therefore, impossible to judge the accuracy of these datasets, since their lack of score may be due to bias in the Gold Standard, dataset inaccuracy, or most likely a combination of the two. Similarly, the datasets which do have positive scores may also be affected by Gold Standard bias to varying extents.

An additional drawback to the use of Gold Standard data are that it is highly likely that some of the data may be derived from the experimental datasets being assessed. In other words, some KEGG PATHWAY annotations may be manually curated from the experimental datasets being scored and

integrated. Unfortunately, in the case of KEGG, references to the original source of annotations are unavailable [1160]. Further, if literature references are available, filtering out annotations directly linked to the source datasets may significantly reduce the size of the Gold Standard. If this is the case accurate assessment of the datasets is impossible due to lack of Gold Standard data. Certainly, given the number of datasets in this study which did not score against the complete KEGG Gold Standard, it is likely that even more datasets would not score if the the Gold Standard were to be filtered in this way, resulting in considerable loss of data.

Despite the obvious drawbacks to the use of Gold Standard data, to date no other reliable method of assessment of dataset quality is available. KEGG was chosen as the Gold Standard in this work, as it is a manually-curated database [277], and also since GO was chosen as the Gold Standard for relevance scoring and evaluation [100]. As the confidence score and relevance scores were based on separate sources, the two scores can be considered independent from one another. It should be noted, however, that there is almost certainly some overlap of the evidence in the literature for annotations between the two databases. Unfortunately this type of overlap is unavoidable given that these biological databases are based on the manual curation of all of the available literature. However, since the KEGG PATHWAYS and GOBP annotations represent different areas of biology, metabolic pathways and biological processes respectively, this overlap is as reduced as possible.

PFINs are more powerful than unweighted networks, since they include a measure of dataset confidence, derived from comparison with Gold Standard data [49, 115, 128]. However, the limitations of Gold Standard databases impose limitations on scoring methods. In fact, despite the development of numerous analysis methods, the size of interactomes and the levels of noise in HTP data remain unknown, with estimates ranging widely even in a well-characterised organism such as *S. cerevisiae* [42, 154, 184, 313–317]. The only accurate measure of a dataset's confidence may be computed by comparison with the interactome itself; a true "Gold Standard" that will not be available for many years, if ever.

However, a rational assumption is that despite these drawbacks network performance is improving over time as more data are gathered and more knowledge of the interactome is gained [1161]. A major factor in this assumption is the effect of data curation. Database design and curation are becoming increasingly important in the field of biology and the number of papers published on these subjects is increasing rapidly. For instance, the journal Database was launched in 2009 as a dedicated forum for biological databases and curation strategies [271].

Highly curated databases are considered to be of high quality, as they are constantly manually updated in an attempt to accurately reflect current knowledge from the literature [266]. The curators of databases such as GO [100], KEGG [99] and SGD [283] add, remove and modify data on a daily

basis to try to keep the databases up-to-date. It therefore follows that integrated networks derived from the most recent database versions should outperform those integrated using previous versions.

Surprisingly this study found that this was not the case. While the newest data generally outperformed older data, network performance does not steadily improve through time but fluctuates as the available data changes and grows (see Section 5.4). In the case of changes to Gold Standard data, fluctuations are very likely to be due to bias in the Gold Standard data, rather than to inaccuracy. As the bias of a Gold Standard changes due to the addition and removal of data, dependent datasets' scores also change, causing a change in network performance. Datasets with high relevance to a [POI](#) may score poorly if the Gold Standard is biased against that process. Therefore, if a Gold Standard bias changes away from a [POI](#), performance will drop despite the high quality of the Gold Standard dataset.

The performance of process-specific networks may also be affected by changes to raw datasets. In this case the fluctuations in network performance are very likely to be attributable to dataset noise. In particular, large, noisy datasets with little relevance to a [POI](#) may mask the data of smaller, more relevant datasets.

Bias and noise remain significant problems in network analysis and are unlikely to be overcome easily, given the nature of biological research: some biological processes will always attract more interest than others [98, 223], and technologies, particularly the [HTP](#) ones, will continue to produce large, valuable, but noisy datasets [68, 69, 430, 673]. Computationally, these problems may be addressed by attempting to remove the noise in [HTP](#) data [690–692] and correct biases [342, 703, 707]. However, noise removal methods are computationally and theoretically non-trivial. Since the true interactome remains unknown it is impossible to determine with any accuracy what is noise or bias and what is true data. Therefore, it is likely that noise removal methods may introduce false negatives to the datasets. In other words, useful true positive data are discarded. The RelCID algorithm overcomes these problems by harnessing dataset bias in order to see past the noise in the data.

Integrated networks are designed to generate new hypotheses in order to build a clearer picture of the interactome. However, while Gold Standard data may be of high quality it is constantly changing. It is clear that the assumption of better data leading to improving network performance over time is incorrect. Rather, network performance changes dynamically, based on the biases in current datasets. As current knowledge becomes closer to the true interactome, fluctuations in network performance are likely to decrease.

Manual curation is important to avoid errors in data, but the curation process is far from perfect. Several significant curation decisions which directly affected network performance were identified

during this project. In BioGRID a large [HTP](#) dataset was removed following community feedback [1159]. The annotations of one KEGG pathway were removed and subsequently re-entered into the database between versions. The reasons for these pathway changes were not recorded [1160], but the removal of the data appears to have possibly been a simple human error. Finally, a large number of annotations were removed from the GO database following a decision by the curators regarding phenotypic data¹.

Each of these changes could directly affect network performance. Therefore, while it is important to choose manually curated databases as source data and Gold Standards for integrated analyses, it is important to remember that these datasets may still contain a level of noise and bias which must be taken into account during interpretation of the results. By focussing on individual processes the RelCID technique overcomes some of this noise and produces a smoother increase in network performance over time.

8.2.5 Novel Hypotheses

[PFIN](#)s are often utilised to infer new knowledge from biological data, for instance in the production of novel functional predictions prior to laboratory analysis. Evaluation of novel hypotheses is non-trivial. While known annotation data can be used to assess network performance by cross-validation, there is usually very little data with which to assess new functional predictions. The most accurate evaluation of new predictions is done using small-scale experimental analysis.

While the hierarchical structure of the GO [DAG](#) can cause problems for many applications, such as the generation of a Gold Standard (see Section 2.5.4.3), this structure provides a means to assess new predictions. Due to the transitivity of the [DAG](#) an annotation to a child term automatically implies annotation to the parents of this term [100]. Conversely, it is logical to assume that a gene annotated to a parent term is more likely to be involved in the child processes of this term than are genes annotated to other processes. Therefore, predictions can be evaluated to some extent by their consistency with known GO annotations by considering whether a prediction is to a child term of a known annotation for the gene. In total, 0.57% of the predictions produced by the RelCID schema are consistent with known annotation data. However, the other 99.43% cannot be considered incorrect, since a lack of supporting data does not in itself refute a novel hypothesis.

A second type computational evaluation is possible using the reviewed computational analysis ([RCA](#)) and inferred from electronic annotation ([IEA](#)) annotations which in this project were omitted from the original analysis. The lower-quality [IEA](#) annotations were excluded from RelCID because they

¹http://wiki.geneontology.org/index.php/SGD_GO-HTP_guidelines

are computational annotations which are not manually curated. [RCA](#) annotations are produced by integrated computational analyses similar to the RelCID method. Naturally, the [RCA](#) annotations were also excluded from the RelCID schema since using [RCA](#) annotations to produce [RCA](#) annotations is counter-intuitive. The [RCA](#) and [IEA](#) annotations therefore provide a means for further evaluation of new functional predictions, since consistency with these annotations would indicate increased confidence in the prediction. In total 26.23% of the new predictions are consistent with these annotation types.

The scores for the novel predictions ranged from 26.11711 to 0.00118 (see Section 7.2). Interestingly, while 46% of the predictions score below 1.0, the majority of the consistent predictions had higher scores, with 68.5% scoring above 1.0. These higher scores indicate that the predictions are of higher confidence, since they are based on higher confidence evidence. However, the higher scores may also be due to the levels of evidence involving the [POI](#), as discussed above. Nevertheless, predictions scoring above 1.0 can be considered higher quality, and are consistent with known annotations. Therefore, this score cut-off was chosen during novel hypothesis filtering.

While the control network produced more predictions, the relevance network predictions are potentially of more interest since they are based on up-weighting of relevant data. The selection of a prediction for experimental validation was, therefore, limited to those made by the relevance networks. The Interaction Relevance network produces the most predictions (93912) and the Edge Relevance the least (24310). This result is consistent with the observations made during the leave-one-out evaluation; the Interaction Relevance network up-weights interactions involving only one node annotated to the [POI](#) and, therefore, has greater scope for annotation transfer and novel hypothesis generation than the Edge Relevance network.

Many nodes in the network had functional predictions to several GO terms. However, these multiple terms arose because the [DAG](#) structure produces overlap between gene annotations. When a group of genes is annotated to the same group of terms, the datasets' relevance scores are similar, leading to similar functional predictions when these terms are used as [POIs](#). Due to this effect the GO structure seems to be closely related to network performance. The performance of the [POIs](#) follows the [DAG](#) structure during evaluation, with directly linked terms having similar performance. Additionally, the parent-child relationships of the [DAG](#) influence the predictions when novel hypotheses are produced. The relationship between [DAG](#) and performance is clearly complex and requires further investigation. It may be possible in future to select the optimum [POI](#) from a group of terms based on these relationships. However, since there was no correlation between GO term specificity and network performance, the development of a new numerical measure may be required to harness this [DAG](#)-based effect.

8.2.5.1 Laboratory Evaluation

A single functional prediction arising from the relevance networks was chosen for laboratory evaluation during this project. The ageing-related **GOBP** term response to oxidative stress provided an ideal **POI** for the laboratory evaluation of predictions for a number of reasons (see Section 7.4). The RelCID networks for *S. cerevisiae* produced 368 functional predictions for response to oxidative stress.

The predictions were filtered in several ways to select the highest confidence and potentially most interesting hypothesis for laboratory evaluation. First, the predictions produced by the control network were discarded, to focus on the predictions that would not be produced in the absence of relevance scoring. Second, predictions to unknown genes were selected, since they represent potentially more interesting hypotheses than predictions to genes with a known function. Next, the prediction scores were used to select the highest-confidence prediction. Finally, since there were two high-scoring predictions, the prediction with the highest level of evidence supporting its associated network edge(s) was selected. The chosen prediction was for the gene *AIM1*. A great deal of evidence supports the involvement of this protein in the ageing process:

- The protein Aim1 has been linked to mitochondrial genome maintenance [1166].
- The evidence for the prediction was transferred to Aim1 from the oxidative stress response proteins Grx3 and Grx4, based on an experimental study by Yu and colleagues [134].
- Grx3 and Grx4 are involved in the glutathione-glutaredoxin system and iron ion homeostasis [1060, 1061].
- The extended neighbourhood of Aim1 in the process-relevant **PFIN** contains several other mitochondrial and oxidative response stress genes, in addition to several genes involved in associated processes such as glutathione biosynthesis, telomere maintenance and the iron response (see Figure 8.1 and Table 7.2).
- The Aim1 neighbourhood also contains several other stress response genes and genes with ageing-related mutant phenotypes (see Table 7.3).
- The Aim1 protein contains a BolA-like domain, a member of a family of domains associated with the general stress response [1167, 1168].
- Two other proteins in the Aim1 neighbourhood, Fra2 and YAL044W-A, contain BolA domains and interact with Grx3 and Grx4 [21, 24, 134, 190, 1065, 1114, 1170].

- Fra2 is involved in the iron regulon, and has also been linked to mitochondrial genome maintenance [1065, 1114, 1166].

Given this evidence, the prediction of the response to oxidative stress as a process in which Aim1 is involved appears plausible. Additionally, the prediction can be extended to include an association with iron homeostasis. The prediction was assessed by spot testing for growth levels under a range of oxidative stress-inducing conditions and varying levels of iron (see Section 3.2.4).

The initial experimental results did not produce any clear link between Aim1 and the oxidative stress or iron responses during stress testing. However due to the time-scale of this project, these experiments only analysed the single Aim1 deletion mutant. In many cases phenotypic changes are not seen in single disruption mutants. In fact, due to the redundancy inherent in biological pathways [250], the disruption of two genes is often required to produce a mutant phenotype [1057]. Therefore, a double mutant may be required to observe an oxidative stress-induced phenotype involving the disruption of *AIM1*. Since two other proteins in the Aim1 neighbourhood contain BolA-like domains

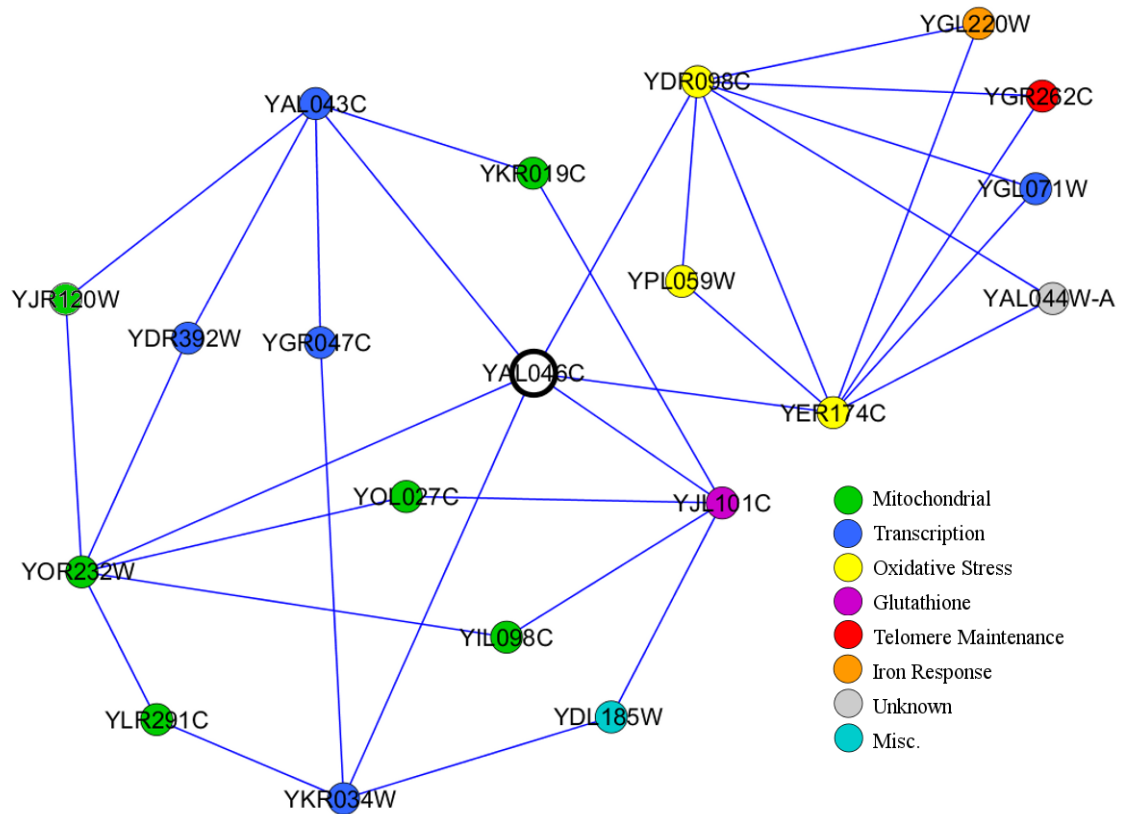


Figure 8.1: The extended neighbourhood of Aim1.

Nodes are coloured by general Gene Ontology biological process groups. It should be noted that there is overlap between the groups displayed. Most notably, the three oxidative stress response genes, YDR098C, YPL059W and YER174C, are also involved in the iron response, and, the iron response gene YGL220W is also involved in transcription.

[1167, 1168] and interact with the glutaredoxins, Grx3 and Grx4, they potentially have roles parallel to that of Aim1.

In addition, spot tests of mutant growth are the simplest method by which phenotypic changes may be identified, and many genes annotated with a particular biological process do not have an associated phenotype. For instance, while the proteins Grx3 and Grx4 are both annotated to the response to oxidative stress, neither gene has a related mutant phenotype [1057]. Consequently, the single *aim1Δ* mutant strain and several double mutant strains are currently being investigated using more sophisticated experimental techniques as a follow-up project.

8.3 Contribution

The RelCID technique provides an integration method which exploits dataset bias in order to tailor PFINs to the investigation of specific biological questions. Unlike previous process-relevant approaches [41, 46, 59, 129, 130, 307, 708, 720, 947], no data are lost, and unannotated areas of the network are not neglected, allowing greater scope for hypothesis generation. RelCID is also extremely flexible and can be applied to all types of biological data.

The final, combined functional prediction results harness all aspects of a dataset's confidence and its relevance to a POI. By combining three forms of relevance-weighted network with a high-confidence control network, functional prediction performance is improved in comparison to the control network for all of the GOBP terms tested. Therefore, the extended RelCID integration and functional prediction schema produces improved performance in relation to a specific biological process.

The RelCID approach also produces novel protein functional predictions which are not produced in the absence of a measure of relevance and have, therefore, not be produced by previous PFIN integration techniques. Many of these predictions are for unannotated, and therefore potentially interesting, genes. Further, many of the predictions are consistent with known biological data. Consequently, RelCID provides a valuable tool for the network-based analysis of specific biological questions and provides considerable scope for the generation of novel hypotheses.

The field of Systems Biology aims to study biology in terms of whole systems in an iterative fashion, with data analyses guiding experimental design, and experimental results, in turn, forming the basis for further analyses and mathematical modelling [26, 34–36]. However, a truly iterative approach to Systems Biology is often difficult. To date, very few network studies have demonstrated feedback between computational and experimental analyses in the iterative fashion suggested. In many cases networks are integrated and evaluated across all biological processes [49, 92, 103, 105–112, 674,

711]. Therefore, while a wealth of novel predictions are often made, the results are generally not applicable to a single biological question.

A process-targeted approach such as the RelCID technique improves the scope for systems-based analyses by overcoming to some extent dataset noise. Networks can be targeted to specific areas of biology and used to guide experimental studies. The data from these studies may, in turn, be used to refine the POI in an iterative fashion.

Importantly, RelCID significantly reduces analysis time compared to traditional methods. In this study the RelCID method produced a high confidence prediction for the GO term *response to oxidative stress* in less than a day. While a large amount of evidence supports the prediction, it could not be produced by a control network integrated without a measure of relevance. In the absence of a network-based approach, such as RelCID, any gene with an interaction involving a POI is a candidate for annotation to that term. Given the large size of most interaction datasets, identifying candidate genes for any given process is non-trivial. There were 51 unannotated candidate genes for the term *response to oxidative stress*. However, for other POIs there may be considerably more candidates. Given that each individual prediction may take 14 hours to assess manually (based on the time taken to assess the Aim1 prediction), assessment of all candidate genes may take considerable time, up to weeks. The RelCID approach drastically reduces this analysis time by providing a numerical score of prediction accuracy. Further, the process-centred approach of RelCID produces more novel predictions to the POI by up-weighting the most relevant data in the network.

8.4 Future Work

During analysis of the results of this project several aspects of the relevance network integration algorithm which may be further investigated in order to optimise performance were identified:

- **Investigation of different Gold Standards.**

KEGG was chosen as the Gold Standard during this study. However, several other manually-curated datasets exist which may be used as Gold Standard datasets. Given the biases inherent in many databases, some Gold Standards may have more relevance to specific biological questions than others. Therefore, the RelCID technique may potentially be extended by applying a measure of relevance to the Gold Standard data in order to further optimise performance.

- **Subnetwork generation.**

As the amount of biological data continues to grow network size and complexity may make many manual and visual network-based analyses impossible. In many cases, particularly where

there is a very specific area of interest, it may therefore become advantageous to produce a process-specific subnetwork prior to analysis. One method of subnetwork generation was demonstrated in Chapter 5. The subnetworks were produced by the integration of datasets scoring above a specified relevance cut-off. Low relevance datasets were discarded, reducing network complexity and improving network functional prediction performance in relation to the POI.

Using the RelCID method, a process-specific subnetwork could be extracted from the whole network using the edge weights, either by selection of high-weighted edges or by selection of up-weighted edges following comparison with a control network. Since the RelCID algorithm upweights edges in the network without regard to the annotation status of the genes involved, the subnetworks will contain both annotated and unannotated genes, unlike most previous subnetwork generation methods.

These two methods of subnetwork generation are a trade-off between dataset confidence and dataset relevance. The high-weighted subnetwork will contain edges with high confidence, some of which have low relevance. Conversely, the up-weighted network will contain edges of high relevance, some of which have low confidence. Potentially, a combination of the two subnetworks may optimise both relevance and confidence.

- **The choice of functional prediction algorithm.**

In this project, the local GBA algorithm Maximum Weight was chosen for network evaluation, since it provides a simple and direct comparison of network performance. More sophisticated algorithms may produce different, and potentially more valuable, results. Protein function prediction is a complex field and subject of ongoing research [1172–1181].

- **Laboratory evaluation.**

While the initial experiential results did not show any link between Aim1 and the oxidative stress or iron responses, the prediction remains plausible with a large amount of supporting evidence. Many genes associated with a particular biological process do not have an equivalent phenotype when disrupted, such as oxidative stress response genes *GRX3* and *GRX4* [1057]. Often two genes must be disrupted together for a phenotypic effect to be seen. Consequently, further investigation of *AIM1* using more sophisticated experimental techniques is currently ongoing.

In addition, several other ageing-related functional predictions produced by RelCID are also being investigated as a follow up project.

Given the versatility of the RelCID method, it is applicable to the construction of integrated networks in virtually any area of biology for which there is sufficient Gold Standard data available. In particular, two areas of biology may be particularly valuable for future investigation via the RelCID technique:

- **Subcellular Location.**

The Cellular Component branch of GO remains the poorest annotated. While many experimental methods exist to tag proteins in order to locate them in the cell, these methods tend to be inaccurate due to the fluid nature of the cell. Put simply, proteins may be found where they do not act. In particular, all proteins will, at some point, be located in the cytoplasm where they are produced. Currently², the majority of yeast proteins are annotated to either the nucleus (2015 genes, 34%), cytoplasm (2116 genes, 36%), or in some cases to both (918 genes, 16%). In order to reconstruct the interactome, more specific subcellular locations must be determined for all yeast proteins.

Given the limitations of experimental determination of protein cellular location, computational methods provide powerful tools to aid our understanding of this aspect of the interactome. Despite the lack of specific **GOCC** annotations, many **GOBP** and **GOMF** terms may be considered to be cellular component specific. For instance the **GOBP** term *telomere maintenance* (GO:0000723) and the **GOMF** term *telomeric DNA binding* (GO:0042162) may be considered indicative of the **GOCC** annotation *telomeric region* (GO:0000781) respectively. Therefore, annotations to these terms allow the **POI** to be extended beyond **GOCC** annotations alone. Consequently, RelCID could be used to construct cellular component-relevant networks allowing the prediction of more specific subcellular locations.

- **Human Disease.**

Identification of the genes associated with human disease is essential. A large number of genes has already been associated with human diseases, and these data are stored in the OMIM³ database. Candidate disease genes can be predicted using **GBA**-based approaches as used for functional prediction, by propagation of annotations along the edges of **PFIN**s. A disease can therefore be considered as a **POI** and the disease-related genes used during relevance scoring to construct a disease-relevant network. Given that *H. sapiens* functional interaction data are of a far greater size and much noisier than that of *S. cerevisiae*, the use of RelCID to produce disease-relevant networks could potentially provide a powerful tool in the understanding of human disease.

²<http://www.yeastgenome.org/>, accessed 2nd July 2011

³<http://www.ncbi.nlm.nih.gov/omim>

8.5 Concluding Remarks

PFINs have been demonstrated to be powerful tools for analysing genome-scale data. However, assessing dataset quality alone ignores dataset content. Data from different sources contain differing levels of bias and noise. While it is possible to attempt to remove bias and noise during network integration, this results in loss of data.

Given the amount of functional data currently being produced and the levels of noise in these data, computational methods, such as network-based analyses, are required to analyse them and generate new hypotheses. However as the amount of available biological data continues to grow, global network analyses become non-trivial. Therefore, process-specific approaches which utilise all the available data will become increasingly important.

While dataset bias can be problematic when analysing cellular biology on a global scale, it can be valuable when investigating specific areas of biology. Network integration may exploit dataset bias in order to tailor PFINs to answer specific questions. By assessing each dataset as a whole, no data are lost and unannotated areas of the network may be treated in the same way as well annotated areas, allowing greater scope for hypothesis generation.

RelCID allows research groups to tailor network analyses to their specific interests without loss of data. Therefore, the networks' performance is increased and more relevant hypotheses may be produced to guide experimental studies. Further, the algorithm is extremely flexible and can be applied to any area of biology using a variety of Gold Standard data. Therefore, RelCID provides a powerful tool to aid in this major aim of Systems Biology: the elucidation of the interactome.

Abbreviations

ATP adenosine triphosphate	82
AUC area-under-the-curve	76
BRET bioluminescence resonance energy transfer	12
CBD calmodulin binding domain	13
COI cluster of interest	128
CP clique percolation	38
DAG directed acyclic graph	29
dSLAM diploid-based synthetic lethality analysis on microarrays	18
E-MAP epistatic miniarray profile	18
FLN functional linkage network	55
FRET fluorescence resonance energy transfer	12
GBA guilt-by-association	74
GFP green fluorescent protein	64
GI genetic interaction	18
GN Girvan-Newman	38

GOBP Gene Ontology biological_process	93
GOCC Gene Ontology cellular_compartment.....	126
GOMF Gene Ontology molecular_function.....	214
GRX glutaredoxin	84
GSH oxidised glutathione.....	83
GSSG reduced glutathione	83
HTP high-throughput.....	1
i2H <i>in silico</i> two hybrid.....	27
IEA inferred from electronic annotation	92
IEP inferred from expression pattern	71
ISS inferred from sequence or structural similarity.....	71
LLS log-likelihood score	161
LTP low-throughput.....	11
MAPPIT Mammalian protein-protein interaction trap	12
MCL Markov clustering algorithm.....	38
MCODE molecular complex detection	38
MS mass spectrometry.....	13
MRF Markov random field.....	5
NADPH nicotinamide adenine dinucleotide phosphate.....	84
NAS non-traceable author statement.....	71

NCBI National Center for Biotechnology Information	16
NP-hard non-deterministic polynomial-time hard	37
NP-complete non-deterministic polynomial-time complete	37
ORF open reading frame	11
PCA protein-fragment complementation assay	12
PCR polymerase chain reaction	99
POI process of interest	93
PFIN Probabilistic Functional Integrated Network	5
PMID PubMed ID	90
PPI protein-protein interaction	3
RF random forest	55
RCA reviewed computational analysis	92
ROC receiver operator characteristic	76
ROS reactive oxygen species	79
RNSC restricted neighbourhood search clustering	38
RRW repeated random walks	42
SGA synthetic genetic array	3
SOD superoxide dismutase	83
SPC super paramagnetic clustering	38
SVM support vector machine	5

TAP tandem affinity purification	9
TAP-MS tandem affinity purification mass spectrometry	13
TEV tobacco etch virus	13
tBOOH tetra-butyl hydroperoxide	85
TRX thioredoxin	84
X-gal bromo-chloro-indolyl-galactopyranoside	9
Y2H yeast two hybrid	9

Appendix A

Graph Theoretic Definitions

Graph theoretic definitions based on those of Cormen and colleagues (2003) [1182]:

- **Undirected Graph**

An undirected graph is a graph, $G = (N, E)$, where N is the set of nodes and $E \subseteq \binom{N}{2}$ is the set of edges.

- **Directed Graph**

A directed graph is a graph, $G = (N, E)$, where N is the set of nodes and $E \subseteq N \times N$ is the set of edges.

- **Path**

A path of length n is a sequence of nodes $v_0, v_1, v_2, \dots, v_n$, where $(v_i, v_{i+1}) \in E$ for $0 \leq i < n$.

- **Cycle**

A cycle is a path, $v_0, v_1, v_2, \dots, v_n$, of length n , where $v_0 = v_n$.

- **Directed Acyclic Graph (DAG)**

A DAG is a directed graph $G = (N, E)$ in which there are no cycles.

- **Bipartite Graph**

A bipartite graph is a graph, $G = (N, E)$, where N consists of two groups, N_1 and N_2 , and $u, v \in E \Rightarrow u \in N_1 \wedge v \in N_2 \vee u \in N_2 \wedge v \in N_1$.

- **Weighted Graph**

A weighted graph is a graph, $G = (N, E)$, where each edge e has an associated weight $w(e)$ given by the function $w : E \rightarrow \mathbb{R}$.

Appendix B

Gene Ontology Evidence Types

Descriptions of the Gene Ontology evidence types as provided by the GO Consortium¹:

B.1 Experimental Evidence Codes

EXP: Inferred from Experiment This code is used in an annotation to indicate that an experimental assay has been located in the cited reference, whose results indicate a gene product's function, process involvement, or subcellular location (indicated by the GO term). The EXP code is the parent code for the IDA, IPI, IMP, IGI and IEP experimental codes.

IDA: Inferred from Direct Assay The IDA evidence code is used to indicate a direct assay was carried out to determine the function, process, or component indicated by the GO term. For instance, enzyme assays, *in vitro* reconstitution (e.g. transcription), immunofluorescence (for cellular component), cell fractionation (for cellular component) and physical interaction/binding assay.

IPI: Inferred from Physical Interaction Covers physical interactions between the gene product of interest and another molecule (such as a protein, ion or complex). IPI can be thought of as a type of IDA, where the actual binding partner or target can be specified, using "with" in the with/from field. For example 2-hybrid interactions, co-purification, co-immunoprecipitation and ion/protein binding experiments.

IMP: Inferred from Mutant Phenotype The IMP evidence code covers those cases when the function, process or cellular localization of a gene product is inferred based on differences in the function, process, or cellular localization between two different alleles of the corresponding gene. The IMP code is used for cases where one allele may be designated 'wild-type' and another as 'mutant'. It is also used in cases where allelic variation occurs naturally and no specific allele is designated as wild-type or mutant. For example:

- Mutations, natural or introduced, that result in partial or complete impairment or alteration of the function of that gene.
- Polymorphism or allelic variation (including where no allele is designated wild-type or mutant).
- Any procedure that disturbs the expression or function of the gene, including RNAi, anti-sense RNAs, antibody depletion, or the use of any molecule or experimental condition that may disturb or affect the normal functioning of the gene, including: inhibitors, blockers, modifiers, any type of antagonists, temperature jumps, changes in pH or ionic strength.
- Overexpression or ectopic expression of wild-type or mutant gene that results in aberrant behavior of the system or aberrant expression where the resulting mutant phenotype is used to make a judgment about the normal activity of that gene product.

¹<http://www.geneontology.org/GO.evidence.shtml>

IGI: Inferred from Genetic Interaction Includes any combination of alterations in the sequence (mutation) or expression of more than one gene/gene product. This code can therefore cover any of the IMP experiments that are done in a non-wild-type background. If there is a single mutation or difference between the two strains compared, use IMP. If there are multiple mutations or differences between the two strains compared, use IGI. When redundant copies of a gene must all be mutated to see an informative phenotype, use IGI. Examples include: "traditional" genetic interactions such as suppressors and synthetic lethals, functional complementation, rescue experiments and inference about one gene drawn from the phenotype of a mutation in a different gene.

IEP: Inferred from Expression Pattern The IEP evidence code covers cases where the annotation is inferred from the timing or location of expression of a gene, particularly when comparing a gene that is not yet characterized with the timing or location of expression of genes known to be involved in a particular process. For instance, transcript levels or timing (e.g. Northern, microarray data) and protein levels (e.g. Western blots).

B.2 Author Statement Evidence Codes

TAS: Traceable Author Statement Any statement in an article where the original evidence (experimental results, sequence comparison, etc.) is not directly shown, but is referenced in the article and therefore can be traced to another source. The TAS evidence code covers author statements that are attributed to a cited source. Typically this type of information comes from review articles. Material from the introductions and discussion sections of non-review papers may also be suitable if another reference is cited as the source of experimental work or analysis.

NAS: Non-traceable Author Statement The NAS evidence code should be used in all cases where the author makes a statement that a curator wants to capture but for which there are neither results presented nor a specific reference cited in the source used to make the annotation. The source of the information may be peer reviewed papers, textbooks, or database records. For some annotations using the NAS code, there will not be an entry in the with/from field. Examples include database entries that don't cite a paper (e.g. UniProt Knowledgebase records, YPD protein reports) and statements in papers (abstract, introduction, or discussion) that a curator cannot trace to another publication.

B.3 Computational Analysis Evidence Codes

ISS: Inferred from Sequence or Structural Similarity The ISS evidence code or one of its sub-categories should be used whenever a sequence-based analysis forms the basis for an annotation and review of the evidence and annotation has been done manually. If the annotation has not been reviewed manually, the correct evidence code is IEA, even if the evidence supporting the annotation is all sequence based. ISS should be used if a combination of sequence-based tools or methods are used. If only one particular type of sequence-based evidence is used then one of the more specific sub-categories of ISS may be more appropriate for the annotation.

ISO: Inferred from Sequence Orthology The ISO code is a sub-category of the ISS code. Orthology is a relationship between genes in different species indicating that the genes derive from a common ancestor. Orthology is established by multiple criteria generally including amino acid and/or nucleotide sequence comparisons and one or more of the following phylogenetic analysis, coincident expression, conserved map location, functional complementation, immunological cross-reaction, similarity in subcellular localization, subunit structure, substrate specificity or response to specific inhibitors.

ISA: Inferred from Sequence Alignment The ISA code is a sub-category of the ISS code. It should be used whenever a sequence alignment is the basis for making an annotation, but only when a curator has manually reviewed the alignment and choice of GO term or if the information is in a published paper, the authors have manually reviewed the evidence. Such alignments may be pairwise alignments (the alignment of two sequences to one another) or multiple alignments (the alignment of 3 or more sequences to one another). For example sequence similarity with experimentally characterized gene products, as determined by alignments, either pairwise or multiple (tools such as BLAST, ClustalW, MUSCLE). BLAST produces pairwise alignments and any annotations based solely on the evaluation of BLAST results should use this code. GO policy states that in order to assert that a query protein has the same function as a match protein, the match protein MUST be experimentally characterized.

ISM: Inferred from Sequence Model The ISM code is a sub-category of the ISS code. The ISM code should be used any time that evidence from some kind of statistical model of a sequence or group of sequences is used to make a prediction about the function of a protein or RNA. Generally, when searching sequences with these modeling tools, the results include statistical scores (such as e values and cutoff scores) that help curators decide when a result is significant enough to warrant making an annotation. If an annotator manually checks these scores and determines if the result makes sense in the context of other information known about the sequence and decides that the evidence warrants a particular annotation, then the evidence code is ISM. For instance, prediction methods for non-coding RNA genes such as tRNAscan-SE, Snoscan, and Rfam, predicted presence of recognized functional domains or membership in protein families, as determined by tools such as profile Hidden Markov Models (HMMs), including Pfam and TIGRFAM, predicted protein features using tools such as TMHMM (transmembrane regions), SignalP (signal peptides on secreted proteins), and TargetP (subcellular localization) and any other kind of domain modeling tool or collections of them such as SMART, PROSITE, PANTHER, InterPro, etc.

IGC: Inferred from Genomic Context - This evidence code can be used whenever information about the genomic context of a gene product forms part of the evidence for a particular annotation. Genomic context includes, but is not limited to, such things as identity of the genes neighboring the gene product in question (i.e. synteny), operon structure, and phylogenetic or other whole genome analysis.

RCA: Inferred from Reviewed Computational Analysis - The RCA code should be used for annotations made from predictions based on computational analyses of large-scale experimental data sets, or on computational analyses that integrate multiple types of data into the analysis. Acceptable experimental data types include protein-protein interaction data (e.g. two-hybrid results, mass spectroscopic identification of proteins identified by affinity tag purifications, etc.) synthetic genetic interactions, microarray expression results. Sequence-based data based on the sequence of the gene product, including structural predictions based on sequence, may be included provided that the analysis included non-sequence-based data as well. Sequence information related to promotor sequence features may also be included as a data type within these analyses. Predictions based on mathematical modelling which attempts to duplicate existing experimental results are also appropriate for use of this evidence code. Examples include predictions based on computational analyses of large-scale experimental data sets and predictions based on computational analyses that integrate datasets of several types, including experimental data (e.g. expression data, protein-protein interaction data, genetic interaction data, etc.), sequence data (e.g. promoter sequence, sequence-based structural predictions, etc.), or mathematical models.

B.4 Computationally-assigned Evidence Codes

IEA: Inferred from Electronic Annotation - Used for annotations that depend directly on computation or automated transfer of annotations from a database, particularly when the analysis is performed internally and not published. A key feature that distinguishes this evidence code from others is that it is not made by a curator; use IEA when no curator has checked the specific annotation to verify its accuracy. The actual method used (BLAST search, Swiss-Prot keyword mapping, etc.) doesn't matter. Examples include annotations based on "matches" in sequence similarity comparisons if they have not been reviewed by a curator, annotations transferred from database records, if not reviewed by a curator and annotations made on the basis of keyword mapping files, if not reviewed by a curator.

Appendix C

Relevance Network Production

The Java code to build the relevance and control networks is available on the attached disc in the folder appendixC.

The folder contains the src folder which, in turn, contains the following packages:

- annotations - SGD file parsers.
- bioGRID - BioGRID file parser.
- go - GO file parser.
- goldStandard/kegg - KEGG file parser and Gold Standard construction.
- llScore - Confidence scoring against the Gold Standard.
- relNet - Network integration.
- relScore - Relevance scoring.

The network build requires the following parameters:

- BioGRID file
- KEGG PATHWAYS file
- SGD annotation file
- GO OBO file
- D-value
- GO term of interest as POI

Networks are built using the following command:

```
java ProduceNetworks BioGRIDfile KEGGfile SGDfile OBOfile GOid Dvalue
```

Four network files are output:

1. Control network integrated without a measure of relevance
2. Node Relevance network
3. Edge Relevance network
4. Interaction Relevance network

File are named: VersionDvalue[POI]Type.txt

for example V52D1.1G00000723NodeRelevance.txt and V52D1.1Control.txt

Network files are in tabbed format: Gene Gene Score

for example YML064C YEL066W 3.4761283869785937

Appendix D

BioGRID Evidence Types

The official BioGRID descriptions of the experimental evidence codes¹:

D.1 Physical interactions

- **Affinity Capture-Luminescence** An interaction is inferred when a bait protein, tagged with luciferase, is enzymatically detected in immunoprecipitates of the prey protein as light emission. The prey protein is affinity captured from cell extracts by either polyclonal antibody or epitope tag.
- **Affinity Capture-MS** An interaction is inferred when a “bait” protein is affinity captured from cell extracts by either polyclonal antibody or epitope tag and the associated interaction partner is identified by mass spectrometric methods.
- **Affinity Capture-RNA** An interaction is inferred when a bait protein is affinity captured from cell extracts by either polyclonal antibody or epitope tag and associated RNA species identified by Northern blot, RT-PCR, affinity labeling, sequencing, or microarray analysis.
- **Affinity Capture-Western** An interaction is inferred when a Bait protein affinity captured from cell extracts by either polyclonal antibody or epitope tag and the associated interaction partner identified by Western blot with a specific polyclonal antibody or second epitope tag. This category is also used if an interacting protein is visualized directly by dye stain or radioactivity. Note that this differs from any co-purification experiment involving affinity capture in that the co-purification experiment involves at least one extra purification step to get rid of potential contaminating proteins.
- **Biochemical Activity** An interaction is inferred from the biochemical effect of one protein upon another, for example, GTP-GDP exchange activity or phosphorylation of a substrate by a kinase. The “bait” protein executes the activity on the substrate “hit” protein. A Modification value is recorded for interactions of this type with the possible values Phosphorylation, Ubiquitination, Sumoylation, Dephosphorylation, Methylation, Prenylation, Acetylation, Deubiquitination, Proteolytic Processing, Glucosylation, Nedd(Rub1)ylation, Deacetylation, No Modification, Demethylation.
- **Co-crystal Structure** Interaction directly demonstrated at the atomic level by X-ray crystallography. Also used for NMR or Electron Microscopy (EM) structures. If a structure is demonstrated between 3 or more proteins, one is chosen as the bait and binary interactions are recorded between that protein and the others.
- **Co-fractionation** Interaction inferred from the presence of two or more protein subunits in a partially purified protein preparation. If co-fractionation is demonstrated between 3 or more proteins, one is chosen as the bait and binary interactions are recorded between that protein and the others.

¹http://wiki.thebiogrid.org/doku.php/experimenta\T1\1_systems

- **Co-localization** An interaction is inferred from co-localization of two proteins in the cell, including co-dependent association of proteins with promoter DNA in chromatin immunoprecipitation experiments.
- **Co-purification** An interaction is inferred from the identification of two or more protein subunits in a purified protein complex, as obtained by classical biochemical fractionation or affinity purification and one or more additional fractionation steps.
- **Far Western** An interaction is detected between a protein immobilized on a membrane and a purified protein probe.
- **FRET** An interaction is inferred when close proximity of interaction partners is detected by fluorescence resonance energy transfer between pairs of fluorophore-labeled molecules, such as occurs between CFP (donor) and YFP (acceptor) fusion proteins.
- **PCA** A Protein-Fragment Complementation Assay (PCA) is a protein-protein interaction assay in which a bait protein is expressed as fusion to one of the either N- or C- terminal peptide fragments of a reporter protein and prey protein is expressed as fusion to the complementary N- or C- terminal fragment of the same reporter protein. Interaction of bait and prey proteins bring together complementary fragments, which can then fold into an active reporter, e.g. the split-ubiquitin assay.
- **Protein-peptide** An interaction is detected between a protein and a peptide derived from an interaction partner. This includes phage display experiments.
- **Protein-RNA** An interaction is detected between and protein and an RNA.
- **Reconstituted Complex** An interaction is detected between purified proteins in vitro.
- **Two-hybrid** Bait protein expressed as a DNA binding domain (DBD) fusion and prey expressed as a transcriptional activation domain (TAD) fusion and interaction measured by reporter gene activation.

D.2 Genetic Interactions

- **Dosage Growth Defect** A genetic interaction is inferred when over expression or increased dosage of one gene causes a growth defect in a strain that is mutated or deleted for another gene.
- **Dosage Lethality** A genetic interaction is inferred when over expression or increased dosage of one gene causes lethality in a strain that is mutated or deleted for another gene.
- **Dosage Rescue** A genetic interaction is inferred when over expression or increased dosage of one gene rescues the lethality or growth defect of a strain that is mutated or deleted for another gene.
- **Negative Genetic** Mutations/deletions in separate genes, each of which alone causes a minimal phenotype, but when combined in the same cell results in a more severe fitness defect or lethality under a given condition.
- **Phenotypic Enhancement** A genetic interaction is inferred when mutation or overexpression of one gene results in enhancement of any phenotype (other than lethality/growth defect) associated with mutation or over expression of another gene.
- **Phenotypic Suppression** A genetic interaction is inferred when mutation or over expression of one gene results in suppression of any phenotype (other than lethality/growth defect) associated with mutation or over expression of another gene.
- **Positive Genetic** Mutations/deletions in separate genes, each of which alone causes a minimal phenotype, but when combined in the same cell results in a less severe fitness defect than expected under a given condition.
- **Synthetic Growth Defect** A genetic interaction is inferred when mutations in separate genes, each of which alone causes a minimal phenotype, result in a significant growth defect under a given condition when combined in the same cell.
- **Synthetic Haploinsufficiency** A genetic interaction is inferred when mutations or deletions in separate genes, at least one of which is hemizygous, cause a minimal phenotype alone but result in lethality when combined in the same cell under a given condition.

- **Synthetic Lethality** A genetic interaction is inferred when mutations or deletions in separate genes, each of which alone causes a minimal phenotype, result in lethality when combined in the same cell under a given condition.
- **Synthetic Rescue** A genetic interaction is inferred when mutations or deletions of one gene rescues the lethality or growth defect of a strain mutated or deleted for another gene.

Appendix E

GO Term Enrichment

GOstats outputs for the datasets analysed in Section 4.1.3.1. The full R output for the datasets analysed in Sections 4.1.3.2-4.1.3.5 are supplied on the attached disk in the folder appendixE.

Table E.1: The top twenty enriched GO Biological Process terms for all the Genetic interactions (138 enriched terms in total).

GOBPID	Pvalue	OddsRatio	ExpCount	Count	Size	Term
GO:0065007	5.84E-72	6.090048	497.30535	704	783	biological regulation
GO:0051641	2.07E-50	5.816283	355.03664	503	559	cellular localization
GO:0046907	4.73E-46	5.650174	330.26664	467	520	intracellular transport
GO:0006996	2.98E-44	3.083049	632.58765	817	996	organelle organization and biogenesis
GO:0044238	1.96E-40	2.370941	996.51608	1208	1569	primary metabolic process
GO:0044260	3.16E-40	2.641645	757.07277	947	1192	cellular macromolecule metabolic process
GO:0051234	1.40E-38	2.91872	592.57457	760	933	establishment of localization
GO:0019538	5.69E-36	2.481291	758.34302	938	1194	protein metabolic process
GO:0031323	4.85E-34	4.518281	291.52383	403	459	regulation of cellular metabolic process
GO:0050896	2.17E-32	5.51017	232.45691	329	366	response to stimulus
GO:0009653	2.19E-29	8.78097	156.87666	231	247	anatomical structure morphogenesis
GO:0045449	4.34E-29	5.059919	222.29486	312	350	regulation of transcription
GO:0016070	1.64E-28	2.780615	455.38689	583	717	RNA metabolic process
GO:0050794	2.79E-25	12.906108	114.32307	172	180	regulation of cellular process
GO:0007001	1.16E-24	4.697779	198.79511	277	313	chromosome organization and biogenesis (sensu Eukaryota)
GO:0022402	7.16E-24	6.305788	152.43076	219	240	cell cycle process
GO:0016192	7.11E-23	8.85	120.03922	177	189	vesicle-mediated transport
GO:0048523	8.18E-22	6.302801	138.45794	199	218	negative regulation of cellular process
GO:0006259	4.14E-21	4.152711	189.26819	260	298	DNA metabolic process
GO:0043283	8.98E-21	3.168457	258.49716	341	407	biopolymer metabolic process
GO:0000723	1.39E-20	4.478048	170.84948	237	269	telomere maintenance

Table E.2: The top twenty enriched GO Biological Process terms for all the Physical interactions (50 enriched terms in total).

GOBPID	Pvalue	OddsRatio	ExpCount	Count	Size	Term
GO:0044260	2.95E-30	5.755886	1071.016	1164	1192	cellular macromolecule metabolic process
GO:0019538	1.34E-29	5.553653	1072.81301	1165	1194	protein metabolic process
GO:0044238	4.40E-25	3.453662	1409.75176	1507	1569	primary metabolic process
GO:0016070	4.55E-21	4.870293	824.82608	894	918	RNA metabolic process
GO:0051234	5.78E-19	4.049084	863.46172	931	961	establishment of localization
GO:0065007	6.93E-19	5.149184	703.52813	764	783	biological regulation
GO:0051641	7.24E-19	9.842371	502.26338	552	559	cellular localization
GO:0044237	3.55E-18	2.77375	1338.77	1420	1490	cellular metabolic process
GO:0006996	6.60E-17	4.774281	660.39997	716	735	organelle organization and biogenesis
GO:0022402	1.38E-13	9.546076	358.50284	394	399	cell cycle process
GO:0044249	2.34E-13	3.195953	755.64132	809	841	cellular biosynthetic process
GO:0006412	2.13E-12	8.842463	334.24325	367	372	translation
GO:0006259	1.10E-11	4.595069	451.94719	490	503	DNA metabolic process
GO:0050794	2.42E-11	4.950508	413.31154	449	460	regulation of cellular process
GO:0000278	7.53E-11	28.740361	219.23482	243	244	mitotic cell cycle
GO:0006325	1.44E-10	27.996992	213.8438	237	238	establishment and/or maintenance of chromatin architecture
GO:0006464	3.52E-10	3.888303	445.65766	481	496	protein modification
GO:0019222	2.53E-09	3.564092	438.46964	472	488	regulation of metabolic process
GO:0006351	2.83E-09	3.666876	423.19508	456	471	transcription, DNA-dependent
GO:0050896	1.03E-08	2.526431	640.63289	680	713	response to stimulus

Table E.3: The top twenty enriched GO Biological Process terms for the Genetic genes (22 enriched terms in total).

GOBPID	Pvalue	OddsRatio	ExpCount	Count	Size	Term
GO:0009987	2.50E-28	3.390005	549.0056769	661	4342	cellular process
GO:0044255	7.27E-13	3.250492	27.6905212	67	219	cellular lipid metabolic process
GO:0009058	3.69E-12	1.96516	115.8197144	183	916	biosynthetic process
GO:0044271	8.71E-12	4.399649	14.0349217	42	111	nitrogen compound biosynthetic process
GO:0009308	1.09E-10	2.919304	27.9434027	63	221	amine metabolic process
GO:0006766	3.18E-10	4.598552	11.0003441	34	87	vitamin metabolic process
GO:0006519	7.25E-10	2.922348	25.1617065	57	199	amino acid and derivative metabolic process
GO:0019752	7.39E-10	2.466605	38.817306	77	307	carboxylic acid metabolic process
GO:0008652	8.81E-10	4.025441	13.0233958	37	103	amino acid biosynthetic process
GO:0006811	1.89E-09	3.790611	13.908481	38	110	ion transport
GO:0015698	1.72E-08	21.053942	2.0230518	12	16	inorganic anion transport
GO:0042364	7.65E-08	5.890443	5.5633924	20	44	water-soluble vitamin biosynthetic process
GO:0051186	1.25E-07	2.735093	21.115603	46	167	cofactor metabolic process
GO:0006066	2.48E-07	2.723789	20.2305178	44	160	alcohol metabolic process
GO:0006733	4.03E-07	4.57751	7.0806812	22	56	oxidoreduction coenzyme metabolic process
GO:0006595	5.04E-07	Inf	0.8850852	7	7	polyamine metabolic process
GO:0046467	7.68E-07	3.851806	8.9772923	25	71	membrane lipid biosynthetic process
GO:0009117	7.70E-07	3.093905	13.7820403	33	109	nucleotide metabolic process
GO:0006769	1.02E-06	5.286611	5.3105109	18	42	nicotinamide metabolic process
GO:0046474	1.43E-06	5.925188	4.4254258	16	35	glycerophospholipid biosynthetic process

Table E.4: The enriched GO Biological Process terms for the Physical only genes (17 enriched terms in total).

GOBPID	Pvalue	OddsRatio	ExpCount	Count	Size	Term
GO:0006412	7.26E-20	3.159723	52.731464	119	372	translation
GO:0030490	1.72E-15	12.962713	6.095304	29	43	processing of 20S pre-rRNA
GO:0022613	3.99E-13	2.701273	45.50215	94	321	ribonucleoprotein complex biogenesis and assembly
GO:0019538	1.21E-12	1.840102	169.250989	248	1194	protein metabolic process
GO:0044249	2.17E-12	1.962804	119.212799	188	841	cellular biosynthetic process
GO:0008152	2.35E-12	1.704595	434.325821	526	3064	metabolic process
GO:0006365	1.02E-09	4.583565	11.056597	33	78	35S primary transcript processing
GO:0044260	1.03E-09	1.693187	168.967487	236	1192	cellular macromolecule metabolic process
GO:0009987	5.33E-09	1.711193	615.483915	680	4342	cellular process
GO:0006364	7.16E-08	4.699056	8.221572	25	58	rRNA processing
GO:0006625	3.45E-07	16.861931	2.126269	11	15	protein targeting to peroxisome
GO:0016070	4.47E-07	1.609983	130.127645	180	918	RNA metabolic process
GO:0042273	4.93E-07	4.076642	8.930329	25	63	ribosomal large subunit biogenesis and assembly
GO:0015992	5.85E-07	8.605679	3.40203	14	24	proton transport
GO:0015986	2.57E-06	9.201355	2.835025	12	20	ATP synthesis coupled proton transport
GO:0006753	2.57E-06	9.201355	2.835025	12	20	nucleoside phosphate metabolic process
GO:0046034	2.57E-06	9.201355	2.835025	12	20	ATP metabolic process

Appendix F

Dataset Integration Rankings

Table F.1: The ranked order of intergration for the datasets to produce the networks described in Section 4.3.1.

Dataset	Confidence Rank	Ageing Rank (A)	Telomere Rank (T)	Combined A & T Rank
Protein-peptide	1	39	17	37
Newman (PubMed:11087867)	2	40	45	41
Ingvarsdottir (PubMed:15657441)	3	19	37	24
Tong (PubMed:11743162)	4	42	36	40
FRET	5	45	44	45
Co-crystal Structure	6	30	21	27
Krogan (PubMed:14759368)	7	31	40	34
Collins (PubMed:17200106)	8	11	25	12
Co-localization	9	22	12	19
Co-purification	10	23	29	25
Co-fractionation	11	34	23	33
Affinity Capture-Western	12	8	7	7
Two-hybrid	13	15	9	13
Reconstituted Complex	14	17	6	15
Phenotypic Enhancement	15	10	1	6
Gavin (PubMed:11805826)	16	20	22	22
Far Western	17	37	38	39
Dosage Growth Defect	18	27	24	26
Biochemical Activity	19	24	13	21
Krogan (PubMed:16554755)	20	14	15	14
Drees (PubMed:11489916)	21	38	28	38
Dosage Lethality	22	25	10	20
Affinity Capture-MS	23	13	32	16
Synthetic Growth Defect	24	3	4	4
Wong (PubMed:17634282)	25	28	27	29
Dosage Rescue	26	12	5	11
Sanders (PubMed:12052880)	27	29	35	31
Phenotypic Suppression	28	4	3	2
Synthetic Rescue	29	9	2	5
Ito (PubMed:11283351)	30	36	26	36
Synthetic Lethality	31	2	8	3
Uetz (PubMed:10688190)	32	26	39	30
Ubersax (PubMed:14574415)	33	41	41	42

Continued on next page

Table F.1: The ranked order of intergration for the datasets to produce the networks described in Section 4.3.1.

Dataset	Confidence Rank	Ageing Rank (A)	Telomere Rank (T)	Combined A & T Rank
Gavin (PubMed:16429126)	34	16	34	18
Schuldiner (PubMed:16269340)	35	43	43	43
Daniel (PubMed:16157669)	36	21	33	23
Tong (PubMed:14764870)	37	5	16	8
Ho (PubMed:11805837)	38	32	11	28
Tong (PubMed:11743205)	39	18	14	17
Lesage (PubMed:15166135)	40	33	31	32
Ye (PubMed:16729061)	41	6	30	9
Collins (PubMed:17314980)	42	7	18	10
Miller (PubMed:16093310)	43	44	42	44
Ptacek (PubMed:16319894)	44	35	19	35
Pan (PubMed:16487579)	45	1	20	1

Appendix G

Dataset Versions

The dataset file versions used in the database curation study of Chapter 5.

Table G.1: File Versions used to produce and evaluate the combined PFINs depicted in 5.13 of the main text.

Version	BioGRID	KEGG	Gene Ontology
17	01/07/06	29/06/06	01/07/06
18	01/08/06	29/06/06	01/08/06
19	01/09/06	29/06/06	01/09/06
20	01/10/06	29/09/06	01/10/06
21	01/11/06	29/09/06	01/11/06
22	01/12/06	29/09/06	01/12/06
23	01/01/07	27/12/06	01/01/07
24	01/02/07	27/12/06	01/02/07
25	01/03/07	27/12/06	01/03/07
26	01/04/07	28/03/07	01/04/07
28	01/05/07	28/03/07	01/05/07
29	01/06/07	28/03/07	01/06/07
30	01/07/07	25/06/07	01/07/07
31	01/08/07	25/06/07	01/08/07
32	01/09/07	25/06/07	01/09/07
33	01/10/07	24/09/07	01/10/07
34	01/11/07	24/09/07	01/11/07
35	01/12/07	24/09/07	01/12/07
36	01/01/08	03/12/07	01/01/08
37	01/02/08	03/12/07	01/02/08
38	01/03/08	03/12/07	01/03/08
39	01/04/08	24/03/08	01/04/08
40	01/05/08	24/03/08	01/05/08
41	01/06/08	24/03/08	01/06/08
42	01/07/08	30/06/08	01/07/08
43	01/08/08	30/06/08	01/08/08
44	01/09/08	30/06/08	01/09/08
45	01/10/08	29/09/08	01/10/08
46	01/11/08	29/09/08	01/11/08
47	01/12/08	29/09/08	01/12/08
48	01/01/09	22/12/08	01/01/09
49	01/02/09	22/12/08	01/02/09
50	01/03/09	22/12/08	01/03/09
Continued on next page			

Table G.1: File Versions used to produce and evaluate the combined PFINs depicted in 5.13 of the main text.

Version	BioGRID	KEGG	Gene Ontology
51	01/04/09	30/03/09	01/04/09
52	01/05/09	30/03/09	01/05/09

Table G.2: File Versions used to produce and evaluate the historic control BioGRID changes in 5.14 of the main text. The static control files are highlighted in bold.

Version	BioGRID	KEGG	Gene Ontology
17	01/07/06	29/06/06	01/07/06
18	01/08/06	29/06/06	01/07/06
19	01/09/06	29/06/06	01/07/06
20	01/10/06	29/06/06	01/07/06
21	01/11/06	29/06/06	01/07/06
22	01/12/06	29/06/06	01/07/06
23	01/01/07	29/06/06	01/07/06
24	01/02/07	29/06/06	01/07/06
25	01/03/07	29/06/06	01/07/06
26	01/04/07	29/06/06	01/07/06
28	01/05/07	29/06/06	01/07/06
29	01/06/07	29/06/06	01/07/06
30	01/07/07	29/06/06	01/07/06
31	01/08/07	29/06/06	01/07/06
32	01/09/07	29/06/06	01/07/06
33	01/10/07	29/06/06	01/07/06
34	01/11/07	29/06/06	01/07/06
35	01/12/07	29/06/06	01/07/06
36	01/01/08	29/06/06	01/07/06
37	01/02/08	29/06/06	01/07/06
38	01/03/08	29/06/06	01/07/06
39	01/04/08	29/06/06	01/07/06
40	01/05/08	29/06/06	01/07/06
41	01/06/08	29/06/06	01/07/06
42	01/07/08	29/06/06	01/07/06
43	01/08/08	29/06/06	01/07/06
44	01/09/08	29/06/06	01/07/06
45	01/10/08	29/06/06	01/07/06
46	01/11/08	29/06/06	01/07/06
47	01/12/08	29/06/06	01/07/06
48	01/01/09	29/06/06	01/07/06
49	01/02/09	29/06/06	01/07/06
50	01/03/09	29/06/06	01/07/06
51	01/04/09	29/06/06	01/07/06
52	01/05/09	29/06/06	01/07/06

Table G.3: File Versions used to produce and evaluate the recent control BioGRID file changes depicted in 5.14 of the main text. The static control files are highlighted in bold.

Version	BioGRID	KEGG	Gene Ontology
17	01/07/06	30/03/09	01/05/09
18	01/08/06	30/03/09	01/05/09
19	01/09/06	30/03/09	01/05/09
20	01/10/06	30/03/09	01/05/09

Continued on next page

Table G.3: File Versions used to produce and evaluate the recent control BioGRID file changes depicted in 5.14 of the main text. The static control files are highlighted in bold.

Version	BioGRID	KEGG	Gene Ontology
21	01/11/06	30/03/09	01/05/09
22	01/12/06	30/03/09	01/05/09
23	01/01/07	30/03/09	01/05/09
24	01/02/07	30/03/09	01/05/09
25	01/03/07	30/03/09	01/05/09
26	01/04/07	30/03/09	01/05/09
28	01/05/07	30/03/09	01/05/09
29	01/06/07	30/03/09	01/05/09
30	01/07/07	30/03/09	01/05/09
31	01/08/07	30/03/09	01/05/09
32	01/09/07	30/03/09	01/05/09
33	01/10/07	30/03/09	01/05/09
34	01/11/07	30/03/09	01/05/09
35	01/12/07	30/03/09	01/05/09
36	01/01/08	30/03/09	01/05/09
37	01/02/08	30/03/09	01/05/09
38	01/03/08	30/03/09	01/05/09
39	01/04/08	30/03/09	01/05/09
40	01/05/08	30/03/09	01/05/09
41	01/06/08	30/03/09	01/05/09
42	01/07/08	30/03/09	01/05/09
43	01/08/08	30/03/09	01/05/09
44	01/09/08	30/03/09	01/05/09
45	01/10/08	30/03/09	01/05/09
46	01/11/08	30/03/09	01/05/09
47	01/12/08	30/03/09	01/05/09
48	01/01/09	30/03/09	01/05/09
49	01/02/09	30/03/09	01/05/09
50	01/03/09	30/03/09	01/05/09
51	01/04/09	30/03/09	01/05/09
52	01/05/09	30/03/09	01/05/09

Table G.4: File Versions used to produce and evaluate the historic control KEGG file changes depicted in 5.16 of the main text. The static control files are highlighted in bold.

Version	BioGRID	KEGG	Gene Ontology
17	01/07/06	29/06/06	01/07/06
20	01/07/06	29/09/06	01/07/06
23	01/07/06	27/12/06	01/07/06
26	01/07/06	28/03/07	01/07/06
30	01/07/06	25/06/07	01/07/06
33	01/07/06	24/09/07	01/07/06
36	01/07/06	03/12/07	01/07/06
39	01/07/06	24/03/08	01/07/06
42	01/07/06	30/06/08	01/07/06
45	01/07/06	29/09/08	01/07/06
48	01/07/06	22/12/08	01/07/06

Table G.5: File Versions used to produce and evaluate the recent control KEGG file changes depicted in 5.16 of the main text. The static control files are highlighted in bold.

Version	BioGRID	KEGG	Gene Ontology
20	01/05/09	29/09/06	01/05/09
23	01/05/09	27/12/06	01/05/09
26	01/05/09	28/03/07	01/05/09
30	01/05/09	25/06/07	01/05/09
33	01/05/09	24/09/07	01/05/09
36	01/05/09	03/12/07	01/05/09
39	01/05/09	24/03/08	01/05/09
42	01/05/09	30/06/08	01/05/09
45	01/05/09	29/09/08	01/05/09
48	01/05/09	22/12/08	01/05/09
52	01/05/09	30/03/09	01/05/09

Table G.6: File versions used to produce and evaluate the historic GO file changes depicted in 5.18 of the main text. The static control files are highlighted in bold.

Version	BioGRID	KEGG	Gene Ontology
17 (HC)	01/07/06	29/06/06	01/07/06
18	01/07/06	29/06/06	01/08/06
19	01/07/06	29/06/06	01/09/06
20	01/07/06	29/06/06	01/10/06
21	01/07/06	29/06/06	01/11/06
22	01/07/06	29/06/06	01/12/06
23	01/07/06	29/06/06	01/01/07
24	01/07/06	29/06/06	01/02/07
25	01/07/06	29/06/06	01/03/07
26	01/07/06	29/06/06	01/04/07
28	01/07/06	29/06/06	01/05/07
29	01/07/06	29/06/06	01/06/07
30	01/07/06	29/06/06	01/07/07
31	01/07/06	29/06/06	01/08/07
32	01/07/06	29/06/06	01/09/07
33	01/07/06	29/06/06	01/10/07
34	01/07/06	29/06/06	01/11/07
35	01/07/06	29/06/06	01/12/07
36	01/07/06	29/06/06	01/01/08
37	01/07/06	29/06/06	01/02/08
38	01/07/06	29/06/06	01/03/08
39	01/07/06	29/06/06	01/04/08
40	01/07/06	29/06/06	01/05/08
41	01/07/06	29/06/06	01/06/08
42	01/07/06	29/06/06	01/07/08
43	01/07/06	29/06/06	01/08/08
44	01/07/06	29/06/06	01/09/08
45	01/07/06	29/06/06	01/10/08
46	01/07/06	29/06/06	01/11/08
47	01/07/06	29/06/06	01/12/08
48	01/07/06	29/06/06	01/01/09
49	01/07/06	29/06/06	01/02/09
50	01/07/06	29/06/06	01/03/09
51	01/07/06	29/06/06	01/04/09
52	01/07/06	29/06/06	01/05/09

Table G.7: File versions used to produce and evaluate the recent control GO file changes depicted in 5.18 of the main text. The static control files are highlighted in bold.

Version	BioGRID	KEGG	Gene Ontology
17	01/05/09	30/03/09	01/07/06
18	01/05/09	30/03/09	01/08/06
19	01/05/09	30/03/09	01/09/06
20	01/05/09	30/03/09	01/10/06
21	01/05/09	30/03/09	01/11/06
22	01/05/09	30/03/09	01/12/06
23	01/05/09	30/03/09	01/01/07
24	01/05/09	30/03/09	01/02/07
25	01/05/09	30/03/09	01/03/07
26	01/05/09	30/03/09	01/04/07
28	01/05/09	30/03/09	01/05/07
29	01/05/09	30/03/09	01/06/07
30	01/05/09	30/03/09	01/07/07
31	01/05/09	30/03/09	01/08/07
32	01/05/09	30/03/09	01/09/07
33	01/05/09	30/03/09	01/10/07
34	01/05/09	30/03/09	01/11/07
35	01/05/09	30/03/09	01/12/07
36	01/05/09	30/03/09	01/01/08
37	01/05/09	30/03/09	01/02/08
38	01/05/09	30/03/09	01/03/08
39	01/05/09	30/03/09	01/04/08
40	01/05/09	30/03/09	01/05/08
41	01/05/09	30/03/09	01/06/08
42	01/05/09	30/03/09	01/07/08
43	01/05/09	30/03/09	01/08/08
44	01/05/09	30/03/09	01/09/08
45	01/05/09	30/03/09	01/10/08
46	01/05/09	30/03/09	01/11/08
47	01/05/09	30/03/09	01/12/08
48	01/05/09	30/03/09	01/01/09
49	01/05/09	30/03/09	01/02/09
50	01/05/09	30/03/09	01/03/09
51	01/05/09	30/03/09	01/04/09
52 (RC)	01/05/09	30/03/09	01/05/09

Appendix H

Functional Predictions

The high confidence (>1.0) functional predictions produced by the RelCID integration method are available on the attached disk in the folder `appendixH`.

The folder contains two files:

1. `NewPredictions.txt` containing the predictions in the tabbed format:

```
term gene score network
```

where C = control, N = node relevance, I = interaction relevance and E = edge relevance.

For example:

```
GO:0000447 YDR440W 1.6424146271356803 I
```

2. `TermDetails.txt` containing the details of the predicted GOBP terms in the tabbed format:

```
GOID description
```

For example:

```
GO:0000041 transition metal ion transport
```


Bibliography

- [1] K Aggarwal and K H Lee. Functional genomics and proteomics as a foundation for systems biology. *Brief Funct Genomic Proteomic*, 2(3):175–84, 2003.
- [2] M R Rose and T H Oakley. The new biology: beyond the Modern Synthesis. *Biol Direct*, 2:30, 2007.
- [3] W K Huh, J V Falvo, L C Gerke, A S Carroll, R W Howson, J S Weissman, and E K O’Shea. Global analysis of protein localization in budding yeast. *Nature*, 425(6959):686–691, Oct 2003.
- [4] H Zhu, M Bilgin, R Bangham, D Hall, A Casamayor, P Bertone, N Lan, R Jansen, S Bidlingmaier, T Houfek, T Mitchell, P Miller, R A Dean, M Gerstein, and M Snyder. Global analysis of protein activities using proteome chips. *Science*, 293(5537):2101–2105, Sep 2001.
- [5] P Ross-Macdonald, P S Coelho, T Roemer, S Agarwal, A Kumar, R Jansen, K H Cheung, A Sheehan, D Symoniatis, L Umansky, M Heidtman, F K Nelson, H Iwasaki, K Hager, M Gerstein, P Miller, G S Roeder, and M Snyder. Large-scale analysis of the yeast genome by transposon tagging and gene disruption. *Nature*, 402(6760):413–418, Nov 1999.
- [6] E A Winzeler, D D Shoemaker, A Astromoff, H Liang, K Anderson, B Andre, R Bangham, R Benito, J D Boeke, H Bussey, A M Chu, C Connelly, K Davis, F Dietrich, S W Dow, M El Bakkoury, F Foury, S H Friend, E Gentalen, G Giaever, J H Hegemann, T Jones, M Laub, H Liao, N Liebundguth, D J Lockhart, A Lucau-Danila, M Lussier, N M’Rabet, P Menard, M Mittmann, C Pai, C Rebischung, J L Revuelta, L Riles, C J Roberts, P Ross-MacDonald, B Scherens, M Snyder, S Sookhai-Mahadeo, R K Storms, S Véronneau, M Voet, G Volckaert, T R Ward, R Wysocki, G S Yen, K Yu, K Zimmermann, P Philippsen, M Johnston, and R W Davis. Functional characterization of the *S. cerevisiae* genome by gene deletion and parallel analysis. *Science*, 285(5429):901–906, Aug 1999.
- [7] D Auerbach, S Thaminy, M O Hottiger, and I Stagljar. The post-genomic era of interactive proteomics: facts and perspectives. *Proteomics*, 2(6):611–623, Jun 2002.
- [8] P Braun and J LaBaer. High throughput protein production for functional proteomics. *Trends Biotechnol*, 21(9):383–388, Sep 2003.
- [9] C L Tucker. High-throughput cell-based assays in yeast. *Drug Discov Today*, 7(18 Suppl):125–130, Sep 2002.
- [10] R B Altman and S Raychaudhuri. Whole-genome expression analysis: challenges beyond clustering. *Curr Opin Struct Biol*, 11(3):340–347, Jun 2001.
- [11] M B Eisen, P T Spellman, P O Brown, and D Botstein. Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci U S A*, 95(25):14863–14868, Dec 1998.
- [12] J Qian, M Dolled-Filhart, J Lin, H Yu, and M Gerstein. Beyond synexpression relationships: local clustering of time-shifted and inverted gene expression profiles identifies new, biologically relevant interactions. *J Mol Biol*, 314(5):1053–1066, Dec 2001.
- [13] T R Hughes, M J Marton, A R Jones, C J Roberts, R Stoughton, C D Armour, H A Bennett, E Coffey, H Dai, Y D He, M J Kidd, A M King, M R Meyer, D Slade, P Y Lum, S B Stepaniants, D D Shoemaker, D Gachotte, K Chakraborty, J Simon, M Bard, and S H Friend. Functional discovery via a compendium of expression profiles. *Cell*, 102(1):109–126, Jul 2000.

- [14] A H Tong, M Evangelista, A B Parsons, H Xu, G D Bader, N Pagé, M Robinson, S Raghbizadeh, C W Hogue, H Bussey, B Andrews, M Tyers, and C Boone. Systematic genetic analysis with ordered arrays of yeast deletion mutants. *Science*, 294(5550):2364–2368, Dec 2001.
- [15] X Pan, D S Yuan, D Xiang, X Wang, S Sookhai-Mahadeo, J S Bader, P Hieter, F Spencer, and J D Boeke. A robust toolkit for functional profiling of the yeast genome. *Mol Cell*, 16(3):487–496, Nov 2004.
- [16] S G Addinall, M Downey, M Yu, M K Zubko, J Dewar, A Leake, J Hallinan, O Shaw, K James, D J Wilkinson, A Wipat, D Durocher, and D Lydall. A genomewide suppressor and enhancer analysis of *cdc13-1* reveals varied cellular processes influencing telomere capping in *Saccharomyces cerevisiae*. *Genetics*, 180(4):2251–2266, Dec 2008.
- [17] M Schuldiner, S R Collins, N J Thompson, V Denic, A Bhamidipati, T Punna, J Ihmels, B Andrews, C Boone, J F Greenblatt, J S Weissman, and N J Krogan. Exploration of the function and organization of the yeast early secretory pathway through an epistatic miniarray profile. *Cell*, 123(3):507–519, Nov 2005.
- [18] A Kumar, S Agarwal, J A Heyman, S Matson, M Heidtman, S Piccirillo, L Umansky, A Drawid, R Jansen, Y Liu, K H Cheung, P Miller, M Gerstein, G S Roeder, and M Snyder. Subcellular localization of the yeast proteome. *Genes Dev*, 16(6):707–719, Mar 2002.
- [19] M R Martzen, S M McCraith, S L Spinelli, F M Torres, S Fields, E J Grayhack, and E M Phizicky. A biochemical genomics approach for identifying genes by the activity of their products. *Science*, 286(5442):1153–1155, Nov 1999.
- [20] J Ptacek, G Devgan, G Michaud, H Zhu, X Zhu, J Fasolo, H Guo, G Jona, A Breitkreutz, R Sopko, R R McCartney, M C Schmidt, N Rachidi, S J Lee, A S Mah, L Meng, M J Stark, D F Stern, C De Virgilio, M Tyers, B Andrews, M Gerstein, B Schweitzer, P F Predki, and M Snyder. Global analysis of protein phosphorylation in yeast. *Nature*, 438(7068):679–684, Dec 2005.
- [21] T Ito, T Chiba, R Ozawa, M Yoshida, M Hattori, and Y Sakaki. A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc Natl Acad Sci U S A*, 98(8):4569–4574, Apr 2001.
- [22] P Uetz, L Giot, G Cagney, T A Mansfield, R S Judson, J R Knight, D Lockshon, V Narayan, M Srinivasan, P Pochart, A Qureshi-Emili, Y Li, B Godwin, D Conover, T Kalbfleisch, G Vijayadamodar, M Yang, M Johnston, S Fields, and J M Rothberg. A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*. *Nature*, 403(6770):623–627, Feb 2000.
- [23] A C Gavin, P Aloy, P Grandi, R Krause, M Boesche, M Marzioch, C Rau, L J Jensen, S Bastuck, B Dümpelfeld, A Edelmann, M A Heurtier, V Hoffman, C Hoefert, K Klein, M Hudak, A M Michon, M Schelder, M Schirle, M Remor, T Rudi, S Hooper, A Bauer, T Bouwmeester, G Casari, G Drewes, G Neubauer, J M Rick, B Kuster, P Bork, R B Russell, and G Superti-Furga. Proteome survey reveals modularity of the yeast cell machinery. *Nature*, 440(7084):631–636, Mar 2006.
- [24] 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 Gandhi, 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, Mar 2006.
- [25] C Auffray, S Imbeaud, M Roux-Rouquié, and L Hood. From functional genomics to systems biology: concepts and practices. *C R Biol*, 326(10-11):879–892, Oct-Nov 2003.
- [26] F J Bruggeman and H V Westerhoff. The nature of systems biology. *Trends Microbiol*, 15(1):45–50, Jan 2007.
- [27] M Y Galperin and G R Cochrane. The 2011 Nucleic Acids Research Database issue and the online Molecular Biology Database Collection. *Nucleic Acids Res*, 39(Database issue):1–6, Jan 2011.

- [28] G R Cochrane and M Y Galperin. The 2010 Nucleic Acids Research Database Issue and online Database Collection: a community of data resources. *Nucleic Acids Res*, 38(Database issue):1–4, Jan 2010.
- [29] M Y Galperin and G R Cochrane. Nucleic Acids Research annual Database Issue and the NAR online Molecular Biology Database Collection in 2009. *Nucleic Acids Res*, 37(Database issue):1–4, Jan 2009.
- [30] M A Andrade and C Sander. Bioinformatics: from genome data to biological knowledge. *Curr Opin Biotechnol*, 8(6):675–683, Dec 1997.
- [31] H Kitano. Computational systems biology. *Nature*, 420(6912):206–210, Nov 2002.
- [32] T Ideker, T Galitski, and L Hood. A new approach to decoding life: systems biology. *Annu Rev Genomics Hum Genet*, 2:343–372, 2001.
- [33] H Kitano. Systems biology: a brief overview. *Science*, 295(5560):1662–1664, Mar 2002.
- [34] M Latterich. Molecular systems biology at the crossroads: to know less about more, or to know more about less? *Proteome Sci*, 3:8, Oct 2005.
- [35] L Hood, L Rowen, D J Galas, and J D Aitchison. Systems biology at the Institute for Systems Biology. *Brief Funct Genomic Proteomic*, 7(4):239–248, Jul 2008.
- [36] N Blow. Systems biology: untangling the protein web. *Nature*, 460(7253):415–418, Jul 2009.
- [37] M E Cusick, N Klitgord, M Vidal, and D E Hill. Interactome: gateway into systems biology. *Hum Mol Genet*, 14 Spec No. 2:171–181, Oct 2005.
- [38] J Goll and P Uetz. The elusive yeast interactome. *Genome Biol*, 7(6):223, 2006.
- [39] A Trewavas. A brief history of systems biology. *Plant Cell*, 18(10):2420–2430, Oct 2006.
- [40] A Adourian, E Jennings, R Balasubramanian, W M Hines, D Damian, T N Plasterer, C B Clish, P Stroobant, R McBurney, E R Verheij, I Bobeldijk, J van der Greef, J Lindberg, K Kenne, U Andersson, H Hellmold, K Nilsson, H Salter, and I Schuppe-Koistinen. Correlation network analysis for data integration and biomarker selection. *Mol Biosyst*, 4(3):249–259, Mar 2008.
- [41] C Li and H Li. Network-constrained regularization and variable selection for analysis of genomic data. *Bioinformatics*, 24(9):1175–1182, May 2008.
- [42] G D Bader and C W Hogue. Analyzing yeast protein-protein interaction data obtained from different sources. *Nat Biotechnol*, 20(10):991–997, Oct 2002.
- [43] S R Collins, P Kemmeren, X C Zhao, J F Greenblatt, F Spencer, F C Holstege, J S Weissman, and N J Krogan. Toward a comprehensive atlas of the physical interactome of *Saccharomyces cerevisiae*. *Mol Cell Proteomics*, 6(3):439–450, Mar 2007.
- [44] M E Futschik, G Chaurasia, and H Herzel. Comparison of human protein-protein interaction maps. *Bioinformatics*, 23(5):605–611, Mar 2007.
- [45] G T Hart, I Lee, and E M Marcotte. A high-accuracy consensus map of yeast protein complexes reveals modular nature of gene essentiality. *BMC Bioinformatics*, 8(1):236–236, Jul 2007.
- [46] C Huttenhower and O G Troyanskaya. Assessing the functional structure of genomic data. *Bioinformatics*, 24(13):330–338, Jul 2008.
- [47] A Beyer, S Bandyopadhyay, and T Ideker. Integrating physical and genetic maps: from genomes to interaction networks. *Nat Rev Genet*, 8(9):699–710, Sep 2007.
- [48] J S Hallinan and A Wipat. Motifs and modules in fractured functional yeast networks. *IEEE Symposium on Computational Intelligence and Bioinformatics and Computational Biology, 2007. CIBCB '07*, pages 189–196, 2007.
- [49] I Lee, S V Date, A T Adai, and E M Marcotte. A probabilistic functional network of yeast genes. *Science*, 306(5701):1555–1558, Nov 2004.

- [50] E Marcotte and S Date. Exploiting big biology: integrating large-scale biological data for function inference. *Brief Bioinform*, 2(4):363–374, Dec 2001.
- [51] A R Joyce and B Ø Palsson. The model organism as a system: integrating ‘omics’ data sets. *Nat Rev Mol Cell Biol*, 7(3):198–210, Mar 2006.
- [52] L H Hartwell, J J Hopfield, S Leibler, and A W Murray. From molecular to modular cell biology. *Nature*, 402(6761 Suppl):47–52, Dec 1999.
- [53] V K Mootha, P Lepage, K Miller, J Bunkenborg, M Reich, M Hjerrild, T Delmonte, A Villeneuve, R Sladek, F Xu, G A Mitchell, C Morin, M Mann, T J Hudson, B Robinson, J D Rioux, and E S Lander. Identification of a gene causing human cytochrome c oxidase deficiency by integrative genomics. *Proc Natl Acad Sci U S A*, 100(2):605–610, Jan 2003.
- [54] H Scheel, S Tomiuk, and K Hofmann. Elucidation of ataxin-3 and ataxin-7 function by integrative bioinformatics. *Hum Mol Genet*, 12(21):2845–2852, Nov 2003.
- [55] R Alfieri, I Merelli, E Mosca, and L Milanesi. A data integration approach for cell cycle analysis oriented to model simulation in systems biology. *BMC Syst Biol*, 1(1):35, Aug 2007.
- [56] V Detours, J E Dumont, H Bersini, and C Maenhaut. Integration and cross-validation of high-throughput gene expression data: comparing heterogeneous data sets. *FEBS Lett*, 546(1):98–102, Jul 2003.
- [57] B Linghu, E S Snitkin, D T Holloway, A M Gustafson, Y Xia, and C Delisi. High-precision high-coverage functional inference from integrated data sources. *BMC Bioinformatics*, 9(1):119, Feb 2008.
- [58] C von Mering, L J Jensen, M Kuhn, S Chaffron, T Doerks, B Krüger, B Snel, and P Bork. STRING 7—recent developments in the integration and prediction of protein interactions. *Nucleic Acids Res*, 35(Database issue):358–362, Jan 2007.
- [59] C L Myers and O G Troyanskaya. Context-sensitive data integration and prediction of biological networks. *Bioinformatics*, 23(17):2322–2330, Sep 2007.
- [60] O G Troyanskaya, K Dolinski, A B Owen, R B Altman, and D Botstein. A Bayesian framework for combining heterogeneous data sources for gene function prediction (in *Saccharomyces cerevisiae*). *Proc Natl Acad Sci U S A*, 100(14):8348–8353, Jul 2003.
- [61] W Zhong and P W Sternberg. Automated data integration for developmental biological research. *Development*, 134(18):3227–3238, Sep 2007.
- [62] A Chatr-Aryamontri, A Ceol, L Licata, and G Cesareni. Protein interactions: integration leads to belief. *Trends Biochem Sci*, 33(6):241–242, Jun 2008.
- [63] M Fellenberg, K Albermann, A Zollner, H W Mewes, and J Hani. Integrative analysis of protein interaction data. *Proc Int Conf Intell Syst Mol Biol*, 8:152–161, 2000.
- [64] T R Hazbun, L Malmström, S Anderson, B J Graczyk, B Fox, M Riffle, B A Sundin, J D Aranda, W H McDonald, C H Chiu, B E Snyderman, P Bradley, E G Muller, S Fields, D Baker, J R Yates, and T N Davis. Assigning function to yeast proteins by integration of technologies. *Mol Cell*, 12(6):1353–1365, Dec 2003.
- [65] H Lu, B Shi, G Wu, Y Zhang, X Zhu, Z Zhang, C Liu, Y Zhao, T Wu, J Wang, and R Chen. Integrated analysis of multiple data sources reveals modular structure of biological networks. *Biochem Biophys Res Commun*, 345(1):302–309, Jun 2006.
- [66] D B Searls. Data integration: challenges for drug discovery. *Nat Rev Drug Discov*, 4(1):45–58, Jan 2005.
- [67] Y Xia, H Yu, R Jansen, M Seringhaus, S Baxter, D Greenbaum, H Zhao, and M Gerstein. Analyzing cellular biochemistry in terms of molecular networks. *Annu Rev Biochem*, 73:1051–1087, 2004.
- [68] X Zhu, M Gerstein, and M Snyder. Getting connected: analysis and principles of biological networks. *Genes Dev*, 21(9):1010–1024, May 2007.

- [69] J De Las Rivas and C Fontanillo. Protein-protein interactions essentials: key concepts to building and analyzing interactome networks. *PLoS Comput Biol*, 6(6):e1000807, June 2010.
- [70] D Fiedler, H Braberg, M Mehta, G Chechik, G Cagney, P Mukherjee, A C Silva, M Shales, S R Collins, S van Wageningen, P Kemmeren, F C Holstege, J S Weissman, M C Keogh, D Koller, K M Shokat, and N J Krogan. Functional organization of the *S. cerevisiae* phosphorylation network. *Cell*, 136(5):952–963, Mar 2009.
- [71] C Y Yang, C H Chang, Y L Yu, T C Lin, S A Lee, C C Yen, J M Yang, J M Lai, Y R Hong, T L Tseng, K M Chao, and C Y Huang. PhosphoPOINT: a comprehensive human kinase interactome and phospho-protein database. *Bioinformatics*, 24(16):14–20, Aug 2008.
- [72] H Zhu and M Snyder. "Omic" approaches for unraveling signaling networks. *Curr Opin Cell Biol*, 14(2):173–179, Apr 2002.
- [73] A Cornish-Bowden, M L Cárdenas, J C Letelier, J Soto-Andrade, and F G Abarzúa. Understanding the parts in terms of the whole. *Biol Cell*, 96(9):713–717, Dec 2004.
- [74] L J Sweetlove and A R Fernie. Regulation of metabolic networks: understanding metabolic complexity in the systems biology era. *New Phytol*, 168(1):9–24, Oct 2005.
- [75] D A Fell and A Wagner. The small world of metabolism. *Nat Biotechnol*, 18(11):1121–1122, Nov 2000.
- [76] H Jeong, B Tombor, R Albert, Z N Oltvai, and A L Barabási. The large-scale organization of metabolic networks. *Nature*, 407(6804):651–654, Oct 2000.
- [77] A Zhang. *Protein Interaction Networks: Computational Analysis*. Cambridge University Press, New York, NY, USA, 2009.
- [78] W Huber, V J Carey, L Long, S Falcon, and R Gentleman. Graphs in molecular biology. *BMC Bioinformatics*, 8(Suppl 6):S8, 2007.
- [79] Y Liu, I Kim, and H Zhao. Protein interaction predictions from diverse sources. *Drug Discov Today*, 13(9-10):409–416, May 2008.
- [80] J Köhler, J Baumbach, J Taubert, M Specht, A Skusa, A Rüegg, C Rawlings, P Verrier, and S Philippi. Graph-based analysis and visualization of experimental results with ONDEX. *Bioinformatics*, 22(11):1383–1390, Jun 2006.
- [81] H David-Eden and Y Mandel-Gutfreund. Revealing unique properties of the ribosome using a network based analysis. *Nucleic Acids Res*, 36(14):4641–4652, Aug 2008.
- [82] R J Williams and N D Martinez. Simple rules yield complex food webs. *Nature*, 404(6774):180–183, Mar 2000.
- [83] J D Han. Understanding biological functions through molecular networks. *Cell Res*, 18(2):224–237, Feb 2008.
- [84] S Khor. Application of graph colouring to biological networks. *IET Syst Biol*, 4(3):185–192, May 2010.
- [85] N Gehlenborg, S I O'Donoghue, N S Baliga, A Goesmann, A Hibbs, Kitano, O Kohlbacher, H Neuweger, R Schneider, D Tenenbaum, and AC Gavin. Visualization of omics data for systems biology. *Nature Methods*, 7:S56–S68, 2010.
- [86] S Orchard and S Kerrien. Molecular interactions and data standardisation. *Methods Mol Biol*, 604:309–318, 2010.
- [87] L Strömbäck, V Jakoniene, H Tan, and P Lambrix. Representing, storing and accessing molecular interaction data: a review of models and tools. *Brief Bioinform*, 7(4):331–338, Dec 2006.
- [88] S Asthana, O D King, F D Gibbons, and F P Roth. Predicting protein complex membership using probabilistic network reliability. *Genome Res*, 14(6):1170–1175, Jun 2004.

- [89] G D Bader and C W Hogue. An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinformatics*, 4:2, Jan 2003.
- [90] C Brun, C Herrmann, and A Guénoche. Clustering proteins from interaction networks for the prediction of cellular functions. *BMC Bioinformatics*, 5:95, Jul 2004.
- [91] H N Chua, W K Sung, and L Wong. Using indirect protein interactions for the prediction of Gene Ontology functions. *BMC Bioinformatics*, 8(Suppl 4):S8, 2007.
- [92] U Karaoz, T M Murali, S Letovsky, Y Zheng, C Ding, C R Cantor, and S Kasif. Whole-genome annotation by using evidence integration in functional-linkage networks. *Proc Natl Acad Sci U S A*, 101(9):2888–2893, Mar 2004.
- [93] C S Goh and F E Cohen. Co-evolutionary analysis reveals insights into protein-protein interactions. *J Mol Biol*, 324(1):177–192, Nov 2002.
- [94] F Pazos and A Valencia. Similarity of phylogenetic trees as indicator of protein-protein interaction. *Protein Eng*, 14(9):609–614, Sep 2001.
- [95] M A Gilchrist, L A Salter, and A Wagner. A statistical framework for combining and interpreting proteomic datasets. *Bioinformatics*, 20(5):689–700, Mar 2004.
- [96] A Clauset, C Moore, and M E Newman. Hierarchical structure and the prediction of missing links in networks. *Nature*, 453(7191):98–101, May 2008.
- [97] F Browne, H Wang, H Zheng, and F Azuaje. GRIP: A web-based system for constructing Gold Standard datasets for protein-protein interaction prediction. *Source Code Biol Med*, 4:2, 2009.
- [98] I Lee, Z Li, and E M Marcotte. An improved, bias-reduced probabilistic functional gene network of baker’s yeast, *Saccharomyces cerevisiae*. *PLoS ONE*, 2(10):e988, 2007.
- [99] M Kanehisa and S Goto. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res*, 28(1):27–30, Jan 2000.
- [100] 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. *Nat Genet*, 25:25–29, 2000.
- [101] J Yu and R L Finley. Combining multiple positive training sets to generate confidence scores for protein-protein interactions. *Bioinformatics*, 25(1):105–111, Jan 2009.
- [102] J Li, X Li, H Su, H Chen, and D W Galbraith. A framework of integrating gene relations from heterogeneous data sources: an experiment on *Arabidopsis thaliana*. *Bioinformatics*, 22(16):2037–2043, Aug 2006.
- [103] S Yellaboina, K Goyal, and S C Mande. Inferring genome-wide functional linkages in *E. coli* by combining improved genome context methods: comparison with high-throughput experimental data. *Genome Res*, 17(4):527–535, Apr 2007.
- [104] M Deng, T Chen, and F Sun. An integrated probabilistic model for functional prediction of proteins. *J Comput Biol*, 11(2-3):463–475, 2004.
- [105] A Jaimovich, G Elidan, H Margalit, and N Friedman. Towards an integrated protein-protein interaction network: a relational Markov network approach. *J Comput Biol*, 13(2):145–164, Mar 2006.
- [106] C L Myers, D Robson, A Wible, M A Hibbs, C Chiriac, C L Theesfeld, K Dolinski, and O G Troyanskaya. Discovery of biological networks from diverse functional genomic data. *Genome Biol*, 6(13):R114, 2005.
- [107] A V Antonov, I V Tetko, and H W Mewes. A systematic approach to infer biological relevance and biases of gene network structures. *Nucleic Acids Res*, 34(1):e6, 2006.
- [108] M Deng, F Sun, and T Chen. Assessment of the reliability of protein-protein interactions and protein function prediction. *Pac Symp Biocomput*, pages 140–151, 2003.

- [109] 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, Oct 2003.
- [110] L J Lu, Y Xia, A Paccanaro, H Yu, and M Gerstein. Assessing the limits of genomic data integration for predicting protein networks. *Genome Res*, 15(7):945–953, Jul 2005.
- [111] D R Rhodes, S A Tomlins, S Varambally, V Mahavisno, T Barrette, S Kalyana-Sundaram, D Ghosh, A Pandey, and A M Chinnaiyan. Probabilistic model of the human protein-protein interaction network. *Nat Biotechnol*, 23(8):951–959, Aug 2005.
- [112] Y Yamanishi, J P Vert, and M Kanehisa. Protein network inference from multiple genomic data: a supervised approach. *Bioinformatics*, 20 Suppl 1:363–370, Aug 2004.
- [113] D Eisenberg, E M Marcotte, I Xenarios, and T O Yeates. Protein function in the post-genomic era. *Nature*, 405(6788):823–826, Jun 2000.
- [114] Y Chen and D Xu. Global protein function annotation through mining genome-scale data in yeast *Saccharomyces cerevisiae*. *Nucleic Acids Res*, 32(21):6414–6424, 2004.
- [115] L Kiemer, S Costa, M Ueffing, and G Cesareni. WI-PHI: a weighted yeast interactome enriched for direct physical interactions. *Proteomics*, 7(6):932–943, Mar 2007.
- [116] W K Kim, C Krumpelman, and E M Marcotte. Inferring mouse gene functions from genomic-scale data using a combined functional network/classification strategy. *Genome Biol*, 9(Suppl 1):S5, 2008.
- [117] Y Guan, C L Myers, R Lu, I R Lemischka, C J Bult, and O G Troyanskaya. A genomewide functional network for the laboratory mouse. *PLoS Comput Biol*, 4(9):e1000165, Sep 2008.
- [118] M G Kann. Protein interactions and disease: computational approaches to uncover the etiology of diseases. *Brief Bioinform*, 8(5):333–346, Jul 2007.
- [119] J Geisler-Lee, N O’Toole, R Ammar, N J Provart, A H Millar, and M Geisler. A predicted interactome for *Arabidopsis*. *Plant Physiol*, 145(2):317–329, Aug 2007.
- [120] X Lin, M Liu, and X Chen. Protein-protein interaction prediction and assessment from model organisms. In *BIBM ’08: Proceedings of the 2008 IEEE International Conference on Bioinformatics and Biomedicine*, pages 187–192, Washington, DC, USA, 2008. IEEE Computer Society.
- [121] R Mrowka, A Patzak, and H Herzel. Is there a bias in proteome research? *Genome Res*, 11(12):1971–1973, Dec 2001.
- [122] 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 Comput Biol*, 3(11):e214, Nov 2007.
- [123] A H Tong, G Lesage, G D Bader, H Ding, H Xu, X Xin, J Young, G F Berriz, R L Brost, M Chang, Y Chen, X Cheng, G Chua, H Friesen, D S Goldberg, J Haynes, C Humphries, G He, S Hussein, L Ke, N Krogan, Z Li, J N Levinson, H Lu, P Ménard, C Munyana, A B Parsons, O Ryan, R Tonikian, T Roberts, A M Sdicu, J Shapiro, B Sheikh, B Suter, S L Wong, L V Zhang, H Zhu, C G Burd, S Munro, C Sander, J Rine, J Greenblatt, M Peter, A Bretscher, G Bell, F P Roth, G W Brown, B Andrews, H Bussey, and C Boone. Global mapping of the yeast genetic interaction network. *Science*, 303(5659):808–813, Feb 2004.
- [124] A M Edwards, B Kus, R Jansen, D Greenbaum, J Greenblatt, and M Gerstein. Bridging structural biology and genomics: assessing protein interaction data with known complexes. *Trends Genet*, 18(10):529–536, Oct 2002.
- [125] C L Myers, D R Barrett, M A Hibbs, C Huttenhower, and O G Troyanskaya. Finding function: evaluation methods for functional genomic data. *BMC Genomics*, 7:187, 2006.
- [126] 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, Aug 2006.

- [127] J Chen, W Hsu, M L Lee, and S K Ng. Discovering reliable protein interactions from high-throughput experimental data using network topology. *Artif Intell Med*, 35(1-2):37–47, Sep-Oct 2005.
- [128] K James, A Wipat, and J Hallinan. Integration of full-coverage probabilistic functional networks with relevance to specific biological processes. *Data Integration in the Life Sciences*, pages 31–46, 2009.
- [129] Z Guo, Y Li, X Gong, C Yao, W Ma, D Wang, Y Li, J Zhu, M Zhang, D Yang, and J Wang. Edge-based scoring and searching method for identifying condition-responsive protein-protein interaction sub-network. *Bioinformatics*, 23(16):2121–2128, Jun 2007.
- [130] Y Li, W Ma, Z Guo, D Yang, D Wang, M Zhang, J Zhu, and Y Li. Characterizing proteins with finer functions: a case study for translational functions of yeast proteins. *Bioinformatics and Biomedical Engineering, 2007. ICBBE 2007*, pages 141–144, 2007.
- [131] S J Wodak, S Pu, J Vlasblom, and B Séraphin. Challenges and rewards of interaction proteomics. *Mol Cell Proteomics*, 8(1):3–18, Jan 2009.
- [132] A J Walhout and M Vidal. Protein interaction maps for model organisms. *Nat Rev Mol Cell Biol*, 2(1):55–62, Jan 2001.
- [133] E Franzosa, B Linghu, and Y Xia. Computational reconstruction of protein-protein interaction networks: algorithms and issues. *Methods Mol Biol*, 541:89–100, 2009.
- [134] 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 Svrvzikapa, 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, Aug 2008.
- [135] J R Bradford, J A Siepen, and D R Westhead. *Fundamentals of protein structure and function*. John Wiley & Sons, Ltd, 2004.
- [136] I M Nooren and J M Thornton. Diversity of protein-protein interactions. *EMBO J*, 22(14):3486–3492, Jul 2003.
- [137] B A Shoemaker and A R Panchenko. Deciphering protein-protein interactions. Part I. Experimental techniques and databases. *PLoS Comput Biol*, 3(3):e42, Mar 2007.
- [138] S Lalonde, D W Ehrhardt, D Loqué, J Chen, S Y Rhee, and W B Frommer. Molecular and cellular approaches for the detection of protein-protein interactions: latest techniques and current limitations. *Plant J*, 53(4):610–635, Feb 2008.
- [139] A R Mendelsohn and R Brent. Protein interaction methods—toward an endgame. *Science*, 284(5422):1948–1950, Jun 1999.
- [140] E M Phizicky and S Fields. Protein-protein interactions: methods for detection and analysis. *Microbiol Rev*, 59(1):94–123, Mar 1995.
- [141] T Berggård, S Linse, and P James. Methods for the detection and analysis of protein-protein interactions. *Proteomics*, 7(16):2833–2842, Aug 2007.
- [142] W P Blackstock and M P Weir. Proteomics: quantitative and physical mapping of cellular proteins. *Trends Biotechnol*, 17(3):121–127, Mar 1999.
- [143] A Pandey and M Mann. Proteomics to study genes and genomes. *Nature*, 405(6788):837–846, Jun 2000.
- [144] S Fields and O Song. A novel genetic system to detect protein-protein interactions. *Nature*, 340(6230):245–246, Jul 1989.
- [145] D Lohr, P Venkov, and J Zlatanova. Transcriptional regulation in the yeast GAL gene family: a complex genetic network. *FASEB J*, 9(9):777–787, Jun 1995.
- [146] L Keegan, G Gill, and M Ptashne. Separation of DNA binding from the transcription-activating function of a eukaryotic regulatory protein. *Science*, 231(4739):699–704, Feb 1986.

- [147] B Suter, S Kittanakom, and I Stagljar. Two-hybrid technologies in proteomics research. *Curr Opin Biotechnol*, 19(4):316–323, Aug 2008.
- [148] A Gurvitz, J G Coe, and I W Dawes. Use of reporter genes for the isolation and characterisation of different classes of sporulation mutants in the yeast *Saccharomyces cerevisiae*. *Curr Genet*, 24(5):451–454, Nov 1993.
- [149] B L Drees, B Sundin, E Brazeau, J P Caviston, G C Chen, W Guo, K G Kozminski, M W Lau, J J Moskow, A Tong, L R Schenkman, A McKenzie, P Brennwald, M Longtine, E Bi, C Chan, P Novick, C Boone, J R Pringle, T N Davis, S Fields, and D G Drubin. A protein interaction map for cell polarity development. *J Cell Biol*, 154(3):549–571, Aug 2001.
- [150] T Ito, K Tashiro, S Muta, R Ozawa, T Chiba, M Nishizawa, K Yamamoto, S Kuhara, and Y Sakaki. Toward a protein-protein interaction map of the budding yeast: a comprehensive system to examine two-hybrid interactions in all possible combinations between the yeast proteins. *Proc Natl Acad Sci U S A*, 97(3):1143–1147, Feb 2000.
- [151] M Fromont-Racine, J C Rain, and P Legrain. Toward a functional analysis of the yeast genome through exhaustive two-hybrid screens. *Nat Genet*, 16(3):277–282, Jul 1997.
- [152] 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, N Klitgord, C Simon, M Boxem, S Milstein, J Rosenberg, D S Goldberg, L V Zhang, S L Wong, G Franklin, S Li, J S Albala, J Lim, C Fraughton, E Llamas, S Cevik, C Bex, P Lamesch, R S Sikorski, J Vandenhaute, H Y Zoghbi, A Smolyar, S Bosak, R Sequerra, L Doucette-Stamm, M E Cusick, D E Hill, F P Roth, and M Vidal. Towards a proteome-scale map of the human protein-protein interaction network. *Nature*, 437(7062):1173–1178, Oct 2005.
- [153] U Stelzl, U Worm, M Lalowski, C Haenig, F H Brembeck, H Goehler, M Stroedicke, M Zenkner, A Schoenherr, S Koeppen, J Timm, S Mintzlauff, C Abraham, N Bock, S Kietzmann, A Goedde, E Toksöz, A Droege, S Krobitsch, B Korn, W Birchmeier, H Lehrach, and E E Wanker. A human protein-protein interaction network: a resource for annotating the proteome. *Cell*, 122(6):957–968, Sep 2005.
- [154] G T Hart, A K Ramani, and E M Marcotte. How complete are current yeast and human protein-interaction networks? *Genome Biol*, 7(11):120, 2006.
- [155] G Chaurasia, Y Iqbal, C Hänig, H Herzel, E E Wanker, and M E Futschik. UniHI: an entry gate to the human protein interactome. *Nucleic Acids Res*, 35(Database issue):590–594, Jan 2007.
- [156] 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, Jan 2004.
- [157] 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 Varone, 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 Carroll, 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, Dec 2003.
- [158] 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 Chanut, 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 Papadopoulou, F Perez, A Plessis, C Rossé, 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 Res*, 15(3):376–384, Mar 2005.

- [159] J R Parrish, J Yu, G Liu, J A Hines, J E Chan, B A Mangiola, H Zhang, S Pacifico, F Fotouhi, V J DiRita, T Ideker, P Andrews, and R L Finley. A proteome-wide protein interaction map for *Campylobacter jejuni*. *Genome Biol*, 8(7):R130, 2007.
- [160] J C Rain, L Selig, H De Reuse, V Battaglia, C Reverdy, S Simon, G Lenzen, F Petel, J Wojcik, V Schächter, Y Chemama, A Labigne, and P Legrain. The protein-protein interaction map of *Helicobacter pylori*. *Nature*, 409(6817):211–215, Jan 2001.
- [161] D J LaCount, M Vignali, R Chettier, A Phansalkar, R Bell, J R Hesselberth, L W Schoenfeld, I Ota, S Sahasrabudhe, C Kurschner, S Fields, and R E Hughes. A protein interaction network of the malaria parasite *Plasmodium falciparum*. *Nature*, 438(7064):103–107, Nov 2005.
- [162] F Colland, X Jacq, V Trouplin, C Mougin, C Groizeleau, A Hamburger, A Meil, J Wojcik, P Legrain, and J M Gauthier. Functional proteomics mapping of a human signaling pathway. *Genome Res*, 14(7):1324–1332, Jul 2004.
- [163] P Uetz, Y A Dong, C Zeretzke, C Atzler, A Baiker, B Berger, S V Rajagopala, M Roupelieva, D Rose, E Fossum, and J Haas. Herpesviral protein networks and their interaction with the human proteome. *Science*, 311(5758):239–242, Jan 2006.
- [164] I Lemmens, S Lievens, and J Tavernier. Strategies towards high-quality binary protein interactome maps. *J Proteomics*, 73(8):1415–1420, Jun 2010.
- [165] D Scholtens, M Vidal, and R Gentleman. Local modeling of global interactome networks. *Bioinformatics*, 21(17):3548–3557, Sep 2005.
- [166] M Dmitrova, G Younès-Cauet, P Oertel-Buchheit, D Porte, M Schnarr, and M Granger-Schnarr. A new LexA-based genetic system for monitoring and analyzing protein heterodimerization in *Escherichia coli*. *Mol Gen Genet*, 257(2):205–212, Jan 1998.
- [167] C A Stanyon, G Liu, B A Mangiola, N Patel, L Giot, B Kuang, H Zhang, J Zhong, and R L Finley. A *Drosophila* protein-interaction map centered on cell-cycle regulators. *Genome Biol*, 5(12):R96, 2004.
- [168] 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, Jun 2008.
- [169] S W Michnick. Protein fragment complementation strategies for biochemical network mapping. *Curr Opin Biotechnol*, 14(6):610–617, Dec 2003.
- [170] I Remy and S W Michnick. Dynamic visualization of expressed gene networks. *J Cell Physiol*, 196(3):419–429, Sep 2003.
- [171] M Damelin and P A Silver. Mapping interactions between nuclear transport factors in living cells reveals pathways through the nuclear pore complex. *Mol Cell*, 5(1):133–140, Jan 2000.
- [172] I Kaganman. FRETting for a more detailed interactome. *Nat Methods*, 4(2):112–113, Feb 2007.
- [173] Y Xu, D W Piston, and C H Johnson. A bioluminescence resonance energy transfer (BRET) system: application to interacting circadian clock proteins. *Proc Natl Acad Sci U S A*, 96(1):151–156, Jan 1999.
- [174] S Eyckerman, I Lemmens, D Catteeuw, A Verhee, J Vandekerckhove, S Lievens, and J Tavernier. Reverse MAPPIT: screening for protein-protein interaction modifiers in mammalian cells. *Nat Methods*, 2(6):427–433, Jun 2005.
- [175] S Lievens, N Vanderroost, J Van der Heyden, V Gesellchen, M Vidal, and J Tavernier. Array MAPPIT: high-throughput interactome analysis in mammalian cells. *J Proteome Res*, 8(2):877–886, Feb 2009.
- [176] H Zhu and M Snyder. Protein chip technology. *Curr Opin Chem Biol*, 7(1):55–63, Feb 2003.
- [177] H Zhu, J F Klemic, S Chang, P Bertone, A Casamayor, K G Klemic, D Smith, M Gerstein, M A Reed, and M Snyder. Analysis of yeast protein kinases using protein chips. *Nat Genet*, 26(3):283–289, Nov 2000.

- [178] H Ge. UPA, a universal protein array system for quantitative detection of protein-protein, protein-DNA, protein-RNA and protein-ligand interactions. *Nucleic Acids Res*, 28(2):e3, Jan 2000.
- [179] R Aebersold and M Mann. Mass spectrometry-based proteomics. *Nature*, 422(6928):198–207, Mar 2003.
- [180] K Ashman, M F Moran, F Sicheri, T Pawson, and M Tyers. Cell signalling - the proteomics of it all. *Sci STKE*, 2001(103):pe33, Oct 2001.
- [181] 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, Jul 2001.
- [182] G Rigaut, A Shevchenko, B Rutz, M Wilm, M Mann, and B Séraphin. A generic protein purification method for protein complex characterization and proteome exploration. *Nat Biotechnol*, 17(10):1030–1032, Oct 1999.
- [183] M Mann, R C Hendrickson, and A Pandey. Analysis of proteins and proteomes by mass spectrometry. *Annu Rev Biochem*, 70:437–473, 2001.
- [184] 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, May 2002.
- [185] M Heo, S Maslov, and E Shakhnovich. Topology of protein interaction network shapes protein abundances and strengths of their functional and nonspecific interactions. *Proc Natl Acad Sci U S A*, 108(10):4258–4263, Mar 2011.
- [186] C Lu, X Hu, G Wang, L J Leach, S Yang, M J Kearsey, and Z W Luo. Why do essential proteins tend to be clustered in the yeast interactome network? *Mol Biosyst*, 6(5):871–877, May 2010.
- [187] L Hakes, D L Robertson, S G Oliver, and S C Lovell. Protein interactions from complexes: a structural perspective. *Comp Funct Genomics*, page 49356, 2007.
- [188] P D’haeseleer and G M Church. Estimating and improving protein interaction error rates. *Proc IEEE Comput Syst Bioinform Conf*, pages 216–223, 2004.
- [189] A C Gavin, M Bosche, R Krause, P Grandi, M Marzioch, A Bauer, J Schultz, J M Rick, A M Michon, C M Cruciat, M Remor, C Hofert, M Schelder, M Brajenovic, H Ruffner, A Merino, K Klein, M Hudak, D Dickson, T Rudi, V Gnau, A Bauch, S Bastuck, B Huhse, C Leutwein, MA Heurtier, RR Copley, A Edelmann, E Querfurth, V Rybin, G Drewes, M Raida, T Bouwmeester, P Bork, B Seraphin, B Kuster, G Neubauer, and G Superti-Furga. Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature*, 415(6868):141–147, 2002.
- [190] Y Ho, A Gruhler, A Heilbut, G D Bader, L Moore, S L Adams, A Millar, P Taylor, K Bennett, K Boutilier, L Yang, C Wolting, I Donaldson, S Schandorff, J Shewnarane, M Vo, J Taggart, M Goudreault, B Muskut, C Alfarano, D Dewar, Z Lin, K Michalickova, A R Willems, H Sassi, P A Nielsen, K J Rasmussen, J R Andersen, L E Johansen, L H Hansen, H Jespersen, A Podtelejnikov, E Nielsen, J Crawford, V Poulsen, B D Sorensen, J Matthiesen, R C Hendrickson, F Gleeson, T Pawson, M F Moran, D Durocher, M Mann, C W Hogue, D Figeys, and M Tyers. Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. *Nature*, 415(6868):180–183, Jan 2002.
- [191] G Butland, J M Peregrín-Alvarez, J Li, W Yang, X Yang, V Canadien, A Starostine, D Richards, B Beattie, N Krogan, M Davey, J Parkinson, J Greenblatt, and A Emili. Interaction network containing conserved and essential protein complexes in *Escherichia coli*. *Nature*, 433(7025):531–537, Feb 2005.
- [192] 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 Res*, 16(5):686–691, May 2006.

- [193] 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. *Mol Syst Biol*, 3:89, 2007.
- [194] J R Hutchins, Y Toyoda, B Hegemann, I Poser, J K Hériché, M M Sykora, M Augsburg, O Hudecz, B A Buschhorn, J Bulkescher, C Conrad, D Comartin, A Schleiffer, M Sarov, A Pozniakovsky, M M Slabicki, S Schloissnig, I Steinmacher, M Leuschner, A Ssykor, S Lawo, L Pelletier, H Stark, K Nasmyth, J Ellenberg, R Durbin, F Buchholz, K Mechtler, A A Hyman, and J M Peters. Systematic analysis of human protein complexes identifies chromosome segregation proteins. *Science*, 328(5978):593–599, Apr 2010.
- [195] G Drewes and T Bouwmeester. Global approaches to protein-protein interactions. *Curr Opin Cell Biol*, 15(2):199–205, Apr 2003.
- [196] Z Dezso, Z N Oltvai, and A L Barabási. Bioinformatics analysis of experimentally determined protein complexes in the yeast *Saccharomyces cerevisiae*. *Genome Res*, 13(11):2450–2454, Nov 2003.
- [197] J Gagneur, L David, and L M Steinmetz. Capturing cellular machines by systematic screens of protein complexes. *Trends Microbiol*, 14(8):336–339, Aug 2006.
- [198] T T Soong, K O Wrzeszczynski, and B Rost. Physical protein-protein interactions predicted from microarrays. *Bioinformatics*, 24(22):2608–2614, Nov 2008.
- [199] F Markowetz. How to understand the cell by breaking it: network analysis of gene perturbation screens. *PLoS Comput Biol*, 6(2):e1000655, 2010.
- [200] Y Qi, Y Suhail, Y Y Lin, J D Boeke, and J S Bader. Finding friends and enemies in an enemies-only network: a graph diffusion kernel for predicting novel genetic interactions and co-complex membership from yeast genetic interactions. *Genome Res*, 18(12):1991–2004, Dec 2008.
- [201] M Schena, D Shalon, R W Davis, and P O Brown. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, 270(5235):467–470, Oct 1995.
- [202] J L DeRisi, V R Iyer, and P O Brown. Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science*, 278(5338):680–686, Oct 1997.
- [203] K P White, S A Rifkin, P Hurban, and D S Hogness. Microarray analysis of *Drosophila* development during metamorphosis. *Science*, 286(5447):2179–2184, Dec 1999.
- [204] C H Jen, I W Manfield, I Michalopoulos, J W Pinney, W G Willats, P M Gilmartin, and D R Westhead. The *Arabidopsis* co-expression tool (ACT): a WWW-based tool and database for microarray-based gene expression analysis. *Plant J*, 46(2):336–348, Apr 2006.
- [205] D J Lockhart, H Dong, M C Byrne, M T Follettie, M V Gallo, M S Chee, M Mittmann, C Wang, M Kobayashi, H Horton, and E L Brown. Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nat Biotechnol*, 14(13):1675–1680, Dec 1996.
- [206] R Edgar, M Domrachev, and A E Lash. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res*, 30(1):207–210, Jan 2002.
- [207] V Di Gesù, R Giancarlo, G Lo Bosco, A Raimondi, and D Scaturro. GenClust: a genetic algorithm for clustering gene expression data. *BMC Bioinformatics*, 6:289, 2005.
- [208] A Prelić, S Bleuler, P Zimmermann, A Wille, P Bühlmann, W Gruissem, L Hennig, L Thiele, and E Zitzler. A systematic comparison and evaluation of biclustering methods for gene expression data. *Bioinformatics*, 22(9):1122–1129, May 2006.
- [209] C Cano, L Adarve, J López, and A Blanco. Possibilistic approach for biclustering microarray data. *Comput Biol Med*, 37(10):1426–1436, Oct 2007.

- [210] K Y Yeung and W L Ruzzo. Principal component analysis for clustering gene expression data. *Bioinformatics*, 17(9):763–774, Sep 2001.
- [211] K Y Yeung, M Medvedovic, and R E Bumgarner. Clustering gene-expression data with repeated measurements. *Genome Biol*, 4(5):R34, 2003.
- [212] M Medvedovic, K Y Yeung, and R E Bumgarner. Bayesian mixture model based clustering of replicated microarray data. *Bioinformatics*, 20(8):1222–1232, May 2004.
- [213] K Y Yeung, D R Haynor, and W L Ruzzo. Validating clustering for gene expression data. *Bioinformatics*, 17(4):309–318, Apr 2001.
- [214] B Samuel Lattimore, S van Dongen, and M J Crabbe. GeneMCL in microarray analysis. *Comput Biol Chem*, 29(5):354–359, Oct 2005.
- [215] R Sharan and R Shamir. CLICK: a clustering algorithm with applications to gene expression analysis. *Proc Int Conf Intell Syst Mol Biol*, 8:307–316, 2000.
- [216] L L Elo, H Järvenpää, M Oresic, R Lahesmaa, and T Aittokallio. Systematic construction of gene co-expression networks with applications to human T helper cell differentiation process. *Bioinformatics*, 23(16):2096–2103, Aug 2007.
- [217] T R Golub, D K Slonim, P Tamayo, C Huard, M Gaasenbeek, J P Mesirov, H Coller, M L Loh, J R Downing, M A Caligiuri, C D Bloomfield, and E S Lander. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, 286(5439):531–537, Oct 1999.
- [218] O Troyanskaya, M Cantor, G Sherlock, P Brown, T Hastie, R Tibshirani, D Botstein, and R B Altman. Missing value estimation methods for DNA microarrays. *Bioinformatics*, 17(6):520–525, Jun 2001.
- [219] P J Park, M Pagano, and M Bonetti. A nonparametric scoring algorithm for identifying informative genes from microarray data. *Pac Symp Biocomput*, pages 52–63, 2001.
- [220] R Aragues, C Sander, and B Oliva. Predicting cancer involvement of genes from heterogeneous data. *BMC Bioinformatics*, 9:172, 2008.
- [221] J Chen and B Yuan. Detecting functional modules in the yeast protein-protein interaction network. *Bioinformatics*, 22(18):2283–2290, Sep 2006.
- [222] Y Chen and D Xu. Genome-scale protein function prediction in yeast *Saccharomyces cerevisiae* through integrating multiple sources of high-throughput data. *Pac Symp Biocomput*, pages 471–482, 2005.
- [223] L Franke, H Bakel, L Fokkens, E D de Jong, M Egmont-Petersen, and C Wijmenga. Reconstruction of a functional human gene network, with an application for prioritizing positional candidate genes. *Am J Hum Genet*, 78(6):1011–1025, Jun 2006.
- [224] T Joshi, Y Chen, J M Becker, N Alexandrov, and D Xu. Genome-scale gene function prediction using multiple sources of high-throughput data in yeast *Saccharomyces cerevisiae*. *OMICS*, 8(4):322–333, 2004.
- [225] I A Maraziotis, K Dimitrakopoulou, and A Bezerianos. An *in silico* method for detecting overlapping functional modules from composite biological networks. *BMC Syst Biol*, 2(1):93, Nov 2008.
- [226] A Bossi and B Lehner. Tissue specificity and the human protein interaction network. *Mol Syst Biol*, 5:260, 2009.
- [227] S Tornow and H W Mewes. Functional modules by relating protein interaction networks and gene expression. *Nucleic Acids Res*, 31(21):6283–6289, Nov 2003.
- [228] K Y Yeung, M Medvedovic, and R E Bumgarner. From co-expression to co-regulation: how many microarray experiments do we need? *Genome Biol*, 5(7):R48, 2004.
- [229] L F Stevenson, B K Kennedy, and E Harlow. A large-scale overexpression screen in *Saccharomyces cerevisiae* identifies previously uncharacterized cell cycle genes. *Proc Natl Acad Sci U S A*, 98(7):3946–3951, Mar 2001.

- [230] A Baudin, O Ozier-Kalogeropoulos, A Denouel, F Lacroute, and C Cullin. A simple and efficient method for direct gene deletion in *Saccharomyces cerevisiae*. *Nucleic Acids Res*, 21(14):3329–3330, Jul 1993.
- [231] S Mnaimneh, A P Davierwala, J Haynes, J Moffat, W T Peng, W Zhang, X Yang, J Pootoolal, G Chua, A Lopez, M Trocheset, D Morse, N J Krogan, S L Hiley, Z Li, Q Morris, J Grigull, N Mitsakakis, C J Roberts, J F Greenblatt, C Boone, C A Kaiser, B J Andrews, and T R Hughes. Exploration of essential gene functions via titratable promoter alleles. *Cell*, 118(1):31–44, Jul 2004.
- [232] A P Davierwala, J Haynes, Z Li, R L Brost, M D Robinson, L Yu, S Mnaimneh, H Ding, H Zhu, Y Chen, X Cheng, G W Brown, C Boone, B J Andrews, and T R Hughes. The synthetic genetic interaction spectrum of essential genes. *Nat Genet*, 37(10):1147–1152, Oct 2005.
- [233] Z Li, F J Vizeacoumar, S Bahr, J Li, J Warringer, F S Vizeacoumar, R Min, B Vandersluis, J Bellay, M Devit, J A Fleming, A Stephens, J Haase, Z Y Lin, A Baryshnikova, H Lu, Z Yan, K Jin, S Barker, A Datti, G Giaever, C Nislow, C Bulawa, C L Myers, M Costanzo, A C Gingras, Z Zhang, A Blomberg, K Bloom, B Andrews, and C Boone. Systematic exploration of essential yeast gene function with temperature-sensitive mutants. *Nat Biotechnol*, 29(4):361–367, Apr 2011.
- [234] G Giaever, A M Chu, L Ni, C Connelly, L Riles, S Véronneau, S Dow, A Lucau-Danila, K Anderson, B André, A P Arkin, A Astromoff, M El-Bakkoury, R Bangham, R Benito, S Brachat, S Campanaro, M Curtiss, K Davis, A Deutschbauer, K D Entian, P Flaherty, F Foury, D J Garfinkel, M Gerstein, D Gotte, U Güldener, J H Hegemann, S Hempel, Z Herman, D F Jaramillo, D E Kelly, S L Kelly, P Kötter, D LaBonte, D C Lamb, N Lan, H Liang, H Liao, L Liu, C Luo, M Lussier, R Mao, P Menard, S L Ooi, J L Revuelta, C J Roberts, M Rose, P Ross-Macdonald, B Scherens, G Schimmack, B Shafer, D D Shoemaker, S Sookhai-Mahadeo, R K Storms, J N Strathern, G Valle, M Voet, G Volckaert, C Y Wang, T R Ward, J Wilhelmy, E A Winzeler, Y Yang, G Yen, E Youngman, K Yu, H Bussey, J D Boeke, M Snyder, P Philippsen, R W Davis, and M Johnston. Functional profiling of the *Saccharomyces cerevisiae* genome. *Nature*, 418(6896):387–391, Jul 2002.
- [235] 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 Troyanskaya, T Ideker, K Dolinski, N N Batada, and M Tyers. Comprehensive curation and analysis of global interaction networks in *Saccharomyces cerevisiae*. *J Biol*, 5(4):11, 2006.
- [236] D M Evans, J Marchini, A P Morris, and L R Cardon. Two-stage two-locus models in genome-wide association. *PLoS Genet*, 2(9):e157, Sep 2006.
- [237] A Bender and J R Pringle. Use of a screen for synthetic lethal and multicopy suppressor mutants to identify two new genes involved in morphogenesis in *Saccharomyces cerevisiae*. *Mol Cell Biol*, 11(3):1295–1305, Mar 1991.
- [238] T Wang and A Bretscher. Mutations synthetically lethal with *tpm1 δ* lie in genes involved in morphogenesis. *Genetics*, 147(4):1595–1607, Dec 1997.
- [239] J R Mullen, V Kaliraman, S S Ibrahim, and S J Brill. Requirement for three novel protein complexes in the absence of the Sgs1 DNA helicase in *Saccharomyces cerevisiae*. *Genetics*, 157(1):103–118, Jan 2001.
- [240] A Baryshnikova, M Costanzo, S Dixon, F J Vizeacoumar, C L Myers, B Andrews, and C Boone. Synthetic genetic array (SGA) analysis in *Saccharomyces cerevisiae* and *Schizosaccharomyces pombe*. *Methods Enzymol*, 470:145–179, 2010.
- [241] P B Meluh, X Pan, D S Yuan, C Tiffany, O Chen, S Sookhai-Mahadeo, X Wang, B D Peyser, R Irizarry, F A Spencer, and J D Boeke. Analysis of genetic interactions on a genome-wide scale in budding yeast: diploid-based synthetic lethality analysis by microarray. *Methods Mol Biol*, 416:221–247, 2008.
- [242] X Pan, D S Yuan, S L Ooi, X Wang, S Sookhai-Mahadeo, P Meluh, and J D Boeke. dSLAM analysis of genome-wide genetic interactions in *Saccharomyces cerevisiae*. *Methods*, 41(2):206–221, Feb 2007.
- [243] X Pan, P Ye, D S Yuan, X Wang, J S Bader, and J D Boeke. A DNA integrity network in the yeast *Saccharomyces cerevisiae*. *Cell*, 124(5):1069–1081, Mar 2006.

- [244] S R Collins, K M Miller, N L Maas, A Roguev, J Fillingham, C S Chu, M Schuldiner, M Gebbia, J Recht, M Shales, H Ding, H Xu, J Han, K Ingvarsdottir, B Cheng, B Andrews, C Boone, S L Berger, P Hieter, Z Zhang, G W Brown, C J Ingles, A Emili, C D Allis, D P Toczyski, J S Weissman, J F Greenblatt, and N J Krogan. Functional dissection of protein complexes involved in yeast chromosome biology using a genetic interaction map. *Nature*, 446(7137):806–810, Apr 2007.
- [245] S L Ooi, D D Shoemaker, and J D Boeke. DNA helicase gene interaction network defined using synthetic lethality analyzed by microarray. *Nat Genet*, 35(3):277–286, Nov 2003.
- [246] S J Dixon, Y Fedyshyn, J L Koh, T S Prasad, C Chahwan, G Chua, K Toufighi, A Baryshnikova, J Hayles, K L Hoe, D U Kim, H O Park, C L Myers, A Pandey, D Durocher, B J Andrews, and C Boone. Significant conservation of synthetic lethal genetic interaction networks between distantly related eukaryotes. *Proc Natl Acad Sci U S A*, 105(43):16653–16658, Oct 2008.
- [247] O Dror, D Schneidman-Duhovny, A Shulman-Peleg, R Nussinov, H J Wolfson, and R Sharan. Structural similarity of genetically interacting proteins. *BMC Syst Biol*, 2(1):69, Jul 2008.
- [248] R Kelley and T Ideker. Systematic interpretation of genetic interactions using protein networks. *Nat Biotechnol*, 23(5):561–566, May 2005.
- [249] C Boone, H Bussey, and B J Andrews. Exploring genetic interactions and networks with yeast. *Nat Rev Genet*, 8(6):437–449, Jun 2007.
- [250] I Ulitsky and R Shamir. Pathway redundancy and protein essentiality revealed in the *Saccharomyces cerevisiae* interaction networks. *Mol Syst Biol*, 3:104, 2007.
- [251] J H Thomas. Thinking about genetic redundancy. *Trends Genet*, 9(11):395–399, Nov 1993.
- [252] M M Garner and A Revzin. A gel electrophoresis method for quantifying the binding of proteins to specific DNA regions: application to components of the *Escherichia coli* lactose operon regulatory system. *Nucleic Acids Res*, 9(13):3047–3060, Jul 1981.
- [253] C Seguin and D H Hamer. Regulation *in vitro* of metallothionein gene binding factors. *Science*, 235(4794):1383–1387, Mar 1987.
- [254] D J Galas and A Schmitz. DNase footprinting: a simple method for the detection of protein-DNA binding specificity. *Nucleic Acids Res*, 5(9):3157–3170, Sep 1978.
- [255] T I Lee, N J Rinaldi, F Robert, D T Odom, Z Bar-Joseph, G K Gerber, N M Hannett, C T Harbison, C M Thompson, I Simon, J Zeitlinger, E G Jennings, H L Murray, D B Gordon, B Ren, J J Wyrick, J B Tagne, T L Volkert, E Fraenkel, D K Gifford, and R A Young. Transcriptional regulatory networks in *Saccharomyces cerevisiae*. *Science*, 298(5594):799–804, Oct 2002.
- [256] M H Kuo and C D Allis. *In vivo* cross-linking and immunoprecipitation for studying dynamic Protein:DNA associations in a chromatin environment. *Methods*, 19(3):425–433, Nov 1999.
- [257] G R Mishra, M Suresh, K Kumaran, N Kannabiran, S Suresh, P Bala, K Shivakumar, N Anuradha, R Reddy, T M Raghavan, S Menon, G Hanumanthu, M Gupta, S Upendran, S Gupta, M Mahesh, B Jacob, P Mathew, P Chatterjee, K S Arun, S Sharma, K N Chandrika, N Deshpande, K Palvankar, R Raghavath, R Krishnakanth, H Karathia, B Rekha, R Nayak, G Vishnupriya, H G Kumar, M Nagini, G S Kumar, R Jose, P Deepthi, S S Mohan, T K Gandhi, H C Harsha, K S Deshpande, M Sarker, T S Prasad, and A Pandey. Human protein reference database–2006 update. *Nucleic Acids Res*, 34(Database issue):411–414, Jan 2006.
- [258] N Daraselia, A Yuryev, S Egorov, S Novichkova, A Nikitin, and I Mazo. Extracting human protein interactions from MEDLINE using a full-sentence parser. *Bioinformatics*, 20(5):604–611, Mar 2004.
- [259] H M Müller, E E Kenny, and P W Sternberg. Textpresso: an ontology-based information retrieval and extraction system for biological literature. *PLoS Biol*, 2(11):e309, Nov 2004.
- [260] A K Ramani, R C Bunescu, R J Mooney, and E M Marcotte. Consolidating the set of known human protein-protein interactions in preparation for large-scale mapping of the human interactome. *Genome Biol*, 6(5):R40, 2005.

- [261] J Saric, L J Jensen, R Ouzounova, I Rojas, and P Bork. Extraction of regulatory gene/protein networks from medline. *Bioinformatics*, 22(6):645–650, Mar 2006.
- [262] B Lehne and T Schlitt. Protein-protein interaction databases: keeping up with growing interactomes. *Hum Genomics*, 3(3):291–297, Apr 2009.
- [263] G D Bader, M P Cary, and C Sander. Pathguide: a pathway resource list. *Nucleic Acids Res*, 34(Database issue):504–506, Jan 2006.
- [264] J D Wren and A Bateman. Databases, data tombs and dust in the wind. *Bioinformatics*, 24(19):2127–2128, Oct 2008.
- [265] S Bureeva, S Zvereva, V Romanov, and T Serebryiskaya. Manual annotation of protein interactions. *Methods Mol Biol*, 563:75–95, 2009.
- [266] D Howe, M Costanzo, P Fey, T Gojobori, L Hannick, W Hide, D P Hill, R Kania, M Schaeffer, S St Pierre, S Twigger, O White, and S Y Rhee. Big data: the future of biocuration. *Nature*, 455(7209):47–50, Sep 2008.
- [267] A Rzhetsky, H Shatkay, and W J Wilbur. How to get the most out of your curation effort. *PLoS Comput Biol*, 5(5):e1000391, May 2009.
- [268] M Shimoyama, G T Hayman, S J Laulederkind, R Nigam, T F Lowry, V Petri, J R Smith, S J Wang, D H Munzenmaier, M R Dwinell, S N Twigger, H J Jacob, and RGD Team. The rat genome database curators: who, what, where, why. *PLoS Comput Biol*, 5(11):e1000582, Nov 2009.
- [269] 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. *Nat Methods*, 6(12):860–861, Dec 2009.
- [270] N Salimi and R Vita. The biocurator: connecting and enhancing scientific data. *PLoS Comput Biol*, 2(10):e125, Oct 2006.
- [271] D Landsman, R Gentleman, J Kelso, and B. F. Francis Ouellette. DATABASE: a new forum for biological databases and curation. *Database*, 2009:bap002, March 2009.
- [272] C Andorf, D Dobbs, and V Honavar. Exploring inconsistencies in genome-wide protein function annotations: a machine learning approach. *BMC Bioinformatics*, 8:284, 2007.
- [273] O Bodenreider and R Stevens. Bio-ontologies: current trends and future directions. *Brief Bioinform*, 7(3):256–274, Sep 2006.
- [274] C Goble, R Stevens, D Hull, K Wolstencroft, and R Lopez. Data curation + process curation=data integration + science. *Brief Bioinform*, 9(6):506–517, Nov 2008.
- [275] M E Cusick, H Yu, A Smolyar, K Venkatesan, A R Carvunis, N Simonis, J F Rual, H Borick, P Braun, M Dreze, J Vandenhoute, M Galli, J Yazaki, D E Hill, J R Ecker, F P Roth, and M Vidal. Literature-curated protein interaction datasets. *Nat Methods*, 6(1):39–46, Jan 2009.
- [276] C Stark, BJ Breitkreutz, T Reguly, L Boucher, A Breitkreutz, and M Tyers. BioGRID: a general repository for interaction datasets. *Nucl Acids Res*, 34(suppl 1):D535–539, 2006.
- [277] H Ogata, S Goto, K Sato, W Fujibuchi, H Bono, and M Kanehisa. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res*, 27(1):29–34, Jan 1999.
- [278] I Xenarios, L Salwinski, X J Duan, P Higney, S M Kim, and D Eisenberg. DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions. *Nucleic Acids Res*, 30(1):303–305, Jan 2002.
- [279] 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. *Nucl Acids Res*, 30:31–34, 2002.

- [280] P V Luc and P Tempst. PINdb: a database of nuclear protein complexes from human and yeast. *Bioinformatics*, 20(9):1413–1415, Jun 2004.
- [281] E Wingender, X Chen, E Fricke, R Geffers, R Hehl, I Liebich, M Krull, V Matys, H Michael, R Ohnhäuser, M Prüss, F Schacherer, S Thiele, and S Urbach. The TRANSFAC system on gene expression regulation. *Nucleic Acids Res*, 29(1):281–283, Jan 2001.
- [282] G Dellaire, R Farrall, and W A Bickmore. The Nuclear Protein Database (NPD): sub-nuclear localisation and functional annotation of the nuclear proteome. *Nucleic Acids Res*, 31(1):328–330, Jan 2003.
- [283] J M Cherry, C Adler, C Ball, S A Chervitz, S S Dwight, E T Hester, Y Jia, G Juvik, T Roe, M Schroeder, S Weng, and D Botstein. SGD: Saccharomyces Genome Database. *Nucleic Acids Res*, 26(1):73–79, Jan 1998.
- [284] J Yu, S Pacifico, G Liu, and R L Finley. DroID: the Drosophila Interactions Database, a comprehensive resource for annotated gene and protein interactions. *BMC Genomics*, 9:461, 2008.
- [285] A Rogers, I Antoshechkin, T Bieri, D Blasiar, C Bastiani, P Canaran, J Chan, W J Chen, P Davis, J Fernandes, T J Fiedler, M Han, T W Harris, R Kishore, R Lee, S McKay, H M Müller, C Nakamura, P Ozersky, A Petcherski, G Schindelman, E M Schwarz, W Spooner, M A Tuli, K Van Auken, D Wang, X Wang, G Williams, K Yook, R Durbin, L D Stein, J Spieth, and P W Sternberg. Wormbase 2007. *Nucleic Acids Res*, 36(Database issue):612–617, Jan 2008.
- [286] S Peri, J D Navarro, R Amanchy, T Z Kristiansen, C K Jonnalagadda, V Surendranath, V Niranjana, B Muthusamy, T K Gandhi, M Gronborg, N Ibarrola, N Deshpande, K Shanker, H N Shivashankar, B P Rashmi, M A Ramya, Z Zhao, K N Chandrika, N Padma, H C Harsha, A J Yatish, M P Kavitha, M Menezes, D R Choudhury, S Suresh, N Ghosh, R Saravana, S Chandran, S Krishna, M Joy, S K Anand, V Madavan, A Joseph, G W Wong, W P Schiemann, S N Constantinescu, L Huang, R Khosravi-Far, H Steen, M Tewari, S Ghaffari, G C Blobe, C V Dang, J G Garcia, J Pevsner, O N Jensen, P Roepstorff, K S Deshpande, A M Chinnaiyan, A Hamosh, A Chakravarti, and A Pandey. Development of Human Protein Reference Database as an initial platform for approaching systems biology in humans. *Genome Res*, 13(10):2363–2371, Oct 2003.
- [287] B J Breitkreutz, C Stark, T Reguly, L Boucher, A Breitkreutz, M Livstone, R Oughtred, D H Lackner, J Bähler, V Wood, K Dolinski, and M Tyers. The BioGRID interaction database: 2008 update. *Nucleic Acids Res*, 36(Database issue):637–640, Jan 2008.
- [288] R Chowdhary, J Zhang, and J S Liu. Bayesian inference of protein-protein interactions from biological literature. *Bioinformatics*, 25(12):1536–1542, Jun 2009.
- [289] P M Kim, A Sboner, Y Xia, and M Gerstein. The role of disorder in interaction networks: a structural analysis. *Mol Syst Biol*, 4:179, 2008.
- [290] T Milenković and N Przulj. Uncovering biological network function via graphlet degree signatures. *Cancer Inform*, 6:257–273, 2008.
- [291] 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, Dec 2009.
- [292] K Voevodski, S H Teng, and Y Xia. Finding local communities in protein networks. *BMC Bioinformatics*, 10:297, 2009.
- [293] B J Breitkreutz, C Stark, and M Tyers. The GRID: the General Repository for Interaction Datasets. *Genome Biol*, 3(12):R0013, 2002.
- [294] 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 Res*, 39(Database issue):698–704, Jan 2011.
- [295] J M Cherry, C Ball, S Weng, G Juvik, R Schmidt, C Adler, B Dunn, S Dwight, L Riles, R K Mortimer, and D Botstein. Genetic and physical maps of *Saccharomyces cerevisiae*. *Nature*, 387(6632 Suppl):67–73, May 1997.

- [296] S S Dwight, M A Harris, K Dolinski, C A Ball, G Binkley, K R Christie, D G Fisk, L Issel-Tarver, M Schroeder, G Sherlock, A Sethuraman, S Weng, D Botstein, and J M Cherry. *Saccharomyces cerevisiae* Database (SGD) provides secondary gene annotation using the Gene Ontology (GO). *Nucleic Acids Res*, 30(1):69–72, Jan 2002.
- [297] 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, J E Hirschman, B C Hitz, C J Krieger, M S Livstone, S R Miyasato, R S Nash, R Oughtred, M S Skrzypek, S Weng, E D Wong, K K Zhu, K Dolinski, D Botstein, and J M Cherry. Gene Ontology annotations at SGD: new data sources and annotation methods. *Nucleic Acids Res*, 36(Database issue):577–581, Jan 2008.
- [298] S R Engel, R Balakrishnan, G Binkley, K R Christie, M C Costanzo, S S Dwight, D G Fisk, J E Hirschman, B C Hitz, E L Hong, C J Krieger, M S Livstone, S R Miyasato, R Nash, R Oughtred, J Park, M S Skrzypek, S Weng, E D Wong, K Dolinski, D Botstein, and J M Cherry. *Saccharomyces* Genome Database provides mutant phenotype data. *Nucleic Acids Res*, 38(Database issue):433–436, Jan 2010.
- [299] C A Ball, K Dolinski, S S Dwight, M A Harris, L Issel-Tarver, A Kasarskis, C R Scafe, G Sherlock, G Binkley, H Jin, M Kaloper, S D Orr, M Schroeder, S Weng, Y Zhu, D Botstein, and J M Cherry. Integrating functional genomic information into the *Saccharomyces* genome database. *Nucleic Acids Res*, 28(1):77–80, Jan 2000.
- [300] S F Altschul, W Gish, W Miller, E W Myers, and D J Lipman. Basic local alignment search tool. *J Mol Biol*, 215(3):403–410, Oct 1990.
- [301] J E Hirschman, R Balakrishnan, K R Christie, M C Costanzo, S S Dwight, S R Engel, D G Fisk, E L Hong, M S Livstone, R Nash, J Park, R Oughtred, M Skrzypek, B Starr, C L Theesfeld, J Williams, R Andrada, G Binkley, Q Dong, C Lane, S Miyasato, A Sethuraman, M Schroeder, M K Thanawala, S Weng, K Dolinski, D Botstein, and J M Cherry. Genome Snapshot: a new resource at the *Saccharomyces* Genome Database (SGD) presenting an overview of the *Saccharomyces cerevisiae* genome. *Nucleic Acids Res*, 34(Database issue):442–445, Jan 2006.
- [302] R Nash, S Weng, B Hitz, R Balakrishnan, K R Christie, M C Costanzo, S S Dwight, S R Engel, D G Fisk, J E Hirschman, E L Hong, M S Livstone, R Oughtred, J Park, M Skrzypek, C L Theesfeld, G Binkley, Q Dong, C Lane, S Miyasato, A Sethuraman, M Schroeder, K Dolinski, D Botstein, and J M Cherry. Expanded protein information at SGD: new pages and proteome browser. *Nucleic Acids Res*, 35(Database issue):468–471, Jan 2007.
- [303] C Sanchez, C Lachaize, F Janody, B Bellon, L Röder, J Euzenat, F Rechenmann, and B Jacq. Grasping at molecular interactions and genetic networks in *Drosophila melanogaster* using FlyNets, an internet database. *Nucleic Acids Res*, 27(1):89–94, Jan 1999.
- [304] J H Hernandez-Toro, C Prieto, and J D Rivas. APID2NET: unified interactome graphic analyzer. *Bioinformatics*, 23(18):2495–2497, Jul 2007.
- [305] M Dreze, D Monachello, C Lurin, M E Cusick, D E Hill, M Vidal, and P Braun. High-quality binary interactome mapping. *Methods Enzymol*, 470:281–315, 2010.
- [306] 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. *Nat Methods*, 6(1):83–90, Jan 2009.
- [307] J R Hughes, A M Meireles, K H Fisher, A Garcia, P R Antrobus, A Wainman, N Zitzmann, C Deane, H Ohkura, and J G Wakefield. A microtubule interactome: complexes with roles in cell cycle and mitosis. *PLoS Biol*, 6(4):e98, Apr 2008.
- [308] T E Shutt and G S Shadel. Expanding the mitochondrial interactome. *Genome Biol*, 8(2):203, 2007.
- [309] M Oeffinger, K E Wei, R Rogers, J A DeGrasse, B T Chait, J D Aitchison, and M P Rout. Comprehensive analysis of diverse ribonucleoprotein complexes. *Nat Methods*, 4(11):951–956, Nov 2007.

- [310] M Brehme and M Vidal. A global protein-lipid interactome map. *Mol Syst Biol*, 6:443, Nov 2010.
- [311] B Andreopoulos, C Winter, D Labudde, and M Schroeder. Triangle network motifs predict complexes by complementing high-error interactomes with structural information. *BMC Bioinformatics*, 10:196, 2009.
- [312] S H Yook, Z N Oltvai, and A L Barabási. Functional and topological characterization of protein interaction networks. *Proteomics*, 4(4):928–942, Apr 2004.
- [313] E Sprinzak, S Sattath, and H Margalit. How reliable are experimental protein-protein interaction data? *J Mol Biol*, 327(5):919–923, Apr 2003.
- [314] C L Tucker, J F Gera, and P Uetz. Towards an understanding of complex protein networks. *Trends Cell Biol*, 11(3):102–106, Mar 2001.
- [315] P Legrain, J Wojcik, and J M Gauthier. Protein-protein interaction maps: a lead towards cellular functions. *Trends Genet*, 17(6):346–352, Jun 2001.
- [316] A Grigoriev. On the number of protein-protein interactions in the yeast proteome. *Nucleic Acids Res*, 31(14):4157–4161, Jul 2003.
- [317] 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(1):605, Dec 2010.
- [318] B A Shoemaker and A R Panchenko. Deciphering protein-protein interactions. Part II. Computational methods to predict protein and domain interaction partners. *PLoS Comput Biol*, 3(4):e43, Apr 2007.
- [319] A Valencia and F Pazos. Computational methods for the prediction of protein interactions. *Curr Opin Struct Biol*, 12(3):368–373, Jun 2002.
- [320] A Goffeau, B G Barrell, H Bussey, R W Davis, B Dujon, H Feldmann, F Galibert, J D Hoheisel, C Jacq, M Johnston, E J Louis, H W Mewes, Y Murakami, P Philippsen, H Tettelin, and S G Oliver. Life with 6000 genes. *Science*, 274(5287):563–567, Oct 1996.
- [321] M Lappe and L Holm. Unraveling protein interaction networks with near-optimal efficiency. *Nat Biotechnol*, 22(1):98–103, Jan 2004.
- [322] L Skrabanek, H K Saini, G D Bader, and A J Enright. Computational prediction of protein-protein interactions. *Mol Biotechnol*, 38(1):1–17, Jan 2008.
- [323] M D McDowall, M S Scott, and G J Barton. PIPs: human protein-protein interaction prediction database. *Nucleic Acids Res*, 37(Database issue):651–656, Jan 2009.
- [324] J Y Chen, S Mamidipalli, and T Huan. HAPPI: an online database of comprehensive human annotated and predicted protein interactions. *BMC Genomics*, 10(Suppl 1):S16, 2009.
- [325] K R Brown and I Jurisica. Online predicted human interaction database. *Bioinformatics*, 21(9):2076–2082, May 2005.
- [326] C von Mering, M Huynen, D Jaeggi, S Schmidt, P Bork, and B Snel. STRING: a database of predicted functional associations between proteins. *Nucleic Acids Res*, 31(1):258–261, Jan 2003.
- [327] T W Huang, A C Tien, W S Huang, Y C Lee, C L Peng, H H Tseng, C Y Kao, and C Y Huang. POINT: a database for the prediction of protein-protein interactions based on the orthologous interactome. *Bioinformatics*, 20(17):3273–3276, Nov 2004.
- [328] J C Mellor, I Yanai, K H Clodfelter, J Mintseris, and C DeLisi. Predictome: a database of putative functional links between proteins. *Nucleic Acids Res*, 30(1):306–309, Jan 2002.
- [329] T Dandekar, B Snel, M Huynen, and P Bork. Conservation of gene order: a fingerprint of proteins that physically interact. *Trends Biochem Sci*, 23(9):324–328, Sep 1998.
- [330] P M Bowers, M Pellegrini, M J Thompson, J Fierro, T O Yeates, and D Eisenberg. Prolinks: a database of protein functional linkages derived from coevolution. *Genome Biol*, 5(5):R35, 2004.

- [331] R Overbeek, M Fonstein, M D'Souza, G D Pusch, and N Maltsev. The use of gene clusters to infer functional coupling. *Proc Natl Acad Sci U S A*, 96(6):2896–2901, Mar 1999.
- [332] M Huynen, B Snel, W Lathe, and P Bork. Predicting protein function by genomic context: quantitative evaluation and qualitative inferences. *Genome Res*, 10(8):1204–1210, Aug 2000.
- [333] E M 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, Jul 1999.
- [334] A J Enright, I Iliopoulos, N C Kyrpides, and C A Ouzounis. Protein interaction maps for complete genomes based on gene fusion events. *Nature*, 402(6757):86–90, Nov 1999.
- [335] I Yanai, A Derti, and C DeLisi. Genes linked by fusion events are generally of the same functional category: a systematic analysis of 30 microbial genomes. *Proc Natl Acad Sci U S A*, 98(14):7940–7945, Jul 2001.
- [336] J Park, M Lappe, and S A Teichmann. Mapping protein family interactions: intramolecular and intermolecular protein family interaction repertoires in the PDB and yeast. *J Mol Biol*, 307(3):929–938, Mar 2001.
- [337] M Deng, S Mehta, F Sun, and T Chen. Inferring domain-domain interactions from protein-protein interactions. *Genome Res*, 12(10):1540–1548, Oct 2002.
- [338] E Sprinzak and H Margalit. Correlated sequence-signatures as markers of protein-protein interaction. *J Mol Biol*, 311(4):681–692, Aug 2001.
- [339] S M Gomez and A Rzhetsky. Towards the prediction of complete protein–protein interaction networks. *Pac Symp Biocomput*, pages 413–424, 2002.
- [340] H Mamitsuka. Mining new protein-protein interactions. using a hierarchical latent-variable model to determine the function of a functionally unknown protein. *IEEE Eng Med Biol Mag*, 24(3):103–108, May-Jun 2005.
- [341] M Hayashida, N Ueda, and T Akutsu. A simple method for inferring strengths of protein-protein interactions. *Genome Inform*, 15(1):56–68, 2004.
- [342] I Lee, B Ambaru, P Thakkar, E M Marcotte, and S Y Rhee. Rational association of genes with traits using a genome-scale gene network for *Arabidopsis thaliana*. *Nat Biotechnol*, 28(2):149–156, Feb 2010.
- [343] X W Chen, M Liu, and R Ward. Protein function assignment through mining cross-species protein-protein interactions. *PLoS ONE*, 3(2):e1562, 2008.
- [344] C Andreini, I Bertini, G Cavallaro, L Decaria, and A Rosato. A simple protocol for the comparative analysis of the structure and occurrence of biochemical pathways across superkingdoms. *J Chem Inf Model*, 51(3):730–738, Mar 2011.
- [345] J R Bock and D A Gough. Predicting protein–protein interactions from primary structure. *Bioinformatics*, 17(5):455–460, May 2001.
- [346] B Snel, V van Noort, and M A Huynen. Gene co-regulation is highly conserved in the evolution of eukaryotes and prokaryotes. *Nucleic Acids Res*, 32(16):4725–4731, 2004.
- [347] 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 Res*, 11(12):2120–2126, Dec 2001.
- [348] M Michaut, S Kerrien, L Montecchi-Palazzi, F Chauvat, C Cassier-Chauvat, J C Aude, and P Legrain. InteroPorc: automated inference of highly conserved protein interaction networks. *Bioinformatics*, 24(14):1625–1631, July 2008.
- [349] M Remm, C E Storm, and E L Sonnhammer. Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *J Mol Biol*, 314(5):1041–1052, Dec 2001.

- [350] B Lehner and A G Fraser. A first-draft human protein-interaction map. *Genome Biol*, 5(9):R63, 2004.
- [351] A M Wiles, M Doderer, J Ruan, T T Gu, D Ravi, B Blackman, and A J Bishop. Building and analyzing protein interactome networks by cross-species comparisons. *BMC Syst Biol*, 4(1):36, Mar 2010.
- [352] S Wuchty and J J Ipsaro. A draft of protein interactions in the malaria parasite *P. falciparum*. *J Proteome Res*, 6(4):1461–1470, Apr 2007.
- [353] D Banky, R Ordog, and V Grolmusz. NASCENT: an automatic protein interaction network generation tool for non-model organisms. *Bioinformatics*, 3(8):361–363, 2009.
- [354] F Pazos, J A Ranea, D Juan, and M J Sternberg. Assessing protein co-evolution in the context of the tree of life assists in the prediction of the interactome. *J Mol Biol*, 352(4):1002–1015, Sep 2005.
- [355] J Wu, S Kasif, and C DeLisi. Identification of functional links between genes using phylogenetic profiles. *Bioinformatics*, 19(12):1524–1530, Aug 2003.
- [356] Y I Wolf, I B Rogozin, A S Kondrashov, and E V Koonin. Genome alignment, evolution of prokaryotic genome organization, and prediction of gene function using genomic context. *Genome Res*, 11(3):356–372, Mar 2001.
- [357] 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, Apr 2002.
- [358] T Gaasterland and M A Ragan. Microbial genespaces: phyletic and functional patterns of ORF distribution among prokaryotes. *Microb Comp Genomics*, 3(4):199–217, 1998.
- [359] M Pellegrini, E M Marcotte, M J Thompson, D Eisenberg, and T O Yeates. Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc Natl Acad Sci U S A*, 96(8):4285–4288, Apr 1999.
- [360] D Juan, F Pazos, and A Valencia. High-confidence prediction of global interactomes based on genome-wide coevolutionary networks. *Proc Natl Acad Sci U S A*, 105(3):934–939, Jan 2008.
- [361] T Sato, Y Yamanishi, M Kanehisa, and H Toh. The inference of protein-protein interactions by co-evolutionary analysis is improved by excluding the information about the phylogenetic relationships. *Bioinformatics*, 21(17):3482–3489, Sep 2005.
- [362] E V Koonin, Y I Wolf, and G P Karev. The structure of the protein universe and genome evolution. *Nature*, 420(6912):218–223, Nov 2002.
- [363] P Aloy and R B Russell. Interrogating protein interaction networks through structural biology. *Proc Natl Acad Sci U S A*, 99(9):5896–5901, Apr 2002.
- [364] J Janin, K Henrick, J Moult, L T Eyck, M J Sternberg, S Vajda, I Vakser, and S J Wodak. CAPRI: a Critical Assessment of PRedicted Interactions. *Proteins*, 52(1):2–9, Jul 2003.
- [365] L Lu, A K Arakaki, H Lu, and J Skolnick. Multimeric threading-based prediction of protein-protein interactions on a genomic scale: application to the *Saccharomyces cerevisiae* proteome. *Genome Res*, 13(6A):1146–1154, Jun 2003.
- [366] P Aloy and R B Russell. InterPreTS: protein interaction prediction through tertiary structure. *Bioinformatics*, 19(1):161–162, Jan 2003.
- [367] M Hue, M Riffle, J P Vert, and W S Noble. Large-scale prediction of protein-protein interactions from structures. *BMC Bioinformatics*, 11(1):144, Mar 2010.
- [368] L Lu, H Lu, and J Skolnick. MULTIPROSPECTOR: an algorithm for the prediction of protein-protein interactions by multimeric threading. *Proteins*, 49(3):350–364, Nov 2002.
- [369] G R Smith and M J Sternberg. Prediction of protein-protein interactions by docking methods. *Curr Opin Struct Biol*, 12(1):28–35, Feb 2002.
- [370] A S Aytuna, A Gursay, and O Keskin. Prediction of protein-protein interactions by combining structure and sequence conservation in protein interfaces. *Bioinformatics*, 21(12):2850–2855, Jun 2005.

- [371] F Pazos and A Valencia. *In silico* two-hybrid system for the selection of physically interacting protein pairs. *Proteins*, 47(2):219–227, May 2002.
- [372] E Andres Leon, I Ezkurdia, B García, A Valencia, and D Juan. EcID. a database for the inference of functional interactions in *E. coli*. *Nucleic Acids Res*, 37(Database issue):629–635, Jan 2009.
- [373] J M Stuart, E Segal, D Koller, and S K Kim. A gene-coexpression network for global discovery of conserved genetic modules. *Science*, 302(5643):249–255, Oct 2003.
- [374] V van Noort, B Snel, and M A Huynen. Predicting gene function by conserved co-expression. *Trends Genet*, 19(5):238–242, May 2003.
- [375] L J Jensen, J Lagarde, C von Mering, and P Bork. ArrayProspector: a web resource of functional associations inferred from microarray expression data. *Nucleic Acids Res*, 32(Web Server issue):445–448, Jul 2004.
- [376] X J Zhou, M C Kao, H Huang, A Wong, J Nunez-Iglesias, M Primig, O M Aparicio, C E Finch, T E Morgan, and W H Wong. Functional annotation and network reconstruction through cross-platform integration of microarray data. *Nat Biotechnol*, 23(2):238–243, Feb 2005.
- [377] C Blaschke, L Hirschman, and A Valencia. Information extraction in molecular biology. *Brief Bioinform*, 3(2):154–165, Jun 2002.
- [378] L Hirschman, J C Park, J Tsujii, L Wong, and C H Wu. Accomplishments and challenges in literature data mining for biology. *Bioinformatics*, 18(12):1553–1561, Dec 2002.
- [379] C Santos, D Eggle, and D J States. Wnt pathway curation using automated natural language processing: combining statistical methods with partial and full parse for knowledge extraction. *Bioinformatics*, 21(8):1653–1658, Apr 2005.
- [380] E M Marcotte, I Xenarios, and D Eisenberg. Mining literature for protein-protein interactions. *Bioinformatics*, 17(4):359–363, Apr 2001.
- [381] B J Stapley and G Benoit. Biobibliometrics: information retrieval and visualization from co-occurrences of gene names in Medline abstracts. *Pac Symp Biocomput*, pages 529–540, 2000.
- [382] J D Wren and H R Garner. Shared relationship analysis: ranking set cohesion and commonalities within a literature-derived relationship network. *Bioinformatics*, 20(2):191–198, Jan 2004.
- [383] T Ono, H Hishigaki, A Tanigami, and T Takagi. Automated extraction of information on protein-protein interactions from the biological literature. *Bioinformatics*, 17(2):155–161, Feb 2001.
- [384] T Sekimizu, H S Park, and J Tsujii. Identifying the interaction between genes and gene products based on frequently seen verbs in medline abstracts. *Genome Inform Ser Workshop Genome Inform*, 9:62–71, 1998.
- [385] A Yuryev, Z Mulyukov, E Kotelnikova, S Maslov, S Egorov, A Nikitin, N Daraselia, and I Mazo. Automatic pathway building in biological association networks. *BMC Bioinformatics*, 7:171, 2006.
- [386] T K Jenssen, A Laegreid, J Komorowski, and E Hovig. A literature network of human genes for high-throughput analysis of gene expression. *Nat Genet*, 28(1):21–28, May 2001.
- [387] T P Mohamed, J G Carbonell, and M K Ganapathiraju. Active learning for human protein-protein interaction prediction. *BMC Bioinformatics*, 11(Suppl 1):S57, 2010.
- [388] Y Qi, Z Bar-Joseph, and J Klein-Seetharaman. Evaluation of different biological data and computational classification methods for use in protein interaction prediction. *Proteins*, 63(3):490–500, May 2006.
- [389] M S Scott and G J Barton. Probabilistic prediction and ranking of human protein-protein interactions. *BMC Bioinformatics*, 8:239, 2007.
- [390] B García-Jiménez, D Juan, I Ezkurdia, E Andrés-León, and A Valencia. Inference of functional relations in predicted protein networks with a machine learning approach. *PLoS ONE*, 5(4):e9969, 2010.

- [391] I Albert and R Albert. Conserved network motifs allow protein-protein interaction prediction. *Bioinformatics*, 20(18):3346–3352, Dec 2004.
- [392] S L Lo, C Z Cai, Y Z Chen, and M C Chung. Effect of training datasets on support vector machine prediction of protein-protein interactions. *Proteomics*, 5(4):876–884, Mar 2005.
- [393] J R Bock and D A Gough. Whole-proteome interaction mining. *Bioinformatics*, 19(1):125–134, Jan 2003.
- [394] A Ben-Hur and W S Noble. Kernel methods for predicting protein-protein interactions. *Bioinformatics*, 21 Suppl 1:38–46, Jun 2005.
- [395] N Lin, B Wu, R Jansen, M Gerstein, and H Zhao. Information assessment on predicting protein-protein interactions. *BMC Bioinformatics*, 5:154, Oct 2004.
- [396] T Sen, A Kloczkowski, and R Jernigan. Functional clustering of yeast proteins from the protein-protein interaction network. *BMC Bioinformatics*, 7:355, 2006.
- [397] T Aho, H Almus, J Matilainen, A Larjo, P Ruusuvuori, K L Aho, T Wilhelm, H Lähdesmäki, A Beyer, M Harju, S Chowdhury, K Leinonen, C Roos, and O Yli-Harja. Reconstruction and validation of RefRec: a global model for the yeast molecular interaction network. *PLoS ONE*, 5(5):e10662, 2010.
- [398] F Markowetz and R Spang. Inferring cellular networks—a review. *BMC Bioinformatics*, 8(Suppl 6):S5, 2007.
- [399] J Dong and S Horvath. Understanding network concepts in modules. *BMC Syst Biol*, 1:24–24, 2007.
- [400] Y Assenov, F Ramírez, S E Schelhorn, T Lengauer, and M Albrecht. Computing topological parameters of biological networks. *Bioinformatics*, 24(2):282–284, Jan 2008.
- [401] R Albert and A L Barabási. Statistical mechanics of complex networks. *Rev. Mod. Phys.*, 74(1):47–97, January 2002.
- [402] S Maslov and K Sneppen. Specificity and stability in topology of protein networks. *Science*, 296(5569):910–913, May 2002.
- [403] S S Shen-Orr, R Milo, S Mangan, and U Alon. Network motifs in the transcriptional regulation network of *Escherichia coli*. *Nat Genet*, 31(1):64–68, May 2002.
- [404] 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, Oct 2002.
- [405] A Vázquez, R Dobrin, D Sergi, J P Eckmann, Z N Oltvai, and A L Barabási. The topological relationship between the large-scale attributes and local interaction patterns of complex networks. *Proc Natl Acad Sci U S A*, 101(52):17940–17945, Dec 2004.
- [406] E Yeger-Lotem, S Sattath, N Kashtan, S Itzkovitz, R Milo, R Y Pinter, U Alon, and H Margalit. Network motifs in integrated cellular networks of transcription-regulation and protein-protein interaction. *Proc Natl Acad Sci U S A*, 101(16):5934–5939, Apr 2004.
- [407] N Przulj, D G Corneil, and I Jurisica. Modeling interactome: scale-free or geometric? *Bioinformatics*, 20(18):3508–3515, Dec 2004.
- [408] R Tanaka. Scale-rich metabolic networks. *Phys Rev Lett*, 94(16):168101, Apr 2005.
- [409] M P Brynildsen, T Y Wu, S S Jang, and J C Liao. Biological network mapping and source signal deduction. *Bioinformatics*, 23(14):1783–1791, May 2007.
- [410] M E Newman. Analysis of weighted networks. *Phys Rev E Stat Nonlin Soft Matter Phys*, 70(5):056131, Nov 2004.
- [411] S Klamt, U U Haus, and F Theis. Hypergraphs and cellular networks. *PLoS Comput Biol*, 5(5):e1000385, May 2009.

- [412] O Mason and M Verwoerd. Graph theory and networks in biology. *IET Syst Biol*, 1(2):89–119, Mar 2007.
- [413] E Chautard, N Thierry-Mieg, and S Ricard-Blum. Interaction networks as a tool to investigate the mechanisms of aging. *Biogerontology*, 11(4):463–473, Mar 2010.
- [414] S H Strogatz. Exploring complex networks. *Nature*, 410(6825):268–276, Mar 2001.
- [415] D Bu, Y Zhao, L Cai, H Xue, X Zhu, H Lu, J Zhang, S Sun, L Ling, N Zhang, G Li, and R Chen. Topological structure analysis of the protein-protein interaction network in budding yeast. *Nucleic Acids Res*, 31(9):2443–2450, May 2003.
- [416] H Jeong, S P Mason, A L Barabási, and Z N Oltvai. Lethality and centrality in protein networks. *Nature*, 411(6833):41–42, May 2001.
- [417] B P Kelley, R Sharan, R M Karp, T Sittler, D E Root, B R Stockwell, and T Ideker. Conserved pathways within bacteria and yeast as revealed by global protein network alignment. *Proc Natl Acad Sci U S A*, 100(20):11394–11399, Sep 2003.
- [418] H Qin, Wu W B Lu, H H S and, and Li W-H. Evolution of the yeast protein interaction network. *Proc Natl Acad Sci U S A*, 100:12820–12824, 2003.
- [419] S Wuchty, Z N Oltvai, and A L Barabási. Evolutionary conservation of motif constituents in the yeast protein interaction network. *Nat Genet*, 35(2):176–179, Oct 2003.
- [420] F Boyer, A Morgat, L Labarre, J Pothier, and A Viari. Syntons, metabolons and interactons: an exact graph-theoretical approach for exploring neighbourhood between genomic and functional data. *Bioinformatics*, 21(23):4209–4215, Dec 2005.
- [421] C E Horak, N M Luscombe, J Qian, P Bertone, S Piccirillo, M Gerstein, and M Snyder. Complex transcriptional circuitry at the G1/S transition in *Saccharomyces cerevisiae*. *Genes Dev*, 16(23):3017–3033, Dec 2002.
- [422] H D Kim, T Shay, E K O’Shea, and A Regev. Transcriptional regulatory circuits: predicting numbers from alphabets. *Science*, 325(5939):429–432, Jul 2009.
- [423] T Pawson and J D Scott. Signaling through scaffold, anchoring, and adaptor proteins. *Science*, 278(5346):2075–2080, Dec 1997.
- [424] D R Hyduke and B Ø Palsson. Towards genome-scale signalling-network reconstructions. *Nat Rev Genet*, 11(4):297–307, Feb 2010.
- [425] A Kamburov, C Wierling, H Lehrach, and R Herwig. ConsensusPathDB—a database for integrating human functional interaction networks. *Nucleic Acids Res*, 37(Database issue):623–628, Jan 2009.
- [426] A Kamburov, K Pentchev, H Galicka, C Wierling, H Lehrach, and R Herwig. ConsensusPathDB: toward a more complete picture of cell biology. *Nucleic Acids Res*, 39(Database issue):712–717, Jan 2011.
- [427] J Rung, T Schlitt, A Brazma, K Freivalds, and J Vilo. Building and analysing genome-wide gene disruption networks. *Bioinformatics*, 18 Suppl 2:202–210, 2002.
- [428] T Milenković, J Lai, and N Przulj. GraphCrunch: a tool for large network analyses. *BMC Bioinformatics*, 9:70, 2008.
- [429] S Brohée and J van Helden. Evaluation of clustering algorithms for protein-protein interaction networks. *BMC Bioinformatics*, 7:488, 2006.
- [430] J D Han, D Dupuy, N Bertin, M E Cusick, and M Vidal. Effect of sampling on topology predictions of protein-protein interaction networks. *Nat Biotechnol*, 23(7):839–844, Jul 2005.
- [431] V Spirin and L A Mirny. Protein complexes and functional modules in molecular networks. *Proc Natl Acad Sci U S A*, 100(21):12123–12128, Oct 2003.

- [432] A L Barabási and Z N Oltvai. Network biology: understanding the cell's functional organization. *Nat Rev Genet*, 5(2):101–113, Feb 2004.
- [433] P Erdős and A Rényi. On the evolution of random graphs. *Publ Math Inst Hung Acad Sci*, 5:17–61, 1960.
- [434] A L Barabasi and R Albert. Emergence of scaling in random networks. *Science*, 286(5439):509–512, Oct 1999.
- [435] L A Amaral, A Scala, M Barthelemy, and H E Stanley. Classes of small-world networks. *Proc Natl Acad Sci U S A*, 97(21):11149–11152, Oct 2000.
- [436] R Albert. Scale-free networks in cell biology. *J Cell Sci*, 118(Pt 21):4947–4957, Nov 2005.
- [437] M E J Newman. The structure and function of complex networks. *SIAM Review*, 45(2):167–256, 2003.
- [438] R Albert, H Jeong, and A L Barabási. Error and attack tolerance of complex networks. *Nature*, 406(6794):378–382, Jul 2000.
- [439] A Wagner. The yeast protein interaction network evolves rapidly and contains few redundant duplicate genes. *Mol Biol Evol*, 18(7):1283–1292, Jul 2001.
- [440] B A Huberman and L A Adamic. Growth dynamics of the World-Wide Web. *Nature*, 401:131, 1999.
- [441] E F Keller. Revisiting "scale-free" networks. *Bioessays*, 27(10):1060–1068, Oct 2005.
- [442] M Girvan and M E Newman. Community structure in social and biological networks. *Proc Natl Acad Sci U S A*, 99(12):7821–7826, Jun 2002.
- [443] M A de Aguiar and Y Bar-Yam. Spectral analysis and the dynamic response of complex networks. *Phys Rev E Stat Nonlin Soft Matter Phys*, 71(1):016106, Jan 2005.
- [444] R Khanin and E Wit. How scale-free are biological networks. *J Comput Biol*, 13(3):810–818, Apr 2006.
- [445] M P Stumpf, C Wiuf, and R M May. Subnets of scale-free networks are not scale-free: sampling properties of networks. *Proc Natl Acad Sci U S A*, 102(12):4221–4224, Mar 2005.
- [446] G Lima-Mendez and J van Helden. The powerful law of the power law and other myths in network biology. *Mol Biosyst*, 5(12):1482–1493, Dec 2009.
- [447] A Clauset, C R Shalizi, and M E J Newman. Power-law distributions in empirical data. *SIAM Review*, 51(4):661–703, 2009.
- [448] V Memisevic, T Milenkovic, and N Przulj. An integrative approach to modeling biological networks. *Journal of Integrative Bioinformatics*, 7(3):120, 2010.
- [449] O Kuchaiev and N Przulj. Learning the structure of protein-protein interaction networks. *Pac Symp Biocomput*, pages 39–50, 2009.
- [450] C C Friedel and R Zimmer. Inferring topology from clustering coefficients in protein-protein interaction networks. *BMC Bioinformatics*, 7:519, 2006.
- [451] Z Wang and J Zhang. In search of the biological significance of modular structures in protein networks. *PLoS Comput Biol*, 3(6):e107, Jun 2007.
- [452] J Rachlin, D D Cohen, C Cantor, and S Kasif. Biological context networks: a mosaic view of the interactome. *Mol Syst Biol*, 2:66, 2006.
- [453] 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, Jun 2005.
- [454] A Thomas, R Cannings, N A Monk, and C Cannings. On the structure of protein-protein interaction networks. *Biochem Soc Trans*, 31(Pt 6):1491–1496, Dec 2003.

- [455] R Tanaka, T M Yi, and J Doyle. Some protein interaction data do not exhibit power law statistics. *FEBS Lett*, 579(23):5140–5144, Sep 2005.
- [456] A Wagner. How the global structure of protein interaction networks evolves. *Proc Biol Sci*, 270(1514):457–466, Mar 2003.
- [457] J Berg, M Lässig, and A Wagner. Structure and evolution of protein interaction networks: a statistical model for link dynamics and gene duplications. *BMC Evol Biol*, 4:51, Nov 2004.
- [458] M Pagel, A Meade, and D Scott. Assembly rules for protein networks derived from phylogenetic-statistical analysis of whole genomes. *BMC Evol Biol*, 7(Suppl 1):S16, 2007.
- [459] D J Watts and S H Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393(6684):440–442, Jun 1998.
- [460] R V Sole, R Pastor-Satorras, E Smith, and T B Kepler. A model of large-scale proteome evolution. *Advances in Complex Systems*, 5:43, 2002.
- [461] D S Goldberg and F P Roth. Assessing experimentally derived interactions in a small world. *Proc Natl Acad Sci U S A*, 100(8):4372–4376, Apr 2003.
- [462] S Milgram. The small-world problem. *Psychology Today*, 1(1):60–67, 1967.
- [463] F Chung and L Lu. The average distances in random graphs with given expected degrees. *Proc Natl Acad Sci U S A*, 99(25):15879–15882, Dec 2002.
- [464] R Cohen and S Havlin. Scale-free networks are ultrasmall. *Phys Rev Lett*, 90(5):058701, Feb 2003.
- [465] L C Freeman. A set of measures of centrality based on betweenness. *Sociometry*, 40:35–41, 1977.
- [466] L C Freeman. Centrality in social networks conceptual clarification. *Social Networks*, 1(3):215 – 239, 1978.
- [467] E Estrada. Virtual identification of essential proteins within the protein interaction network of yeast. *Proteomics*, 6(1):35–40, Jan 2006.
- [468] G Sabidussi. The centrality of a graph. *Psychometrika*, 31(4):581–603, Dec 1966.
- [469] U Brandes. A faster algorithm for betweenness centrality. *Journal of Mathematical Sociology*, 25:163–177, 2001.
- [470] H Yu, P M Kim, E Sprecher, V Trifonov, and M Gerstein. The importance of bottlenecks in protein networks: correlation with gene essentiality and expression dynamics. *PLoS Comput Biol*, 3(4):e59, Apr 2007.
- [471] M P Joy, A Brock, D E Ingber, and S Huang. High-betweenness proteins in the yeast protein interaction network. *J Biomed Biotechnol*, 2005(2):96–103, Jun 2005.
- [472] M E Newman and M Girvan. Finding and evaluating community structure in networks. *Phys Rev E Stat Nonlin Soft Matter Phys*, 69(2):026113, Feb 2004.
- [473] J Yoon, A Blumer, and K Lee. An algorithm for modularity analysis of directed and weighted biological networks based on edge-betweenness centrality. *Bioinformatics*, 22(24):3106–3108, Dec 2006.
- [474] J Goni, F J Esteban, N Velez de Mendizabal, J Sepulcre, S Ardanza-Treijano, I Agirrezabal, and P Villoslada. A computational analysis of the protein-protein interaction networks in neurodegenerative diseases. *BMC Syst Biol*, 2(1):52, Jun 2008.
- [475] W C Hwang, A Zhang, and M Ramanathan. Identification of information flow-modulating drug targets: a novel bridging paradigm for drug discovery. *Clin Pharmacol Ther*, 84(5):563–572, Nov 2008.
- [476] E Estrada and J A Rodriguez-Velazquez. Subgraph centrality in complex networks. *Physical Review E*, 71(5):056103, 2005.

- [477] P Bonacich. Power and centrality: a family of measures. *The American Journal of Sociology*, 92(5):1170–1182, March 1987.
- [478] K Stephenson. Rethinking centrality: methods and applications. *Social Networks*, 11:1–37, 1989.
- [479] A Wagner. Robustness against mutations in genetic networks of yeast. *Nat Genet*, 24(4):355–361, Apr 2000.
- [480] N Przulj, D A Wigle, and I Jurisica. Functional topology in a network of protein interactions. *Bioinformatics*, 20(3):340–348, Feb 2004.
- [481] O Ozier, N Amin, and T Ideker. Global architecture of genetic interactions on the protein network. *Nat Biotechnol*, 21(5):490–491, May 2003.
- [482] M W Hahn and A D Kern. Comparative genomics of centrality and essentiality in three eukaryotic protein-interaction networks. *Mol Biol Evol*, 22(4):803–806, Apr 2005.
- [483] I Lee. Probabilistic functional gene societies. *Prog Biophys Mol Biol*, 106(2):435–442, Jan 2011.
- [484] X He and J Zhang. Why do hubs tend to be essential in protein networks? *PLoS Genet*, 2(6):e88, Jun 2006.
- [485] M C Palumbo, A Colosimo, A Giuliani, and L Farina. Functional essentiality from topology features in metabolic networks: a case study in yeast. *FEBS Lett*, 579(21):4642–4646, Aug 2005.
- [486] K V Brinda and S Vishveshwara. Oligomeric protein structure networks: insights into protein-protein interactions. *BMC Bioinformatics*, 6:296, 2005.
- [487] F A Rodrigues and L d F Costa. Protein lethality investigated in terms of long range dynamical interactions. *Mol Biosyst*, 5(4):385–390, Apr 2009.
- [488] L D Costa, F A Rodrigues, and G Travieso. Protein domain connectivity and essentiality. *Appl Phys Lett*, 89:174101–174103, 2006.
- [489] J D Han, N Bertin, T Hao, D S Goldberg, G F Berriz, L V 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(6995):88–93, Jul 2004.
- [490] K L Tew, X L Li, and S H Tan. Functional centrality: detecting lethality of proteins in protein interaction networks. *Genome Inform*, 19:166–177, 2007.
- [491] S Wuchty. Interaction and domain networks of yeast. *Proteomics*, 2(12):1715–1723, Dec 2002.
- [492] H Yu, D Greenbaum, H Xin Lu, X Zhu, and M Gerstein. Genomic analysis of essentiality within protein networks. *Trends Genet*, 20(6):227–231, Jun 2004.
- [493] 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 Biol*, 4(10):e317, Sep 2006.
- [494] K Pang, H Sheng, and X Ma. Understanding gene essentiality by finely characterizing hubs in the yeast protein interaction network. *Biochem Biophys Res Commun*, 401(1):112–116, Oct 2010.
- [495] M R Wilkins and S K Kummerfeld. Sticking together? Falling apart? Exploring the dynamics of the interactome. *Trends Biochem Sci*, 33(5):195–200, May 2008.
- [496] N Bertin, N Simonis, D Dupuy, M E Cusick, J D Han, H B Fraser, F P Roth, and M Vidal. Confirmation of organized modularity in the yeast interactome. *PLoS Biol*, 5(6):e153, Jun 2007.
- [497] M Vidal. A biological atlas of functional maps. *Cell*, 104(3):333–339, Feb 2001.
- [498] 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, Aug 2002.
- [499] C Zhang, S Liu, and Y Zhou. Fast and accurate method for identifying high-quality protein-interaction modules by clique merging and its application to yeast. *J Proteome Res*, 5(4):801–807, Apr 2006.

- [500] J Ihmels, G Friedlander, S Bergmann, O Sarig, Y Ziv, and N Barkai. Revealing modular organization in the yeast transcriptional network. *Nat Genet*, 31(4):370–377, Aug 2002.
- [501] B Alberts. The cell as a collection of protein machines: preparing the next generation of molecular biologists. *Cell*, 92(3):291–294, Feb 1998.
- [502] T Wilhelm, H P Nasheuer, and S Huang. Physical and functional modularity of the protein network in yeast. *Mol Cell Proteomics*, 2(5):292–298, May 2003.
- [503] E Segal, M Shapira, A Regev, D Pe’er, D Botstein, D Koller, and N Friedman. Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nat Genet*, 34(2):166–176, Jun 2003.
- [504] R Dunn, F Dudbridge, and C M Sanderson. The use of edge-betweenness clustering to investigate biological function in protein interaction networks. *BMC Bioinformatics*, 6:39, 2005.
- [505] A Lancichinetti, M Kivelä, J Saramäki, and S Fortunato. Characterizing the community structure of complex networks. *PLoS ONE*, 5(8):e11976, 2010.
- [506] C H Chin, S H Chen, C W Ho, M T Ko, and C Y Lin. A hub-attachment based method to detect functional modules from confidence-scored protein interactions and expression profiles. *BMC Bioinformatics*, 11(Suppl 1):S25, 2010.
- [507] S Letovsky and S Kasif. Predicting protein function from protein/protein interaction data: a probabilistic approach. *Bioinformatics*, 19 Suppl 1:197–204, 2003.
- [508] E M Marcotte, M Pellegrini, M J Thompson, T O Yeates, and D Eisenberg. A combined algorithm for genome-wide prediction of protein function. *Nature*, 402(6757):83–86, Nov 1999.
- [509] R Sharan, I Ulitsky, and R Shamir. Network-based prediction of protein function. *Mol Syst Biol*, 3:88, 2007.
- [510] P Bogdanov and A K Singh. Molecular function prediction using neighborhood features. *IEEE/ACM Trans Comput Biol Bioinform*, 7(2):208–217, Apr-Jun 2010.
- [511] T Schlitt, K Palin, J Rung, S Dietmann, M Lappe, E Ukkonen, and A Brazma. From gene networks to gene function. *Genome Res*, 13(12):2568–2576, Dec 2003.
- [512] R Milo, S Itzkovitz, N Kashtan, R Levitt, S Shen-Orr, I Ayzenshtat, M Sheffer, and U Alon. Superfamilies of evolved and designed networks. *Science*, 303(5663):1538–1542, Mar 2004.
- [513] R Sharan, S Suthram, R M Kelley, T Kuhn, S McCuine, P Uetz, T Sittler, R M Karp, and T Ideker. Conserved patterns of protein interaction in multiple species. *Proc Natl Acad Sci U S A*, 102(6):1974–1979, Feb 2005.
- [514] J F Poyatos and L D Hurst. How biologically relevant are interaction-based modules in protein networks? *Genome Biol*, 5(11):R93, 2004.
- [515] R Guimerà and L A Nunes Amaral. Functional cartography of complex metabolic networks. *Nature*, 433(7028):895–900, Feb 2005.
- [516] R M Karp. Reducibility among combinatorial problems. In R E Miller and J W Thatcher, editors, *Complexity of Computer Computations*, pages 85–103. Plenum Press, 1972.
- [517] G Liu, L Wong, and H N Chua. Complex discovery from weighted PPI networks. *Bioinformatics*, 25(15):1891–1897, Aug 2009.
- [518] R L Wang, Z Tang, and Q P Cao. An efficient approximation algorithm for finding a maximum clique using Hopfield network learning. *Neural Comput*, 15(7):1605–1619, Jul 2003.
- [519] B Balasundaram, S Butenko, and S Trukhanov. Novel approaches for analyzing biological networks. *Journal of Combinatorial Optimization*, 10:23–39, 2005.
- [520] S B Seidman. Network structure and minimum degree. *Social Networks*, 5(3):269 – 287, 1983.

- [521] M Altaf-Ul-Amin, Y Shinbo, K Mihara, K Kurokawa, and S Kanaya. Development and implementation of an algorithm for detection of protein complexes in large interaction networks. *BMC Bioinformatics*, 7:207, 2006.
- [522] C von Mering, E M Zdobnov, S Tsoka, F D Ciccarelli, J B Pereira-Leal, C A Ouzounis, and P Bork. Genome evolution reveals biochemical networks and functional modules. *Proc Natl Acad Sci U S A*, 100(26):15428–15433, Dec 2003.
- [523] I J Farkas, C Wu, C Chennubhotla, I Bahar, and Z N Oltvai. Topological basis of signal integration in the transcriptional-regulatory network of the yeast, *Saccharomyces cerevisiae*. *BMC Bioinformatics*, 7:478, 2006.
- [524] J Hallinan. Gene duplication and hierarchical modularity in intracellular interaction networks. *Biosystems*, 74(1-3):51–62, Apr-Jun 2004.
- [525] F Radicchi, C Castellano, F Cecconi, V Loreto, and D Parisi. Defining and identifying communities in networks. *Proc Natl Acad Sci U S A*, 101(9):2658–2663, Mar 2004.
- [526] C Wang, C Ding, Q Yang, and S R Holbrook. Consistent dissection of the protein interaction network by combining global and local metrics. *Genome Biol*, 8(12):R271, 2007.
- [527] K C Kao and J Y Huang. Accurate and fast computational method for identifying protein function using protein-protein interaction data. *Mol Biosyst*, 6(5):830–839, May 2010.
- [528] J A Hartigan. *Clustering Algorithms*. John Wiley & Sons, Inc., 1975.
- [529] S Fortunato. Community detection in graphs. *Physics Reports*, 486(3-5):75 – 174, 2010.
- [530] F D Gibbons and F P Roth. Judging the quality of gene expression-based clustering methods using gene annotation. *Genome Res*, 12(10):1574–1581, Oct 2002.
- [531] A Lancichinetti and S Fortunato. Community detection algorithms: a comparative analysis. *Phys Rev E Stat Nonlin Soft Matter Phys*, 80(5 Pt 2):056117, Nov 2009.
- [532] A D King, N Przulj, and I Jurisica. Protein complex prediction via cost-based clustering. *Bioinformatics*, 20(17):3013–3020, Nov 2004.
- [533] M Blatt, S Wiseman, and E Domany. Superparamagnetic clustering of data. *Phys Rev Lett*, 76(18):3251–3254, Apr 1996.
- [534] A J Enright, S Van Dongen, and C A Ouzounis. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res*, 30(7):1575–1584, Apr 2002.
- [535] C Brun, F Chevenet, D Martin, J Wojcik, A Guénoche, and B Jacq. Functional classification of proteins for the prediction of cellular function from a protein-protein interaction network. *Genome Biol*, 5(1):R6, 2003.
- [536] M P Samanta and S Liang. Predicting protein functions from redundancies in large-scale protein interaction networks. *Proc Natl Acad Sci U S A*, 100(22):12579–12583, Oct 2003.
- [537] M Mete, F Tang, X Xu, and N Yuruk. A structural approach for finding functional modules from large biological networks. *BMC Bioinformatics*, 9(Suppl 9):S19, 2008.
- [538] S Navlakha, M C Schatz, and C Kingsford. Revealing biological modules via graph summarization. *J Comput Biol*, 16(2):253–264, Feb 2009.
- [539] M Habibi, C Eslahchi, and L Wong. Protein complex prediction based on k-connected subgraphs in protein interaction network. *BMC Syst Biol*, 4:129, 2010.
- [540] J Gagneur, R Krause, T Bouwmeester, and G Casari. Modular decomposition of protein-protein interaction networks. *Genome Biol*, 5(8):R57, 2004.
- [541] V Arnau, S Mars, and I Marín. Iterative cluster analysis of protein interaction data. *Bioinformatics*, 21(3):364–378, Feb 2005.

- [542] B Snel, P Bork, and M A Huynen. The identification of functional modules from the genomic association of genes. *Proc Natl Acad Sci U S A*, 99(9):5890–5895, Apr 2002.
- [543] P Holme, M Huss, and H Jeong. Subnetwork hierarchies of biochemical pathways. *Bioinformatics*, 19(4):532–538, Mar 2003.
- [544] D M Wilkinson and B A Huberman. A method for finding communities of related genes. *Proc Natl Acad Sci U S A*, 101 Suppl 1:5241–5248, Apr 2004.
- [545] M E Newman. Fast algorithm for detecting community structure in networks. *Phys Rev E Stat Nonlin Soft Matter Phys*, 69(6 Pt 2):066133, Jun 2004.
- [546] J R Tyler, D M Wilkinson, and B A Huberman. Email as spectroscopy: automated discovery of community structure within organizations. pages 81–96. Kluwer Academic Publishers, Mar 2003.
- [547] J Rattigan, M Maier, and D Jensen. Graph clustering with network structure indices. In *Proceedings of the 24th international conference on Machine learning, ICML '07*, pages 783–790, New York, NY, USA, 2007. ACM.
- [548] Q Yang and S Lonardi. A parallel edge-betweenness clustering tool for protein-protein interaction networks. *Int J Data Min Bioinform*, 1(3):241–247, 2007.
- [549] P Pei and A Zhang. A "seed-refine" algorithm for detecting protein complexes from protein interaction data. *IEEE Trans Nanobioscience*, 6(1):43–50, Mar 2007.
- [550] A Clauset, M E Newman, and C Moore. Finding community structure in very large networks. *Phys Rev E Stat Nonlin Soft Matter Phys*, 70(6 Pt 2):066111–066111, Dec 2004.
- [551] A W Rives and T Galitski. Modular organization of cellular networks. *Proc Natl Acad Sci U S A*, 100(3):1128–1133, Feb 2003.
- [552] W Hwang, Y R Cho, A Zhang, and M Ramanathan. A novel functional module detection algorithm for protein-protein interaction networks. *Algorithms Mol Biol*, 1:24, 2006.
- [553] C G Rivera, R Vakil, and J S Bader. NeMo: Network module identification in Cytoscape. *BMC Bioinformatics*, 11(Suppl 1):S61, 2010.
- [554] M Wu, X Li, C K Kwok, and S K Ng. A core-attachment based method to detect protein complexes in PPI networks. *BMC Bioinformatics*, 10:169, 2009.
- [555] H C Leung, Q Xiang, S M Yiu, and F Y Chin. Predicting protein complexes from PPI data: a core-attachment approach. *J Comput Biol*, 16(2):133–144, Feb 2009.
- [556] 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, Jan 2007.
- [557] M E Newman. Modularity and community structure in networks. *Proc Natl Acad Sci U S A*, 103(23):8577–8582, Jun 2006.
- [558] G Palla, A L Barabási, and T Vicsek. Quantifying social group evolution. *Nature*, 446(7136):664–667, Apr 2007.
- [559] M Sales-Pardo, R Guimerà, A A Moreira, and L A Amaral. Extracting the hierarchical organization of complex systems. *Proc Natl Acad Sci U S A*, 104(39):15224–15229, Sep 2007.
- [560] J Hallinan and A Wipat. Clustering and cross-talk in a yeast functional interaction network. *Computational Intelligence and Bioinformatics and Computational Biology, 2006. CIBCB '06*, pages 1–8, 2006.
- [561] E Zotenko, K S Guimarães, R Jothi, and T M Przytycka. Decomposition of overlapping protein complexes: a graph theoretical method for analyzing static and dynamic protein associations. *Algorithms Mol Biol*, 1(1):7, 2006.
- [562] Y Y Ahn, J P Bagrow, and S Lehmann. Link communities reveal multiscale complexity in networks. *Nature*, 466(7307):761–764, Aug 2010.

- [563] T S Evans and R Lambiotte. Line graphs, link partitions, and overlapping communities. *Phys Rev E Stat Nonlin Soft Matter Phys*, 80(1 Pt 2):016105, Jul 2009.
- [564] J B Pereira-Leal, A J Enright, and C A Ouzounis. Detection of functional modules from protein interaction networks. *Proteins*, 54(1):49–57, Jan 2004.
- [565] I Derényi, G Palla, and T Vicsek. Clique percolation in random networks. *Phys Rev Lett*, 94(16):160202, Apr 2005.
- [566] B Adamcsek, G Palla, I J Farkas, I Derényi, and T Vicsek. CFinder: locating cliques and overlapping modules in biological networks. *Bioinformatics*, 22(8):1021–1023, Apr 2006.
- [567] J M Kumpula, M Kivelä, K Kaski, and J Saramäki. Sequential algorithm for fast clique percolation. *Phys Rev E Stat Nonlin Soft Matter Phys*, 78(2 Pt 2):026109, Aug 2008.
- [568] A Lancichinetti and S Fortunato. Benchmarks for testing community detection algorithms on directed and weighted graphs with overlapping communities. *Phys Rev E Stat Nonlin Soft Matter Phys*, 80(1 Pt 2):016118, Jul 2009.
- [569] I Ispolatov, I Mazo, and A Yuryev. Finding mesoscopic communities in sparse networks. *J Stat Mech*, 9:p09014, Sep 2006.
- [570] T Wittkop, D Emig, S Lange, S Rahmann, M Albrecht, J H Morris, S Böcker, J Stoye, and J Baumbach. Partitioning biological data with transitivity clustering. *Nat Methods*, 7(6):419–420, Jun 2010.
- [571] J Vlasblom and S J Wodak. Markov clustering versus affinity propagation for the partitioning of protein interaction graphs. *BMC Bioinformatics*, 10:99, 2009.
- [572] T Theodosiou, N Darzentas, L Angelis, and C A Ouzounis. PuReD-MCL: a graph-based PubMed document clustering methodology. *Bioinformatics*, 24(17):1935–1941, Sep 2008.
- [573] S Srihari, K Ning, and H W Leong. MCL-CAw: a refinement of MCL for detecting yeast complexes from weighted PPI networks by incorporating core-attachment structure. *BMC Bioinformatics*, 11:504, 2010.
- [574] S Srihari, K Ning, and H W Leong. Refining Markov Clustering for protein complex prediction by incorporating core-attachment structure. *Genome Inform*, 23(1):159–168, Oct 2009.
- [575] K Macropol, T Can, and A K Singh. RRW: repeated random walks on genome-scale protein networks for local cluster discovery. *BMC Bioinformatics*, 10:283, 2009.
- [576] I Ulitsky and R Shamir. Identification of functional modules using network topology and high-throughput data. *BMC Syst Biol*, 1:8, 2007.
- [577] D Hanisch, A Zien, R Zimmer, and T Lengauer. Co-clustering of biological networks and gene expression data. *Bioinformatics*, 18 Suppl 1:145–154, 2002.
- [578] J A Parkkinen and S Kaski. Searching for functional gene modules with interaction component models. *BMC Syst Biol*, 4:4, 2010.
- [579] I Ulitsky and R Shamir. Identifying functional modules using expression profiles and confidence-scored protein interactions. *Bioinformatics*, 25(9):1158–1164, May 2009.
- [580] A Vazquez, A Flammini, A Maritan, and A Vespignani. Global protein function prediction from protein-protein interaction networks. *Nat Biotechnol*, 21(6):697–700, Jun 2003.
- [581] Z Lubovac, J Gamalielsson, and B Olsson. Combining functional and topological properties to identify core modules in protein interaction networks. *Proteins*, 64(4):948–959, Sep 2006.
- [582] S Bandyopadhyay, R Kelley, N J Krogan, and T Ideker. Functional maps of protein complexes from quantitative genetic interaction data. *PLoS Comput Biol*, 4(4):e1000065, Apr 2008.
- [583] Y Ozawa, R Saito, S Fujimori, H Kashima, M Ishizaka, H Yanagawa, E Miyamoto-Sato, and M Tomita. Protein complex prediction via verifying and reconstructing the topology of domain-domain interactions. *BMC Bioinformatics*, 11:350, 2010.

- [584] L Kiemer and G Cesareni. Comparative interactomics: comparing apples and pears? *Trends Biotechnol*, 25(10):448–454, Oct 2007.
- [585] J Berg and M Lässig. Local graph alignment and motif search in biological networks. *Proc Natl Acad Sci U S A*, 101(41):14689–14694, Oct 2004.
- [586] L Parida. Discovering topological motifs using a compact notation. *J Comput Biol*, 14(3):300–323, Apr 2007.
- [587] N Przulj and D J Higham. Modelling protein-protein interaction networks via a stickiness index. *J R Soc Interface*, 3(10):711–716, Oct 2006.
- [588] N Przulj. Biological network comparison using graphlet degree distribution. *Bioinformatics*, 23(2):177–183, Jan 2007.
- [589] E J Deeds, O Ashenberg, and E I Shakhnovich. A simple physical model for scaling in protein-protein interaction networks. *Proc Natl Acad Sci U S A*, 103(2):311–316, Jan 2006.
- [590] F Schreiber and H Schwöbbermeyer. MAVisto: a tool for the exploration of network motifs. *Bioinformatics*, July 2005.
- [591] X Wu, Q Liu, and R Jiang. Align human interactome with phenome to identify causative genes and networks underlying disease families. *Bioinformatics*, 25(1):98–104, Jan 2009.
- [592] S Brohée, K Faust, G Lima-Mendez, O Sand, R Janky, G Vanderstocken, Y Deville, and J van Helden. NeAT: a toolbox for the analysis of biological networks, clusters, classes and pathways. *Nucleic Acids Res*, 36 (Web Server issue):W444–451, Jul 2008.
- [593] R Singh, J Xu, and B Berger. Global alignment of multiple protein interaction networks with application to functional orthology detection. *Proc Natl Acad Sci U S A*, 105(35):12763–12768, Sep 2008.
- [594] 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):253–258, Jun 2009.
- [595] O Kuchaiev, T Milenkovic, V Memisevic, W Hayes, and N Przulj. Topological network alignment uncovers biological function and phylogeny. *Journal of the Royal Society*, 7(50):1341–1354, September 2010.
- [596] 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 Res*, 14(6):1107–1118, Jun 2004.
- [597] S Wernicke and F Rasche. Simple and fast alignment of metabolic pathways by exploiting local diversity. *Bioinformatics*, 23(15):1978–1985, Aug 2007.
- [598] K Plaimas, R Eils, and R Konig. Identifying essential genes in bacterial metabolic networks with machine learning methods. *BMC Syst Biol*, 4(1):56, May 2010.
- [599] R Y Pinter, O Rokhlenko, E Yeger-Lotem, and M Ziv-Ukelson. Alignment of metabolic pathways. *Bioinformatics*, 21(16):3401–3408, Aug 2005.
- [600] R Sharan and T Ideker. Modeling cellular machinery through biological network comparison. *Nat Biotechnol*, 24(4):427–433, Apr 2006.
- [601] S Bandyopadhyay, R Sharan, and T Ideker. Systematic identification of functional orthologs based on protein network comparison. *Genome Res*, 16(3):428–435, 2006.
- [602] R Sharan, T Ideker, B Kelley, R Shamir, and R M Karp. Identification of protein complexes by comparative analysis of yeast and bacterial protein interaction data. *J Comput Biol*, 12:835–46, 2005.
- [603] T K Gandhi, J Zhong, S Mathivanan, L Karthick, K N Chandrika, S S Mohan, S Sharma, S Pinkert, S Nagaraju, B Periaswamy, G Mishra, K Nandakumar, B Shen, N Deshpande, R Nayak, M Sarker, J D Boeke, G Parmigiani, J Schultz, J S Bader, and A Pandey. Analysis of the human protein interactome and comparison with yeast, worm and fly interaction datasets. *Nat Genet*, 38(3):285–293, Mar 2006.

- [604] E Hirsh and R Sharan. Identification of conserved protein complexes based on a model of protein network evolution. *Bioinformatics*, 23(2):170–176, Jan 2007.
- [605] M Koyutürk, A Grama, and W Szpankowski. Pairwise local alignment of protein interaction networks guided by models of evolution. In *Research in Computational Molecular Biology*, volume 3500 of *Lecture Notes in Computer Science*, pages 48–65. Springer Berlin / Heidelberg, 2005.
- [606] Z Liang, M Xu, M Teng, and L Niu. NetAlign: a web-based tool for comparison of protein interaction networks. *Bioinformatics*, 22(17):2175–2177, Sep 2006.
- [607] M Hayashida and T Akutsu. Comparing biological networks via graph compression. *BMC Syst Biol*, 4(Suppl 2):S13, 2010.
- [608] L Peshkin. Structure induction by lossless graph compression. In *In Proc. 2007 Data Compression Conference*, pages 53–62, 2007.
- [609] M Koyutürk, Y Kim, S Subramaniam, W Szpankowski, and A Grama. Detecting conserved interaction patterns in biological networks. *J Comput Biol*, 13(7):1299–1322, Sep 2006.
- [610] M Suderman and M Hallett. Tools for visually exploring biological networks. *Bioinformatics*, 23(20):2651–2659, Oct 2007.
- [611] C Huttenhower, S O Mehmood, and O G Troyanskaya. Graphle: Interactive exploration of large, dense graphs. *BMC Bioinformatics*, 10:417, 2009.
- [612] Z Hu, J Mellor, J Wu, and C DeLisi. VisANT: an online visualization and analysis tool for biological interaction data. *BMC Bioinformatics*, 5:17, Feb 2004.
- [613] G A Pavlopoulos, E Pafilis, M Kuhn, S D Hooper, and R Schneider. OnTheFly: a tool for automated document-based text annotation, data linking and network generation. *Bioinformatics*, 25(7):977–978, Apr 2009.
- [614] L A Flórez, C R Lammers, R Michna, and J Stülke. CellPublisher: a web platform for the intuitive visualization and sharing of metabolic, signalling and regulatory pathways. *Bioinformatics*, 26(23):2997–2999, Dec 2010.
- [615] U Dogrusoz, E Z Erson, E Giral, E Demir, O Babur, A Cetintas, and R Colak. PATIKAwed: a Web interface for analyzing biological pathways through advanced querying and visualization. *Bioinformatics*, 22(3):374–375, Feb 2006.
- [616] H J Chung, M Kim, C H Park, J Kim, and J H Kim. ArrayXPath: mapping and visualizing microarray gene-expression data with integrated biological pathway resources using Scalable Vector Graphics. *Nucleic Acids Res*, 32(Web Server issue):460–464, Jul 2004.
- [617] M Baitaluk, X Qian, S Godbole, A Raval, A Ray, and A Gupta. PathSys: integrating molecular interaction graphs for systems biology. *BMC Bioinformatics*, 7:55, 2006.
- [618] P D Karp, S Paley, and P Romero. The Pathway Tools software. *Bioinformatics*, 18 Suppl 1:225–232, 2002.
- [619] P Shannon, A Markiel, O Ozier, N S Baliga, J T Wang, D Ramage, N Amin, B Schwikowski, and T Ideker. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res*, 13(11):2498–2504, Nov 2003.
- [620] V Batagelj and A Mrvar. Pajek - program for large network analysis. *Connections*, 21:47–57, 1998.
- [621] B J Breitkreutz, C Stark, and M Tyers. Osprey: a network visualization system. *Genome Biol*, 4(3):R22, 2003.
- [622] S M Paley and P D Karp. The Pathway Tools cellular overview diagram and Omics Viewer. *Nucleic Acids Res*, 34(13):3771–3778, 2006.
- [623] A Nikitin, S Egorov, N Daraselia, and I Mazo. Pathway studio—the analysis and navigation of molecular networks. *Bioinformatics*, 19(16):2155–2157, Nov 2003.

- [624] L Goldovsky, I Cases, A J Enright, and C A Ouzounis. BioLayout(Java): versatile network visualisation of structural and functional relationships. *Appl Bioinformatics*, 4(1):71–74, 2005.
- [625] K R Brown, D Otasek, M Ali, M J McGuffin, W Xie, B Devani, I L Toch, and I Jurisica. NAVi-GaTOR: Network Analysis, Visualization and Graphing Toronto. *Bioinformatics*, 25(24):3327–3329, Dec 2009.
- [626] F Iragne, M Nikolski, B Mathieu, D Auber, and D Sherman. ProViz: protein interaction visualization and exploration. *Bioinformatics*, 21(2):272–274, Jan 2005.
- [627] M Weniger, J C Engelmann, and J Schultz. Genome expression pathway analysis tool—analysis and visualization of microarray gene expression data under genomic, proteomic and metabolic context. *BMC Bioinformatics*, 8:179, 2007.
- [628] E Demir, O Babur, U Dogrusoz, A Gursoy, G Nisanci, R Cetin-Atalay, and M Ozturk. PATIKA: an integrated visual environment for collaborative construction and analysis of cellular pathways. *Bioinformatics*, 18(7):996–1003, Jul 2002.
- [629] I Letunic, T Yamada, M Kanehisa, and P Bork. iPath: interactive exploration of biochemical pathways and networks. *Trends Biochem Sci*, 33(3):101–103, Mar 2008.
- [630] W J Longabaugh, E H Davidson, and H Bolouri. Computational representation of developmental genetic regulatory networks. *Dev Biol*, 283(1):1–16, Jul 2005.
- [631] M Kuhn, D Szklarczyk, A Franceschini, M Campillos, C von Mering, L J Jensen, A Beyer, and P Bork. STITCH 2: an interaction network database for small molecules and proteins. *Nucleic Acids Res*, 38(Database issue):552–556, Jan 2010.
- [632] T B Hashimoto, M Nagasaki, K Kojima, and S Miyano. BFL: a node and edge betweenness based fast layout algorithm for large scale networks. *BMC Bioinformatics*, 10(1):19, Jan 2009.
- [633] T Hebert and R Leahy. A generalized EM algorithm for 3-D Bayesian reconstruction from Poisson data using Gibbs priors. *IEEE Trans Med Imaging*, 8(2):194–202, 1989.
- [634] T Fruchterman and E Reingold. Graph drawing by force-directed placement. *Software -Practice and Experience (Wiley)*, 21 (11):1129–1164, 1991.
- [635] P Eades. A heuristic for graph drawing. *Congressus Numerantium*, 42:149–160, 1984.
- [636] K Kojima, M Nagasaki, and S Miyano. Fast grid layout algorithm for biological networks with sweep calculation. *Bioinformatics*, 24(12):1433–1441, Jun 2008.
- [637] P Minguéz, S Götz, D Montaner, F Al-Shahrour, and J Dopazo. SNOW, a web-based tool for the statistical analysis of protein-protein interaction networks. *Nucleic Acids Res*, 37(Web Server issue):109–114, Jul 2009.
- [638] H Yu, X Zhu, D Greenbaum, J Karro, and M Gerstein. TopNet: a tool for comparing biological sub-networks, correlating protein properties with topological statistics. *Nucleic Acids Res*, 32(1):328–337, 2004.
- [639] S Wernicke and F Rasche. FANMOD: a tool for fast network motif detection. *Bioinformatics*, 22(9):1152–1153, May 2006.
- [640] N Kashtan, S Itzkovitz, R Milo, and U Alon. Efficient sampling algorithm for estimating subgraph concentrations and detecting network motifs. *Bioinformatics*, 20(11):1746–1758, Jul 2004.
- [641] M Kohl, S Wiese, and B Warscheid. Cytoscape: software for visualization and analysis of biological networks. *Methods Mol Biol*, 696:291–303, 2011.
- [642] C T Lopes, M Franz, F Kazi, S L Donaldson, Q Morris, and G D Bader. Cytoscape web: an interactive web-based network browser. *Bioinformatics*, 26(18):2347–2348, Sep 2010.
- [643] J S Luciano. PAX of mind for pathway researchers. *Drug Discov Today*, 10(13):937–942, Jul 2005.

- [644] H Hermjakob, L Montecchi-Palazzi, G Bader, J Wojcik, L Salwinski, A Ceol, S Moore, S Orchard, U Sarkans, C von Mering, B Roechert, S Poux, E Jung, H Mersch, P Kersey, M Lappe, Y Li, R Zeng, D Rana, M Nikolski, H Husi, C Brun, K Shanker, S G Grant, C Sander, P Bork, W Zhu, A Pandey, A Brazma, B Jacq, M Vidal, D Sherman, P Legrain, G Cesareni, I Xenarios, D Eisenberg, B Steipe, C Hogue, and R Apweiler. The HUPO PSI's molecular interaction format—a community standard for the representation of protein interaction data. *Nat Biotechnol*, 22(2):177–183, Feb 2004.
- [645] A Bauer-Mehren, M Rautschka, F Sanz, and L I Furlong. DisGeNET: a Cytoscape plugin to visualize, integrate, search and analyze gene-disease networks. *Bioinformatics*, 26(22):2924–2926, Nov 2010.
- [646] D Emig, N Salomonis, J Baumbach, T Lengauer, B R Conklin, and M Albrecht. AltAnalyze and DomainGraph: analyzing and visualizing exon expression data. *Nucleic Acids Res*, 38(Web Server issue):755–762, Jul 2010.
- [647] J Gao, V G Tarcea, A Karnovsky, B R Mirel, T E Weymouth, C W Beecher, J D Cavalcoli, B D Athey, G S Omenn, C F Burant, and H V Jagadish. Metscape: a Cytoscape plug-in for visualizing and interpreting metabolomic data in the context of human metabolic networks. *Bioinformatics*, 26(7):971–973, Feb 2010.
- [648] J Garcia-Garcia, E Guney, R Aragues, J Planas-Iglesias, and B Oliva. Biana: a software framework for compiling biological interactions and analyzing networks. *BMC Bioinformatics*, 11:56, 2010.
- [649] A Joshi, T Van Parys, Y V Peer, and T Michoel. Characterizing regulatory path motifs in integrated networks using perturbational data. *Genome Biol*, 11(3):R32, 2010.
- [650] K Laukens, J Hollunder, T H Dang, G De Jaeger, M Kuiper, E Witters, A Verschoren, and K Van Leemput. Flexible network reconstruction from relational databases with Cytoscape and CytoSQL. *BMC Bioinformatics*, 11:360, 2010.
- [651] F Li, P Li, W Xu, Y Peng, X Bo, and S Wang. PerturbationAnalyzer: a tool for investigating the effects of concentration perturbation on protein interaction networks. *Bioinformatics*, 26(2):275–277, Jan 2010.
- [652] A Martin, M Ochagavia, L Rabasa, J Miranda, J F de Cossio, and R Bringas. BisoGenet: a new tool for gene network building, visualization and analysis. *BMC Bioinformatics*, 11(1):91, 2010.
- [653] D Merico, R Isserlin, O Stueker, A Emili, and G D Bader. Enrichment Map: a network-based method for gene-set enrichment visualization and interpretation. *PLoS One*, 5(11):e13984, 2010.
- [654] J Montojo, K Zuberi, H Rodriguez, F Kazi, G Wright, S L Donaldson, Q Morris, and G D Bader. GeneMANIA Cytoscape plugin: fast gene function predictions on the desktop. *Bioinformatics*, 26(22):2927–2928, Nov 2010.
- [655] K Pentchev, K Ono, R Herwig, T Ideker, and A Kamburov. Evidence mining and novelty assessment of protein-protein interactions with the ConsensusPathDB plugin for Cytoscape. *Bioinformatics*, 26(21):2796–2797, Nov 2010.
- [656] H Wang, H Zheng, and F Azuaje. Ontology- and graph-based similarity assessment in biological networks. *Bioinformatics*, 26(20):2643–2644, Oct 2010.
- [657] G Warsow, B Greber, S S Falk, C Harder, M Siatkowski, S Schordan, A Som, N Endlich, H Schöler, D Reipsilber, K Endlich, and G Fuellen. ExprEssence—revealing the essence of differential experimental data in the context of an interaction/regulation network. *BMC Syst Biol*, 4:164, 2010.
- [658] M A Westenberg, J B Roerdink, O P Kuipers, and S A van Hijum. SpotXplore: a Cytoscape plugin for visual exploration of hotspot expression in gene regulatory networks. *Bioinformatics*, 26(22):2922–2923, Nov 2010.
- [659] M Woźniak, J Tiuryn, and J Dutkowski. MODEVO: exploring modularity and evolution of protein interaction networks. *Bioinformatics*, 26(14):1790–1791, Jul 2010.
- [660] J Zheng, D Zhang, P F Przytycki, R Zielinski, J Capala, and T M Przytycka. SimBoolNet—a Cytoscape plugin for dynamic simulation of signaling networks. *Bioinformatics*, 26(1):141–142, Jan 2010.

- [661] T Xia, J V Hemert, and J A Dickerson. OmicsAnalyzer: a Cytoscape plug-in suite for modeling omics data. *Bioinformatics*, 26(23):2995–2996, Dec 2010.
- [662] I Avila-Campillo, K Drew, J Lin, D J Reiss, and R Bonneau. BioNetBuilder: automatic integration of biological networks. *Bioinformatics*, 23(3):392–393, Feb 2007.
- [663] 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 8—a global view on proteins and their functional interactions in 630 organisms. *Nucleic Acids Res*, 37(Database issue):412–416, Jan 2009.
- [664] S J Cockell, J Weile, P Lord, C Wipat, D Andriychenko, M Pocock, D Wilkinson, M Young, and A Wipat. An integrated dataset for *in silico* drug discovery. *J Integr Bioinform*, 7(3):116, 2010.
- [665] K Hassani-Pak, R Legaie, C Canevet, H A van den Berg, J D Moore, and C J Rawlings. Enhancing data integration with text analysis to find proteins implicated in plant stress response. *J Integr Bioinform*, 7(3):121, 2010.
- [666] J Weile, M Pocock, S J Cockell, P Lord, J M Dewar, E Holstein, D Wilkinson, D Lydall, J Hallinan, and A Wipat. Customisable views on semantically integrated networks for systems biology. *Bioinformatics*, 27(9):1299–1306, Mar 2011.
- [667] K C Gunsalus, H Ge, A J Schetter, D S Goldberg, J D Han, T Hao, G F Berriz, N Bertin, J Huang, L S Chuang, N Li, R Mani, A A Hyman, B Sönnichsen, C J Echeverri, F P Roth, M Vidal, and F Piano. Predictive models of molecular machines involved in *Caenorhabditis elegans* early embryogenesis. *Nature*, 436(7052):861–865, Aug 2005.
- [668] G R Lanckriet, M Deng, N Cristianini, M I Jordan, and W S Noble. Kernel-based data fusion and its application to protein function prediction in yeast. *Pac Symp Biocomput*, pages 300–311, 2004.
- [669] A G Fraser and E M Marcotte. A probabilistic view of gene function. *Nat Genet*, 36(6):559–564, Jun 2004.
- [670] M W Covert, E M Knight, J L Reed, M J Herrgard, and B Ø Palsson. Integrating high-throughput and computational data elucidates bacterial networks. *Nature*, 429(6987):92–96, May 2004.
- [671] P. Bordron, D. Eveillard, and I. Rusu. SIPPER: A flexible method to integrate heterogeneous data into a metabolic network. In *Computational Advances in Bio and Medical Sciences (ICABS), 2011 IEEE 1st International Conference on*, pages 40–45, Feb 2011.
- [672] M R Wilkins. Hares and tortoises: the high- versus low-throughput proteomic race. *Electrophoresis*, 30 Suppl 1:150–155, Jun 2009.
- [673] V van Noort, B Snel, and M A Huynen. Exploration of the omics evidence landscape: adding qualitative labels to predicted protein-protein interactions. *Genome Biol*, 8(9):R197, Sep 2007.
- [674] B Titz, M Schlesner, and P Uetz. What do we learn from high-throughput protein interaction data? *Expert Rev Proteomics*, 1(1):111–121, Jun 2004.
- [675] J Yu and F Fotouhi. Computational approaches for predicting protein-protein interactions: a survey. *J Med Syst*, 30(1):39–44, Feb 2006.
- [676] Y Zhang, J Xuan, B G de los Reyes, R Clarke, and H W Ressom. Network motif-based identification of transcription factor-target gene relationships by integrating multi-source biological data. *BMC Bioinformatics*, 9:203, 2008.
- [677] S L Wong, L V Zhang, A H Tong, Z Li, D S Goldberg, O D King, G Lesage, M Vidal, B Andrews, H Bussey, C Boone, and F P Roth. Combining biological networks to predict genetic interactions. *Proc Natl Acad Sci U S A*, 101(44):15682–15687, Nov 2004.
- [678] Y Zhang, J Xuan, B G de los Reyes, R Clarke, and H W Ressom. Reconstruction of gene regulatory modules in cancer cell cycle by multi-source data integration. *PLoS ONE*, 5(4):e10268, 2010.
- [679] F Ramírez, A Schlicker, Y Assenov, T Lengauer, and M Albrecht. Computational analysis of human protein interaction networks. *Proteomics*, 7(15):2541–2552, Aug 2007.

- [680] I Yanai and C DeLisi. The society of genes: networks of functional links between genes from comparative genomics. *Genome Biol*, 3(11):research0064, Oct 2002.
- [681] A Patil and H Nakamura. Filtering high-throughput protein-protein interaction data using a combination of genomic features. *BMC Bioinformatics*, 6:100, 2005.
- [682] M Gerstein, N Lan, and R Jansen. Integrating interactomes. *Science*, 295(5553):284–287, Jan 2002.
- [683] R Jansen, N Lan, J Qian, and M Gerstein. Integration of genomic datasets to predict protein complexes in yeast. *J Struct Funct Genomics*, 2(2):71–81, 2002.
- [684] M R Said, T J Begley, A V Oppenheim, D A Lauffenburger, and L D Samson. Global network analysis of phenotypic effects: protein networks and toxicity modulation in *Saccharomyces cerevisiae*. *Proc Natl Acad Sci U S A*, 101(52):18006–18011, Dec 2004.
- [685] J P Mackay, M Sunde, J A Lowry, M Crossley, and J M Matthews. Protein interactions: is seeing believing? *Trends Biochem Sci*, 32(12):530–531, Dec 2007.
- [686] S Suthram, T Shlomi, E Ruppin, R Sharan, and T Ideker. A direct comparison of protein interaction confidence assignment schemes. *BMC Bioinformatics*, 7:360, 2006.
- [687] A Grigoriev. A relationship between gene expression and protein interactions on the proteome scale: analysis of the bacteriophage T7 and the yeast *Saccharomyces cerevisiae*. *Nucleic Acids Res*, 29(17):3513–3519, Sep 2001.
- [688] C M Deane, Ł Salwiński, I Xenarios, and D Eisenberg. Protein interactions: two methods for assessment of the reliability of high throughput observations. *Mol Cell Proteomics*, 1(5):349–356, May 2002.
- [689] S V Rajagopala, B Titz, J Goll, J R Parrish, K Wohlbold, M T McKeivitt, T Palzkill, H Mori, R L Finley, and P Uetz. The protein network of bacterial motility. *Mol Syst Biol*, 3:128, 2007.
- [690] M Albers, H Kranz, I Kober, C Kaiser, M Klink, J Suckow, R Kern, and M Koegl. Automated yeast two-hybrid screening for nuclear receptor-interacting proteins. *Mol Cell Proteomics*, 4(2):205–213, Feb 2005.
- [691] P Legrain, J L Jestin, and V Schächter. From the analysis of protein complexes to proteome-wide linkage maps. *Curr Opin Biotechnol*, 11(4):402–407, Aug 2000.
- [692] B Titz, S V Rajagopala, J Goll, R Häuser, M T McKeivitt, T Palzkill, and P Uetz. The binary protein interactome of *Treponema pallidum*—the syphilis spirochete. *PLoS ONE*, 3(5):e2292, 2008.
- [693] O Kuchaiev, M Rasajski, D J Higham, and N Przulj. Geometric de-noising of protein-protein interaction networks. *PLoS Comput Biol*, 5(8):e1000454, Aug 2009.
- [694] M A Mahdavi and Y H Lin. False positive reduction in protein-protein interaction predictions using Gene Ontology annotations. *BMC Bioinformatics*, 8:262, 2007.
- [695] N Sugaya, K Ikeda, T Tashiro, S Takeda, J Otomo, Y Ishida, A Shiratori, A Toyoda, H Noguchi, T Takeda, S Kuhara, Y Sakaki, and T Iwayanagi. An integrative *in silico* approach for discovering candidates for drug-targetable protein-protein interactions in interactome data. *BMC Pharmacol*, 7(1):10, Aug 2007.
- [696] 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 Res*, 33(Database issue):433–437, Jan 2005.
- [697] B Aranda, H Blankenburg, S Kerrien, F S Brinkman, A Ceol, E Chautard, J M Dana, J De Las Rivas, M Dumousseau, E Galeota, A Gaulton, J Goll, R E Hancock, R Isserlin, R C Jimenez, J Kerssemakers, J Khadake, D J Lynn, M Michaut, G O’Kelly, K Ono, S Orchard, C Prieto, S Razick, O Rigina, L Salwinski, M Simonovic, S Velankar, A Winter, G Wu, G D Bader, G Cesareni, I M Donaldson, D Eisenberg, G J Kleywegt, J Overington, S Ricard-Blum, M Tyers, M Albrecht, and H Hermjakob. PSICQUIC and PSISCORE: accessing and scoring molecular interactions. *Nat Methods*, 8(7):528–529, 2011.

- [698] P Bork, L J Jensen, C von Mering, A K Ramani, I Lee, and E M Marcotte. Protein interaction networks from yeast to human. *Curr Opin Struct Biol*, 14(3):292–299, Jun 2004.
- [699] R Jansen and M Gerstein. Analyzing protein function on a genomic scale: the importance of gold-standard positives and negatives for network prediction. *Curr Opin Microbiol*, 7(5):535–545, Oct 2004.
- [700] I V Tetko, B Brauner, I Dunger-Kaltenbach, G Frishman, C Montrone, G Fobo, A Ruepp, A V Antonov, D Surmeli, and H W Mewes. MIPS bacterial genomes functional annotation benchmark dataset. *Bioinformatics*, 21(10):2520–2521, May 2005.
- [701] P Smialowski, P Pagel, P Wong, B Brauner, I Dunger, G Fobo, G Frishman, C Montrone, T Rattei, D Frishman, and A Ruepp. The Negatome database: a reference set of non-interacting protein pairs. *Nucleic Acids Res*, 38(Database issue):540–544, Jan 2010.
- [702] B S Srinivasan, A Novak, J A Flannick, S Batzoglou, and H H Mcadams. Integrated protein interaction networks for 11 microbes. In *In Proceedings of the 10th Annual International Conference on Research in Computational Molecular Biology (RECOMB)*, pages 1–14, 2006.
- [703] I Lee, B Lehner, C Crombie, W Wong, A G Fraser, and E M Marcotte. A single gene network accurately predicts phenotypic effects of gene perturbation in *Caenorhabditis elegans*. *Nat Genet*, 40(2):181–188, Feb 2008.
- [704] M Deng, K Zhang, S Mehta, T Chen, and F Sun. Prediction of protein function using protein-protein interaction data. *J Comput Biol*, 10(6):947–960, 2003.
- [705] A Ben-Hur and W S Noble. Choosing negative examples for the prediction of protein-protein interactions. *BMC Bioinformatics*, 7(Suppl 1):S2, 2006.
- [706] K L McGary, I Lee, and E M Marcotte. Broad network-based predictability of *Saccharomyces cerevisiae* gene loss-of-function phenotypes. *Genome Biol*, 8(12):R258, Dec 2007.
- [707] I Lee, B Lehner, T Vavouri, J Shin, A G Fraser, and E M Marcotte. Predicting genetic modifier loci using functional gene networks. *Genome Res*, 20(8):1143–1153, Aug 2010.
- [708] J C Costello, M M Dalkilic, S M Beason, J R Gehlhausen, R Patwardhan, S Middha, B D Eads, and J R Andrews. Gene networks in *Drosophila melanogaster*: integrating experimental data to predict gene function. *Genome Biol*, 10(9):R97, 2009.
- [709] D J Wilkinson. Bayesian methods in bioinformatics and computational systems biology. *Brief Bioinform*, 8(2):109–116, Mar 2007.
- [710] K Y Yip and M Gerstein. Training set expansion: an approach to improving the reconstruction of biological networks from limited and uneven reliable interactions. *Bioinformatics*, 25(2):243–250, Jan 2009.
- [711] Y Tao, L Sam, J Li, C Friedman, and Y A Lussier. Information theory applied to the sparse Gene Ontology annotation network to predict novel gene function. *Bioinformatics*, 23(13):529–538, Jul 2007.
- [712] J S Bader, A Chaudhuri, J M Rothberg, and J Chant. Gaining confidence in high-throughput protein interaction networks. *Nat Biotechnol*, 22(1):78–85, Jan 2004.
- [713] G R Lanckriet, T De Bie, N Cristianini, M I Jordan, and W S Noble. A statistical framework for genomic data fusion. *Bioinformatics*, 20(16):2626–2635, Nov 2004.
- [714] Y Qi, J Klein-Seetharaman, and Z Bar-Joseph. Random forest similarity for protein-protein interaction prediction from multiple sources. *Pac Symp Biocomput*, pages 531–542, 2005.
- [715] T Jiang and A E Keating. AVID: an integrative framework for discovering functional relationships among proteins. *BMC Bioinformatics*, 6:136, 2005.
- [716] S S Ray, S Bandyopadhyay, and S K Pal. Combining multisource information through functional-annotation-based weighting: gene function prediction in yeast. *IEEE Trans Biomed Eng*, 56(2):229–236, Feb 2009.

- [717] H N Chua, W K Sung, and L Wong. An efficient strategy for extensive integration of diverse biological data for protein function prediction. *Bioinformatics*, 23(24):3364–3373, Dec 2007.
- [718] W Zhong and P W Sternberg. Genome-wide prediction of *C. elegans* genetic interactions. *Science*, 311(5766):1481–1484, Mar 2006.
- [719] S V Date and C J Stoeckert. Computational modeling of the *Plasmodium falciparum* interactome reveals protein function on a genome-wide scale. *Genome Res*, 16(4):542–549, Apr 2006.
- [720] C Huttenhower, E M Haley, M A Hibbs, V Dumeaux, D R Barrett, H A Collier, and O G Troyanskaya. Exploring the human genome with functional maps. *Genome Res*, 19(6):1093–1106, Jun 2009.
- [721] Y Chen and D Xu. Computational analyses of high-throughput protein-protein interaction data. *Curr Protein Pept Sci*, 4(3):159–181, Jun 2003.
- [722] I Lee and E M Marcotte. Effects of functional bias on supervised learning of a gene network model. *Methods Mol Biol*, 541:463–475, 2009.
- [723] M Tyers and M Mann. From genomics to proteomics. *Nature*, 422(6928):193–197, Mar 2003.
- [724] D Devos and R B Russell. A more complete, complexed and structured interactome. *Curr Opin Struct Biol*, 17(3):370–377, Jun 2007.
- [725] B Schwikowski, P Uetz, and S Fields. A network of protein-protein interactions in yeast. *Nat Biotechnol*, 18(12):1257–1261, Dec 2000.
- [726] R D Sleator and P Walsh. An overview of *in silico* protein function prediction. *Arch Microbiol*, 192(3):151–155, Mar 2010.
- [727] B Rost, J Liu, R Nair, K O Wrzeszczynski, and Y Ofra. Automatic prediction of protein function. *Cell Mol Life Sci*, 60(12):2637–2650, Dec 2003.
- [728] P Bork, T Dandekar, Y Diaz-Lazcoz, F Eisenhaber, M Huynen, and Y Yuan. Predicting function: from genes to genomes and back. *J Mol Biol*, 283(4):707–725, Nov 1998.
- [729] W R Pearson and D J Lipman. Improved tools for biological sequence comparison. *Proc Natl Acad Sci U S A*, 85(8):2444–2448, Apr 1988.
- [730] S F Altschul, T L Madden, A A Schäffer, J Zhang, Z Zhang, W Miller, and D J Lipman. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*, 25(17):3389–3402, Sep 1997.
- [731] R L Tatusov, E V Koonin, and D J Lipman. A genomic perspective on protein families. *Science*, 278(5338):631–637, Oct 1997.
- [732] E A Koppel, K P van Gisbergen, T B Geijtenbeek, and Y van Kooyk. Distinct functions of DC-SIGN and its homologues L-SIGN (DC-SIGNR) and mSIGNR1 in pathogen recognition and immune regulation. *Cell Microbiol*, 7(2):157–165, Feb 2005.
- [733] C P Ponting. Issues in predicting protein function from sequence. *Brief Bioinform*, 2(1):19–29, Mar 2001.
- [734] D Devos and A Valencia. Practical limits of function prediction. *Proteins*, 41(1):98–107, Oct 2000.
- [735] I Friedberg. Automated protein function prediction—the genomic challenge. *Brief Bioinform*, 7(3):225–242, Sep 2006.
- [736] D Lee, O Redfern, and C Orengo. Predicting protein function from sequence and structure. *Nat Rev Mol Cell Biol*, 8(12):995–1005, Dec 2007.
- [737] P D Karp. What we do not know about sequence analysis and sequence databases. *Bioinformatics*, 14(9):753–754, 1998.
- [738] S E Brenner. Errors in genome annotation. *Trends Genet*, 15(4):132–133, Apr 1999.

- [739] S S Krishna, R I Sadreyev, and N V Grishin. A tale of two ferredoxins: sequence similarity and structural differences. *BMC Struct Biol*, 6:8, 2006.
- [740] R Das and M Gerstein. A method using active-site sequence conservation to find functional shifts in protein families: application to the enzymes of central metabolism, leading to the identification of an anomalous isocitrate dehydrogenase in pathogens. *Proteins*, 55(2):455–463, May 2004.
- [741] J Pei and N V Grishin. AL2CO: calculation of positional conservation in a protein sequence alignment. *Bioinformatics*, 17(8):700–712, Aug 2001.
- [742] J D Thompson, T J Gibson, and D G Higgins. Multiple sequence alignment using ClustalW and ClustalX. *Curr Protoc Bioinformatics*, Chapter 2:Unit 2.3, Aug 2002.
- [743] S S Hannehalli and R B Russell. Analysis and prediction of functional sub-types from protein sequence alignments. *J Mol Biol*, 303(1):61–76, Oct 2000.
- [744] M Y Galperin and E V Koonin. Who’s your neighbor? New computational approaches for functional genomics. *Nat Biotechnol*, 18(6):609–613, Jun 2000.
- [745] J Tamames, G Casari, C Ouzounis, and A Valencia. Conserved clusters of functionally related genes in two bacterial genomes. *J Mol Evol*, 44(1):66–73, Jan 1997.
- [746] M Strong, P Mallick, M Pellegrini, M J Thompson, and D Eisenberg. Inference of protein function and protein linkages in *Mycobacterium tuberculosis* based on prokaryotic genome organization: a combined computational approach. *Genome Biol*, 4(9):R59, 2003.
- [747] C S Goh, A A Bogan, M Joachimiak, D Walther, and F E Cohen. Co-evolution of proteins with their interaction partners. *J Mol Biol*, 299(2):283–293, Jun 2000.
- [748] A E Lobley, T Nugent, C A Orengo, and D T Jones. FFPred: an integrated feature-based function prediction server for vertebrate proteomes. *Nucleic Acids Res*, 36(Web Server issue):297–302, Jul 2008.
- [749] B Rost. Marrying structure and genomics. *Structure*, 6(3):259–263, Mar 1998.
- [750] R A Laskowski, J D Watson, and J M Thornton. From protein structure to biochemical function? *J Struct Funct Genomics*, 4(2-3):167–177, 2003.
- [751] C Zhang and S H Kim. Overview of structural genomics: from structure to function. *Curr Opin Chem Biol*, 7(1):28–32, Feb 2003.
- [752] T I Zarembinski, L W Hung, H J Mueller-Dieckmann, K K Kim, H Yokota, R Kim, and S H Kim. Structure-based assignment of the biochemical function of a hypothetical protein: a test case of structural genomics. *Proc Natl Acad Sci U S A*, 95(26):15189–15193, Dec 1998.
- [753] D L Wild and M A S Saqi. Structural proteomics: inferring function from protein structure. *Current Proteomics*, 1:59–65, 2004.
- [754] L Wang, L Y Wu, Y Wang, X S Zhang, and L Chen. SANA: an algorithm for sequential and non-sequential protein structure alignment. *Amino Acids*, 39(2):417–425, Jul 2010.
- [755] J D Watson, R A Laskowski, and J M Thornton. Predicting protein function from sequence and structural data. *Curr Opin Struct Biol*, 15(3):275–284, Jun 2005.
- [756] K Mizuguchi, C M Deane, T L Blundell, and J P Overington. HOMSTRAD: a database of protein structure alignments for homologous families. *Protein Sci*, 7(11):2469–2471, Nov 1998.
- [757] J Casbon and M A Saqi. S4: structure-based sequence alignments of SCOP superfamilies. *Nucleic Acids Res*, 33(Database issue):219–222, Jan 2005.
- [758] H Berman, K Henrick, H Nakamura, and J L Markley. The worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data. *Nucleic Acids Res*, 35(Database issue):301–303, Jan 2007.

- [759] 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. *J Mol Biol*, 247(4):536–540, Apr 1995.
- [760] M Menke, B Berger, and L Cowen. Matt: local flexibility aids protein multiple structure alignment. *PLoS Comput Biol*, 4(1):e10, Jan 2008.
- [761] C Kim, C H Tai, and B Lee. Iterative refinement of structure-based sequence alignments by seed extension. *BMC Bioinformatics*, 10:210–210, 2009.
- [762] A R Ortiz, C E Strauss, and O Olmea. MAMMOTH (matching molecular models obtained from theory): an automated method for model comparison. *Protein Sci*, 11(11):2606–2621, Nov 2002.
- [763] I Ilinkin, J Ye, and R Janardan. Multiple structure alignment and consensus identification for proteins. *BMC Bioinformatics*, 11:71, 2010.
- [764] I N Shindyalov and P E Bourne. Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Eng*, 11(9):739–747, Sep 1998.
- [765] Y Ye and A Godzik. FATCAT: a web server for flexible structure comparison and structure similarity searching. *Nucleic Acids Res*, 32(Web Server issue):582–585, Jul 2004.
- [766] M S Madhusudhan, B M Webb, M A Marti-Renom, N Eswar, and A Sali. Alignment of multiple protein structures based on sequence and structure features. *Protein Eng Des Sel*, 22(9):569–574, Sep 2009.
- [767] X Miao, P J Waddell, and H Valafar. TALI: local alignment of protein structures using backbone torsion angles. *J Bioinform Comput Biol*, 6(1):163–181, Feb 2008.
- [768] J H Chiang and H C Yu. MeKE: discovering the functions of gene products from biomedical literature via sentence alignment. *Bioinformatics*, 19(11):1417–1422, Jul 2003.
- [769] C Blaschke, E A Leon, M Krallinger, and A Valencia. Evaluation of BioCreAtIvE assessment of task 2. *BMC Bioinformatics*, 6(Suppl 1):S16, 2005.
- [770] S Raychaudhuri, J T Chang, P D Sutphin, and R B Altman. Associating genes with Gene Ontology codes using a maximum entropy analysis of biomedical literature. *Genome Res*, 12(1):203–214, Jan 2002.
- [771] F M Couto, M J Silva, V Lee, E Dimmer, E Camon, R Apweiler, H Kirsch, and D Rebholz-Schuhmann. GOAnnotator: linking protein GO annotations to evidence text. *J Biomed Discov Collab*, 1:19, 2006.
- [772] C E Crangle, J M Cherry, E L Hong, and A Zbyslaw. Mining experimental evidence of molecular function claims from the literature. *Bioinformatics*, 23(23):3232–3240, Dec 2007.
- [773] N Daraselia, A Yuryev, S Egorov, I Mazo, and I Ispolatov. Automatic extraction of Gene Ontology annotation and its correlation with clusters in protein networks. *BMC Bioinformatics*, 8:243, 2007.
- [774] A Cakmak and G Ozsoyoglu. Discovering gene annotations in biomedical text databases. *BMC Bioinformatics*, 9:143, 2008.
- [775] S Jaeger, S Gaudan, U Leser, and D Rebholz-Schuhmann. Integrating protein-protein interactions and text mining for protein function prediction. *BMC Bioinformatics*, 9(Suppl 8):S2, 2008.
- [776] A Clare and R D King. Predicting gene function in *Saccharomyces cerevisiae*. *Bioinformatics*, 19 Suppl 2:42–49, Oct 2003.
- [777] A Mateos, J Dopazo, R Jansen, Y Tu, M Gerstein, and G Stolovitzky. Systematic learning of gene functional classes from dna array expression data by using multilayer perceptrons. *Genome Res*, 12(11):1703–1715, Nov 2002.
- [778] K Tsuda, H Shin, and B Schölkopf. Fast protein classification with multiple networks. *Bioinformatics*, 21 Suppl 2:S9–65, Sep 2005.

- [779] J Xiong, S Rayner, K Luo, Y Li, and S Chen. Genome wide prediction of protein function via a generic knowledge discovery approach based on evidence integration. *BMC Bioinformatics*, 7:268, 2006.
- [780] P Pavlidis, J Weston, J Cai, and W S Noble. Learning gene functional classifications from multiple data types. *J Comput Biol*, 9(2):401–411, 2002.
- [781] G Schatz and B Dobberstein. Common principles of protein translocation across membranes. *Science*, 271(5255):1519–1526, Mar 1996.
- [782] R K Niedenthal, L Riles, M Johnston, and J H Hegemann. Green fluorescent protein as a marker for gene expression and subcellular localization in budding yeast. *Yeast*, 12(8):773–786, Jun 1996.
- [783] J L Gardy and F S Brinkman. Methods for predicting bacterial protein subcellular localization. *Nat Rev Microbiol*, 4(10):741–751, Oct 2006.
- [784] Z P Feng. An overview on predicting the subcellular location of a protein. *In Silico Biol*, 2(3):291–303, 2002.
- [785] K N Pandey. Small peptide recognition sequence for intracellular sorting. *Curr Opin Biotechnol*, 21(5):611–620, Oct 2010.
- [786] F Eisenhaber and P Bork. Wanted: subcellular localization of proteins based on sequence. *Trends Cell Biol*, 8(4):169–170, Apr 1998.
- [787] K Nakai and P Horton. PSORT: a program for detecting sorting signals in proteins and predicting their subcellular localization. *Trends Biochem Sci*, 24(1):34–36, Jan 1999.
- [788] K Nakai. Protein sorting signals and prediction of subcellular localization. *Adv Protein Chem*, 54:277–344, 2000.
- [789] J Fang, R J Haasl, Y Dong, and G H Lushington. Discover protein sequence signatures from protein-protein interaction data. *BMC Bioinformatics*, 6:277, 2005.
- [790] L J Zhao and R Padmanabhan. Nuclear transport of adenovirus DNA polymerase is facilitated by interaction with preterminal protein. *Cell*, 55(6):1005–1015, Dec 1988.
- [791] J Garcia-Bustos, J Heitman, and M N Hall. Nuclear protein localization. *Biochim Biophys Acta*, 1071(1):83–101, Mar 1991.
- [792] J Cedano, P Aloy, J A Pérez-Pons, and E Querol. Relation between amino acid composition and cellular location of proteins. *J Mol Biol*, 266(3):594–600, Feb 1997.
- [793] H Nakashima and K Nishikawa. Discrimination of intracellular and extracellular proteins using amino acid composition and residue-pair frequencies. *J Mol Biol*, 238(1):54–61, Apr 1994.
- [794] Z P Feng. Prediction of the subcellular location of prokaryotic proteins based on a new representation of the amino acid composition. *Biopolymers*, 58(5):491–499, Apr 2001.
- [795] K C Chou. Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes. *Bioinformatics*, 21(1):10–19, Jan 2005.
- [796] E M Marcotte, I Xenarios, A M van Der Blik, and D Eisenberg. Localizing proteins in the cell from their phylogenetic profiles. *Proc Natl Acad Sci U S A*, 97(22):12115–12120, Oct 2000.
- [797] A Reinhardt and T Hubbard. Using neural networks for prediction of the subcellular location of proteins. *Nucleic Acids Res*, 26(9):2230–2236, May 1998.
- [798] O Emanuelsson, H Nielsen, and G von Heijne. ChloroP, a neural network-based method for predicting chloroplast transit peptides and their cleavage sites. *Protein Sci*, 8(5):978–984, May 1999.
- [799] S Hua and Z Sun. Support vector machine approach for protein subcellular localization prediction. *Bioinformatics*, 17(8):721–728, Aug 2001.
- [800] C S Yu, Y C Chen, C H Lu, and J K Hwang. Prediction of protein subcellular localization. *Proteins*, 64(3):643–651, Aug 2006.

- [801] J Y Shi, S W Zhang, Q Pan, Y M Cheng, and J Xie. Prediction of protein subcellular localization by support vector machines using multi-scale energy and pseudo amino acid composition. *Amino Acids*, 33(1):69–74, Jul 2007.
- [802] Q Cui, T Jiang, B Liu, and S Ma. Esub8: a novel tool to predict protein subcellular localizations in eukaryotic organisms. *BMC Bioinformatics*, 5:66, May 2004.
- [803] T Tamura and T Akutsu. Subcellular location prediction of proteins using support vector machines with alignment of block sequences utilizing amino acid composition. *BMC Bioinformatics*, 8:466, 2007.
- [804] A Garg, M Bhasin, and G P Raghava. Support vector machine-based method for subcellular localization of human proteins using amino acid compositions, their order, and similarity search. *J Biol Chem*, 280(15):14427–14432, Apr 2005.
- [805] K J Park and M Kanehisa. Prediction of protein subcellular locations by support vector machines using compositions of amino acids and amino acid pairs. *Bioinformatics*, 19(13):1656–1663, Sep 2003.
- [806] P Horton, K J Park, T Obayashi, N Fujita, H Harada, C J Adams-Collier, and K Nakai. WoLF PSORT: protein localization predictor. *Nucleic Acids Res*, 35(Web Server issue):585–587, Jul 2007.
- [807] P Horton and K Nakai. Better prediction of protein cellular localization sites with the k nearest neighbors classifier. *Proc Int Conf Intell Syst Mol Biol*, 5:147–152, 1997.
- [808] Q Xu, D H Hu, H Xue, W Yu, and Q Yang. Semi-supervised protein subcellular localization. *BMC Bioinformatics*, 10(Suppl 1):S47, 2009.
- [809] A Drawid and M Gerstein. A Bayesian system integrating expression data with sequence patterns for localizing proteins: comprehensive application to the yeast genome. *J Mol Biol*, 301(4):1059–1075, Aug 2000.
- [810] K C Chou and H B Shen. Predicting protein subcellular location by fusing multiple classifiers. *J Cell Biochem*, 99(2):517–527, Oct 2006.
- [811] K C Chou and H B Shen. Hum-PLoc: a novel ensemble classifier for predicting human protein subcellular localization. *Biochem Biophys Res Commun*, 347(1):150–157, Aug 2006.
- [812] R Nair and B Rost. Inferring sub-cellular localization through automated lexical analysis. *Bioinformatics*, 18 Suppl 1:78–86, 2002.
- [813] K van Auken, J Jaffery, J Chan, H M Müller, and P W Sternberg. Semi-automated curation of protein subcellular localization: a text mining-based approach to Gene Ontology (GO) Cellular Component curation. *BMC Bioinformatics*, 10:228, 2009.
- [814] M Zhu and S Zhao. Candidate gene identification approach: progress and challenges. *Int J Biol Sci*, 3(7):420–427, 2007.
- [815] A D Roses. Pharmacogenetics and the practice of medicine. *Nature*, 405(6788):857–865, Jun 2000.
- [816] S Aerts, D Lambrechts, S Maity, P Van Loo, B Coessens, F De Smet, L C Tranchevent, B De Moor, P Marynen, B Hassan, P Carmeliet, and Y Moreau. Gene prioritization through genomic data fusion. *Nat Biotechnol*, 24(5):537–544, May 2006.
- [817] C Perez-Iratxeta, P Bork, and M A Andrade-Navarro. Update of the G2D tool for prioritization of gene candidates to inherited diseases. *Nucleic Acids Res*, 35(Web Server issue):212–216, Jul 2007.
- [818] E A Adie, R R Adams, K L Evans, D J Porteous, and B S Pickard. SUSPECTS: enabling fast and effective prioritization of positional candidates. *Bioinformatics*, 22(6):773–774, Mar 2006.
- [819] E A Adie, R R Adams, K L Evans, D J Porteous, and B S Pickard. Speeding disease gene discovery by sequence based candidate prioritization. *BMC Bioinformatics*, 6:55, 2005.
- [820] F S Turner, D R Clutterbuck, and C A Semple. POCUS: mining genomic sequence annotation to predict disease genes. *Genome Biol*, 4(11):R75, 2003.

- [821] N López-Bigas and C A Ouzounis. Genome-wide identification of genes likely to be involved in human genetic disease. *Nucleic Acids Res*, 32(10):3108–3114, 2004.
- [822] M Oti, J van Reeuwijk, M A Huynen, and H G Brunner. Conserved co-expression for candidate disease gene prioritization. *BMC Bioinformatics*, 9:208, 2008.
- [823] M Oti and H G Brunner. The modular nature of genetic diseases. *Clin Genet*, 71(1):1–11, Jan 2007.
- [824] M A van Driel, J Bruggeman, G Vriend, H G Brunner, and J A Leunissen. A text-mining analysis of the human phenome. *Eur J Hum Genet*, 14(5):535–542, May 2006.
- [825] K Lage, E O Karlberg, Z M Storling, P I Olason, A G Pedersen, O Rigina, A M Hinsby, Z Tümer, F Pociot, N Tommerup, Y Moreau, and S Brunak. A human phenome-interactome network of protein complexes implicated in genetic disorders. *Nat Biotechnol*, 25(3):309–316, Mar 2007.
- [826] U Ala, R M Piro, E Grassi, C Damasco, L Silengo, M Oti, P Provero, and F Di Cunto. Prediction of human disease genes by human-mouse conserved coexpression analysis. *PLoS Comput Biol*, 4(3):e1000043, Mar 2008.
- [827] S B Davidson, C Overton, and P Buneman. Challenges in integrating biological data sources. *J Comput Biol*, 2(4):557–572, 1995.
- [828] B Smith, M Ashburner, C Rosse, J Bard, W Bug, W Ceusters, L J Goldberg, K Eilbeck, A Ireland, C J Mungall, OBI Consortium, N Leontis, P Rocca-Serra, A Ruttenberg, S A Sansone, R H Scheuermann, N Shah, P L Whetzel, and S Lewis. The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nat Biotechnol*, 25(11):1251–1255, Nov 2007.
- [829] A Ruepp, A Zollner, D Maier, K Albermann, J Hani, M Mokrejs, I Tetko, U Güldener, G Mannhaupt, M Münsterkötter, and H W Mewes. The FunCat, a functional annotation scheme for systematic classification of proteins from whole genomes. *Nucleic Acids Res*, 32(18):5539–5545, 2004.
- [830] E C Webb. *Enzyme nomenclature 1992 : recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology on the nomenclature and classification of enzymes*. Published for the International Union of Biochemistry and Molecular Biology by Academic Press, San Diego, 1992.
- [831] A Hamosh, A F Scott, J S Amberger, C A Bocchini, and V A McKusick. Online Mendelian Inheritance in Man (omim), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res*, 33(Database issue):514–517, Jan 2005.
- [832] K Hashimoto, S Goto, S Kawano, K F Aoki-Kinoshita, N Ueda, M Hamajima, T Kawasaki, and M Kanehisa. KEGG as a glycome informatics resource. *Glycobiology*, 16(5):63R–70R, May 2006.
- [833] M Kanehisa, M Araki, S Goto, M Hattori, M Hirakawa, M Itoh, T Katayama, S Kawashima, S Okuda, T Tokimatsu, and Y Yamanishi. KEGG for linking genomes to life and the environment. *Nucleic Acids Res*, 36(Database issue):480–484, Jan 2008.
- [834] M Kanehisa, S Goto, M Furumichi, M Tanabe, and M Hirakawa. KEGG for representation and analysis of molecular networks involving diseases and drugs. *Nucleic Acids Res*, 38(Database issue):355–360, Jan 2010.
- [835] M Kanehisa, S Goto, M Hattori, K F Aoki-Kinoshita, M Itoh, S Kawashima, T Katayama, M Araki, and M Hirakawa. From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Res*, 34(Database issue):354–357, Jan 2006.
- [836] M Kanehisa, S Goto, S Kawashima, Y Okuno, and M Hattori. The KEGG resource for deciphering the genome. *Nucleic Acids Res*, 32(Database issue):277–280, Jan 2004.
- [837] A Masoudi-Nejad, S Goto, T R Endo, and M Kanehisa. KEGG bioinformatics resource for plant genomics research. *Methods Mol Biol*, 406:437–458, 2007.
- [838] S Okuda, T Yamada, M Hamajima, M Itoh, T Katayama, P Bork, S Goto, and M Kanehisa. KEGG Atlas mapping for global analysis of metabolic pathways. *Nucleic Acids Res*, 36(Web Server issue):423–426, Jul 2008.

- [839] A V Antonov, E E Schmidt, S Dietmann, M Krestyaninova, and H Hermjakob. R spider: a network-based analysis of gene lists by combining signaling and metabolic pathways from Reactome and KEGG databases. *Nucleic Acids Res*, 38(Web Server issue):78–83, Jul 2010.
- [840] K F Aoki, A Yamaguchi, N Ueda, T Akutsu, H Mamitsuka, S Goto, and M Kanehisa. KCaM (KEGG Carbohydrate Matcher): a software tool for analyzing the structures of carbohydrate sugar chains. *Nucleic Acids Res*, 32(Web Server issue):267–272, Jul 2004.
- [841] F Le Fèvre, S Smidtas, C Combe, M Durot, F d’Alché Buc, and V Schachter. CycSim—an online tool for exploring and experimenting with genome-scale metabolic models. *Bioinformatics*, 25(15):1987–1988, Aug 2009.
- [842] I Medina, J Carbonell, L Pulido, S C Madeira, S Goetz, A Conesa, J Tárraga, A Pascual-Montano, R Nogales-Cadenas, J Santoyo, F García, M Marbà, D Montaner, and J Dopazo. Babelomics: an integrative platform for the analysis of transcriptomics, proteomics and genomic data with advanced functional profiling. *Nucleic Acids Res*, 38(Web Server issue):210–213, Jul 2010.
- [843] Y Moriya, M Itoh, S Okuda, A C Yoshizawa, and M Kanehisa. KAAS: an automatic genome annotation and pathway reconstruction server. *Nucleic Acids Res*, 35(Web Server issue):182–185, Jul 2007.
- [844] K Moutselos, I Kanaris, A Chatziioannou, I Maglogiannis, and F N Kolisis. KEGGconverter: a tool for the *in-silico* modelling of metabolic networks of the KEGG Pathways database. *BMC Bioinformatics*, 10:324, 2009.
- [845] C H Sun, M S Kim, Y Han, and G S Yi. COFECO: composite function annotation enriched by protein complex data. *Nucleic Acids Res*, 37(Web Server issue):350–355, Jul 2009.
- [846] A B Tchagang, A Gawronski, H Bérubé, S Phan, F Famili, and Y Pan. GOAL: a software tool for assessing biological significance of genes groups. *BMC Bioinformatics*, 11:229, 2010.
- [847] T Wylie, J Martin, S Abubucker, Y Yin, D Messina, Z Wang, J P McCarter, and M Mitreva. NemaPath: online exploration of KEGG-based metabolic pathways for nematodes. *BMC Genomics*, 9:525, 2008.
- [848] F Azuaje, F Al-Shahrour, and J Dopazo. Ontology-driven approaches to analyzing data in functional genomics. *Methods Mol Biol*, 316:67–86, 2006.
- [849] J Zhu, J Wang, Z Guo, M Zhang, D Yang, Y Li, D Wang, and G Xiao. GO-2D: identifying 2-dimensional cellular-localized functional modules in Gene Ontology. *BMC Genomics*, 8:30, 2007.
- [850] Gene Ontology Consortium. The Gene Ontology in 2010: extensions and refinements. *Nucleic Acids Res*, 38(Database issue):331–335, Jan 2010.
- [851] S Myhre, H Tveit, T Mollestad, and A Laegreid. Additional Gene Ontology structure for improved biological reasoning. *Bioinformatics*, 22(16):2020–2027, Aug 2006.
- [852] M Dejongh, P Van Dort, and B Ramsay. Linking molecular function and biological process terms in the ontology for gene expression data analysis. *Conf Proc IEEE Eng Med Biol Soc*, 4:2984–2986, 2004.
- [853] D Pal. On Gene Ontology and function annotation. *Bioinformation*, 1(3):97–98, 2006.
- [854] D P Hill, J A Blake, J E Richardson, and M Ringwald. Extension and integration of the Gene Ontology (GO): combining GO vocabularies with external vocabularies. *Genome Res*, 12(12):1982–1991, Dec 2002.
- [855] K R Christie, E L Hong, and J M Cherry. Functional annotations for the *Saccharomyces cerevisiae* genome: the knowns and the known unknowns. *Trends Microbiol*, 17(7):286–294, Jul 2009.
- [856] D Barrell, E Dimmer, R P Huntley, D Binns, C O’Donovan, and R Apweiler. The GOA database in 2009—an integrated Gene Ontology Annotation resource. *Nucleic Acids Res*, 37(Database issue):396–403, Jan 2009.

- [857] L du Plessis, N Skunca, and C Dessimoz. The what, where, how and why of Gene Ontology—a primer for bioinformaticians. *Brief Bioinform*, 12(6):723–735, Feb 2011.
- [858] D Devos and A Valencia. Intrinsic errors in genome annotation. *Trends Genet*, 17(8):429–431, Aug 2001.
- [859] W Zhang, Q D Morris, R Chang, O Shai, M A Bakowski, N Mitsakakis, N Mohammad, M D Robinson, R Zirngibl, E Somogyi, N Laurin, E Eftekharpour, E Sat, J Grigull, Q Pan, W T Peng, N Krogan, J Greenblatt, M Fehlings, D van der Kooy, J Aubin, B G Bruneau, J Rossant, B J Blencowe, B J Frey, and T R Hughes. The functional landscape of mouse gene expression. *J Biol*, 3(5):21, 2004.
- [860] H Wu, Z Su, F Mao, V Olman, and Y Xu. Prediction of functional modules based on comparative genome analysis and Gene Ontology application. *Nucleic Acids Res*, 33(9):2822–2837, 2005.
- [861] D P Hill, B Smith, M S McAndrews-Hill, and J A Blake. Gene Ontology annotations: what they mean and where they come from. *BMC Bioinformatics*, 9(Suppl 5):S2, 2008.
- [862] S Falcon and R Gentleman. Using GOstats to test gene lists for GO term association. *Bioinformatics*, 23(2):257–258, Jan 2007.
- [863] A Schlicker, F S Domingues, J Rahnenführer, and T Lengauer. A new measure for functional similarity of gene products based on Gene Ontology. *BMC Bioinformatics*, 7:302, 2006.
- [864] X Wu, L Zhu, J Guo, D Y Zhang, and K Lin. Prediction of yeast protein-protein interaction network: insights from the Gene Ontology and annotations. *Nucleic Acids Res*, 34(7):2137–2150, 2006.
- [865] F Pazos and M J Sternberg. Automated prediction of protein function and detection of functional sites from structure. *Proc Natl Acad Sci U S A*, 101(41):14754–14759, Oct 2004.
- [866] S Mostafavi and Q Morris. Fast integration of heterogeneous data sources for predicting gene function with limited annotation. *Bioinformatics*, 26(14):1759–1765, May 2010.
- [867] P Resnik. Using information content to evaluate semantic similarity in a taxonomy. In *IJCAI’95: Proceedings of the 14th international joint conference on Artificial intelligence*, pages 448–453, San Francisco, CA, USA, 1995. Morgan Kaufmann Publishers Inc.
- [868] J Wang, X Zhou, J Zhu, C Zhou, and Z Guo. Revealing and avoiding bias in semantic similarity scores for protein pairs. *BMC Bioinformatics*, 11(1):290, May 2010.
- [869] C Pesquita, D Faria, H Bastos, A E Ferreira, A O Falcão, and F M Couto. Metrics for GO based protein semantic similarity: a systematic evaluation. *BMC Bioinformatics*, 9(Suppl 5):S4, 2008.
- [870] T Xu, L Du, and Y Zhou. Evaluation of GO-based functional similarity measures using *S. cerevisiae* protein interaction and expression profile data. *BMC Bioinformatics*, 9(1):472, Nov 2008.
- [871] Z Du, L Li, C F Chen, P S Yu, and J Z Wang. G-SESAME: web tools for GO-term-based gene similarity analysis and knowledge discovery. *Nucleic Acids Res*, 37(Web Server issue):345–349, Jul 2009.
- [872] J Z Wang, Z Du, R Payattakool, P S Yu, and C F Chen. A new method to measure the semantic similarity of GO terms. *Bioinformatics*, 23(10):1274–1281, Mar 2007.
- [873] M Mistry and P Pavlidis. Gene Ontology term overlap as a measure of gene functional similarity. *BMC Bioinformatics*, 9:327, 2008.
- [874] J Chabalier, J Mosser, and A Burgun. A transversal approach to predict gene product networks from ontology-based similarity. *BMC Bioinformatics*, 8:235, 2007.
- [875] Y R Cho, W Hwang, M Ramanathan, and A Zhang. Semantic integration to identify overlapping functional modules in protein interaction networks. *BMC Bioinformatics*, 8:265, 2007.
- [876] K Ovaska, M Laakso, and S Hautaniemi. Fast Gene Ontology based clustering for microarray experiments. *BioData Min*, 1(1):11, 2008.

- [877] 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, Jul 2003.
- [878] J L Chen, Y Liu, L T Sam, J Li, and Y A Lussier. Evaluation of high-throughput functional categorization of human disease genes. *BMC Bioinformatics*, 8(Suppl 3):S7, 2007.
- [879] 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, 2009.
- [880] T Joshi and D Xu. Quantitative assessment of relationship between sequence similarity and function similarity. *BMC Genomics*, 8:222, 2007.
- [881] J L Sevilla, V Segura, A Podhorski, E Guruceaga, J M Mato, L A Martínez-Cruz, F J Corrales, and A Rubio. Correlation between gene expression and GO semantic similarity. *IEEE/ACM Trans Comput Biol Bioinform*, 2(4):330–338, Oct-Dec 2005.
- [882] H Wang and F Azuaje. An ontology-driven clustering method for supporting gene expression analysis. In *CBMS '05: Proceedings of the 18th IEEE Symposium on Computer-Based Medical Systems*, pages 389–394, Washington, DC, USA, 2005. IEEE Computer Society.
- [883] H K Lee, A K Hsu, J Sajdak, J Qin, and P Pavlidis. Coexpression analysis of human genes across many microarray data sets. *Genome Res*, 14(6):1085–1094, Jun 2004.
- [884] P Khatri and S Drăghici. Ontological analysis of gene expression data: current tools, limitations, and open problems. *Bioinformatics*, 21(18):3587–3595, Sep 2005.
- [885] U Yu, Y J Choi, J K Choi, and S Kim. TO-GO: a Java-based Gene Ontology navigation environment. *Bioinformatics*, 21(17):3580–3581, Sep 2005.
- [886] R S Sealfon, M A Hibbs, C Huttenhower, C L Myers, and O G Troyanskaya. GOLEM: an interactive graph-based gene-ontology navigation and analysis tool. *BMC Bioinformatics*, 7:443, 2006.
- [887] O Garcia, C Saveanu, M Cline, M Fromont-Racine, A Jacquier, B Schwikowski, and T Aittokallio. GOLORize: a Cytoscape plug-in for network visualization with Gene Ontology-based layout and coloring. *Bioinformatics*, 23(3):394–396, Feb 2007.
- [888] D W Huang, B T Sherman, and R A Lempicki. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res*, 37(1):1–13, Jan 2009.
- [889] D Martin, C Brun, E Remy, P Mouren, D Thieffry, and B Jacq. GOToolBox: functional analysis of gene datasets based on Gene Ontology. *Genome Biol*, 5(12):R101, 2004.
- [890] S W Doniger, N Salomonis, K D Dahlquist, K Vranizan, S C Lawlor, and B R Conklin. MAPPFinder: using Gene Ontology and GenMAPP to create a global gene-expression profile from microarray data. *Genome Biol*, 4(1):R7, 2003.
- [891] S Bauer, S Grossmann, M Vingron, and P N Robinson. Ontologizer 2.0—a multifunctional tool for GO term enrichment analysis and data exploration. *Bioinformatics*, 24(14):1650–1651, Jul 2008.
- [892] S Maere, K Heymans, and M Kuiper. BiNGO: a Cytoscape plugin to assess overrepresentation of Gene Ontology categories in biological networks. *Bioinformatics*, 21(16):3448–3449, Aug 2005.
- [893] G Bindea, B Mlecnik, H Hackl, P Charoentong, M Tosolini, A Kirilovsky, W H Fridman, F Pagès, Z Trajanoski, and J Galon. ClueGO: a Cytoscape plug-in to decipher functionally grouped Gene Ontology and pathway annotation networks. *Bioinformatics*, 25(8):1091–1093, Apr 2009.
- [894] N Massjouni, C G Rivera, and T M Murali. VIRGO: computational prediction of gene functions. *Nucleic Acids Res*, 34(Web Server issue):340–344, Jul 2006.
- [895] J R Bradford, C J Needham, P Tedder, M A Care, A J Bulpitt, and D R Westhead. GO-At: *In silico* prediction of gene function in *Arabidopsis thaliana* by combining heterogeneous data. *Plant J*, 61(4):713–721, Nov 2009.

- [896] G Dennis, B T Sherman, D A Hosack, J Yang, W Gao, H C Lane, and R A Lempicki. DAVID: Database for annotation, visualization, and integrated discovery. *Genome Biol*, 4(5):P3, 2003.
- [897] P Törönen, P Pehkonen, and L Holm. Generation of Gene Ontology benchmark datasets with various types of positive signal. *BMC Bioinformatics*, 10:319, 2009.
- [898] S Oliver. Guilt-by-association goes global. *Nature*, 403(6770):601–603, Feb 2000.
- [899] I Lee, R Narayanaswamy, and E M Marcotte. Bioinformatic prediction of yeast gene function. In Ian Stansfield and Michael JR Stark, editors, *Yeast Gene Analysis - Second Edition*, volume 36 of *Methods in Microbiology*, chapter 24, pages 597 – 628. Academic Press, 2007.
- [900] A Droit, G G Poirier, and J M Hunter. Experimental and bioinformatic approaches for interrogating protein-protein interactions to determine protein function. *J Mol Endocrinol*, 34(2):263–280, Apr 2005.
- [901] R Jansen, D Greenbaum, and M Gerstein. Relating whole-genome expression data with protein-protein interactions. *Genome Res*, 12(1):37–46, Jan 2002.
- [902] E Nabieva, K Jim, A Agarwal, B Chazelle, and M Singh. Whole-proteome prediction of protein function via graph-theoretic analysis of interaction maps. *Bioinformatics*, 21 Suppl 1:302–310, Jun 2005.
- [903] Z Barutcuoglu, R E Schapire, and O G Troyanskaya. Hierarchical multi-label prediction of gene function. *Bioinformatics*, 22(7):830–836, Apr 2006.
- [904] L Aravind. Guilt by association: contextual information in genome analysis. *Genome Res*, 10(8):1074–1077, Aug 2000.
- [905] C J Wolfe, I S Kohane, and A J Butte. Systematic survey reveals general applicability of "guilt-by-association" within gene coexpression networks. *BMC Bioinformatics*, 6:227, 2005.
- [906] P Hu, G Bader, D A Wigle, and A Emili. Computational prediction of cancer-gene function. *Nat Rev Cancer*, 7(1):23–34, Jan 2007.
- [907] M Deng, Z Tu, F Sun, and T Chen. Mapping Gene Ontology to proteins based on protein-protein interaction data. *Bioinformatics*, 20(6):895–902, Apr 2004.
- [908] Y Wu and S Lonardi. A linear-time algorithm for predicting functional annotations from PPI networks. *J Bioinform Comput Biol*, 6(6):1049–1065, Dec 2008.
- [909] N Nariai, E D Kolaczyk, and S Kasif. Probabilistic protein function prediction from heterogeneous genome-wide data. *PLoS ONE*, 2:e337, 2007.
- [910] J McDermott, R Bumgarner, and R Samudrala. Functional annotation from predicted protein interaction networks. *Bioinformatics*, 21(15):3217–3226, Aug 2005.
- [911] 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, Jul 2006.
- [912] R Llewellyn and D S Eisenberg. Annotating proteins with generalized functional linkages. *Proc Natl Acad Sci U S A*, 105(46):17700–17705, Nov 2008.
- [913] H Hishigaki, K Nakai, T Ono, A Tanigami, and T Takagi. Assessment of prediction accuracy of protein function from protein–protein interaction data. *Yeast*, 18(6):523–531, Apr 2001.
- [914] V Farutin, K Robison, E Lightcap, V Dancik, A Ruttenberg, S Letovsky, and J Pradines. Edge-count probabilities for the identification of local protein communities and their organization. *Proteins*, 62(3):800–818, Mar 2006.
- [915] M P Brown, W N Grundy, D Lin, N Cristianini, C W Sugnet, T S Furey, M Ares, and D Haussler. Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proc Natl Acad Sci U S A*, 97(1):262–267, Jan 2000.

- [916] A Tanay, R Sharan, M Kupiec, and R Shamir. Revealing modularity and organization in the yeast molecular network by integrated analysis of highly heterogeneous genomewide data. *Proc Natl Acad Sci U S A*, 101(9):2981–2986, Mar 2004.
- [917] M Leone and A Pagnani. Predicting protein functions with message passing algorithms. *Bioinformatics*, 21(2):239–247, Jan 2005.
- [918] S Sun, Y Zhao, Y Jiao, Y Yin, L Cai, Y Zhang, H Lu, R Chen, and D Bu. Faster and more accurate global protein function assignment from protein interaction networks using the MFGO algorithm. *FEBS Lett*, 580(7):1891–1896, Mar 2006.
- [919] Y R Cho, L Shi, M Ramanathan, and A Zhang. A probabilistic framework to predict protein function from interaction data integrated with semantic knowledge. *BMC Bioinformatics*, 9:382, 2008.
- [920] A R Henderson. Assessing test accuracy and its clinical consequences: a primer for receiver operating characteristic curve analysis. *Ann Clin Biochem*, 30 (Pt 6):521–539, Nov 1993.
- [921] J A Hanley and B J McNeil. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, 143(1):29–36, Apr 1982.
- [922] D R Rhodes, J Yu, K Shanker, N Deshpande, R Varambally, D Ghosh, T Barrette, A Pandey, and A M Chinnaiyan. ONCOMINE: a cancer microarray database and integrated data-mining platform. *Neoplasia*, 6(1):1–6, Jan-Feb 2004.
- [923] P Hernandez, X Sole, J Valls, V Moreno, G Capella, A Urruticoechea, and M A Pujana. Integrative analysis of a cancer somatic mutome. *Mol Cancer*, 6:13, 2007.
- [924] M A van Driel, K Cuelenaere, P P Kemmeren, J A Leunissen, H G Brunner, and G Vriend. GeneSeeker: extraction and integration of human disease-related information from web-based genetic databases. *Nucleic Acids Res*, 33(Web Server issue):758–761, Jul 2005.
- [925] M Oti, B Snel, M A Huynen, and H G Brunner. Predicting disease genes using protein-protein interactions. *J Med Genet*, 43(8):691–698, Aug 2006.
- [926] R A George, J Y Liu, L L Feng, R J Bryson-Richardson, D Fatkin, and M A Wouters. Analysis of protein sequence and interaction data for candidate disease gene prediction. *Nucleic Acids Res*, 34(19):e130, 2006.
- [927] J Chen, E E Bardes, B J Aronow, and A G Jegga. ToppGene Suite for gene list enrichment analysis and candidate gene prioritization. *Nucleic Acids Res*, 37(Web Server issue):305–311, Jul 2009.
- [928] L C Tranchevent, R Barriot, S Yu, S Van Vooren, P Van Loo, B Coessens, B De Moor, S Aerts, and Y Moreau. ENDEAVOUR update: a web resource for gene prioritization in multiple species. *Nucleic Acids Res*, 36(Web Server issue):377–384, Jul 2008.
- [929] H B Fraser and J B Plotkin. Using protein complexes to predict phenotypic effects of gene mutation. *Genome Biol*, 8(11):R252, 2007.
- [930] X Wu, R Jiang, M Q Zhang, and S Li. Network-based global inference of human disease genes. *Mol Syst Biol*, 4:189, 2008.
- [931] T Ideker and R Sharan. Protein networks in disease. *Genome Res*, 18(4):644–652, Apr 2008.
- [932] S Rossi, D Masotti, C Nardini, E Bonora, G Romeo, E Macii, L Benini, and S Volinia. TOM: a web-based integrated approach for identification of candidate disease genes. *Nucleic Acids Res*, 34(Web Server issue):285–292, Jul 2006.
- [933] H Y Chuang, E Lee, Y T Liu, D Lee, and T Ideker. Network-based classification of breast cancer metastasis. *Mol Syst Biol*, 3:140, 2007.
- [934] S E Calvano, W Xiao, D R Richards, R M Felciano, H V Baker, R J Cho, R O Chen, B H Brownstein, J P Cobb, S K Tschoeke, C Miller-Graziano, L L Moldawer, M N Mindrinos, R W Davis, R G Tompkins, S F Lowry, and Inflamm and Host Response to Injury Large Scale Collab. Res. Program. A network-based analysis of systemic inflammation in humans. *Nature*, 437(7061):1032–1037, Oct 2005.

- [935] J Chen, B J Aronow, and A G Jegga. Disease candidate gene identification and prioritization using protein interaction networks. *BMC Bioinformatics*, 10:73, 2009.
- [936] K I Goh, M E Cusick, D Valle, B Childs, M Vidal, and A L Barabási. The human disease network. *Proc Natl Acad Sci U S A*, 104(21):8685–8690, May 2007.
- [937] M G Walker, W Volkmuth, E Sprinzak, D Hodgson, and T Klingler. Prediction of gene function by genome-scale expression analysis: prostate cancer-associated genes. *Genome Res*, 9(12):1198–1203, Dec 1999.
- [938] M A Pujana, J D Han, L M Starita, K N Stevens, M Tewari, J S Ahn, G Rennert, V Moreno, T Kirchhoff, B Gold, V Assmann, W M Elshamy, J F Rual, D Levine, L S Rozek, R S Gelman, K C Gunsalus, R A Greenberg, B Sobhian, N Bertin, K Venkatesan, N Ayivi-Guedehoussou, X Solé, P Hernández, C Lázaro, K L Nathanson, B L Weber, M E Cusick, D E Hill, K Offit, D M Livingston, S B Gruber, J D Parvin, and M Vidal. Network modeling links breast cancer susceptibility and centrosome dysfunction. *Nat Genet*, 39(11):1338–1349, Nov 2007.
- [939] S Köhler, S Bauer, D Horn, and P N Robinson. Walking the interactome for prioritization of candidate disease genes. *Am J Hum Genet*, 82(4):949–958, Apr 2008.
- [940] O Vanunu and R Sharan. A propagation-based algorithm for inferring gene-disease associations. In *Proceeding of the German Conference on Bioinformatics*, pages 54–63, 2008.
- [941] J Xu and Y Li. Discovering disease-genes by topological features in human protein-protein interaction network. *Bioinformatics*, 22(22):2800–2805, Nov 2006.
- [942] B Linghu, E S Snitkin, Z Hu, Y Xia, and C Delisi. Genome-wide prioritization of disease genes and identification of disease-disease associations from an integrated human functional linkage network. *Genome Biol*, 10(9):R91, 2009.
- [943] W Pan. Network-based model weighting to detect multiple loci influencing complex diseases. *Hum Genet*, 124(3):225–234, Oct 2008.
- [944] R L Ho and C A Lieu. Systems biology: an evolving approach in drug discovery and development. *Drugs R D*, 9(4):203–216, 2008.
- [945] M Kuhn, M Campillos, P González, L J Jensen, and P Bork. Large-scale prediction of drug-target relationships. *FEBS Lett*, 582(8):1283–1290, Feb 2008.
- [946] L Hood. Systems biology: integrating technology, biology, and computation. *Mech Ageing Dev*, 124(1):9–16, Jan 2003.
- [947] L Paris, E Bononi, and G Bazzoni. Network analysis of cell adhesion: adhesomes as context-defined subnetworks. *Commun Integr Biol*, 2(1):20–22, 2009.
- [948] D Botstein, S A Chervitz, and J M Cherry. Yeast as a model organism. *Science*, 277(5330):1259–1260, Aug 1997.
- [949] D Botstein and G R Fink. Yeast: an experimental organism for modern biology. *Science*, 240(4858):1439–1443, Jun 1988.
- [950] K Altmann, M Dürr, and B Westermann. *Saccharomyces cerevisiae* as a model organism to study mitochondrial biology. *Methods in Molecular Biology*, 372:81–90, June 2007.
- [951] D E Bassett, M S Boguski, and P Hieter. Yeast genes and human disease. *Nature*, 379(6566):589–590, Feb 1996.
- [952] G M Rubin, M D Yandell, J R Wortman, G L Gabor Miklos, C R Nelson, I K Hariharan, M E Fortini, P W Li, R Apweiler, W Fleischmann, J M Cherry, S Henikoff, M P Skupski, S Misra, M Ashburner, E Birney, M S Boguski, T Brody, P Brokstein, S E Celniker, S A Chervitz, D Coates, A Cravchik, A Gabrielian, R F Galle, W M Gelbart, R A George, L S Goldstein, F Gong, P Guan, N L Harris, B A Hay, R A Hoskins, J Li, Z Li, R O Hynes, S J Jones, P M Kuehl, B Lemaitre, J T Littleton, D K Morrison, C Mungall, P H O’Farrell, O K Pickeral, C Shue, L B VossHall, J Zhang, Q Zhao, X H Zheng, and S Lewis. Comparative genomics of the eukaryotes. *Science*, 287(5461):2204–2215, Mar 2000.

- [953] L H Hartwell. Yeast and cancer. *Biosci Rep*, 24(4-5):523–544, Aug-Oct 2004.
- [954] K P O'Brien, M Remm, and E L Sonnhammer. Inparanoid: a comprehensive database of eukaryotic orthologs. *Nucleic Acids Res*, 33(Database issue):476–480, Jan 2005.
- [955] R Kucharczyk and J Rytka. *Saccharomyces cerevisiae*—a model organism for the studies on vacuolar transport. *Acta Biochim Pol*, 48(4):1025–1042, 2001.
- [956] F Perocchi, E Mancera, and L M Steinmetz. Systematic screens for human disease genes, from yeast to human and back. *Mol Biosyst*, 4(1):18–29, Jan 2008.
- [957] B K Kennedy and L Guarente. Genetic analysis of aging in *Saccharomyces cerevisiae*. *Trends Genet*, 12(9):355–359, Sep 1996.
- [958] J C Game. New genome-wide methods bring more power to yeast as a model organism. *Trends Pharmacol Sci*, 23(10):445–447, Oct 2002.
- [959] B Suter, D Auerbach, and I Stagljar. Yeast-based functional genomics and proteomics technologies: the first 15 years and beyond. *Biotechniques*, 40(5):625–644, May 2006.
- [960] N Burns, B Grimwade, P B Ross-Macdonald, E Y Choi, K Finberg, G S Roeder, and M Snyder. Large-scale analysis of gene expression, protein localization, and gene disruption in *Saccharomyces cerevisiae*. *Genes Dev*, 8(9):1087–1105, May 1994.
- [961] T Kirkwood. Ageing: too fast by mistake. *Nature*, 444(7122):1015–1017, Dec 2006.
- [962] M MacLean, N Harris, and P W Piper. Chronological lifespan of stationary phase yeast cells; a model for investigating the factors that might influence the ageing of postmitotic tissues in higher organisms. *Yeast*, 18(6):499–509, Apr 2001.
- [963] K U Fröhlich, H Fussi, and C Ruckenstein. Yeast apoptosis—from genes to pathways. *Semin Cancer Biol*, 17(2):112–121, Apr 2007.
- [964] L Guarente and C Kenyon. Genetic pathways that regulate ageing in model organisms. *Nature*, 408(6809):255–262, Nov 2000.
- [965] P W Piper. Long-lived yeast as a model for ageing research. *Yeast*, 23(3):215–226, Feb 2006.
- [966] M D Gray, J C Shen, A S Kamath-Loeb, A Blank, B L Sopher, G M Martin, J Oshima, and L A Loeb. The Werner syndrome protein is a DNA helicase. *Nat Genet*, 17(1):100–103, Sep 1997.
- [967] D A Sinclair, K Mills, and L Guarente. Accelerated aging and nucleolar fragmentation in yeast *sgs1* mutants. *Science*, 277(5330):1313–1316, Aug 1997.
- [968] P M Watt, I D Hickson, R H Borts, and E J Louis. *SGS1*, a homologue of the Bloom's and Werner's syndrome genes, is required for maintenance of genome stability in *Saccharomyces cerevisiae*. *Genetics*, 144(3):935–945, Nov 1996.
- [969] K Yamagata, J Kato, A Shimamoto, M Goto, Y Furuichi, and H Ikeda. Bloom's and Werner's syndrome genes suppress hyperrecombination in yeast *sgs1* mutant: implication for genomic instability in human diseases. *Proc Natl Acad Sci U S A*, 95(15):8733–8738, Jul 1998.
- [970] L Ye, J Nakura, A Morishima, and T Miki. Transcriptional activation by the Werner syndrome gene product in yeast. *Exp Gerontol*, 33(7-8):805–812, Nov-Dec 1998.
- [971] J Viña, M C Gomez-Cabrera, C Borrás, T Froio, F Sanchis-Gomar, V E Martínez-Bello, and F V Pallardo. Mitochondrial biogenesis in exercise and in ageing. *Adv Drug Deliv Rev*, 61(14):1369–1374, Nov 2009.
- [972] P D Sozou and T B Kirkwood. A stochastic model of cell replicative senescence based on telomere shortening, oxidative stress, and somatic mutations in nuclear and mitochondrial DNA. *J Theor Biol*, 213(4):573–586, Dec 2001.
- [973] F L Muller, M S Lustgarten, Y Jang, A Richardson, and H Van Remmen. Trends in oxidative aging theories. *Free Radic Biol Med*, 43(4):477–503, Aug 2007.

- [974] J F Passos, G Nelson, C Wang, T Richter, C Simillion, C J Proctor, S Miwa, S Olijslagers, J Hallinan, A Wipat, G Saretzki, K L Rudolph, T B Kirkwood, and T von Zglinicki. Feedback between p21 and reactive oxygen production is necessary for cell senescence. *Mol Syst Biol*, 6:347, 2010.
- [975] R K Moyzis, J M Buckingham, L S Cram, M Dani, L L Deaven, M D Jones, J Meyne, R L Ratliff, and J R Wu. A highly conserved repetitive DNA sequence, (TTAGGG)_n, present at the telomeres of human chromosomes. *Proc Natl Acad Sci U S A*, 85(18):6622–6626, Sep 1988.
- [976] E H Blackburn. Structure and function of telomeres. *Nature*, 350(6319):569–573, Apr 1991.
- [977] E H Blackburn. Switching and signaling at the telomere. *Cell*, 106(6):661–673, Sep 2001.
- [978] S H Askree, T Yehuda, S Smolikov, R Gurevich, J Hawk, C Coker, A Krauskopf, M Kupiec, and M J McEachern. A genome-wide screen for *Saccharomyces cerevisiae* deletion mutants that affect telomere length. *Proc Natl Acad Sci U S A*, 101(23):8658–8663, Jun 2004.
- [979] C M Counter, W C Hahn, W Wei, S D Caddle, R L Beijersbergen, P M Lansdorp, J M Sedivy, and R A Weinberg. Dissociation among *in vitro* telomerase activity, telomere maintenance, and cellular immortalization. *Proc Natl Acad Sci U S A*, 95(25):14723–14728, Dec 1998.
- [980] C W Greider and E H Blackburn. Identification of a specific telomere terminal transferase activity in Tetrahymena extracts. *Cell*, 43(2 Pt 1):405–413, Dec 1985.
- [981] D Shore and A Bianchi. Telomere length regulation: coupling DNA end processing to feedback regulation of telomerase. *EMBO J*, 28(16):2309–2322, Aug 2009.
- [982] A M Olovnikov. Telomeres, telomerase, and aging: origin of the theory. *Exp Gerontol*, 31(4):443–448, Jul-Aug 1996.
- [983] M Z Levy, R C Allsopp, A B Futcher, C W Greider, and C B Harley. Telomere end-replication problem and cell aging. *J Mol Biol*, 225(4):951–960, Jun 1992.
- [984] R D Portugal, M G Land, and B F Svaiter. A computational model for telomere-dependent cell-replicative aging. *Biosystems*, 91(1):262–267, Jan 2008.
- [985] M J McEachern, A Krauskopf, and E H Blackburn. Telomeres and their control. *Annu Rev Genet*, 34:331–358, 2000.
- [986] E H Blackburn. Telomere states and cell fates. *Nature*, 408(6808):53–56, Nov 2000.
- [987] J Berman, C Y Tachibana, and B K Tye. Identification of a telomere-binding activity from yeast. *Proc Natl Acad Sci U S A*, 83(11):3713–3717, Jun 1986.
- [988] A Bertuch and V Lundblad. Telomeres and double-strand breaks: trying to make ends meet. *Trends Cell Biol*, 8(9):339–342, Sep 1998.
- [989] A K Adams and C Holm. Specific DNA replication mutations affect telomere length in *Saccharomyces cerevisiae*. *Mol Cell Biol*, 16(9):4614–4620, Sep 1996.
- [990] R J Craven, P W Greenwell, M Dominska, and T D Petes. Regulation of genome stability by TEL1 and MEC1, yeast homologs of the mammalian ATM and ATR genes. *Genetics*, 161(2):493–507, Jun 2002.
- [991] D Lydall. Taming the tiger by the tail: modulation of DNA damage responses by telomeres. *EMBO J*, 28(15):2174–2187, Aug 2009.
- [992] T Gathbonton, M Imbesi, M Nelson, J M Akey, D M Ruderfer, L Kruglyak, J A Simon, and A Bedalov. Telomere length as a quantitative trait: genome-wide survey and genetic mapping of telomere length-control genes in yeast. *PLoS Genet*, 2(3):e35, Mar 2006.
- [993] A Greenall, G Lei, D C Swan, K James, L Wang, H Peters, A Wipat, D J Wilkinson, and D Lydall. A genome wide analysis of the response to uncapped telomeres in budding yeast reveals a novel role for the NAD⁺ biosynthetic gene BNA2 in chromosome end protection. *Genome Biol*, 9(10):R146, 2008.

- [994] A G Bodnar, M Ouellette, M Frolkis, S E Holt, C P Chiu, G B Morin, C B Harley, J W Shay, S Lichtsteiner, and W E Wright. Extension of life-span by introduction of telomerase into normal human cells. *Science*, 279(5349):349–352, Jan 1998.
- [995] S E Artandi, S Chang, S L Lee, S Alson, G J Gottlieb, L Chin, and R A DePinho. Telomere dysfunction promotes non-reciprocal translocations and epithelial cancers in mice. *Nature*, 406(6796):641–645, Aug 2000.
- [996] C M Counter, A A Avilion, C E LeFeuvre, N G Stewart, C W Greider, C B Harley, and S Bacchetti. Telomere shortening associated with chromosome instability is arrested in immortal cells which express telomerase activity. *EMBO J*, 11(5):1921–1929, May 1992.
- [997] L Hayflick. The limited *in vitro* lifetime of human diploid cell strains. *Exp Cell Res*, 37:614–636, Mar 1965.
- [998] K K Steffen, B K Kennedy, and M Kaeberlein. Measuring replicative life span in the budding yeast. *J Vis Exp*, 28:1209, Jun 2009.
- [999] M Collado, M A Blasco, and M Serrano. Cellular senescence in cancer and aging. *Cell*, 130(2):223–233, Jul 2007.
- [1000] J W Shay and W E Wright. Telomeres and telomerase: implications for cancer and aging. *Radiat Res*, 155(1 Pt 2):188–193, Jan 2001.
- [1001] L Hayflick and P S Moorhead. The serial cultivation of human diploid cell strains. *Exp Cell Res*, 25:585–621, Dec 1961.
- [1002] C B Harley, H Vaziri, C M Counter, and R C Allsopp. The telomere hypothesis of cellular aging. *Exp Gerontol*, 27(4):375–382, Jul-Aug 1992.
- [1003] I A Rodriguez-Brenes and C S Peskin. Quantitative theory of telomere length regulation and cellular senescence. *Proc Natl Acad Sci U S A*, 107(12):5387–5392, Mar 2010.
- [1004] T de Lange, L Shiue, R M Myers, D R Cox, S L Naylor, A M Killery, and H E Varmus. Structure and variability of human chromosome ends. *Mol Cell Biol*, 10(2):518–527, Feb 1990.
- [1005] N W Kim, M A Piatyszek, K R Prowse, C B Harley, M D West, P L Ho, G M Coviello, W E Wright, S L Weinrich, and J W Shay. Specific association of human telomerase activity with immortal cells and cancer. *Science*, 266(5193):2011–2015, Dec 1994.
- [1006] M A Blasco. Telomeres and human disease: ageing, cancer and beyond. *Nat Rev Genet*, 6(8):611–622, Aug 2005.
- [1007] M A Blasco. Telomere length, stem cells and aging. *Nat Chem Biol*, 3(10):640–649, Oct 2007.
- [1008] W Klapper, R Parwaresch, and G Krupp. Telomere biology in human aging and aging syndromes. *Mech Ageing Dev*, 122(7):695–712, May 2001.
- [1009] H Vaziri, F Schächter, I Uchida, L Wei, X Zhu, R Effros, D Cohen, and C B Harley. Loss of telomeric DNA during aging of normal and trisomy 21 human lymphocytes. *Am J Hum Genet*, 52(4):661–667, Apr 1993.
- [1010] T Finkel, M Serrano, and M A Blasco. The common biology of cancer and ageing. *Nature*, 448(7155):767–774, Aug 2007.
- [1011] A S Multani and S Chang. WRN at telomeres: implications for aging and cancer. *J Cell Sci*, 120(Pt 5):713–721, Mar 2007.
- [1012] L G Ball and W Xiao. Molecular basis of Ataxia Telangiectasia and related diseases. *Acta Pharmacol Sin*, 26(8):897–907, Aug 2005.
- [1013] D J Jamieson. Oxidative stress responses of the yeast *Saccharomyces cerevisiae*. *Yeast*, 14(16):1511–1527, Dec 1998.

- [1014] S O Yoon, C H Yun, and A S Chung. Dose effect of oxidative stress on signal transduction in aging. *Mech Ageing Dev*, 123(12):1597–1604, Nov 2002.
- [1015] I W Dawes. Molecular mechanism of the sensing of the hydrogen peroxide stress response in *Escherichia coli*. *Redox Rep*, 3(5-6):255–256, Oct-Dec 1997.
- [1016] J L Martindale and N J Holbrook. Cellular response to oxidative stress: signaling for suicide and survival. *J Cell Physiol*, 192(1):1–15, Jul 2002.
- [1017] G G Perrone, S X Tan, and I W Dawes. Reactive oxygen species and yeast apoptosis. *Biochim Biophys Acta*, 1783(7):1354–1368, Jul 2008.
- [1018] P Moradas-Ferreira, V Costa, P Piper, and W Mager. The molecular defences against reactive oxygen species in yeast. *Mol Microbiol*, 19(4):651–658, Feb 1996.
- [1019] M B Toledano, A Delaunay, B Biteau, D Spector, and D Azevedo. *Topics in Current Genetics*, chapter 6 - Oxidative stress responses in yeast, pages 241–303. Springer, 2003.
- [1020] C M Grant. Role of the glutathione/glutaredoxin and thioredoxin systems in yeast growth and response to stress conditions. *Mol Microbiol*, 39(3):533–541, Feb 2001.
- [1021] V I Lushchak. Budding yeast *Saccharomyces cerevisiae* as a model to study oxidative modification of proteins in eukaryotes. *Acta Biochim Pol*, 53(4):679–684, 2006.
- [1022] V D Longo and P Fabrizio. Regulation of longevity and stress resistance: a molecular strategy conserved from yeast to humans? *Cell Mol Life Sci*, 59(6):903–908, Jun 2002.
- [1023] T Drakulic, M D Temple, R Guido, S Jarolim, M Breitenbach, P V Attfield, and I W Dawes. Involvement of oxidative stress response genes in redox homeostasis, the level of reactive oxygen species, and ageing in *Saccharomyces cerevisiae*. *FEMS Yeast Res*, 5(12):1215–1228, Dec 2005.
- [1024] B Halliwell and J M C Gutteridge. *Free Radicals in Biology and Medicine*. Oxford University Press, Great Clarendon St. Oxford OX2 6DP, 4th edition, 2007.
- [1025] G Gille and K Sigler. Oxidative stress and living cells. *Folia Microbiologica*, 40(2):131–152, 1995.
- [1026] B M Babior. Oxygen-dependent microbial killing by phagocytes (first of two parts). *N Engl J Med*, 298(12):659–668, Mar 1978.
- [1027] B M Babior. Oxygen-dependent microbial killing by phagocytes (second of two parts). *N Engl J Med*, 298(13):721–725, Mar 1978.
- [1028] M Valko, C J Rhodes, J Moncol, M Izakovic, and M Mazur. Free radicals, metals and antioxidants in oxidative stress-induced cancer. *Chem Biol Interact*, 160(1):1–40, Mar 2006.
- [1029] S G Rhee. Cell signaling. H_2O_2 , a necessary evil for cell signaling. *Science*, 312(5782):1882–1883, Jun 2006.
- [1030] E A Veal, A M Day, and B A Morgan. Hydrogen peroxide sensing and signaling. *Mol Cell*, 26(1):1–14, Apr 2007.
- [1031] J Aguirre, M Ríos-Momberg, D Hewitt, and W Hansberg. Reactive oxygen species and development in microbial eukaryotes. *Trends Microbiol*, 13(3):111–118, Mar 2005.
- [1032] D C Chan. Mitochondria: dynamic organelles in disease, aging, and development. *Cell*, 125(7):1241–1252, Jun 2006.
- [1033] S Papa. Mitochondrial oxidative phosphorylation changes in the life span. Molecular aspects and physiopathological implications. *Biochim Biophys Acta*, 1276(2):87–105, Sep 1996.
- [1034] A L Jackson and L A Loeb. The contribution of endogenous sources of DNA damage to the multiple mutations in cancer. *Mutat Res*, 477(1-2):7–21, Jun 2001.

- [1035] C Dupuy, A Virion, R Ohayon, J Kaniewski, D Dème, and J Pommier. Mechanism of hydrogen peroxide formation catalyzed by NADPH oxidase in thyroid plasma membrane. *J Biol Chem*, 266(6):3739–3743, Feb 1991.
- [1036] P Niethammer, C Grabher, A T Look, and T J Mitchison. A tissue-scale gradient of hydrogen peroxide mediates rapid wound detection in zebrafish. *Nature*, 459(7249):996–999, Jun 2009.
- [1037] M Valko, H Morris, and M T Cronin. Metals, toxicity and oxidative stress. *Curr Med Chem*, 12(10):1161–1208, 2005.
- [1038] M J Koivula and T Eeva. Metal-related oxidative stress in birds. *Environ Pollut*, 158(7):2359–2370, Jul 2010.
- [1039] B Halliwell and J M Gutteridge. Biologically relevant metal ion-dependent hydroxyl radical generation. An update. *FEBS Lett*, 307(1):108–112, Jul 1992.
- [1040] E Herrero, J Ros, G Bellí, and E Cabiscol. Redox control and oxidative stress in yeast cells. *Biochim Biophys Acta*, 1780(11):1217–1235, Nov 2008.
- [1041] C E Outten, R L Falk, and V C Culotta. Cellular factors required for protection from hyperoxia toxicity in *Saccharomyces cerevisiae*. *Biochem J*, 388(Pt 1):93–101, May 2005.
- [1042] J M McCord and I Fridovich. Superoxide dismutase. an enzymic function for erythrocuprein (hemocuprein). *J Biol Chem*, 244(22):6049–6055, Nov 1969.
- [1043] P Chelikani, I Fita, and P C Loewen. Diversity of structures and properties among catalases. *Cell Mol Life Sci*, 61(2):192–208, Jan 2004.
- [1044] G Cohen, W Rapatz, and H Ruis. Sequence of the *Saccharomyces cerevisiae* CTA1 gene and amino acid sequence of catalase A derived from it. *Eur J Biochem*, 176(1):159–163, Sep 1988.
- [1045] C M Grant, G Perrone, and I W Dawes. Glutathione and catalase provide overlapping defenses for protection against hydrogen peroxide in the yeast *Saccharomyces cerevisiae*. *Biochem Biophys Res Commun*, 253(3):893–898, Dec 1998.
- [1046] O Carmel-Harel and G Storz. Roles of the glutathione- and thioredoxin-dependent reduction systems in the *Escherichia coli* and *Saccharomyces cerevisiae* responses to oxidative stress. *Annu Rev Microbiol*, 54:439–461, 2000.
- [1047] C M Grant, F H MacIver, and I W Dawes. Glutathione is an essential metabolite required for resistance to oxidative stress in the yeast *Saccharomyces cerevisiae*. *Curr Genet*, 29(6):511–515, May 1996.
- [1048] U H Dormer, J Westwater, D W Stephen, and D J Jamieson. Oxidant regulation of the *Saccharomyces cerevisiae* *GSH1* gene. *Biochim Biophys Acta*, 1576(1-2):23–29, Jun 2002.
- [1049] F Q Schafer and G R Buettner. Redox environment of the cell as viewed through the redox state of the glutathione disulfide/glutathione couple. *Free Radic Biol Med*, 30(11):1191–1212, Jun 2001.
- [1050] D P Jones. Redox potential of GSH/GSSG couple: assay and biological significance. *Methods Enzymol*, 348:93–112, 2002.
- [1051] M Kistler, K H Summer, and F Eckardt. Isolation of glutathione-deficient mutants of the yeast *Saccharomyces cerevisiae*. *Mutat Res*, 173(2):117–120, Feb 1986.
- [1052] D W Stephen and D J Jamieson. Glutathione is an important antioxidant molecule in the yeast *Saccharomyces cerevisiae*. *FEMS Microbiol Lett*, 141(2-3):207–212, Aug 1996.
- [1053] Q Ran, H Liang, Y Ikeno, W Qi, T A Prolla, L J Roberts, N Wolf, H Van Remmen, H VanRemmen, and A Richardson. Reduction in glutathione peroxidase 4 increases life span through increased sensitivity to apoptosis. *J Gerontol A Biol Sci Med Sci*, 62(9):932–942, Sep 2007.
- [1054] H Z Chae, S J Chung, and S G Rhee. Thioredoxin-dependent peroxide reductase from yeast. *J Biol Chem*, 269(44):27670–27678, Nov 1994.

- [1055] P Y Lee, K H Bae, C W Kho, S Kang, d o H Lee, S Cho, S Kang, S C Lee, B C Park, and S G Park. Interactome analysis of yeast glutathione peroxidase 3. *J Microbiol Biotechnol*, 18(8):1364–1367, Aug 2008.
- [1056] Y Inoue, T Matsuda, K Sugiyama, S Izawa, and A Kimura. Genetic analysis of glutathione peroxidase in oxidative stress response of *Saccharomyces cerevisiae*. *J Biol Chem*, 274(38):27002–27009, Sep 1999.
- [1057] M T Rodríguez-Manzanque, J Ros, E Cabisco, A Sorribas, and E Herrero. Grx5 glutaredoxin plays a central role in protection against protein oxidative damage in *Saccharomyces cerevisiae*. *Mol Cell Biol*, 19(12):8180–8190, Dec 1999.
- [1058] M M Molina, G Bellí, M A de la Torre, M T Rodríguez-Manzanque, and E Herrero. Nuclear monothiol glutaredoxins of *Saccharomyces cerevisiae* can function as mitochondrial glutaredoxins. *J Biol Chem*, 279(50):51923–51930, Dec 2004.
- [1059] J R Pedrajas, E Kosmidou, A Miranda-Vizuet, J A Gustafsson, A P Wright, and G Spyrou. Identification and functional characterization of a novel mitochondrial thioredoxin system in *Saccharomyces cerevisiae*. *J Biol Chem*, 274(10):6366–6373, Mar 1999.
- [1060] N Pujol-Carrion, G Belli, E Herrero, A Nogues, and M A de la Torre-Ruiz. Glutaredoxins Grx3 and Grx4 regulate nuclear localisation of Aft1 and the oxidative stress response in *Saccharomyces cerevisiae*. *J Cell Sci*, 119(Pt 21):4554–4564, Nov 2006.
- [1061] N Rouhier, J Couturier, M K Johnson, and J P Jacquot. Glutaredoxins: roles in iron homeostasis. *Trends Biochem Sci*, 35(1):43–52, Jan 2010.
- [1062] L Ojeda, G Keller, U Muhlenhoff, J C Rutherford, R Lill, and D R Winge. Role of glutaredoxin-3 and glutaredoxin-4 in the iron regulation of the Aft1 transcriptional activator in *Saccharomyces cerevisiae*. *J Biol Chem*, 281(26):17661–17669, Jun 2006.
- [1063] C C Philpott and O Protchenko. Response to iron deprivation in *Saccharomyces cerevisiae*. *Eukaryot Cell*, 7(1):20–27, Jan 2008.
- [1064] M T Rodríguez-Manzanque, J Tamarit, G Bellí, J Ros, and E Herrero. Grx5 is a mitochondrial glutaredoxin required for the activity of iron/sulfur enzymes. *Mol Biol Cell*, 13(4):1109–1121, Apr 2002.
- [1065] A Kumánovics, O S Chen, L Li, D Bagley, E M Adkins, H Lin, N N Dingra, C E Outten, G Keller, D Winge, D M Ward, and J Kaplan. Identification of FRA1 and FRA2 as genes involved in regulating the yeast iron regulon in response to decreased mitochondrial iron-sulfur cluster synthesis. *J Biol Chem*, 283(16):10276–10286, Apr 2008.
- [1066] T Draculic, I W Dawes, and C M Grant. A single glutaredoxin or thioredoxin gene is essential for viability in the yeast *Saccharomyces cerevisiae*. *Mol Microbiol*, 36(5):1167–1174, Jun 2000.
- [1067] G Monteiro, B B Horta, D C Pimenta, O Augusto, and L E Netto. Reduction of 1-Cys peroxiredoxins by ascorbate changes the thiol-specific antioxidant paradigm, revealing another function of vitamin C. *Proc Natl Acad Sci U S A*, 104(12):4886–4891, Mar 2007.
- [1068] R D Hancock, J R Galpin, and R Viola. Biosynthesis of L-ascorbic acid (vitamin C) by *Saccharomyces cerevisiae*. *FEMS Microbiol Lett*, 186(2):245–250, May 2000.
- [1069] P Raspor, S Plesnicar, Z Gazdag, M Pesti, M Miklavcic, B Lah, R Logar-Marinsek, and B Poljsak. Prevention of intracellular oxidation in yeast: the role of vitamin E analogue, Trolox (6-hydroxy-2,5,7,8-tetramethylkroman-2-carboxyl acid). *Cell Biol Int*, 29(1):57–63, Jan 2005.
- [1070] H Sies. Oxidative stress: oxidants and antioxidants. *Exp Physiol*, 82(2):291–295, Mar 1997.
- [1071] K B Beckman and B N Ames. Oxidative decay of DNA. *J Biol Chem*, 272(32):19633–19636, Aug 1997.
- [1072] D Wang, D A Kreutzer, and J M Essigmann. Mutagenicity and repair of oxidative DNA damage: insights from studies using defined lesions. *Mutat Res*, 400(1-2):99–115, May 1998.

- [1073] K B Beckman and B N Ames. Endogenous oxidative damage of mtDNA. *Mutat Res*, 424(1-2):51–58, Mar 1999.
- [1074] C Richter, J W Park, and B N Ames. Normal oxidative damage to mitochondrial and nuclear DNA is extensive. *Proc Natl Acad Sci U S A*, 85(17):6465–6467, Sep 1988.
- [1075] M H Goyns. Genes, telomeres and mammalian ageing. *Mech Ageing Dev*, 123(7):791–799, Apr 2002.
- [1076] C J Proctor and T B Kirkwood. Modelling telomere shortening and the role of oxidative stress. *Mech Ageing Dev*, 123(4):351–363, Feb 2002.
- [1077] B S Berlett and E R Stadtman. Protein oxidation in aging, disease, and oxidative stress. *J Biol Chem*, 272(33):20313–20316, Aug 1997.
- [1078] T Nyström. Role of oxidative carbonylation in protein quality control and senescence. *EMBO J*, 24(7):1311–1317, Apr 2005.
- [1079] C Jacob, I Knight, and P G Winyard. Aspects of the biological redox chemistry of cysteine: from simple redox responses to sophisticated signalling pathways. *Biol Chem*, 387(10-11):1385–1397, Oct-Nov 2006.
- [1080] I Dalle-Donne, D Giustarini, R Colombo, R Rossi, and A Milzani. Protein carbonylation in human diseases. *Trends Mol Med*, 9(4):169–176, Apr 2003.
- [1081] H Mirzaei and F Regnier. Protein:protein aggregation induced by protein oxidation. *J Chromatogr B Analyt Technol Biomed Life Sci*, 873(1):8–14, Sep 2008.
- [1082] H W Gardner. Oxygen radical chemistry of polyunsaturated fatty acids. *Free Radic Biol Med*, 7(1):65–86, 1989.
- [1083] B Halliwell and C E Cross. Oxygen-derived species: their relation to human disease and environmental stress. *Environ Health Perspect*, 102 Suppl 10:5–12, Dec 1994.
- [1084] S Mizukami-Murata, H Iwahashi, S Kimura, K Nojima, Y Sakurai, T Saitou, N Fujii, Y Murata, S Suga, K Kitagawa, K Tanaka, S Endo, and M Hoshi. Genome-wide expression changes in *Saccharomyces cerevisiae* in response to high-LET ionizing radiation. *Appl Biochem Biotechnol*, 162(3):855–870, Oct 2010.
- [1085] R Gniadecki, T Thorn, J Vicanova, A Petersen, and H C Wulf. Role of mitochondria in ultraviolet-induced oxidative stress. *J Cell Biochem*, 80(2):216–222, Oct 2000.
- [1086] J Flattery-O’Brien, L P Collinson, and I W Dawes. *Saccharomyces cerevisiae* has an inducible response to menadione which differs from that to hydrogen peroxide. *J Gen Microbiol*, 139(3):501–507, Mar 1993.
- [1087] N S Kosower and E M Kosower. Diamide: an oxidant probe for thiols. *Methods Enzymol*, 251:123–133, 1995.
- [1088] J S Bus and J E Gibson. Paraquat: model for oxidant-initiated toxicity. *Environ Health Perspect*, 55:37–46, Apr 1984.
- [1089] J Lee, D Spector, C Godon, J Labarre, and M B Toledano. A new antioxidant with alkyl hydroperoxide defense properties in yeast. *J Biol Chem*, 274(8):4537–4544, Feb 1999.
- [1090] R J Brennan and R H Schiestl. Cadmium is an inducer of oxidative stress in yeast. *Mutat Res*, 356(2):171–178, Sep 1996.
- [1091] M Thorsen, G Lagniel, E Kristiansson, C Junot, O Nerman, J Labarre, and M J Tamás. Quantitative transcriptome, proteome, and sulfur metabolite profiling of the *Saccharomyces cerevisiae* response to arsenite. *Physiol Genomics*, 30(1):35–43, Jun 2007.
- [1092] B Almeida, A Silva, A Mesquita, B Sampaio-Marques, F Rodrigues, and P Ludovico. Drug-induced apoptosis in yeast. *Biochim Biophys Acta*, 1783(7):1436–1448, Jul 2008.

- [1093] R Chowdhury, R Chatterjee, A K Giri, C Mandal, and K Chaudhuri. Arsenic-induced cell proliferation is associated with enhanced ROS generation, Erk signaling and CyclinA expression. *Toxicol Lett*, 198(2):263–271, Oct 2010.
- [1094] G Gobe and D Crane. Mitochondria, reactive oxygen species and cadmium toxicity in the kidney. *Toxicol Lett*, 198(1):49–55, Sep 2010.
- [1095] D J Jamieson. *Saccharomyces cerevisiae* has distinct adaptive responses to both hydrogen peroxide and menadione. *J Bacteriol*, 174(20):6678–6681, Oct 1992.
- [1096] G W Thorpe, C S Fong, N Alic, V J Higgins, and I W Dawes. Cells have distinct mechanisms to maintain protection against different reactive oxygen species: oxidative-stress-response genes. *Proc Natl Acad Sci U S A*, 101(17):6564–6569, Apr 2004.
- [1097] D J Jamieson, S L Rivers, and D W Stephen. Analysis of *Saccharomyces cerevisiae* proteins induced by peroxide and superoxide stress. *Microbiology*, 140 (Pt 12):3277–3283, Dec 1994.
- [1098] C Godon, G Lagniel, J Lee, J M Buhler, S Kieffer, M Perrot, H Boucherie, M B Toledano, and J Labarre. The H_2O_2 stimulon in *Saccharomyces cerevisiae*. *J Biol Chem*, 273(35):22480–22489, Aug 1998.
- [1099] I Kim, H Yun, and I Jin. Comparative proteomic analyses of the yeast *Saccharomyces cerevisiae* KNU5377 strain against menadione-induced oxidative stress. *J Microbiol Biotechnol*, 17(2):207–217, Feb 2007.
- [1100] A Ikner and K Shiozaki. Yeast signaling pathways in the oxidative stress response. *Mutat Res*, 569(1-2):13–27, Jan 2005.
- [1101] D W Stephen, S L Rivers, and D J Jamieson. The role of the YAP1 and YAP2 genes in the regulation of the adaptive oxidative stress responses of *Saccharomyces cerevisiae*. *Mol Microbiol*, 16(3):415–423, May 1995.
- [1102] M R Fernando, H Nanri, S Yoshitake, K Nagata-Kuno, and S Minakami. Thioredoxin regenerates proteins inactivated by oxidative stress in endothelial cells. *Eur J Biochem*, 209(3):917–922, Nov 1992.
- [1103] K S Doris. *The regulation of the cell division cycle in response to oxidative stress in Saccharomyces cerevisiae*. PhD thesis, Newcastle University, 2008.
- [1104] L A Rowe, N Degtyareva, and P W Doetsch. DNA damage-induced reactive oxygen species (ROS) stress response in *Saccharomyces cerevisiae*. *Free Radic Biol Med*, 45(8):1167–1177, Oct 2008.
- [1105] T Bender, C Leidhold, T Ruppert, S Franken, and W Voos. The role of protein quality control in mitochondrial protein homeostasis under oxidative stress. *Proteomics*, 10(7):1426–1443, Apr 2010.
- [1106] X Ouyang, Q T Tran, S Goodwin, R S Wible, C H Sutter, and T R Sutter. Yap1 activation by H_2O_2 or thiol-reactive chemicals elicits distinct adaptive gene responses. *Free Radic Biol Med*, 50(1):1–13, Nov 2010.
- [1107] O V Lushchak and V I Lushchak. Possible pathways involved in activation of catalase and superoxide dismutase with sodium nitroprusside in yeast *Saccharomyces cerevisiae*. *Ukr Biokhim Zh*, 81(2):34–39, Mar-Apr 2009.
- [1108] J Lee, C Godon, G Lagniel, D Spector, J Garin, J Labarre, and M B Toledano. Yap1 and Skn7 control two specialized oxidative stress response regulons in yeast. *J Biol Chem*, 274(23):16040–16046, Jun 1999.
- [1109] X J He, K E Mulford, and J S Fassler. Oxidative stress function of the *Saccharomyces cerevisiae* Skn7 receiver domain. *Eukaryot Cell*, 8(5):768–778, May 2009.
- [1110] S Boissnard, G Lagniel, C Garmendia-Torres, M Molin, E Boy-Marcotte, M Jacquet, M B Toledano, J Labarre, and S Chédin. H_2O_2 activates the nuclear localization of Msn2 and Maf1 through thioredoxins in *Saccharomyces cerevisiae*. *Eukaryot Cell*, 8(9):1429–1438, Sep 2009.

- [1111] A P Gasch, P T Spellman, C M Kao, O Carmel-Harel, M B Eisen, G Storz, D Botstein, and P O Brown. Genomic expression programs in the response of yeast cells to environmental changes. *Mol Biol Cell*, 11(12):4241–4257, Dec 2000.
- [1112] R Kelley and T Ideker. Genome-wide fitness and expression profiling implicate Mga2 in adaptation to hydrogen peroxide. *PLoS Genet*, 5(5):e1000488, May 2009.
- [1113] C V Lowry and R S Zitomer. Oxygen regulation of anaerobic and aerobic genes mediated by a common factor in yeast. *Proc Natl Acad Sci U S A*, 81(19):6129–6133, Oct 1984.
- [1114] H Li, D T Mapolelo, N N Dingra, S G Naik, N S Lees, B M Hoffman, P J Riggs-Gelasco, B H Huynh, M K Johnson, and C E Outten. The yeast iron regulatory proteins Grx3/4 and Fra2 form heterodimeric complexes containing a [2Fe-2S] cluster with cysteinyl and histidyl ligation. *Biochemistry*, 48(40):9569–9581, Oct 2009.
- [1115] F Estruch. Stress-controlled transcription factors, stress-induced genes and stress tolerance in budding yeast. *FEMS Microbiol Rev*, 24(4):469–486, Oct 2000.
- [1116] G M Martin, S N Austad, and T E Johnson. Genetic analysis of ageing: role of oxidative damage and environmental stresses. *Nat Genet*, 13(1):25–34, May 1996.
- [1117] D Harman. Aging: a theory based on free radical and radiation chemistry. *J Gerontol*, 11(3):298–300, Jul 1956.
- [1118] D Harman. The biologic clock: the mitochondria? *J Am Geriatr Soc*, 20(4):145–147, Apr 1972.
- [1119] L Gil del Valle. Oxidative stress in aging: theoretical outcomes and clinical evidences in humans. *Biomed Pharmacother*, Sep 2010. DOI = "10.1016/j.biopha.2010.09.010",.
- [1120] T Finkel and N J Holbrook. Oxidants, oxidative stress and the biology of ageing. *Nature*, 408(6809):239–247, Nov 2000.
- [1121] M K Shigenaga, T M Hagen, and B N Ames. Oxidative damage and mitochondrial decay in aging. *Proc Natl Acad Sci U S A*, 91(23):10771–10778, Nov 1994.
- [1122] V Valls, C Peiro, P Muñiz, and G T Saez. Age-related changes in antioxidant status and oxidative damage to lipids and DNA in mitochondria of rat liver. *Process Biochemistry*, 40(2):903 – 908, 2005.
- [1123] P Laun, A Pichova, F Madeo, J Fuchs, A Ellinger, S Kohlwein, I Dawes, K U Fröhlich, and M Breitenbach. Aged mother cells of *Saccharomyces cerevisiae* show markers of oxidative stress and apoptosis. *Mol Microbiol*, 39(5):1166–1173, Mar 2001.
- [1124] G Reverter-Branchat, E Cabiscol, J Tamarit, and J Ros. Oxidative damage to specific proteins in replicative and chronological-aged *Saccharomyces cerevisiae*: common targets and prevention by calorie restriction. *J Biol Chem*, 279(30):31983–31989, Jul 2004.
- [1125] E R Stadtman. Protein oxidation and aging. *Free Radic Res*, 40(12):1250–1258, Dec 2006.
- [1126] T Toda, M Nakamura, H Morisawa, M Hirota, R Nishigaki, and Y Yoshimi. Proteomic approaches to oxidative protein modifications implicated in the mechanism of aging. *Geriatr Gerontol Int*, 10 Suppl 1:25–31, Jul 2010.
- [1127] V D Longo, L L Liou, J S Valentine, and E B Gralla. Mitochondrial superoxide decreases yeast survival in stationary phase. *Arch Biochem Biophys*, 365(1):131–142, May 1999.
- [1128] P Fabrizio, L Li, and V D Longo. Analysis of gene expression profile in yeast aging chronologically. *Mech Ageing Dev*, 126(1):11–16, Jan 2005.
- [1129] V D Longo, L M Ellerby, D E Bredesen, J S Valentine, and E B Gralla. Human Bcl-2 reverses survival defects in yeast lacking superoxide dismutase and delays death of wild-type yeast. *J Cell Biol*, 137(7):1581–1588, Jun 1997.
- [1130] M A Sorolla, G Reverter-Branchat, J Tamarit, I Ferrer, J Ros, and E Cabiscol. Proteomic and oxidative stress analysis in human brain samples of Huntington disease. *Free Radic Biol Med*, 45(5):667–678, Sep 2008.

- [1131] F V Pallardó, A Lloret, M Lebel, M d'Ischia, V C Cogger, D G Le Couteur, M N Gadaleta, G Castello, and G Pagano. Mitochondrial dysfunction in some oxidative stress-related genetic diseases: Ataxia-Telangiectasia, Down Syndrome, Fanconi Anaemia and Werner Syndrome. *Biogerontology*, 11(4):401–419, Aug 2010.
- [1132] M D Evans, M Dizdaroglu, and M S Cooke. Oxidative DNA damage and disease: induction, repair and significance. *Mutat Res*, 567(1):1–61, Sep 2004.
- [1133] J L Scott, C Gabrielides, R K Davidson, T E Swingle, I M Clark, G A Wallis, R P Boot-Handford, T B Kirkwood, R W Talyor, and D A Young. Superoxide dismutase downregulation in osteoarthritis progression and end-stage disease. *Ann Rheum Dis*, 69(8):1502–1510, Aug 2010.
- [1134] S C Bondy. The relation of oxidative stress and hyperexcitation to neurological disease. *Proc Soc Exp Biol Med*, 208(4):337–345, Apr 1995.
- [1135] D Dreher and A F Junod. Role of oxygen free radicals in cancer development. *Eur J Cancer*, 32A(1):30–38, Jan 1996.
- [1136] C R Burtner, C J Murakami, B K Kennedy, and M Kaeberlein. A molecular mechanism of chronological aging in yeast. *Cell Cycle*, 8(8):1256–1270, Apr 2009.
- [1137] A G Wiese, R E Pacifici, and K J Davies. Transient adaptation of oxidative stress in mammalian cells. *Arch Biochem Biophys*, 318(1):231–240, Apr 1995.
- [1138] M Ristow and K Zarse. How increased oxidative stress promotes longevity and metabolic health: the concept of mitochondrial hormesis (mitohormesis). *Exp Gerontol*, 45(6):410–418, Jun 2010.
- [1139] D P Shanley and T B Kirkwood. Caloric restriction does not enhance longevity in all species and is unlikely to do so in humans. *Biogerontology*, 7(3):165–168, Jun 2006.
- [1140] P Resnik. Semantic similarity in a taxonomy: an information-based measure and its application to problems of ambiguity in natural language. *Journal of Artificial Intelligence Research*, 11:95–130, 1999.
- [1141] S J Lycett. Interaction network integration using Bayesian data fusion methods. Master's thesis, School of Computing Science, Newcastle University, 2007.
- [1142] E W Dijkstra. A note on two problems in connexion with graphs. *Numerische Mathematik*, 1(1):269–271–271, December 1959.
- [1143] C. S. Hoffman. *Current Protocols in Molecular Biology: Preparation of Yeast DNA*, volume 13.11.1–13.11.4. John Wiley & Sons, Inc., 2001.
- [1144] S L Sanders, J Jennings, A Canutescu, A J Link, and P A Weil. Proteomics of the eukaryotic transcription machinery: identification of proteins associated with components of yeast TFIID by multidimensional mass spectrometry. *Mol Cell Biol*, 22(13):4723–4738, Jul 2002.
- [1145] R Zhao, M Davey, Y C Hsu, P Kaplanek, A Tong, A B Parsons, N Krogan, G Cagney, D Mai, J Greenblatt, C Boone, A Emili, and W A Houry. Navigating the chaperone network: an integrative map of physical and genetic interactions mediated by the Hsp90 chaperone. *Cell*, 120(5):715–727, Mar 2005.
- [1146] A E Frazier, R D Taylor, D U Mick, B Warscheid, N Stoepel, H E Meyer, M T Ryan, B Guiard, and P Rehling. Mdm38 interacts with ribosomes and is a component of the mitochondrial protein export machinery. *J Cell Biol*, 172(4):553–564, Feb 2006.
- [1147] V Measday, K Baetz, J Guzzo, K Yuen, T Kwok, B Sheikh, H Ding, R Ueta, T Hoac, B Cheng, I Pot, A Tong, Y Yamaguchi-Iwai, C Boone, P Hieter, and B Andrews. Systematic yeast synthetic lethal and synthetic dosage lethal screens identify genes required for chromosome segregation. *Proc Natl Acad Sci U S A*, 102(39):13956–13961, Sep 2005.
- [1148] N J Krogan, K Baetz, M C Keogh, N Datta, C Sawa, T C Kwok, N J Thompson, M G Davey, J Pootoolal, T R Hughes, A Emili, S Buratowski, P Hieter, and J F Greenblatt. Regulation of chromosome stability by the histone H2A variant Htz1, the Swr1 chromatin remodeling complex, and the histone acetyltransferase NuA4. *Proc Natl Acad Sci U S A*, 101(37):13513–13518, Sep 2004.

- [1149] K Ingvarsdottir, N J Krogan, N C Emre, A Wyce, N J Thompson, A Emili, T R Hughes, J F Greenblatt, and S L Berger. H2B ubiquitin protease Ubp8 and Sgf11 constitute a discrete functional module within the *Saccharomyces cerevisiae* SAGA complex. *Mol Cell Biol*, 25(3):1162–1172, Feb 2005.
- [1150] A H Tong, B Drees, G Nardelli, G D Bader, B Brannetti, L Castagnoli, M Evangelista, S Ferracuti, B Nelson, S Paoluzi, M Quondam, A Zucconi, C W Hogue, S Fields, C Boone, and G Cesareni. A combined experimental and computational strategy to define protein interaction networks for peptide recognition modules. *Science*, 295(5553):321–324, Jan 2002.
- [1151] N J Krogan, M C Keogh, N Datta, C Sawa, O W Ryan, H Ding, R A Haw, J Pootoolal, A Tong, V Canadien, D P Richards, X Wu, A Emili, T R Hughes, S Buratowski, and J F Greenblatt. A Snf2 family ATPase complex required for recruitment of the histone H2A variant Htz1. *Mol Cell*, 12(6):1565–1576, Dec 2003.
- [1152] D L Lindstrom, S L Squazzo, N Muster, T A Burckin, K C Wachter, C A Emigh, J A McCleery, J R Yates, and G A Hartzog. Dual roles for Spt5 in pre-mRNA processing and transcription elongation revealed by identification of Spt5-associated proteins. *Mol Cell Biol*, 23(4):1368–1378, Feb 2003.
- [1153] S E Kong, M S Kobor, N J Krogan, B P Somesh, T M Sogaard, J F Greenblatt, and J Q Svejstrup. Interaction of Fcp1 phosphatase with elongating RNA polymerase II holoenzyme, enzymatic mechanism of action, and genetic interaction with elongator. *J Biol Chem*, 280(6):4299–4306, Feb 2005.
- [1154] J T Hannich, A Lewis, M B Kroetz, S J Li, H Heide, A Emili, and M Hochstrasser. Defining the SUMO-modified proteome by multiple approaches in *Saccharomyces cerevisiae*. *J Biol Chem*, 280(6):4102–4110, Feb 2005.
- [1155] S H Millson, A W Truman, V King, C Prodromou, L H Pearl, and P W Piper. A two-hybrid screen of the yeast proteome for Hsp90 interactors uncovers a novel Hsp90 chaperone requirement in the activity of a stress-activated mitogen-activated protein kinase, Slt2p (Mpk1p). *Eukaryot Cell*, 4(5):849–860, May 2005.
- [1156] N P Allen, L Huang, A Burlingame, and M Rexach. Proteomic analysis of nucleoporin interacting proteins. *J Biol Chem*, 276(31):29268–29274, Aug 2001.
- [1157] 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. *Proc Natl Acad Sci U S A*, 102(34):12123–12128, Aug 2005.
- [1158] R Srivastava and J Varner. Emerging technologies: systems biology. *Biotechnol Prog*, 23(1):24–27, Jan-Feb 2007.
- [1159] BioGRID Administration Team. biogridadmin@gmail.com *Personnal Communication*. 28th September 2009.
- [1160] GenomeNet Support Team. <http://www.genome.jp/feedback/>. *Personnal Communication*. 19th May 2009.
- [1161] K James, A Wipat, and J Hallinan. Is newer better? an evaluation of the effects of data curation on integrated analyses in *Saccharomyces cerevisiae*. *Integr. Biol.*, 2011 (in press).
- [1162] A M Dudley, D M Janse, A Tanay, R Shamir, and G M Church. A global view of pleiotropy and phenotypically derived gene function in yeast. *Mol Syst Biol*, 1:2005.0001, 2005.
- [1163] A Ando, T Nakamura, Y Murata, H Takagi, and J Shima. Identification and classification of genes required for tolerance to freeze-thaw stress revealed by genome-wide screening of *Saccharomyces cerevisiae* deletion strains. *FEMS Yeast Res*, 7(2):244–253, Mar 2007.
- [1164] P Kolodziej and R A Young. RNA polymerase II subunit RPB3 is an essential component of the mRNA transcription apparatus. *Mol Cell Biol*, 9(12):5387–5394, Dec 1989.
- [1165] GOA curators (2000). Gene Ontology annotation based on Swiss-Prot keyword mapping. GO_REF:0000004 (<http://www.yeastgenome.org/cgi-bin/reference/reference.pl?dbid=S000124038>).

- [1166] D C Hess, C L Myers, C Huttenhower, M A Hibbs, A P Hayes, J Paw, J J Clore, R M Mendoza, B S Luis, C Nislow, G Giaever, M Costanzo, O G Troyanskaya, and A A Caudy. Computationally driven, quantitative experiments discover genes required for mitochondrial biogenesis. *PLoS Genet*, 5(3):e1000407, Mar 2009.
- [1167] J M Santos, P Freire, M Vicente, and C M Arraiano. The stationary-phase morphogene *bolA* from *Escherichia coli* is induced by stress during early stages of growth. *Mol Microbiol*, 32(4):789–798, May 1999.
- [1168] J M Santos, M Lobo, A P Matos, M A De Pedro, and C M Arraiano. The gene *bolA* regulates *dacA* (PBP5), *dacC* (PBP6) and *ampC* (AmpC), promoting normal morphology in *Escherichia coli*. *Mol Microbiol*, 45(6):1729–1740, Sep 2002.
- [1169] M Aldea, T Garrido, C Hernández-Chico, M Vicente, and S R Kushner. Induction of a growth-phase-dependent promoter triggers transcription of *bolA*, an *Escherichia coli* morphogene. *EMBO J*, 8(12):3923–3931, Dec 1989.
- [1170] H Li, D T Mapolelo, N N Dingra, G Keller, P J Riggs-Gelasco, D R Winge, M K Johnson, and C E Outten. Histidine 103 in Fra2 is an iron-sulfur cluster ligand in the [2Fe-2S] Fra2-Grx3 complex and is required for *in vivo* iron signaling in yeast. *J Biol Chem*, 286(1):867–876, Jan 2011.
- [1171] M Costanzo, A Baryshnikova, J Bellay, Y Kim, E D Spear, C S Sevier, H Ding, J L Koh, K Toufighi, S Mostafavi, J Prinz, R P St Onge, B VanderSluis, T Makhnevych, F J Vizeacoumar, S Alizadeh, S Bahr, R L Brost, Y Chen, M Cokol, R Deshpande, Z Li, Z Y Lin, W Liang, M Marback, J Paw, B J San Luis, E Shuteriqi, A H Tong, N van Dyk, I M Wallace, J A Whitney, M T Weirauch, G Zhong, H Zhu, W A Houry, M Brudno, S Ragibizadeh, B Papp, C Pál, F P Roth, G Giaever, C Nislow, O G Troyanskaya, H Bussey, G D Bader, A C Gingras, Q D Morris, P M Kim, C A Kaiser, C L Myers, B J Andrews, and C Boone. The genetic landscape of a cell. *Science*, 327(5964):425–431, Jan 2010.
- [1172] A Gómez, J Cedano, I Amela, A Planas, J Piñol, and E Querol. Gene Ontology function prediction in mollicutes using protein-protein association networks. *BMC Syst Biol*, 5:49, 2011.
- [1173] J Gillis and P Pavlidis. The role of indirect connections in gene networks in predicting function. *Bioinformatics*, 27(13):1860–1866, Jul 2011.
- [1174] S Erdin, A M Lisewski, and O Lichtarge. Protein function prediction: towards integration of similarity metrics. *Curr Opin Struct Biol*, 21(2):180–188, Apr 2011.
- [1175] I Sendiña-Nadal, Y Ofran, J A Almendral, J M Buldú, I Leyva, D Li, S Havlin, and S Boccaletti. Unveiling protein functions through the dynamics of the interaction network. *PLoS One*, 6(3):e17679, 2011.
- [1176] S Pinkert, J Schultz, and J Reichardt. Protein interaction networks—more than mere modules. *PLoS Comput Biol*, 6(1):e1000659, Jan 2010.
- [1177] X Jiang, D Gold, and E D Kolaczyk. Network-based auto-probit modeling for protein function prediction. *Biometrics*, 67(3):958–966, Dec 2010.
- [1178] Z Chen, Z Cai, M Li, and B Liu. Using search engine technology for protein function prediction. *Int J Bioinform Res Appl*, 7(1):101–113, 2011.
- [1179] P G Sun, L Gao, and S Han. Prediction of human disease-related gene clusters by clustering analysis. *Int J Biol Sci*, 7(1):61–73, 2011.
- [1180] G K Mazandu and N J Mulder. Scoring protein relationships in functional interaction networks predicted from sequence data. *PLoS One*, 6(4):e18607, 2011.
- [1181] K S Ahmed, N H Saloma, and Y M Kadah. Improving the prediction of yeast protein function using weighted protein-protein interactions. *Theor Biol Med Model*, 8:11, 2011.
- [1182] T H Cormen, C E Leiserson, R L Rivest, and C Stein. *Introduction to algorithms*, chapter VI and Appendix B. MIT Press, 2003.