UNIVERSITY OF CALIFORNIA, SAN DIEGO

Understanding cellular function through the analysis of protein interaction networks

A Dissertation submitted in partial satisfaction of the
requirements for the degree of Doctor of Philosophy

in

Bioinformatics

by

Silpa Suthram

Committee in charge:

Professor Trey G. Ideker, Chair
Professor Vineet Bafna
Professor Philip Bourne
Professor Jeff Hasty
Professor Daniel O'Connor
`

2008

The Dissertation of Silpa Suthram is approved, and it is acceptable in quality and form for publication on microfilm:

Chair

University of California, San Diego

2008

To Sanjit

I couldn't have done it without you.

Aum

Tamaso Ma Jyotir Gamaya ||

(Brhadaranyaka Upanishad — I.iii.28)

Translation: Lead me from darkness (ignorance) to light (knowledge).

Table of Contents

List of Figures

List of Tables

Acknowledgements

First and foremost, I would like to thank my advisor Dr. Trey Ideker. This dissertation would not have been possible without his guidance and inspiration. His energy, lucid thinking and work ethic have always served as inspiration. Often, discussions with him helped me overcome disappointments, ready to tackle the obstacles afresh. He also facilitated many collaborations which have enriched my experience in research.

I would also like to thank all the members of my committee: Dr. Vineet Bafna, Dr. Jeff Hasty, Dr. Phil Bourne and, Dr. Dan O'Connor. Their comments and suggestions during various committee and individual meetings have been instrumental in improving my dissertation. I would like to thank the Bioinformatics program Steering committee, especially Dr. Shankar Subramaniam and Dr. Pavel Pevzner, for accepting me into the program and, suggesting classes and rotations to take in my first year.

During the length of my graduate studies, I have worked with various collaborators, and my thesis has benefited greatly from their insights. Specifically, Taylor Sittler and Dr. Andreas Beyer for always being available for long discussions and, the ensuing exchange of ideas has been vital to my research. In addition, I also want to thank Dr. Roded Sharan for his expertise in algorithmic and statistical techniques which have inspired me to work towards improving my understanding of them.

I am grateful to all my colleagues in the Ideker lab for their suggestions and criticisms during weekly lab meetings. In particular, Dr. Kai Tan and Sourav Bandyopadhyay, who were my “neighbors” in the lab for chats on systems biology research, and also for willingly reviewing the initial manuscripts of my papers and providing valuable comments.

I also thank my classmates in the Bioinformatics program, especially Leah Barrera and Thuy Vo, for their research insights and also for lending a ready ear whenever I had some questions and also for being great friends. I also want to thank Han-Yu Chuang, Alice Lee, Samantha Smeraglia, Barbara Mattson, Sandeep Kanumuri and Sourja Ray for their friendship. I thank the Association of Indian Graduate Students (AIGS) for organizing various social events over the years. I am also grateful to Mrs. Pushpa Reddy, Dr. Sekar Krishna, Debashish Panigrahi, Drs. Ranjita and Kanishka Lahiri who helped me adjust to a new environment when I first moved to San Diego.

Lastly, I would like to thank all my family for their support. Specially, my parents, Radha and Suresh, for instilling in me an appreciation for knowledge and education from an early age and encouraging me in all my pursuits. My younger brother Sagar has always been a source of inspiration for me. Finally, I thank my husband Sanjit, for always believing in me and encouraging me to try my best.

Chapter 2 in part quotes sections from the Supplementary Methods of Sharan R, Suthram S, Kelley RM, Kuhn T, McCuine S, Uetz P, Sittler T, Karp RM, Ideker T (2005) Conserved patterns of protein interaction in multiple species. *Proc Natl Acad Sci U S A* **102:** 1974-1979. I was the second author in that work and was responsible for the generation of interaction probabilities which is quoted in this chapter. This chapter also contains the complete reprint of the work Suthram S, Shlomi T, Ruppin E, Sharan R, Ideker T (2006) A direct comparison of protein interaction confidence assignment schemes. *BMC Bioinformatics* **7:** 360. I was the primary author of this paper.

Chapter 3 contains the complete reprint of the paper Sharan R, Suthram S, Kelley RM, Kuhn T, McCuine S, Uetz P, Sittler T, Karp RM, Ideker T (2005) Conserved patterns of protein interaction in multiple species. *Proc Natl Acad Sci U S A* **102**: 1974-1979. I was the second author in that work and was responsible for the generation of interaction probabilities and analysis of the algorithm. I was also involved in the critical reading of the manuscript and analyzing the biological significance of the conserved complexes. Roded

Sharan was responsible for the conception and implementation of the NetworkBLAST algorithm. Ryan Kelley implemented the layout algorithm.

Chapter 4 contains the complete reprint of the work Suthram S, Sittler T, Ideker T. The Plasmodium protein network diverges from those of other eukaryotes. *Nature* 2005;438:108-12. I was the primary co-author of this paper.

Chapter 5 contains the complete reprint of the work Suthram S, Beyer A., Ideker T. eQED: An efficient method for interpreting eQTL associations using protein networks. Proceedings of 3rd Annual RECOMB Satellite Conference on Systems Biology, 2007. I was the primary author of this paper.

Vita

2001	B.Sc (Hons.)	Indian Institute of Technology, Kharagpur (India)
2008	Ph.D.	University of California, San Diego, (CA)

Publications

- Suthram S.**,** Beyer A, and Ideker T. *eQED: An efficient method for interpreting eQTL associations using protein networks*. **Proceedings of 3rd RECOMB Satellite Workshop on Systems Biology**, 2007.

- Suthram S., Shlomi T., Ruppin E., Sharan R., and Ideker T. *A direct comparison of protein interaction confidence assignment schemes*. **BMC Bioinformatics,** 2006 July, p. 7:360.

- Suthram S.**,** Shlomi T., Ruppin E., Sharan R., and Ideker T. *Comparison of protein-protein interaction confidence assignment schemes***.** **Proceedings of 1st RECOMB Satellite Workshop on Systems Biology**, 2005.

- Suthram S*., Sittler T.* and Ideker T. *The Plasmodium protein network diverges from those of other eukaryotes*. **Nature**. 2005;438(7064):108-12.

- Sharan R., Suthram S., Kelley RM, Kuhn T, McCuine S, Uetz P, Sittler T, Karp RM, and Ideker T. *Conserved Patterns of Protein Interaction in Multiple Species*. **Proceedings of National Academy of Sciences, USA (PNAS)**, 2005;102(6):1974-9.

- Suguna C., Maithreye R., Suthram S., and Sinha S. *Dynamics of cellular processes: From single cell to collective behaviour*. Recent Research Developments in Biophysical Chemistry, edited by C.A Condat and A. Baruzzi, **Research Signpost** (2002).

(*) Authors contributed equally

ABSTRACT OF THE DISSERTATION

Understanding cellular function through the analysis of protein interaction networks

by

Silpa Suthram

Doctor of Philosophy in Bioinformatics

University of California, 2008

Professor Trey Ideker, Chair

A major challenge of post-genomic biology is understanding the complex networks of interacting genes, proteins and small molecules that give rise to biological form and function. Advances in whole-genome approaches are now enabling us to characterize these networks systematically, using procedures such as the two-hybrid assay and protein co-immunoprecipitation to screen for protein-protein interactions (PPI). Large protein networks are now available for many species like the baker's yeast, worm, fruit fly and the malaria parasite *P. falciparum.* These data also introduce a number of technical challenges: how to separate true protein-protein interactions from false positives; how to annotate interactions with functional roles; and, ultimately, how to organize large-scale interaction data into models of cellular signaling and machinery. Further, as protein interactions form the backbone of cellular function, they can potentially be used in conjunction with other large-scale data types to get more insights into the functioning of the cell. In this dissertation, I try to address some the above questions that arise during the analysis of protein networks.

First, I describe a new method to assign confidence scores to protein interactions derived from large-scale studies. Subsequently, I perform a benchmarking analysis to compare its performance with other existing methods. Next, I extend the network comparison algorithm, NetworkBLAST, to compare protein networks across multiple species. In particular, to elucidate cellular machinery on a global scale, I performed a multiple comparison of the protein-protein interaction networks of *C. elegans*, *D.*

melanogaster and *S. cerevisiae*. This comparison integrated protein interaction and sequence information to reveal 71 network regions that were conserved across all three species and many exclusive to the metazoans. I then applied this technique to the analysis of the protein network of the malaria pathogen *Plasmodium falciparum* and showed that its patterns of interaction, like its genome sequence, set it apart from other species.

Finally, I integrated the PPI network data with expression Quantitative Loci (eQTL) data in yeast to efficiently interpret them. I present an efficient method, called 'eQTL Electrical Diagrams' (eQED), that integrates eQTLs with protein interaction networks by modeling the two data sets as a wiring diagram of current sources and resistors. eQED achieved a 79% accuracy in recovering a reference set of regulator-target pairs in yeast, which is significantly higher performance than three competing methods. eQED also annotates 368 protein-protein interactions with their directionality of information flow with an accuracy of approximately 75%.

1 Introduction

One of the most important consequences of the Human Genome Project[1,2] is the advent of a new field of biology called *Systems Biology*. The "systems" approach involves understanding and analyzing all aspects of a system (e.g whole cell system) simultaneously rather than studying them one at a time. It has also led to the development of technologies that allow us to measure levels of the underlying components on a very large scale[3,4]. For instance, advances in the microarray technologies let us determine the mRNA expression levels of all the genes in any given species at the same time[5]. New technologies such as chromatin immunoprecipitation with microarrays (ChIP-chip) and protein binding microarrays (PBMs)[6,7] have enabled us to discover the transcription factor DNA interactions on a large scale. The availability of such large quantities of data requires the generation of new methods to store, analyze and understand them. As a result, inputs from various disciplines such as computer science, statistics and mathematics are essential to a "systems" approach of analyzing data. For example, the computer science concept of "database" has been implemented to create the GenBank[8] to not only store large amounts of genome sequence data that is being generated for many species, but also provide a free publicly available resource that everyone can access.

A major data set that has originated from the advances in high-throughput technology is the large scale discovery of protein interactions within a cell. Proteins regulate and mediate many of the processes in the cell. In most cases, they act in concert with other proteins as part of pathways or larger molecular assemblies called complexes. Thus to understand the function of a protein, it is essential to learn all its associated interactions. Extrapolating this idea, it is necessary to obtain an aggregate of all the protein interactions in the cell to understand cellular behavior. Until recently, protein interactions were mainly discovered by small-scale methods like GST pull down[9] and FRET microscopy[10]. These reveal only a small number of protein interactions in one experiment. Now, high-throughput techniques like yeast two-hybrid[11] and tandem affinity purification (TAP) followed by mass-spectrometry[12], reveal protein interactions at the level of the whole species. Large protein networks are now available for many species including, most recently, the malaria pathogen *P. falciparum*[13]. These data also introduce a number of technical challenges: how to separate true protein-protein interactions from false positives[14]; how to annotate interactions with functional roles[15]; and, ultimately, how to organize large-scale interaction data into models of cellular signaling and machinery[16]. In this dissertation, I try to address some of these questions that arise in the analysis of protein interaction data.

1.1 Generation and visualization of large-scale protein interaction data

There are many methods to explore the enormous number protein interactions within a cell. Here, I discuss two technologies, yeast-two hybrid and Tandem Affinity Purification (TAP) followed by mass-spectrometry, that have been used to generate the various large protein-protein interaction (PPI) networks that are presently available[12,17-20].

1.1.1 Yeast-two hybrid assay

The yeast two-hybrid (Y2H) assay utilizes the yeast cell to check whether two proteins interact or not[11]. The main principle is to test whether a downstream reporter gene is activated when a reconstituted transcription factor binds to its upstream activation sequence (UAS). Generally, the transcription factor consists of a DNA-binding domain (BD) and an activation domain (AD) and both are essential for the activation of the reporter gene. For many eukaryotic transcription factors, these two domains (AD and BD) are modular. In other words, they only need to be in close proximity to function properly and not necessarily be part of the same protein. Using this hypothesis, in the Y2H assay, one of the proteins (*A*) to be tested is genetically engineered to have the AD of the transcription factor and the other protein (*B*) has the BD. Therefore, only when proteins *A* and *B* interact with each other, the AD and BD will be in close proximity and the downstream reported gene will be activated. Generally, the GAL4 AD and BD are used in the Y2H assay and the reporter gene is LacZ. Figure 1.1 shows a schematic of this process.

1.1.2 Tandem Affinity Purification (TAP) followed by mass-spectrometry

The method of TAP followed by mass-spectrometry (I shall refer to this method as mass-spectrometry from here on) discovers protein complexes rather than binary protein interactions as in the case of Y2H[12]. Specifically, a TAP tag is inserted at the 3′ end of the protein of interest (bait protein). The protein is then immunoprecipitated *invivo* using an antibody against the TAP tag, purified and checked for binding partners. Next, these protein assemblies are separated using a denaturing gel electrophoresis and digested using trypsin. The resulting peptides are then put through a mass-spectrometer in order to identify the component proteins. Thus, the technique of mass-spectrometry identifies groups of proteins bound to the bait protein.

1.1.3 Visualization of protein interaction networks

Protein-protein interaction networks are usually represented as a network of nodes and links (see Figure 1.2) and, I will use this representation throughout this dissertation. The nodes correspond to proteins, while the links between them correspond to protein interactions. The protein interactions obtained in an Y2H assay can easily be adapted to form a protein network explained above. However, the mass-spectrometry experiment provides a list of protein complexes rather than individual interactions. Therefore, for each protein complex, I assume that there exist interactions between all members of the complex. After this interpolation, it is easy to abstract the protein complex as part of the protein network.

Many software tools have been developed to enable the visualization of large-scale protein networks[21-25]. Cytoscape (www.cytoscape.org)[25] is one such tool for modeling and comparing large-scale networks of molecular interactions and combining them with expression and cellular perturbation data. Originally developed as an open-source project by Dr. Trey Ideker and colleagues, it combines the ability to view and manipulate genome-sized networks of cellular pathways and an extensible architecture such that new analysis tools can be added dynamically. I use Cytoscape for visualizing all the networks and results in the rest of the dissertation.

1.2 Overview of the dissertation

A major challenge of post-genomic biology is to understand how the complex networks of interacting genes, proteins and small molecules give rise to biological form and function. This issue is of immediate importance as the amount of data on protein interactions is increasing rapidly (see Figure 1.3). The wide variety and number of PPI networks motivate many questions like: How do the networks organize into regulatory pathways and complexes to accomplish various cellular functions? Are some of these modules also present in other species and how does this translate to conservation of function? Can the conservation of protein interactions be used to make functional predictions? We are also interested in knowing if there exist modules that are functionally important, but unique to a given species. These modules are especially insightful when picking drug targets against pathogenic species like *Plasmodium falciparum*. Moreover, as protein interactions form the backbone of cellular function, they can potentially be used in conjunction with other large-scale data types to get more insights into the functioning of the cell. All the above questions have prompted two different approaches to analyze protein interaction data: comparative network analysis and network integration. In this thesis, I present two studies implementing the comparative network analysis approach. Later, I also discuss another work incorporating the strategy of network integration. Overviews of the three studies are explained in the following sections.

All the aforementioned questions can be addressed only if the underlying protein interaction network is of good quality. Unfortunately, large-scale protein networks, like other high-throughput data, has a lot of false-positives[14]. Therefore, I begin this thesis by addressing the issue of noise in large-scale protein networks.

1.2.1 Noise in the data

All large-scale measurements and particularly new technologies suffer from a high level of noise. The high-throughput techniques of Y2H and mass-spectrometry result in a considerable number of false-positive interactions[14]. These errors arise due to bias in the experimental conditions. For instance, in a Y2H assay, if one of the proteins activates transcription on its own (auto-activation), it will invariably lead to the inference that the interaction being tested is true. Hence, tools are required to distinguish the false interactions from the true ones. Recent years have also seen an increase in the accumulation of other sources of biological data such as whole genome sequence, mRNA expression, protein expression and functional annotation[13]. This is particularly advantageous as some of these data sets can be utilized to reinforce true protein interactions while downgrading others. For instance, true protein interactions have been shown to have high mRNA expression correlation for the corresponding genes[26]. In Chapter 2, I propose a new method to assign confidence scores to protein interactions. In addition, I also benchmark this method with other existing methods.

1.2.2 Comparative network analysis

The first method of protein network analysis, called *Comparative network analysis*, is akin to the genome sequence comparison methods. Evolutionary conservation is a fundamental principle in biology that is widely used to infer functional relationships among species. Conservation of protein/gene sequences across species is used to make function and domain assignments[27]. In the same vein, comparing protein interaction networks across species will highlight evolutionarily conserved pathways and modules. I address these questions in Chapter 3. First, I explain an extension of a previous method[28], NetworkBLAST, to align the PPI networks of two or more species at a time. The method is then implemented to compare the protein networks of yeast, worm and the fruit fly. The conserved modules obtained are also used to make new functional predictions in the three species. Moreover, this comparative method allowed us to make new protein interaction predictions. In Chapter 4, I compared the protein network of the malaria parasite, *Plasmodium falciparum*, across three other eukaryotes and not only determined *Plasmodium* protein complexes that were conserved across the species, but also modules that were unique to *Plasmodium*. This distinct set of *Plasmodium* modules is an important resource in understanding the *Plasmodium* system and could probably be used to find new drug targets against the parasite.

1.2.3 Network integration

Next, I attempt to analyze the PPI networks using the second approach called *Network Integration*. While comparative network analysis dealt with protein interactions spread across multiple species, network integration tries to understand cellular function only in one species by combining additional sources of information on the same set of proteins (genes). Network integration of different sources of data has been used previously to discover cellular machinery. For instance, Kelley *et al.*[29] has integrated protein-protein interactions with genetic interactions in yeast to produce modules that were enriched in both types of interactions. They found that most modules contained genetic interactions across two pathways which shared complementary function. This discovery underscores the value of combined analysis of protein and genetic interactions.

In particular, in Chapter 5, I integrate the PPI network data in yeast with expression Quantitative Loci (eQTL) data also in yeast to efficiently interpret them. Genetic variation gives rise to changes in many quantitative traits including gene expression. The technique of expression quantitative trait loci (eQTL) investigates the interactions between genetic loci and the changes in gene expression[30,31]. Specifically, a collection of genetically diverse strains is used to establish correlations between a quantitative phenotype (such as gene expression) and polymorphisms at a specific genetic locus. This method allows the detection of loci and consequently, the genes contained in them that regulate the expression of the gene downstream. eQTLs can be mainly divided into two categories, namely *cis*- and *trans*-eQTLs (or 'local' versus 'distant'[32]. The *cis*-eQTLs correspond to DNA variation in close proximity to the target gene. It is assumed that in most cases, the *cis*-eQTLs can be found in the transcriptional regulatory regions of the target gene. On the other hand, *trans*-eQTLs are located far away from the target gene and are more difficult to ascertain. I focus on understanding the nature of these *trans*-eQTLs. A *trans*-eQTL imposes directionality of information flow from the locus to the affected target gene. One might assume that transcription factors (TFs) result in the strongest eQTL, since their effect on down-stream genes is most immediate. However, in many cases eQTLs cannot be explained by such simple TF – target relationships[33]. Hence, in order to fully comprehend the mechanisms underlying significant *trans*-eQTLs one has to take into account more complex, indirect interactions via sequences of protein-protein and protein-DNA interactions[34-36].

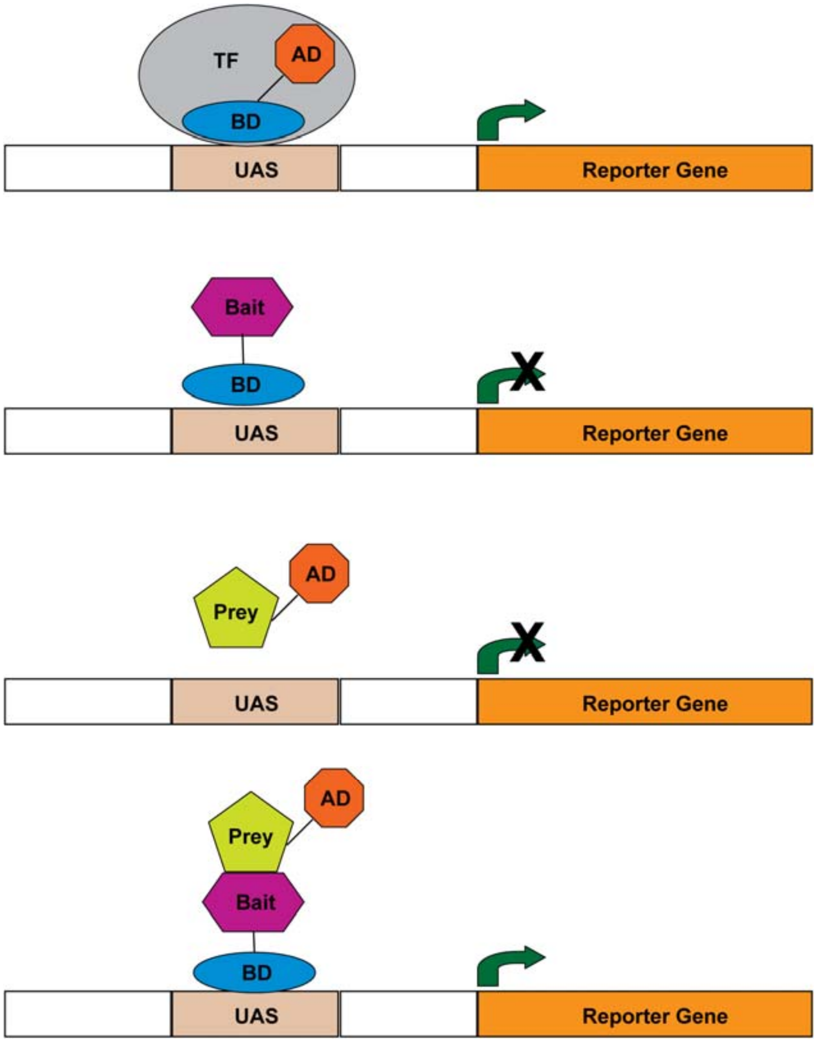


Figure 1.1: Schematic of the yeast two-hybrid assay.
The reporter gene is activated only when the Activating Domain (AD) and Binding Domain (BD) are in close proximity, which happens only when the bait and the prey protein interact with each other.

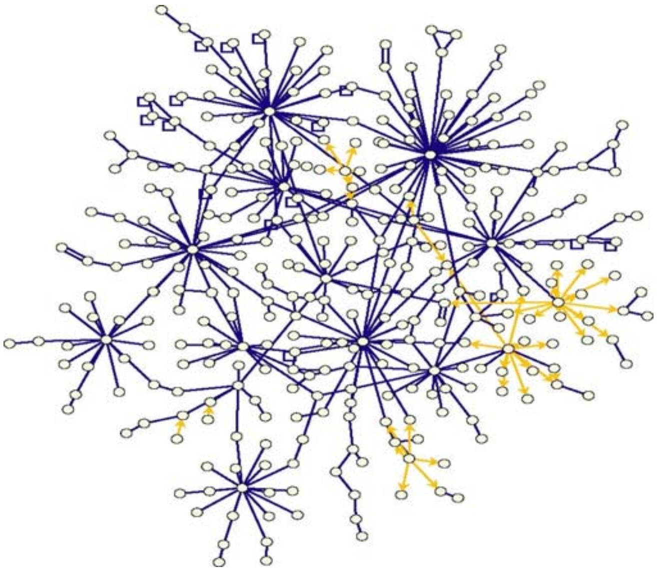


Figure 1.2: Network abstraction of large-scale protein interactions.
The nodes correspond to proteins and the links between nodes corresponds to protein interactions. The figure has been drawn using the Cytoscape software[25].

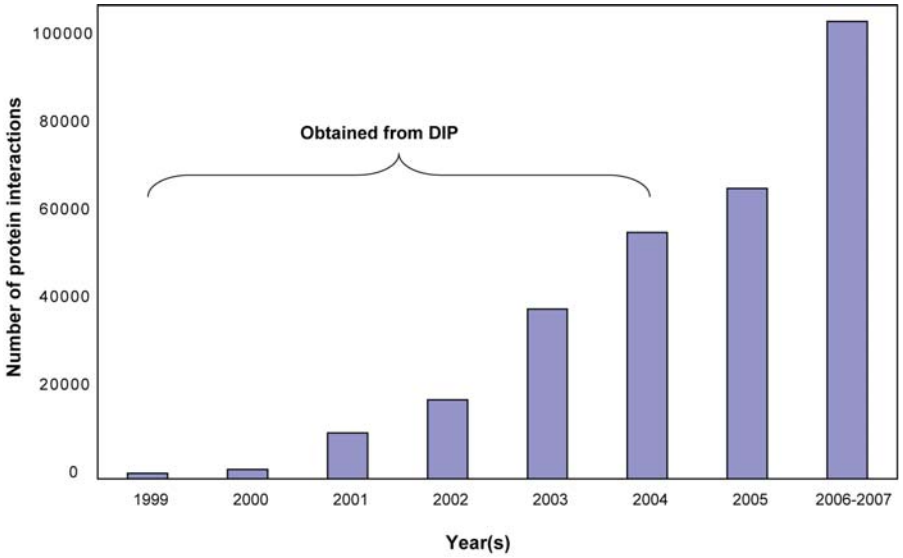


Figure 1.3: Increase in the number of protein-protein interactions.
The figure shows the rapid increase in the number of protein interactions over the past few years. The data for the years 1999-2004 was collected from the Database of Interacting Proteins (DIP)[37]. The data for the year 2005 included the protein-protein interaction network generated for *Plasmodium falciparum* and human. The data for 2007 included the protein interactions curated by the Human Protein Reference Database (HPRD)[38].

2 Assigning confidence scores to protein interactions

Systematic elucidation of protein-protein interaction networks will be essential for understanding how different behaviors and protein functions are integrated within the cell. Recently, the advent of high-throughput experimental techniques like yeast two-hybrid (Y2H) assays[11] and co-immunoprecipitation (co-IP) screens[12] has led to the elucidation of large-scale protein interaction networks in different species, including *S. cerevisiae* (yeast)[12,17,18,20], *D. melanogaster* (fly)[39], *C. elegans* (worm)[40] and *H. sapiens* (human)[41,42]. These networks, while incorporating thousands or tens of thousands of measured interactions, have so far only partially covered the complete repertoire of protein interactions in an organism, and they have been determined to contain a significant number of false-positive interactions depending on the study[14]. However, recent years have also seen an increase in the accumulation of other sources of biological data such as whole genome sequence, mRNA expression, protein expression and functional annotation. This is particularly advantageous as some of these data sets can be utilized to reinforce true (physical) protein interactions while downgrading others. For instance, biologically relevant protein interactions have been shown to have high mRNA expression correlation for the proteins involved[26].

As a result, many integrative bioinformatic approaches have been developed to unearth true protein-protein interactions. These can be mainly divided into two categories: (1) methods that assign reliability measurements to previously observed interactions; and (2) methods that predict interactions *ab initio*. For category (1), Deane *et al.*[43] and Deng *et al.*[44] introduced methods to tackle the problem of assigning reliabilities to interactions using similarity in mRNA expression profiles. Subsequently, Bader *et al.*[45] used additional features of interacting proteins, including functional similarity and high network clustering[46], to assign confidence scores to protein interactions. For category (2), Marcotte *et al.*[47], von Mering *et al.*[48], Myers *et al.*[49] and Jansen *et al.*[50] were among the first to predict new protein interactions by incorporating a combination of different features like high mRNA expression correlation, functional similarity, co-essentiality, and co-evolution. These schemes calculate a log-likelihood score for each interaction. As yet another approach in this category, Qi *et al.*[51] predicted new protein interactions using a method based on random forests. Presumably, the relative performance of each of these approaches versus the others involves a combination of factors such as the types of evidence used as inputs, the efficacy of each classification algorithms, and the sets of true and false interactions used as gold standards during training. Very recently, a second work by Qi *et al.*[52] studied the effect of the underlying classification algorithm by comparing the accuracies of different classifiers such as naïve Bayes, logistic regression, and decision trees.

In section 2.1, we first propose a new method to assign confidence scores to protein-protein interactions obtained from high-throughput screens. In the remaining sections, we perform a benchmarking analysis to evaluate the published interaction confidence schemes versus one another. Rather than isolate every factor that could influence a scheme's performance, we take a practical approach and evaluate the overall accuracy of each set of confidence scores as reported in the literature and available from the authors' websites. We limit ourselves to works that have assigned confidence scores to a common set of experimentally-observed interactions in yeast; this includes all of the category (1) schemes above, as well as the Qi. *et al.* scheme from category (2). The remaining *ab initio* schemes are concerned with predicting new interactions and do not assign confidences to those interactions that have already been experimentally observed. We also assess the performance of a "null hypothesis", a uniform scheme that considers the same probability for all interactions. To compare the quantitative accuracy of the methods, we examine the correlations between the confidence estimates and different biological attributes such as function and expression. As a further comparison criterion, we apply the signal processing concept of 'Signal-to-Noise Ratio' (SNR) to evaluate the significance of protein complexes identified in the network based on the different schemes[53]. The discovery of these complexes depends on the connectivity of the interaction network which, in turn, is influenced by the underlying interaction probabilities[53,54].

2.1 Estimation of interaction probabilities

We assign confidence values to protein interactions using a novel logistic regression model. For a given species, our model represents the probability of true interaction as a function of three observed random variables on a pair of proteins. First, the number of times an interaction between the proteins was observed experimentally. The number of observations of a given interaction is an indicator of the reproducibility of an interaction and is a good measure of the "truth" of an interaction. It has also been used as a measure to predict interaction confidence scores previously[44]. Second, the Pearson correlation coefficient of expression patterns of the corresponding genes. The similarity in the mRNA expression profiles has been associated with biologically relevant PPIs[26]. Specifically, let x and y be two m-long vectors of expression levels for two genes. The *Pearson correlation coefficient* between the two vectors is defined as $\rho = \frac{\frac{1}{m}\sum_{i=1}^{m} x_i y_i - \overline{xy}}{\sigma_x \sigma_y}$ where $\bar{x}, \bar{y}$ are the sample means and σ_x, σ_y are the standard deviations of x and y, respectively. Third, the small world clustering coefficient of the two proteins. For proteins, v and w denote the sets of proteins that interact with them by $N(v)$ and $N(w)$, respectively. Let N be the total number of proteins in the network. The *small-world clustering coefficient* for v and w is:

$$C_{vw} = -\log \sum_{i=|N(v)\cap N(w)|}^{\min\{|N(v)|,|N(w)|\}} \frac{\binom{|N(v)|}{i}\binom{N-|N(v)|}{|N(w)|-i}}{\binom{N}{|N(w)|}}$$

The clustering coefficient was suggested by Goldberg *et al.*[46] to account for similarity in network connections. They also showed that higher values for small-world clustering coefficient corresponded to biologically relevant protein interactions.

According to the logistic distribution, the probability of a true interaction T_{uv} given the three input variables, represented by $X=(X_1,X_2,X_3)$, is :

$\Pr(T_{uv} \mid X) = \dfrac{1}{1 + \exp(-\beta_0 - \sum_{i=1}^{3} \beta_i X_i)}$, where $\beta_0, \ldots, \beta_3$ are the parameters of the distribution. Given training data, one can optimize the distribution parameters so as to maximize the likelihood of the data. To this end we used the *glmfit* function of MATLAB[55]. The training data is determined based on the species under consideration. Generally, two sets of gold standard training data are required. The positive training data is a set of protein interactions that are known to be true, while the negative training data is a set of protein interactions that are known to be false.

2.2 Benchmarking analysis

Although large-scale protein interaction networks are being generated for a number of species, *S. cerevisiae* is perhaps the best studied among them and is associated with the largest variety and quantity of protein interaction data. Hence, most of the interaction probability schemes have been developed using the yeast protein interaction network as a guide. As the probability schemes were previously computed for different subsets of yeast protein-protein interactions, we compiled a test set of 11,883 yeast interactions common to all schemes. These yeast interactions were derived from both yeast two-hybrid[18,20] and mass-spectrometry-based[12,17] screens.

In total, we considered seven interaction probability assignment schemes, including Bader *et al.*[45], Deane *et al.*[43], Deng *et al.*[44], Qi *et al.*[51] and our model, descibed in the previous section called Sharan *et al.*[54]. Bader *et al.*, Sharan *et al.* and Qi *et al.* have assigned specific probabilities to every yeast interaction, while Deane *et al.* and Deng *et al.* have grouped yeast interactions into high/medium/low confidence groups. All of the above schemes define and use some set of gold standard positive and negative interaction examples for the probability estimation.

Bader et al. (BADER_LOW / BADER_HIGH)

As a gold standard positive training data set, Bader *et al.*[45] used interactions determined by co-IP, in which the proteins were also one or two links apart in the Y2H network. The negative training data set was selected from interactions reported either by co-IP or Y2H, but whose distance (after excluding the interaction) was larger than the median distance in Y2H or co-IP respectively. Using these training data, they constructed a logistic regression model that computes the probability of each interaction based on explanatory variables including data source, number of interacting partners, and other topological features like network clustering. We refer to this scheme as Bader *et al.* (low) or BADER_LOW in our analysis.

Initially, the authors used measures based on Gene Ontology (GO)[56] annotations, co-expression, and presence of genetic interactions as measures to validate their data. However, they also combined these measurements into the probability score to bolster their confidence of true interactions. We consider these new confidence scores in our analysis as Bader *et al.* (high) or BADER_HIGH.

Deane et al. (DEANE)

Deane *et al.*[43] estimated the reliability of protein-protein interactions using the expression profiles of the interacting partners. Protein interactions observed in small-scale experiments that were also curated in the Database of Interacting Proteins (DIP)[37] were considered as the gold standard positive interactions. As a gold standard negative, they randomly picked protein pairs from the yeast proteome that were not reported in DIP. The authors used this information to compute the reliabilities of groups of interactions (obtained from an experiment or a database). Higher reliabilities were assigned to groups whose combined expression profile was closer to the gold standard positive than the gold standard negative interactions. Specifically, reliabilities were assigned to the whole DIP database, the set of all protein interactions generated in any high-throughput genome screen, and protein interactions generated by Ito *et al.*[18].

Deng et al. (DENG)

Deng *et al.*[44] estimated the reliabilities of different interaction data sources in a manner similar to Deane *et al.*[43]. They separately considered experiments that report pair-wise interactions like Y2H and those that report complex membership like mass spectrometry. Curated pair-wise interactions from the literature and membership in protein complexes from Munich Information center for Protein Sequences (MIPS)[57] were used as the gold standard positive set in each case. Randomly chosen protein pairs formed the gold standard negative data set. Reliabilities for each data source were computed using a maximum likelihood scheme based on the expression profiles of each data set. The authors evaluated reliabilities for Y2H data sources like Uetz *et al.*[20] and Ito *et al.*[18], and protein complex data sources like Tandem Affinity Purification (TAP)[12] and High-throughput Mass Spectrometric Protein Complex Identification (HMS-PCI)[17]. In addition to assigning reliabilities to each dataset, the authors also provided a conditional probability scheme to compute probabilities for groups of interactions observed in two or more data sources. This calculation results in assigning a high probability (0.99) to yeast interactions observed in more than 1 data source. We use the probabilities generated by this method for our comparative analysis.

Sharan et al. (SHARAN)

Recently, Sharan *et al.*[54] also implemented an interaction probability assignment scheme similar to the one proposed by Bader *et al.* The scheme assigned probabilities to interactions using a logistic regression model based on mRNA expression, interaction clustering and number of times an interaction was observed in independent experiments. Here, we use a modification of this scheme, assigning probabilities to interactions based only on direct experimental evidence. Specifically, interactions with at least two literature references or those that had a distance ≤ 2 in both the co-IP and Y2H networks were defined as the gold standard positives. Conversely, proteins at a distance > 4 in the entire network (after removing the interaction in question) were defined as the gold standard negatives. Binary variables were used to denote whether the interaction was reported in a co-IP data set, Y2H data set, a small-scale experiment or a large–scale experiment. Interaction probabilities were then estimated using logistic regression on the predictor parameters similarly to Bader *et al*.

Qi et al. (QI)

In this study, the authors used interactions that were observed in small-scale experiments and reported by either DIP or Bader *et al.* as their gold standard positive training data[51]. Randomly picked protein pairs were used as the gold standard negative training data. The method incorporates direct evidence such as the type of experiment used to generate the data and indirect evidence like gene expression, existence of synthetic lethal interactions, and domain-domain interactions to construct a random forest (a collection of decision trees). The resulting forest is then used to calculate the probability that two proteins interact.

Equal Probabilities (EQUAL)

Finally, we also considered the case in which all observed interactions were considered to be equally true. We refer to this case as EQUAL in the analysis.

A summary of all attributes used as inputs to the different probability schemes is provided in Table 2.1. It should be noted that even though the different probability schemes utilize some of the same types of inputs (e.g., experiment type, expression similarity), they may incorporate these inputs in different ways. For instance, both SHARAN and DENG use "experiment type" as input, but SHARAN explicitly includes each type of experiment as a separate indicator variable in its logistic regression function, while DENG pools data from each experimental type and assigns a single confidence level to the interactions in each pool.

We also compared global statistics such as the average and median probability assigned by each scheme (see Table 2.2). We found that most probability schemes had an average probability in the range of [0.3 - 0.5]. In contrast, Deane *et al.* (DEANE) had a very high average and median probability: over half of the interactions in the test set were assigned a probability of 1. We also computed Spearman correlations among the different probability schemes to measure their levels of inter-dependency (Table 2.3). The maximum correlation was seen between BADER_LOW and BADER_HIGH, as might be expected since both schemes were reported in the same study and BADER_HIGH was derived from BADER_LOW. On the other hand, Qi *et al.* (QI) had very low Spearman correlation with any of the probability schemes. The low correlation may reflect an inherent difference between schemes that assign probabilities to experimentally observed interactions and ones that predict protein interactions *ab initio*. The probabilities assigned by the schemes can be obtained from the Supplementary website[58].

2.3 Quality assessment

One of the most objective ways to assess the performance of the different confidence assignment schemes would be to compare their success at correctly classifying a gold standard set of true protein interactions. However, all of the schemes considered in this analysis had already used the available gold standard sets of known yeast interactions in the training phase of their classifiers and, consequently, assigned high confidence scores to them. As an alternative approach, we employed five measures that had been shown to associate with true protein interactions[26,53,59,60] to gauge the performance of the schemes. One caveat of this approach is that, in some cases, one of the measures used to assess a scheme's performance had already been used (in full or in part) as an input to assigning its probabilities. To avoid circularity, this measure was used only for gauging the performance of the remaining schemes. For each of the five measures, two ways were used to estimate the level of association: Spearman correlation and weighted average (see

Methods). Importantly, by using the Spearman correlation coefficient, we are in fact comparing how the schemes rank the interactions, not the absolute scores that are assigned. Note that the EQUAL probability scheme results in Spearman correlation of 0, by definition.

2.3.1 Presence of conserved interactions in other species

Presence of conserved interactions across species is believed to be associated with biologically meaningful interactions[60]. As our benchmark, we used yeast protein interactions that were conserved with measured *C. elegans* and *D. melanogaster* interactions obtained from the Database of Interacting Proteins (DIP)[37]. An interaction was considered conserved if homologs of the interacting yeast proteins were also interacting in another species. Homologs were based on amino-acid sequence similarity computed using BLAST[27], thus allowing a protein to possibly match with multiple proteins in the opposite species (if interacting yeast proteins were homologous to any pair of homologs with an observed interaction, the yeast interaction was counted as conserved). In particular, we allow interactions whose interacting proteins are themselves homologs, but filter cases where both the interacting proteins pointed to the same protein in the other species. We evaluated the weighted average and Spearman correlation between the probability assignment for each yeast interaction and the number of conserved interactions across worm and fly (0, 1, or 2). We used an E-value cut-off of $1X10^{-10}$ to make the homology assignments (Table 2.4). We observed that SHARAN and BADER_HIGH had the highest weighted average and Spearman correlation. Not surprisingly, EQUAL had the lowest weighted average. Note that the conserved interactions test is a very strong filter for true interactions as it heavily depends on the level of completeness of the interaction networks of other species being considered. However, as the underlying set of interactions is the same across the different probability schemes, this filter affects all schemes similarly.

2.3.2 Expression correlation

Yeast expression data for ~790 conditions were obtained from the Stanford Microarray Database (SMD)[61]. For every pair of interacting proteins, we computed the Pearson correlation coefficient of expression. We then calculated the Spearman correlation and weighted average between the expression correlation coefficients of interacting proteins and their corresponding probability assignments in the different schemes (see Table 2.4

and 2.5). We found significant association between expression correlations and probabilities in the case of BADER_HIGH, BADER_LOW, QI and DENG. This result is expected as these schemes, with the exception of BADER_LOW, utilize expression similarity for interaction probability calculation. Surprisingly, DEANE probabilities showed very little correlation with expression, even though mRNA expression profiles were used as input in the prediction process reflecting the difference in the way expression similarity is incorporated in this method. In particular, DEANE is the only method where expression similarity between two interacting proteins is taken into account as the Euclidean distance between their expression profiles versus other methods which incorporated the Pearson correlation coefficient of expression. On the other hand, BADER_LOW had a higher Spearman correlation than SHARAN, though both had very similar weighted averages and did not utilize expression data in the training phase.

2.3.3 Gene Ontology (GO) similarity

As a first measure, we adopted the common notion that two interacting proteins are frequently involved in the same process and hence should have similar GO assignments[56]. The Gene Ontology terms are represented using a directed acyclic graph data structure in which an edge from term 'a' to term 'b' indicates that term 'b' is either a more specific functional type than term 'a', or is a part of term 'a'. As a result, terms that appear deeper in the graph are more specific. Moreover, specific terms also have fewer proteins assigned to them or their descendants.

Let "P_i" and "P_j" be two proteins that have been observed to interact with each other. To measure their functional similarity, we evaluated the size (number of proteins assigned to the term), represented as "S_{ij}", of the deepest common GO term assignment (deepest common ancestor in graph) shared between them. Thus, a smaller value of S_{ij} indicates a greater functional similarity between P_i and P_j. In addition, we also found that known yeast interactions generally have lower values for S_{ij} than random background (see Figure 2.1). To ensure that higher values of our GO measure correspond to higher performance (as is the case for other quality assessment metrics below), we use the negative of S_{ij} (or $-S_{ij}$) to represent the overall GO similarity.

Table 2.4 shows the relationship between GO similarity and the interaction probabilities for each scheme. Of the schemes that did not use functional annotations as inputs, DENG and SHARAN both had a very high Spearman correlation with GO (with DENG slightly higher than SHARAN).

However, one potential concern was that GO functional assignments could incorporate evidence of co-expression which was used as an input by the DENG scheme. This potential circularity can be addressed by use of the partial correlation coefficient to factor out the dependency of GO on co-expression (see Table 2.6). However, the partial correlation is almost certainly an overcorrection since GO similarity and co-expression (and in fact any two lines of evidence) are expected to have some correlation if they are both predictive of true interactions. Regardless, with or without the correction, DENG and SHARAN scored within 2% of each other; thus the two schemes are practically indistinguishable by the GO metric.

2.3.4 Signal-to-Noise Ratio of protein complexes

Most cellular processes involve proteins that act together by assembling into functional complexes. Several methods[28,54,62-64] have been developed to identify complexes embedded within a protein interaction network, in which a complex is typically modeled as a densely-connected protein sub-network. Recently, we showed that the quality of these identified protein complexes could be estimated by computing their signal-to-noise ratio (SNR), a standard measure used in information theory and signal processing to assess data quality (see Methods)[53]. Essentially, SNR evaluates the density of complexes found in the protein interaction network against a randomized version of the same network.

As the SNR is independent of the number of complexes reported, its value can be directly compared across the different probability schemes. For discovery of protein complexes, we applied a previously-published algorithm[54] which includes interaction probabilities in the complex identification process. SNR was then computed on the set of complexes identified by each probability scheme. Results are shown in Table 2.7; out of all of the schemes, DENG had the highest SNR of protein complex detection.

2.3.5 Conservation rate coherency

Interacting proteins have been shown to evolve at similar rates, probably due to selection pressure to maintain the interaction over time[59]. For every pair of interacting proteins, P_i and P_j, let "r_i" and "r_j" be their respective rates of evolution. We then computed a "conservation rate coherency score" (CR_{ij}) as the negative absolute value of the difference between the evolutionary rates of the two corresponding genes: $CR_{ij} = -| r_i - r_j |$. The negative absolute value was used to ensure that higher values represent higher performance, consistent with other metrics.

Evolutionary rates were obtained from Fraser *et al.*[65] and estimated using nucleotide substitution frequencies. We calculated the Spearman correlation between the values of CR for the interacting proteins and their corresponding probability assignments in the different schemes (see Table 2.7). For all probability assignment schemes we obtained a statistically significant correlation (p-value < 0.05) between the conservation rate coherency scores and the corresponding probabilities, indicating that proteins with high probability interactions tend to have similar conservation rates. The highest correlation was obtained for DENG.

2.4 Discussion

A brief review of the performance results suggests that the DENG method (Deng *et al.*) emerges as the clear winner, with top scores in three out of four non-circular quality metrics. Comprising a 'second tier' are BADER_HIGH, BADER_LOW (the two Bader methods) and SHARAN, which perform very similarly across most metrics with some differences in conservation coherency or gene expression (for which SHARAN performs better or worse, respectively). BADER_LOW, which considers experiment type and interaction clustering as inputs, has a higher expression score than SHARAN, which considers experiment type only, implying that interaction clustering helps capture expression similarity. Interestingly, BADER_HIGH, which incorporates more input attributes than BADER_LOW or SHARAN, does not have substantially higher rankings. Thus, in this case, adding more inputs to a probability assignment scheme does not appear to strongly enhance its quality.

As for the remaining schemes with lower overall performance (DEANE and QI), it is interesting to note that these were arguably the least and most sophisticated schemes, respectively. The DEANE method relied on only a single evidence type for assigning confidences, that of gene expression, whereas it appears that other factors may have been more informative (Table 2.1). In contrast to DEANE, QI had the largest number of inputs for assigning confidences and, among these, included data on both co-expression and experiment type. However, it is well known that classifier accuracy can be degraded by including many irrelevant input variables[66], and perhaps this is the reason for QI's lower performance. As an alternative explanation, in Qi *et al.*'s evaluation of classification schemes, they concluded that their method was very successful in predicting co-complex membership, but performed poorly when considering physical interactions[52]. In our analysis, all interactions (even co-complex membership) were treated as pair-wise protein interactions, and this assumption may have contributed to the poor performance of Qi *et al.* Certainly, their classification method was among the most sophisticated of the schemes that we evaluated, and as such it is worthy of future exploration (perhaps with different sources of input data) regardless of its performance in the present study.

Finally, EQUAL almost always scored lowest, regardless of quality metric. Thus, utilizing any probability scheme is better than considering all observed interactions to be true or equally probable.

Beyond these broad rankings, is it possible to synthesize data from five largely independent metrics to arrive at an overall quantitative index of performance? As one approach, we normalized the scores for each metric as a fraction of the best score achieved within that metric over all confidence assignment schemes (i.e., for each metric, the highest score was fixed to 1 and the scores of the remaining schemes were converted to fractional values between 0 and 1). Table 2.8 summarizes the fractional scores for the six probability schemes and five quality assessment measures. Note that expressing scores as fractional values is an intermediate normalization which preserves the score distribution but compresses its range; although potentially more informative than the non-parametric analysis above based only on ranks, it must also be interpreted with more caution. However, in this case, the fractional scores reinforce the findings reported above based on rank.

We have compared and contrasted seven probability assignment schemes for yeast protein interactions. Surprisingly, Deng *et al.* performs significantly better than others while being one of the least sophisticated. It assigns discrete probability scores to large groups of interactions rather than to individuals, and it inputs just two lines of evidence, experiment type and expression similarity, rather than many. Generalizing these observations, more complex approaches are so far unable to outperform simpler variants. Thus, we arrive at a somewhat unexpected conclusion: At least in interaction confidence assignment, sometimes less means more.

2.5 Methods

2.5.1 GO databases

The Gene Ontology annotations for yeast proteins were obtained from the July 5th, 2005 download of the Saccharomyces Genome Database (SGD)[67];

the graph of relations between terms was obtained from the Gene Ontology consortium (http://www.geneontology.org/).

2.5.2 Signal to noise ratio (SNR)

To compute SNR, a search for dense interaction complexes is initiated from each node (protein) and the highest scoring complex from each is reported. This yields a distribution of complex scores over all nodes in the network. A score distribution is also generated for 100 randomized networks, which have identical degree distribution to the original network. SNR is computed using these original and random score distributions (representing signal and noise, respectively) according to the standard formula[68] using the root-mean-square (rms):

$$\mathrm{SNR} = \log_{10}\frac{\mathrm{rms(original\ complex\ scores)}}{\mathrm{rms(random\ complex\ scores)}}, \quad \text{where } \mathrm{rms}\left(x_1 \cdots x_M\right) = \sqrt{\frac{1}{M}\sum_{i=1}^{M} x_i^2}$$

where M denotes the total number of complexes (in this case, equal to the number of nodes) and x_i represents the score of an individual complex.

2.5.3 Weighted average

The weighted average is given by $WA = \frac{\sum_{i=1}^{N} p_i * m_i}{\sum_{i=1}^{N} p_i}$, where p_i is the probability of a given interaction and m_i is the value of one of the five measures for the interaction.

Acknowledgements

Chapter 2 in part quotes sections from the Supplementary Methods of Sharan R, Suthram S, Kelley RM, Kuhn T, McCuine S, Uetz P, Sittler T, Karp RM, Ideker T (2005) Conserved patterns of protein interaction in multiple species. *Proc Natl Acad Sci U S A* **102:** 1974-1979. I was the second author in that work and was responsible for the generation of interaction probabilities which is quoted in this chapter. This chapter also contains the complete reprint of the work Suthram S, Shlomi T, Ruppin E, Sharan R, Ideker T (2006) A direct comparison of protein interaction confidence assignment schemes. *BMC Bioinformatics* **7:** 360. I was the primary author of this paper.

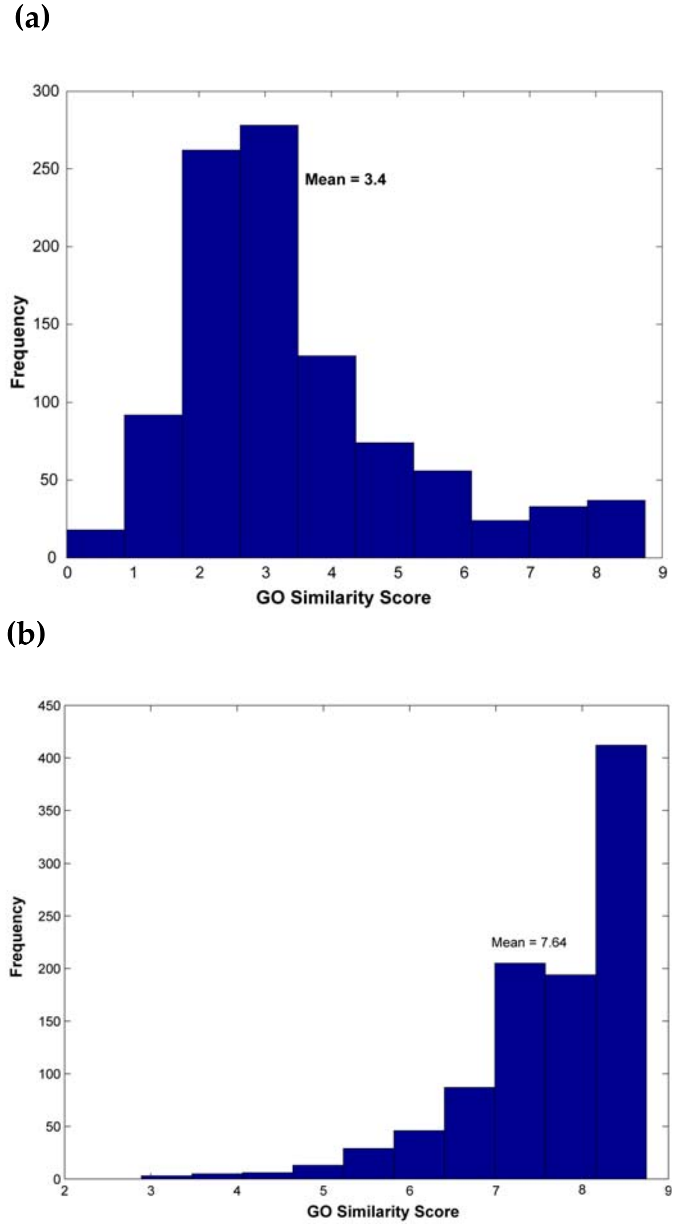
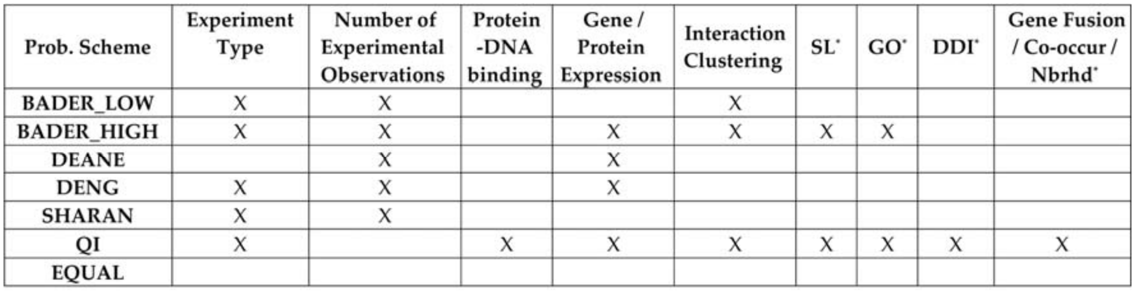


Figure 2.1 Histogram of GO similarity scores.

(a) for the known yeast interactions reported in the MIPS database. **(b)** for 1,000 interactions generated randomly. These random interactions were generated by picking pairs of proteins randomly from the set of interacting proteins in yeast. It is evident from the two figures that true proteins interactions (i.e known yeast interactions reported in MIPS) generally have lower GO similarity scores than the background.

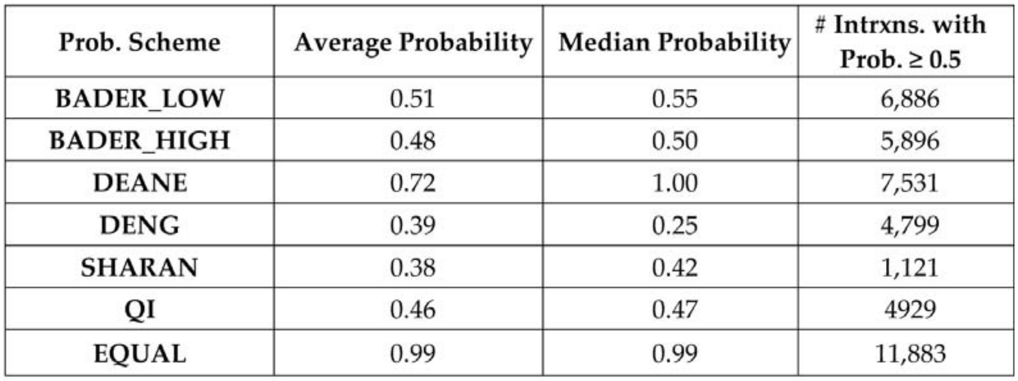
Table 2.1: Summary of input attributes for the different probability schemes.



Prob. Scheme	Experiment Type	Number of Experimental Observations	Protein -DNA binding	Gene / Protein Expression	Interaction Clustering	SL*	GO*	DDI*	Gene Fusion / Co-occur / Nbrhd*
BADER_LOW	X	X			X				
BADER_HIGH	X	X		X	X	X	X		
DEANE		X		X					
DENG	X	X		X					
SHARAN	X	X							
QI	X		X	X	X	X	X	X	X
EQUAL									

*SL: Synthetic Lethal; GO: Gene Ontology; DDI: Domain-domain Interactions; Nbrhd: Neighborhood

Table 2.2: Global properties of the probability assignment schemes.



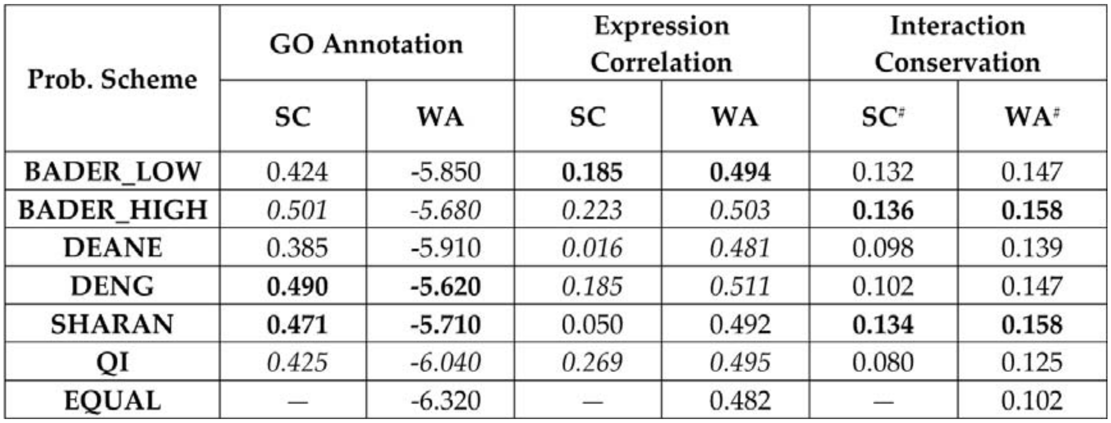
Prob. Scheme	Average Probability	Median Probability	# Intrxns. with Prob. ≥ 0.5
BADER_LOW	0.51	0.55	6,886
BADER_HIGH	0.48	0.50	5,896
DEANE	0.72	1.00	7,531
DENG	0.39	0.25	4,799
SHARAN	0.38	0.42	1,121
QI	0.46	0.47	4929
EQUAL	0.99	0.99	11,883

Table 2.3: Correlation of different probability schemes.

	BADER_HIGH	DEANE	DENG	SHARAN	QI
BADER_LOW	0.923	0.655	0.633	0.626	0.095
	BADER_HIGH	0.672	0.644	0.665	0.151
		DEANE	0.718	0.847	-0.090
			DENG	0.680	0.185
				SHARAN	-0.013

**p*-values of all correlation measurements were significant (*p*-value ≤ 2×10^{-16}).

Table 2.4: Correlation of interaction probabilities with the GO similarity measure, mRNA expression correlation and interaction conservation.*



Prob. Scheme	GO Annotation		Expression Correlation		Interaction Conservation	
	SC	WA	SC	WA	SC$^{\#}$	WA$^{\#}$
BADER_LOW	0.424	-5.850	**0.185**	**0.494**	0.132	0.147
BADER_HIGH	*0.501*	*-5.680*	*0.223*	*0.503*	**0.136**	**0.158**
DEANE	0.385	-5.910	*0.016*	*0.481*	0.098	0.139
DENG	**0.490**	**-5.620**	*0.185*	*0.511*	0.102	0.147
SHARAN	**0.471**	**-5.710**	0.050	0.492	**0.134**	**0.158**
QI	*0.425*	*-6.040*	*0.269*	*0.495*	0.080	0.125
EQUAL	—	-6.320	—	0.482	—	0.102

*Bold values indicate the scheme that performs the best. Italicized values indicate potential circularity, i.e., schemes that use GO annotations or mRNA expression profiles for confidence scoring that are similar to those used here for comparative assessment. *P*-values for all the Spearman correlation measurements are significant. SC: Spearman Correlation; WA: Weighted Average.

$^{\#}$ All measurements were done at an E-value cut-off of 1X10^{-10}.

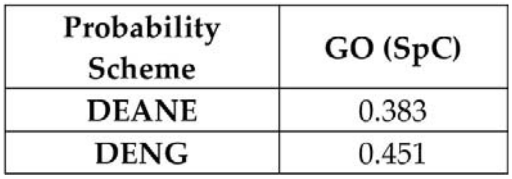
Table 2.5: Correlation of interaction probabilities with mRNA expression correlation.*

Prob. Scheme	Expression Correlation	
	SC	WA
BADER_LOW	**0.185 (0.187)**	**0.494 (0.497)**
BADER_HIGH	*0.223 (0.221)*	*0.503 (0.505)*
DEANE	*0.016 (0.010)*	*0.481 (0.483)*
DENG	*0.185 (0.185)*	*0.511 (0.514)*
SHARAN	0.050 (0.045)	0.492 (0.495)
QI	*0.269 (0.274)*	*0.495 (0.499)*
EQUAL	—	0.482 (0.485)

*Ribosomal components are among the most co-expressed genes, and could potentially lead to the observed relative importance of co-expression data. To check for the effect of ribosomal proteins, we filtered the yeast interaction set in our analysis to remove all ribosomal proteins and calculated the correlation between co-expression and interaction probability. The values in brackets correspond to the yeast interaction set which is filtered for ribosomal proteins.

Bold values indicate the scheme that performs the best. Italicized values indicate potential circularity, i.e., schemes that use mRNA expression profiles for confidence scoring that are similar to those used here for comparative assessment. *P*-values for all the Spearman correlation measurements are significant. SC: Spearman Correlation; WA: Weighted Average.

Table 2.6: Effect of correlation between GO and mRNA expression.



Probability Scheme	GO (SpC)
DEANE	0.383
DENG	0.451

* SpC = Spearman partial correlation. Both schemes used expression as input to assign confidence scores to protein interactions. The Spearman partial rank correlation coefficient between two random variables A and X, given the fact that both A and X are correlated to random variable Y, denotes the correlation between A and X, when Y is kept constant. It is calculated as follows:

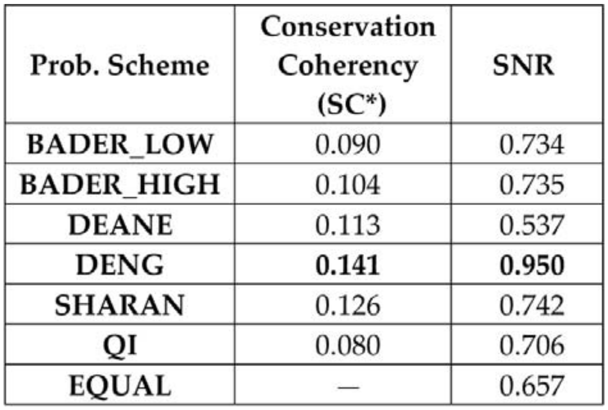
$$r_{AX,Y} = \frac{r_{AX} - r_{XY} r_{AY}}{\sqrt{(1 - r_{XY}^2)(1 - r_{AY}^2)}}$$

Here, r_{AX}, r_{XY} and r_{AY} represent the Spearman correlation coefficients between A and X, X and Y, and, A and Y respectively. The significance level is given by

$$D_{AX,Y} = 1/2\sqrt{N-4}\ln\left(\frac{1 + r_{AX,Y}}{1 - r_{AX,Y}}\right)$$

$D_{AX,Y}$ has a normal distribution with zero mean and variance one. N represents the size of the data set.

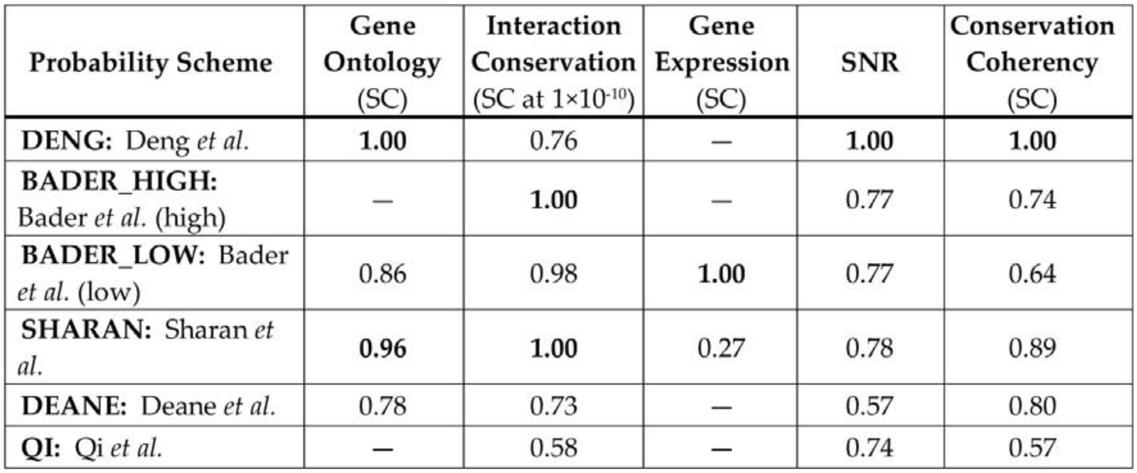
Table 2.7: Associations of conservation rate coherency scores and SNR with interaction probabilities.



Prob. Scheme	Conservation Coherency (SC*)	SNR
BADER_LOW	0.090	0.734
BADER_HIGH	0.104	0.735
DEANE	0.113	0.537
DENG	**0.141**	**0.950**
SHARAN	0.126	0.742
QI	0.080	0.706
EQUAL	—	0.657

* SC: Spearman Correlation. Bold values indicate the scheme which performs the best. Note that conservation scores based on weighted averages were omitted as they were very similar across the different confidence assignment schemes.

Table 2.8: Fractional scores of the confidence assignment schemes in each of the five quality measures.*



Probability Scheme	Gene Ontology (SC)	Interaction Conservation (SC at 1×10^{-10})	Gene Expression (SC)	SNR	Conservation Coherency (SC)
DENG: Deng *et al.*	**1.00**	0.76	—	**1.00**	**1.00**
BADER_HIGH: Bader *et al.* (high)	—	**1.00**	—	0.77	0.74
BADER_LOW: Bader *et al.* (low)	0.86	0.98	**1.00**	0.77	0.64
SHARAN: Sharan *et al.*	**0.96**	**1.00**	0.27	0.78	0.89
DEANE: Deane *et al.*	0.78	0.73	—	0.57	0.80
QI: Qi *et al.*	—	0.58	—	0.74	0.57

*Fractional scores are between [0,1] with 1 performing the best (indicated in bold for each measure). Cells with a dash (-) indicate circularity, i.e., the measures used as (full or partial) input to the corresponding probability schemes. SC: Spearman Correlation; SNR: Signal to Noise Ratio.

3 Comparative network analysis

A major challenge of post-genomic biology is to understand the complex networks of interacting genes, proteins and small molecules that give rise to biological form and function. Advances in whole-genome approaches are now enabling us to characterize these networks systematically, using procedures such as the two-hybrid assay[11] and protein co-immunoprecipitation[69] to screen for protein-protein interactions. To date, these technologies have generated large interaction networks for bacteria[70], yeast [12,17,18,20] and, recently, fruit fly [39] and nematode worm [40].

The large amount of protein interaction data now available presents new opportunities and challenges in understanding evolution and function. Such challenges involve assigning functional roles to interactions[15]; separating true protein-protein interactions from false positives[14]; and, ultimately, organizing large-scale interaction data into models of cellular signaling and regulatory machinery. We addressed the issue of distinguishing true interactions from false in the previous chapter. In this chapter, we try to concentrate on the questions of organizing the PPIs in regulatory modules. As is often the case in biology, an approach based on evolutionary cross-species comparisons provides a valuable framework for addressing these challenges. However, while methods for comparing DNA and protein sequences have been a mainstay of bioinformatics over the past 30 years, development of similar tools at other levels of biological information—protein interactions[28,71,72], metabolic networks[73-75] or gene expression data[76-78] —is just beginning.

Recently, Kelley *et al.* devised a method called PathBLAST[28] for comparing the protein interaction networks of two species. Just as BLAST performs rapid pairwise alignment of protein sequences[27], PathBLAST is based on efficient alignment of two protein networks to identify conserved network regions. In the rest of the chapter, we extend this approach to present the first computational framework for alignment and comparison of more than two protein networks. We apply this multiple network alignment strategy to compare the newly-available protein networks for worm, fly and yeast, and show that while any single network contains false-positive interactions, embedded beneath this noise are a repertoire of protein interaction complexes and pathways conserved across all three species.

3.1 NetworkBLAST Algorithm

3.1.1 *Network alignment graph*

We developed a general framework for comparison and analysis of multiple protein networks. The main goal was to identify protein sub-networks that approximate a given structure and are conserved across a group of *k* species of interest, where in the present study we focused on *k*=2,3. A structure is specified as a property on the protein networks, e.g., being a path (sequence of protein interactions modeling signaling pathways) or a clique (a dense cluster of protein interactions modeling protein complexes), and sets our expectations with respect to a sub-network that approximated that structure. For instance, a sub-network that corresponds to a clique should involve densely interacting proteins.

Conservation of a network structure requires the fulfillment of two conditions: (1) the set of sub-network interactions within each species should approximate the desired structure; and (2) there should exist a (many-to-many) correspondence between the sets of proteins exhibiting the structure in the different species, so that groups of *k* proteins, one from each species, induced by this correspondence, represent *k* putatively homologous proteins.

To capture these conservation requirements and to allow efficient search for conserved sub-networks we define a *network alignment graph* that integrates interactions with sequence information. Each node in the graph consists of a group of sequence-similar proteins, one from each species; each link between a pair of nodes represents conserved protein interactions between the corresponding protein groups (Figure 3.1). The proteins are considered homologous if their BLAST E-value[27] is smaller than 10^{-7} (corresponding to an adjusted *p*-value of 0.01), and each is among the 10 best BLAST matches of the other. A group of *k* distinct proteins, one from each species, comprise a node, if the group cannot be split into two parts with no homology between them. For *k*=2,3 this condition translates to the requirement that every protein in the group has at least one homolog in the group. Two nodes $(p_1 \ldots p_k)$ and $(q_1 \ldots q_k)$ in the graph are connected by an edge if and only if one of the following conditions is true w.r.t. the protein pairs (p_i, q_i): (1) one pair of proteins directly interacts and all other pairs include proteins with distance at most two (indirectly connected through another protein) in the corresponding interaction maps; (2) all protein pairs are of distance exactly two in the corresponding interaction maps; or (3) at least max {2,*k*-1} protein pairs directly interact. Note that it may be the case that for some *i*, $p_i = q_i$; we then consider the pair (p_i, q_i) to have a distance 0.

A subgraph of the network alignment graph corresponds to a conserved sub-network. For each species *S*, the set of proteins included in the nodes of the subgraph defines the sub-network that is induced on *S*. The node memberships define the homology relationships between the sets of proteins of the different species.

3.1.2 *A probabilistic model of protein sub-networks*

In order to detect structured sub-networks, we score subgraphs of the alignment graph which corresponds to collections of conserved sub-networks in different species. Our score is based on a likelihood ratio model for the fit of a single sub-network to the given structure. The log-likelihood ratios are summed over all species to produce the score of the collection. In the following we describe the likelihood ratio model.

Let *G* be the interaction graph of a given species on a set of proteins *P*. Suppose at first, that we have perfect interaction data i.e., each edge in the interaction graph represents a true interaction and each non-edge represents a true non-interaction. To score the fit of a subgraph to a predefined structure we formulate a log-likelihood ratio model that is additive over the edges and non-edges of G, such that high-scoring subgraphs would correspond to likely structured sub-networks. Such a model requires specifying a null model and a protein sub-network model for the node pairs. These models extend those presented in Sharan *et al.*[72] to account for any target structure, although in the discussion below we concentrate on monotone graph properties: if a graph satisfies it then it continues to satisfy it after adding any set of edges to it.

Let s be a target monotone graph property (e.g., being a clique), let $P' \subseteq P$ be a subset of the proteins, and let H be a labeled graph on P' that satisfies s. We define the two models as follows: the *sub-network model*, M_s, corresponding to the target graph H, assumes that every two proteins that are connected in H are also connected in G with some high probability β. In contrast, the *null model*, M_n, assumes that each edge is present with a probability that one would expect if the edges of G were randomly distributed but respected the degrees of the nodes. More precisely, we let F^G be the family of all graphs having the same node set as G and the same degree sequence, and define the probability of observing the edge (u,v) to be the fraction of graphs in F^G that include the edge. Note that in this way, edges incident on nodes with higher degrees have higher probability. We estimate these probabilities using a Mote-Carlo approach as described in Sharan *et al.*[72]

Next, we refine the above models to the realistic case in which we are give partial, noisy observations of the true interaction data. In this case, the probabilistic model must distinguish between observed interactions and true interactions. For ease of presentation we concentrate on the case that the target structure is a clique (corresponding to a protein complex), but the models generalizes to other structures as well. Let us denote by T_{uv} the event that two proteins u,v interact, and by F_{uv} the event that they do not interact. Denote by O_{uv} the (possibly empty) set of available observations on the proteins u and v, that is, the set of experiments in which an interaction between u and v was or was not observed. Given a subset U of the nodes, we wish to compute the likelihood of U under a sub-network model and under a null model. Denote by O_u the collection of all observations on node pairs in U. Under the assumption that all pairwise interactions are independent we have:

$$
\begin{aligned}
&\Pr((O_U \mid M_s) = \prod_{(u,v)\in U\times U} \Pr(O_{uv} \mid M_s) \\
&= \prod_{(u,v)\in U\times U} \left[\Pr(O_{uv} \mid T_{uv}, M_s)\Pr(T_{uv} \mid M_s) + \Pr(O_{uv} \mid F_{uv}, M_s)\Pr(F_{uv} \mid M_s)\right] \\
&= \prod_{(u,v)\in U\times U} \left[\beta \Pr(O_{uv} \mid T_{uv}) + (1-\beta)\Pr(O_{uv} \mid F_{uv})\right]
\end{aligned}
$$

To compute *Pr(Ou|Mn)* we must update the null model, which depends on knowing the degree sequence of the interaction graph. We overcome this difficulty by approximating the degree of each node *I* in the hidden interaction graph by its expected degree, *di*. This refined model assumes that *G* is drawn uniformly at random from the collection of all graphs whose degree sequence is $d_1,\ldots d_n$. This induces a probability *puv* for every node pair *(u,v)*. Thus, we have:

$$\Pr(O_U \mid M_n) = \prod_{(u,v)\in U\times U} [p_{uv}\Pr(O_{uv} \mid T_{uv})] + (1-p_{uv})\Pr(O_{uv} \mid F_{uv})$$

Finally, the log-likelihood ratio that we assign to a subset of nodes *U* is

$$L(U) = \log\frac{\Pr(O_U \mid M_s)}{\Pr(O_U \mid M_n)} = \sum_{(u,v)\in U\times U} \log\frac{\beta\Pr(O_{uv} \mid T_{uv}) + (1-\beta)\Pr(O_{uv} \mid F_{uv})}{p_{uv}\Pr(O_{uv} \mid T_{uv}) + (1-p_{uv})\Pr(O_{uv} \mid F_{uv})}$$

3.1.3 Search algorithm

Using the above model, the problem of identifying conserved protein networks reduces to the problem of identifying high-scoring subgraphs of the network alignment graph. This problem is computationally hard[72]. Thus, we present a heuristic strategy for the search problem.

We perform a bottom-up search for high-scoring subgraphs in the alignment graph. The highest scoring paths with four nodes are identified using an exhaustive search. For dense subgraphs, we start from high-scoring nodes (called seeds), refine them, and then expand them using local search. Similar approaches based on local search were shown to work well in analyzing high-throughput genomic data[72,79].

In the first phase of the search, we compute a seed around each node *v* in the alignment graph using two seeding methods. The first method greedily adds *p* other nodes (*p*=3), one at a time, such that the added node maximally increases the score of the current seed. Next, we enumerate all subsets of the seed of size at least 3 that contain *v*. Each such subset serves as a refined seed. The second seeding method computes the highest scoring path of four nodes that includes *v*, and these four nodes serve as a refined seed.

In the second phase, we apply a local search heuristic on each refined seed. During the local search we iteratively add a node, whose contribution to the score of the current seed is maximum, or remove a node, whose contribution to the current seed is minimum (and negative), as long as this operation increases the overall score of the seed. Throughout the process we preserve the original seed and do not delete nodes from it. For practical considerations, we limit the size of the discovered subgraphs to 15 nodes. For each node in the alignment graph we record up to four highest scoring subgraphs that were discovered around that node.

In the third phase, we use a greedy algorithm to filter subgraphs with a high degree of overlap. We define two subgraphs as overlapping if one of the following two conditions is satisfied: (1) their node intersection size over size of the node union is greater than 80%; or (2) for each species separately, the intersection over the union, computed on the subset of proteins from that species that take part in at least one of the two subgraphs, is greater than 80%. The algorithm iteratively finds the highest scoring subgraph, adds it to the final output list, and removes all other overlapping subgraphs.

Finally, in order to evaluate the statistical significance of the identified sub-networks, we compute a p-value that is based on the distribution of top scores obtained by applying our method to randomized data. The randomized data are produced by random shuffling of each of the input interaction networks, preserving the degrees of the vertices. We also randomize the homology relationships between the different proteins, preserving the number of homologs for each protein. For each randomized dataset, we build a network alignment graph and search for the highest scoring sub-network of a given size. This process is then repeated a large number of times (usually 100 times). We then estimate the *p*-value of a suggested sub-network of the same size, as the fraction of random runs in which the output sub-network had larger score. We retain only sub-networks at a 0.01 significance level.

3.2 Complexes conserved across multiple species

We applied the multiple network alignment framework (Figure 3.1) to perform a three-way alignment of the protein-protein interaction networks of *Caenorhabditis elegans*, *Drosophila melanogaster* and *Saccharomyces cerevisiae*. These species span the largest sets of protein interactions in the public databases to-date and, along with mouse, comprise the major model organisms used to study cellular physiology, development and disease. Protein interaction data were obtained from the Database of Interacting Proteins[37] (February 2004 download) and contained 14,319 interactions among 4,389 proteins in yeast; 3,926 interactions among 2,718 proteins in worm; and 20,720 interactions among 7,038 proteins in fly. Protein sequences obtained from Saccharomyces Genome Database[67], WormBase[80], and FlyBase[81] were combined with the protein interaction data to generate a network alignment of

9,011 protein similarity groups and 49,688 conserved interactions for the three networks.

The NetworkBLAST algorithm was then applied to find conserved complexes across the three species. Since the search is guided by reliability estimates for each protein interaction, we computed them for the protein interactions in each species' network using the method described in the previous chapter. For gold standard positive training data, we used the protein interactions from the Munich Information Center for Protein Sequences (MIPS)[57] for yeast. These protein interactions are generally considered to be true. On the other hand, there exist no accepted set of gold standard set of protein interactions. Hence, we considered as interaction between two proteins in worm or fly to be true if there exists an interaction between their homologs in yeast in the MIPS dataset. To determine homology between a pair of proteins, we used a strict threshold of 10^{-10} on their BLAST E-value. We tried two choices of negative training data. The first considers random pairs of proteins; the second, motivated by the abundance of false-positives in the protein interaction data, considers random observed interactions as true negatives. We performed five-fold cross-validation experiments to evaluate the two choices. The latter generalized better to the test data and was used in the rest of the study. We treated the chosen negative data as noisy indications that the corresponding interactions are false, and assigned those interactions a probability of 0.1397 for being true, where the value of this parameter was optimized using cross-validation. Altogether we collected 1006 postive examples and 1006 negative examples for yeast; 92 positive and 92 negative examples for fly; and 24 positive and 50 negative examples for worm. Histograms of the interaction probabilities learned for each species are presented in Figure 3.5.

Subsequently, a search over the network alignment identified 183 protein clusters and 240 paths conserved at a significance level of $p<0.01$. These covered a total of 649 proteins among yeast, worm and fly. Representative examples of conserved clusters and paths are shown in Figure 3.3. A database of all identified conserved clusters and paths, along with their graphical layouts, is available at http://www.cellcircuits.org/Sharan2004/.

Figure 3.4 shows a global map of all clusters and paths conserved among the yeast, worm and fly protein networks. The map shows evidence of modular structure—groups of conserved clusters overlap to define 71 distinct network regions, most enriched for one or more well-defined biological functions. The largest numbers of conserved clusters were involved in protein degradation (green boxes at lower right), RNA polyadenylation and splicing (blue boxes at lower left) and protein phosphorylation and signal transduction (red boxes at upper right). Other significant conserved clusters were involved in DNA synthesis, nuclear-cytoplasmic transport and protein folding. The map also reveals conserved links between different biological processes, for instance linking kinase signaling (red) to protein catabolism (green; lower right) or to regulation of transcription (yellow; upper middle).

To validate our results, we compared these conserved clusters to known complexes in yeast as annotated by the MIPS[57]. We only considered MIPS complexes that were manually annotated independently from the DIP interaction data (i.e., excluding complexes in MIPS category 550 that are based on high-throughput experiments). Overall, the network alignment contained 486 annotated yeast proteins spanning 57 categories at level 3 of the MIPS hierarchy. We defined a cluster to be *pure* if it contained three or more annotated proteins and at least half of these shared the same annotation. Ninety-four percent of the conserved clusters were pure, indicating the high specificity of our approach, compared to a lower percentage of 83% when applying a non-comparative variant of our method to data from yeast only (i.e., applying the same methodology to search for high-scoring clusters within the yeast network only).

We further checked whether the conserved clusters were biased by spurious interactions, resulting from 'sticky' proteins that lead to positive two-hybrid tests without interaction. Of 39 proteins with more than 50 network neighbors, only ten were included in conserved clusters. These ten proteins were involved in 60 intra-cluster interactions, 85% of which were supported by co-immunoprecipitation experiments. This indicates that the clusters were not biased due to artifacts of the yeast two-hybrid assays.

3.2.1 Three-way versus two-way network alignment

In addition to the three-way comparison, we also performed all possible pairwise network alignments: yeast/worm, yeast/fly and worm/fly. This process identified 220 significant conserved clusters for yeast/worm, 835 for yeast/fly and 132 for worm/fly. Several examples of these are shown in Figure 3.9. Global overviews of the pairwise conserved clusters (similar to Figure 3.2) are provided in Figures 3.5-3.8.

Analysis of the proteins shared among the different pairwise and three-way network comparisons led to two general findings. First, the density and number of conserved clusters found in the yeast/fly comparison were considerably greater than for the other comparisons, due to the large amounts of interaction data for these species relative to worm (see Table 3.6 and Figure 3.10). Second, the worm/fly conserved clusters were largely distinct from the clusters arising from the other analyses. For example, only 29% of the proteins in the worm/fly clusters were assigned to conserved clusters in the three-way analysis (135 out of 462). This observation is consistent with the closer taxonomic relationship of worm and fly compared to yeast and the particular selection of protein "baits" for the *C. elegans* protein-protein interaction screen: roughly one quarter were specifically chosen to be metazoan specific, and almost two-thirds had no clear yeast ortholog[40].

3.2.2 Prediction of new protein functions

Conserved sub-networks that contain many proteins of the same known function suggest that their remaining proteins also have that function. Based on this concept, we predicted new protein functions whenever the set of proteins in a conserved cluster or path (combined over all species) was significantly enriched for a particular Gene Ontology (GO)[82] annotation ($p<0.01$) and at least half of the annotated proteins in the cluster or path had that annotation. When these criteria were met, all remaining proteins in the sub-network were predicted to have the enriched GO annotation (see Methods).

This process resulted in 4,669 predictions of new GO Biological Process annotations spanning 1,442 distinct proteins in yeast, worm and fly; and 3,221 predictions of novel GO Molecular Function annotations spanning 1,120 proteins. We estimated the specificity of these predictions using the technique of cross validation, in which one hides part of the data, uses the rest of the data for prediction, and tests the prediction success using the held-out data (see Methods). As shown in Table 3.1, depending on the species, 58-63% of our predictions of GO Processes agreed with the known annotations (see also Tables 3.3 and 3.4). This analysis outperformed a sequence-based method of annotating proteins based on the known functions of their best sequence matches, for which the accuracy ranged between 37 and 53% (see Methods).

3.2.3 Prediction of new protein interactions

We also used the multiple network alignment to predict new protein-protein physical interactions. We predicted an interaction between a pair of proteins based on [1] evidence that proteins with similar sequences interact within other species (directly, or via a common network neighbor) and, optionally, [2] co-occurrence of these proteins in the same conserved cluster or path. The accuracy of these predictions was evaluated using five-fold cross validation, as described in the Methods section. In cross validation, strategy [1] achieved 77-84% specificity and 23-50% sensitivity, depending on the species for which the predictions were made (see Table 3.2 and 3.5). These results were highly significant for the three species. Combining both strategies [1] and [2] resulted in eliminating virtually all false positive predictions (specificity>99%), while greatly reducing the number of true positives, yielding sensitivities of 10% and lower (see Table 3.2). Given the elevated specificity of the combined strategies, we were able to predict 176 new interactions for yeast, 1,139 for worm and 1,294 for fly with high confidence. Thus, although protein interactions have been used previously to predict interactions among the orthologous proteins of other species[40,60], screening these against conserved paths and clusters markedly improves the specificity of prediction. The complete list of predicted protein interactions is provided on our website.

To further evaluate the utility of protein interaction prediction based on network conservation, we tested experimentally 65 of the interactions that were predicted for yeast using the combined strategies [1] and [2] above (Figure 3.4a). The tests were performed using two-hybrid assays[11,20], which are based on a reporter gene that is transcriptionally activated if the two tested proteins (bait and prey) can physically interact (see Methods and Figure 3.4b). Five of the tests involved baits that induced reporter activity in the absence of any prey (Figure 3.4c). Of the remaining 60 putative interactions, 31 tested positive (more conservatively, 19 out of 48—see Figure 3.4) yielding an overall success rate in the range of 40-52%.

3.3 Discussion

3.3.1 Comparison to existing methods

Kelley *et al.*[28] previously developed pairwise network alignment algorithms that were used to detect linear paths and Sharan *et al.*[72] found dense clusters that are conserved between yeast and the bacteria *H. pylori*. The multiple network alignment scheme that we have presented here is an extension of these earlier approaches to handle more than two species. Additional advantages of the current approach over the previous ones are: [1] a unified method to detect both paths and clusters, which generalizes to other network structures; [2] incorporation of a refined probabilistic model for protein interaction data; and [3] an automatic system for laying out and visualizing the resulting conserved sub-networks. A related method that uses cross-species data for predicting protein interactions is the interolog approach[71,78]: a pair of proteins in one species is predicted to interact if their best sequence matches in another species were reported to interact. In comparison, our proposed scheme can associate proteins that are not necessarily each other's best sequence match. This confers increased flexibility in detecting conserved function by allowing for paralogous family expansion and contraction, or gene loss. Since conservation is evaluated in the context of a protein interaction sub-network and not independently for each interaction, the high specificity of the resulting predictions can be maintained (see below section "Validation of predicted interactions").

3.3.2 Best BLAST hits may not imply functional conservation

Frequently, the network alignment associates sequence-similar proteins between species even though they are not each other's best sequence match. For instance, the conserved network region in Figure 3.2[h] suggests that the worm protein exc-7 plays the same functional role as yeast Pab1 and fly CG33070 (BLAST E-value $\approx 10^{-42}$) based on the conserved interactions with

Asc1/F08G12.2/Rack1 (yeast/worm/fly), Rna15/Unc-75 (yeast/worm) and T01D1.2/Tbph (worm/fly). However, CG33070 is only the fifth best BLAST match in fly overall (the best being CG3151 at *E*-value ≈ $10^{\circledcirc 70}$). Overall, out of 679 protein triples aligned at the same position within a three-way conserved cluster, only 177 contained at least one pair of best sequence matches; out of 129 additional triples in conserved paths, only 31 contained best sequence matches. Clearly, in some cases the best matches are not present within conserved clusters due to missing interactions in the protein networks of one or more species. However, it is unlikely that true interactions with the best-matching proteins would be missed repeatedly across multiple proteins in a cluster and across multiple species. These observations suggest that protein network comparisons provide essential information about function conservation.

3.3.3 Functional links within conserved networks

Conserved network regions enriched for several functions point to cellular processes that may work together in a coordinated fashion. Due to the appreciable error rates inherent in measurements of protein-protein interactions, an interaction in a single species linking two previously unrelated processes would typically be ignored as a false positive. However, an observation that two or three networks reinforce this interaction is considerably more compelling, especially when the interaction is embedded in a densely-connected conserved network region. For example, Figure 3.2[h] links protein degradation to the process of poly-A RNA elongation. Although these two processes are not connected in this region of the yeast network, several protein interactions link them in the networks of worm and fly (e.g., Pros25-Rack1-Msi or Pros25-Rack1-Tbph). These findings are consistent with previously-documented association of proteasomes with mRNA binding proteins, although the exact nature of this association has been controversial[83,84]. A related functional link between the proteasome and nucleic acid synthesis was detected in our earlier network comparison of yeast and the bacteria *H. pylori*[28].

As another example, Figure 3.9[l] shows a worm/fly conserved cluster for which ~40% of the proteins have no significant yeast ortholog (BLAST *E*-value > 0.01). The cluster links functions such as proteolysis (Pros25, Pros28.1, Pas-1-7), actin binding (Cher,W04D2.1), ion transport (CG32810, C40A11.7, C52B11.2) and axon guidance (Fra). Taken together, these functions suggest a role for this cluster in growth cone formation during axon guidance.

Guidance of axons to their synaptic targets is an initial step in the development of the central nervous system[85] and is mediated by special compartments called growth cones at the tips of the extending neurites. Formation of growth cones is induced by elevated levels of Ca^{2+} ions, which trigger local proteolysis and restructuring of the actin cytoskeleton[86]. Thus, as implicated by our findings, axon guidance requires synergy between proteolysis, actin binding and ion transport within an intricate network of protein interactions.

3.3.4 Validation of predicted interactions

Our two-hybrid tests of predicted interactions yielded a success rate in the range of 40-52%. These results are satisfactory for three reasons. First, the performance is clearly significant compared to the chance of identifying protein interactions at random (0.024%, estimated from an earlier two-hybrid screen[20] of 192 baits × 6000 preys that yielded 281 interacting pairs). Second, two-hybrid analysis is known to miss a substantial portion of true interactions[14]; this is particularly likely in our case where protein pairs were checked in only one orientation of bait and prey. For instance, two of the pairs that tested negative (YJR068W-YOR217W; YBL105C-YHR030C) have been shown to interact genetically in synthetic-lethal screens[57], suggesting a possible physical interaction as well. Third, predicting interactions using a multiple network alignment approach compares favorably to previous approaches based on conservation of individual protein interactions. For instance, in Mathews *et al.*[71] the interolog approach was applied to a set of 72 reported interactions in yeast, predicting 71 new interactions in worm. Seven of the predicted worm interactions tested positive using a two-hybrid assay (10%), while 19 of the previously-reported yeast interactions (26%) retested positive. Considering only the worm interactions that were predicted based on the 19 confirmed interactions in yeast, six of these tested positive, upper bounding the prediction accuracy at 31%. In tests of 145 additional predictions, 28 were confirmed, obtaining an overall accuracy of 16%. Similar results were obtained in a subsequent study by Yu *et al.*[60], where the accuracies of the interolog approach and an extension of it were estimated at 30-31%.

3.3.5 Conclusion

Nearly all comparative genomic studies of multiple species have been based on DNA and protein sequence analysis. Here, we transcend that framework by presenting a comparative study of the protein-protein interaction networks of three model eukaryotes. These comparisons show that many circuits embedded within the protein networks are conserved over evolution, and that these circuits cover a variety of well-defined functional categories. Since measurements of protein interactions tend to be noisy and incomplete, it would have been difficult if not impossible to find these mechanisms by looking at only a single species. Moreover, many of these similarities and the network connections they imply would not have been suggested by sequence similarity alone, as the proteins involved are frequently not best sequence matches. The multiple network alignment also allows us to ascribe new functions to many proteins and predict previously unobserved protein-protein interactions. Comparative network analysis is thus a powerful approach for elucidating network organization and function.

3.4 Methods

3.4.1 Scoring functional enrichment

Protein pathways and complexes were associated with known biological functions using the Gene Ontology annotations (GO; May 2004 version)[56]. Since the GO terms are not independent but connected by an ontology of parent-child relationships, we computed the enrichment of each term conditioned on the enrichment of its parent terms as follows. Define a protein to be below a GO term *t* if it is assigned or any other term that is a descendent of *t* in the GO hierarchy. For each pathway or complex (specifying a set of proteins) and candidate GO term we recorded the following quantities: (1) the number of proteins in the sub-network that are below the GO term; (2) the total number of proteins below the GO term; (3) the number of proteins in the sub-network that are below all parents of the GO term; and (4) the total number of proteins below all parents of the GO term. Given these quantities, we computed a *p*-value of significance using a hypergeometric test. The *p*-value was further Bonferroni corrected for multiple testing. All terms assigned to at least one protein in the set were tested.

3.4.2 Prediction of protein functions

We used the inferred pathways and complexes for predicting novel protein functions. A conserved complex or pathway in which many proteins are of the same known function predicts that the remaining proteins in the sub-network will also have this function. Based on this concept, we predicted new protein functions whenever the following four conditions were satisfied: (1) the set of proteins in a conserved complex or pathway (combined across all species) was significantly enriched for a particular GO annotation (p<0.01); (2) at least five of the proteins in the sub-network has this significant annotation; (3) these proteins accounted for at least half of the annotated proteins in the sub-network overall; and (4) the annotation was sufficiently specific (at GO level four or higher). For every species, all remaining proteins in the sub-network were then predicted to have the enriched GO annotation, provided that at least one protein from that species has the enriched annotation.

This process resulted in 4,669 predictions of new GO Biological Process annotations spanning 1,442 distinct proteins in yeast, worm and fly; and 3,221 predictions of novel GO Molecular Function annotations covering 1,120 proteins across the three species. We tested the accuracy of our predictions using the technique of cross-validation: we partitioned the set of known protein annotations into 10 parts of equal size. We then iterated over those parts, where at each iteration we hid the annotations that were included in the current part, and used the remaining annotations to predict the held-out annotations. For each protein, we predicted at most one function-that with the lowest p-value. The prediction was considered correct if the protein has some true annotation that lies on a path in the gene otology tree from the root to a leaf that visits the predicted annotation. As shown in Tables 3.4 and 3.5, depending on the networks and species being compared, 33-63% of our predictions were correct. In particular, our predictions of GO Biological Process using the three-way complexes and pathways achieved success rates of 58% for yeast, 59% for worm and 63% for fly.

We further compared the performance of our function prediction procedure to a simpler prediction process, in which a protein with one or more known functions predicts that its best sequence match in another species has at least one of those functions. For each pair of species yeast/worm, worm/fly and yeast/fly, we used proteins in the first species to predict the function of their best BLAST matches in the second species; the success rates achieved in this process were 36.5%, 40% and 53%, respectively. Even though the annotation using best BLAST matched predicted multiple functions per protein, only one of which had to match a true annotation, the results achieved in the process were comparable to those achieved using pairwaise alignment graphs and inferior to those achieved with thee-way alignment (see Table 3.4). This comparison demonstrated the superiority of an approach that takes into account the interaction data.

3.4.3 Prediction of protein interactions

We also used the alignment graph and the computed sub-networks to predict protein interactions. We experimented with several ways of predicting interactions. The simplest criterion that we tested is to predict as interaction between two proteins whenever there were two nodes in the alignment graph that contained them, such that for at least *l* of the species, the two respective proteins included in those nodes has distance at most 2 within that species' interaction graph. We tried both *l*=1 and *l*=2 and tested our predictions using 5-fold cross-validation.

We defined the training interaction data for the cross-validation experiments as follows: we considered the *n* highest scoring interactions in each species as positive examples, and the *n* lowest scoring interactions as negative examples. To avoid bias toward interactions within dense network regions due to their high clustering coefficient, we recomputed the reliabilities of the protein interactions excluding the clustering coefficient from the model. We removed from the training data interactions that were used for estimating the interaction probabilities; we also removed protein pairs that were not included in the alignment graph being analyzed. At each iteration of the cross-validation experiments we hid one fifth of the interactions (both positives and negatives) and used the remaining data for prediction. Since yeast and fly networks were considerably richer we used *n*=1500 for these two species and *n*=500 for worm.

We applied this strategy to the three-way alignment graph and to the three pairwise graphs. For yeast, *l*=2 gave the highest success rates (percents of correct predictions) in cross-validation; for worm and fly *l*=1 yielded the highest success rates. Denote by TP, FP, TN and FN the numbers of true positives, false positives, true negatives and false negatives, respectively. The sensitivity of the predictions, which is defined as TP/(TP+FN), varied between 19-50%; the specificity of the predictions, TN/(TN+FP), varied between 78-94%. In addition, we also computed the hypergeometric *p*-value for the results, defined as the probability of choosing at random (without replacement) (TP+FP) balls that are labeled negative, so that at least TP balls are positive. In all cases our prediction accuracy was highly significant. The results of the cross-validation experiments are summarized in Table 3.3.

Next, we tested the utility of using information on inferred complexes and pathways in improving the accuracy of the predictions. By adding the requirement that two proteins in a predicted interaction are included in an inferred complex or a pathway, we eliminated virtually all false positives, although at the price of greatly reducing the percent of true positive predictions. The performance of this inference strategy for three-way alignment graph is summarized in Table 3.3.

Based on the high specificity achieved in the cross-validation experiments, we applied our approach to predict novel protein-protein interactions using more stringent criteria described above. We computed a ranked list of predictions by collecting evidence on each predicted interaction as follows: we evaluated the probability that the interaction map of each species induces on the predicted interaction, and combined these into an OR probability for the interaction. That is, if the probability assigned by species *I* to the interaction is p_i, then the probability that we assign to the interaction if $p = 1 - \prod_i (1 - p_i)$. The computation of the probability p_i assigned to an interaction in a single species was based on the two proteins that participated in the nodes of the alignment graph w.r.t. which the interaction was predicted. If these proteins were observed to interact, we used their probability of interaction as p_i. If the distance of the two proteins in species *I* interaction map was 2, we computed p_i using an OR probability on all paths of length 2 between the two proteins, similar to Kelley *et al.*[28]. Otherwise, we assigned $p_i =$ 0.

3.4.4 Automatic layout of conserved complexes

We developed a plug-in for cytoscape[25] to automatically layout collections of conserved complexes for visual inspection. An ideal layout has two properties: (1) within a given complex, nodes do not overlap; and (2) nodes that are connected by an edge are located in close proximity. Laying out several conserved complexes imposes as additional constraint, namely, homologous proteins should be located in analogous positions in their respective species' complexes. The first two constraints are well addressed by existing graph layout strategies. One such strategy is Kamada and Kawai's layout algorithm[87]. In this scheme, each edge is modeled as a spring which exerts a force attracting its endpoint nodes. In addition, all nodes exert a repulsive force to discourage overlap. Given this framework, an ideal layout is one with the lowest possible energy as determined by the forces exerted in the system. In order to satisfy the additional constraint imposed by the conserved complexes, we modify the basic scheme. First, edges are added between all pairs of homologous proteins. Then, the repulsive forces between nodes in distinct complexes are eliminated. After applying the force directed layout, each complex is overlaid with homologous proteins in similar locations. These individual complexes are then separated to yield side-by-side layouts of conserved complexes.

Acknowledgements

Chapter 3 contains the complete reprint of the paper Sharan R, Suthram S, Kelley RM, Kuhn T, McCuine S, Uetz P, Sittler T, Karp RM, Ideker T (2005) Conserved patterns of protein interaction in multiple species. *Proc Natl Acad Sci U S A* **102:** 1974-1979. I was the second author in that work and was responsible for the generation of interaction probabilities and analysis of the algorithm. I was also involved in the critical reading of the manuscript and analyzing the biological significance of the conserved complexes. Roded Sharan was responsible for the conception and implementation of the NetworkBLAST algorithm. Ryan Kelley implemented the layout algorithm.

Table 3.1: Cross-validation for protein cellular process prediction.

Species	#Correct	#Predictions	Success rate
Yeast	114	198	58
Worm	57	95	60
Fly	115	184	63

For each species the table lists the number of correct predictions, the total number of predictions and the success rate in ten-fold cross-validation.

Table 3.2: Cross-validation results for protein interaction predictions.

Species	Sensitivity	Specificity	*p*-value	Strate
Yeast	50	77	1.1e-25	[1]
Worm	43	82	1e-13	[1]
Fly	23	84	5.3e-5	[1]
Yeast	9	99	1.2e-6	[1]+[2]
Worm	10	100	6e-4	[1]+[2]
Fly	0.4	100	0.5	[1]+[2]

For each species the table lists the specificity and sensitivity of the predictions in five-fold cross validation, the significance of the results and the prediction strategy (see text).

Table 3.3: Expanded cross-validation results for protein interaction predictions.

Species	TP	FN	TN	FP	Sensitivity	Specificity	*p*-value
Yeast-Worm	51	149	287	18	26%	94%	1.30E-13
Yeast-**Worm**	42	48	29	5	47%	85%	1.30E-13
Yeast-Fly	77	311	624	42	20%	94%	1.30E-14
Yeast-**Fly**	67	88	101	28	43%	78%	3.90E-14
Worm-Fly	37	161	133	16	19%	89%	3.00E-04
Worm-**Fly**	49	141	126	15	26%	89%	1.40E-08
Yeast-Worm-Fly	117	117	262	80	50%	77%	1.10E-25
Yeast-**Worm**-Fly	54	72	53	12	43%	82%	1.00E-13
Yeast-Worm-**Fly**	54	182	178	33	23%	84%	5.30E-05
(*)**Yeast**-Worm-Fly	20	214	339	3	9%	99%	1.20E-06
(*)Yeast-**Worm**-Fly	13	113	65	0	10%	100%	6.00E-04
(*)Yeast-Worm-**Fly**	1	235	211	0	0.40%	100%	5.00E-01

Entries are: the alignment graph used for predicting interactions for the species that appears in bold-type: overall numbers of true positives (TP), false negatives (FN), true negatives (TN), and false positive (FP) predictions; specificity and sensitivity of the predictions; and a hypergeometric *p*-value of the results. An asterisk denotes that the predictions were made by further requiring the two proteins to be included in a conserved sub-network.

Table 3.4: Expanded cross-validation results for predicting GO Biological Processes.

Species	**#Correct**	**#Predictions**	**Success rate**
Yeast-Worm	93	216	43%
Yeast-**Worm**	54	121	45%
Yeast-Fly	280	637	44%
Yeast-**Fly**	208	517	40%
Worm-Fly	22	55	40%
Worm-**Fly**	34	67	51%
Yeast-Worm-Fly	114	198	58%
Yeast-**Worm**-Fly	57	95	60%
Yeast-Worm-**Fly**	115	184	63%

Entries are the alignment graph used for predicting functions for the species that appears in bold-type; the number of correct predictions; the total number of predictions; and the success rate.

Table 3.5: Cross-validation results for predicting protein GO Molecular Functions.

Species	**# Correct**	**# Predictions**	**Success rate**
Yeast-Worm	61	179	34%
Yeast-**Worm**	40	118	33%
Yeast-Fly	171	488	35%
Yeast-**Fly**	156	402	39%
Worm-Fly	37	64	58%
Worm-**Fly**	31	61	51%
Yeast-Worm-Fly	79	162	49%
Yeast-**Worm**-Fly	51	103	49.50%
Yeast-Worm-**Fly**	77	149	52%

Entries are: the alignment graph used for predicting functions for the species that appears in bold-type; the number of correct predictions; the total number of predictions; and the success rate.

Table 3.6: Protein coverage by complexes and pathways.

Species	# **Proteins**	# **Proteins in sub-networks**	**Coverage**
Yeast-Worm	765	271	35%
Yeast-**Worm**	536	204	38%
Yeast-Fly	1,494	790	53%
Yeast-**Fly**	1,559	778	50%
Worm-Fly	852	246	29%
Worm-**Fly**	1,131	291	26%
Yeast-Worm-Fly	801	219	27%
Yeast-**Worm**-Fly	551	190	34%
Yeast-Worm-**Fly**	911	240	26%

For each alignment graph and each species (appearing in bold-type), given are the number of distinct proteins for this species in the corresponding alignment graph, the number of proteins that are covered by significant complexes and pathways, and the percent of coverage.

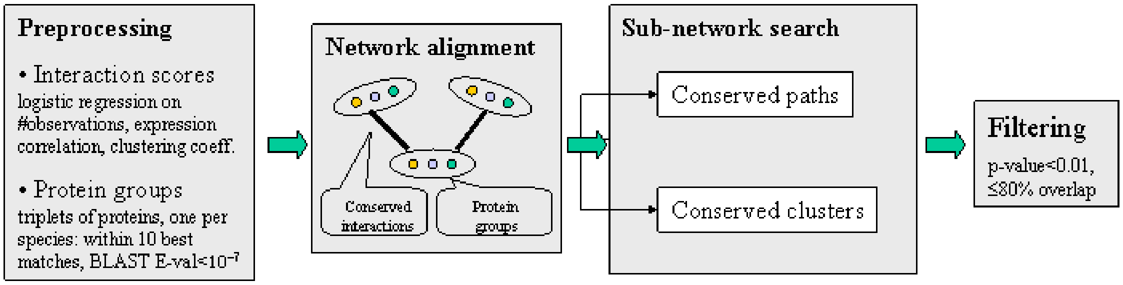

Preprocessing
• Interaction scores
logistic regression on #observations, expression correlation, clustering coeff.
• Protein groups
triplets of proteins, one per species: within 10 best matches, BLAST E-val<10^-7
Network alignment
Conserved interactions
Protein groups
Sub-network search
Conserved paths
Conserved clusters
Filtering
p-value<0.01, ≤80% overlap

Figure 3.1: Schematic of the multiple network comparison pipeline.
Raw data are preprocessed to estimate the reliability of the available protein interactions and identify groups of sequence-similar proteins. A protein group contains one protein from each species and requires that each protein has a significant sequence match to at least one other protein in the group (BLAST E-value $<10^{-7}$; considering the ten best matches only). Next, protein networks are combined to produce a *network alignment,* which connects protein similarity groups whenever the two proteins within each species directly interact or are connected via a common network neighbor. Conserved paths and clusters identified within the network alignment are compared to those computed from randomized data, and those at a significance level of $p<0.01$ are retained. A final filtering step removes paths and clusters with greater than 80% overlap.

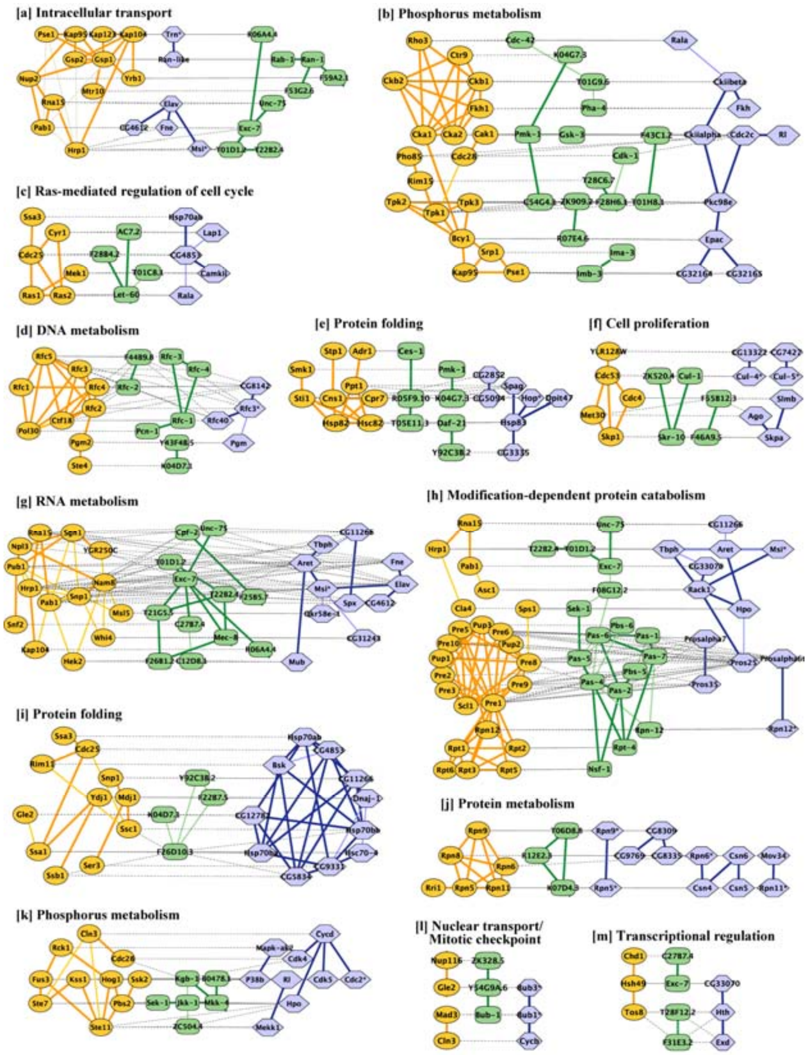


Figure 3.2: Representative conserved network regions.

Shown are conserved clusters **[a-k]** and paths **[l-m]** identified within the networks of yeast, worm and fly. Each region contains one or more overlapping clusters or paths (see Figure 3.3). Proteins from yeast (orange ovals), worm (green rectangles) or fly (blue hexagons) are connected by direct (thick link) or indirect (connection via a common network neighbor; thin link) protein interactions. Horizontal dotted gray links indicate cross-species sequence similarity between proteins (similar proteins are typically placed on the same row of the alignment). Automated layout of network alignments was performed using a specialized plug-in to the Cytoscape software[25] as described in Methods.

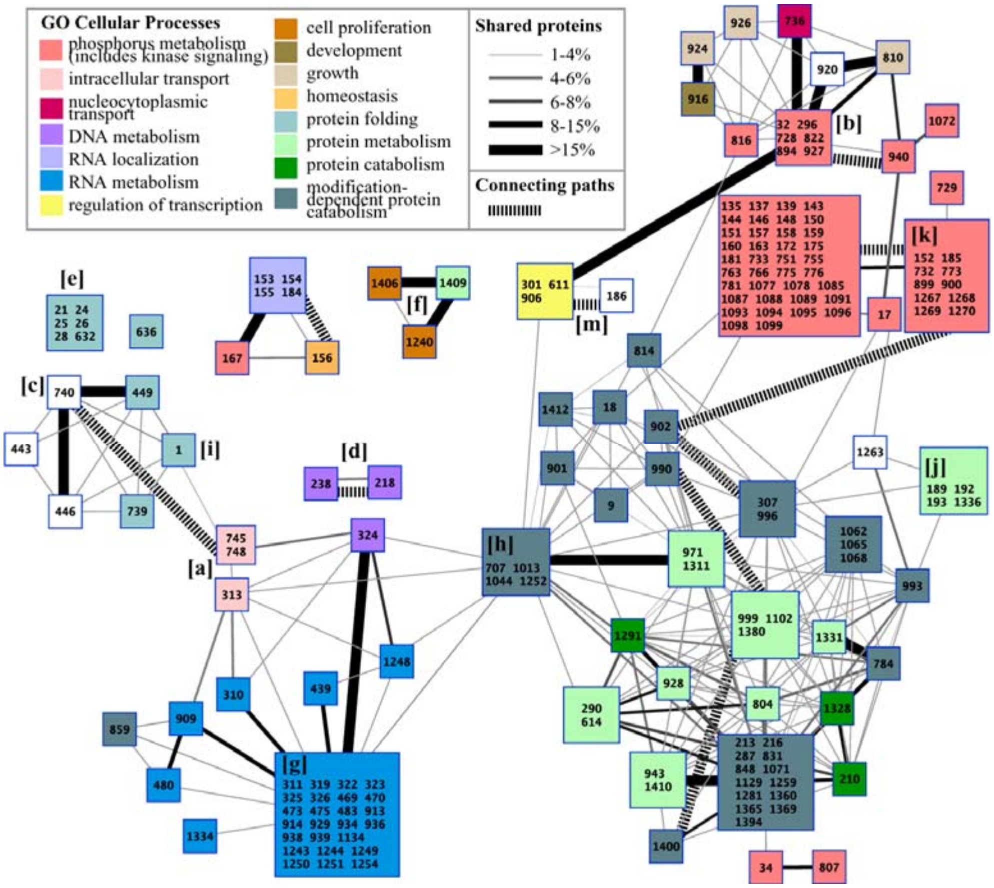
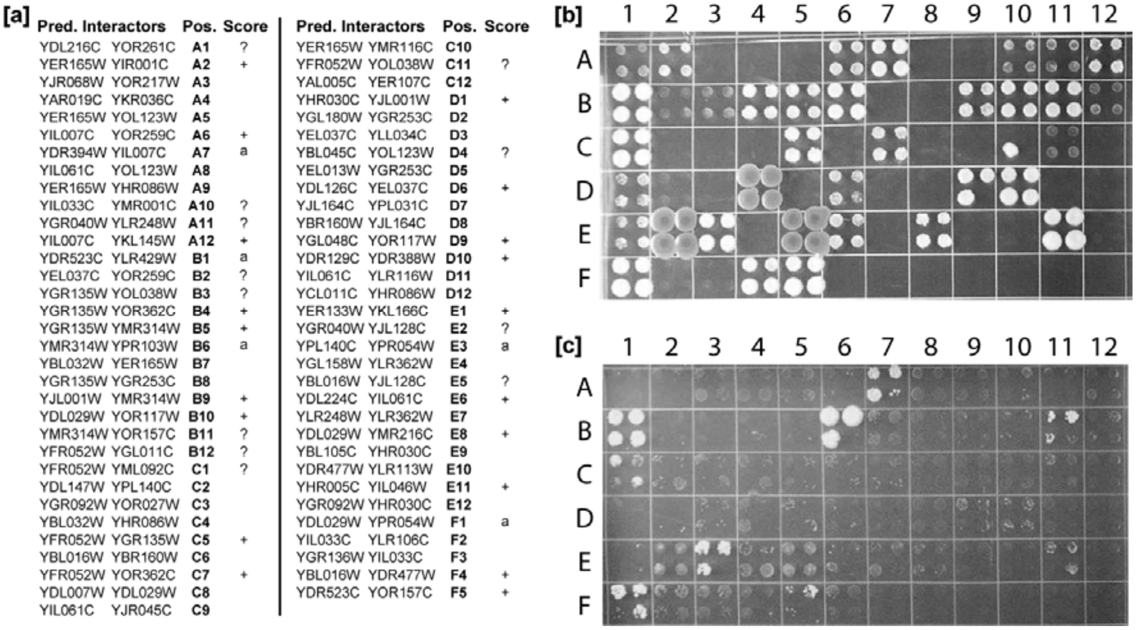


Figure 3.3: Modular structure of conserved clusters among yeast, worm and fly.

Multiple network alignment revealed 183 conserved clusters, organized into 71 network regions represented by colored squares. Regions group together clusters that share >15% overlap with at least one other cluster in the group and are all enriched for the same GO cellular process ($p<0.05$ with the enriched processes indicated by color). Cluster ID numbers are given within each square; numbers are not sequential due to filtering. Solid links indicate overlaps between different regions, where thickness is proportional to the percentage of shared proteins (intersection / union). Hashed links indicate conserved paths that connect clusters together. Labels [a-k],[m] mark the network regions exemplified in Figure 3.2.



Pred. Interactors		Pos.	Score
YDL216C	YOR261C	A1	?
YER165W	YIR001C	A2	+
YJR068W	YOR217W	A3	
YAR019C	YKR036C	A4	
YER165W	YOL123W	A5	
YIL007C	YOR259C	A6	+
YDR394W	YIL007C	A7	a
YIL061C	YOL123W	A8	
YER165W	YHR086W	A9	
YIL033C	YMR001C	A10	?
YGR040W	YLR248W	A11	?
YIL007C	YKL145W	A12	+
YDR523C	YLR429W	B1	a
YEL037C	YOR259C	B2	?
YGR135W	YOL038W	B3	?
YGR135W	YOR362C	B4	+
YGR135W	YMR314W	B5	+
YMR314W	YPR103W	B6	a
YBL032W	YER165W	B7	
YGR135W	YGR253C	B8	
YJL001W	YMR314W	B9	+
YDL029W	YOR117W	B10	+
YMR314W	YOR157C	B11	?
YFR052W	YGL011C	B12	?
YFR052W	YML092C	C1	?
YDL147W	YPL140C	C2	
YGR092W	YOR027W	C3	
YBL032W	YHR086W	C4	
YFR052W	YGR135W	C5	+
YBL016W	YBR160W	C6	
YFR052W	YOR362C	C7	+
YDL007W	YDL029W	C8	
YIL061C	YJR045C	C9	
YER165W	YMR116C	C10	
YFR052W	YOL038W	C11	?
YAL005C	YER107C	C12	
YHR030C	YJL001W	D1	+
YGL180W	YGR253C	D2	
YEL037C	YLL034C	D3	
YBL045C	YOL123W	D4	?
YEL013W	YGR253C	D5	
YDL126C	YEL037C	D6	+
YJL164C	YPL031C	D7	
YBR160W	YJL164C	D8	
YGL048C	YOR117W	D9	+
YDR129C	YDR388W	D10	+
YIL061C	YLR116W	D11	
YCL011C	YHR086W	D12	
YER133W	YKL166C	E1	+
YGR040W	YJL128C	E2	?
YPL140C	YPR054W	E3	a
YGL158W	YLR362W	E4	
YBL016W	YJL128C	E5	?
YDL224C	YIL061C	E6	+
YLR248W	YLR362W	E7	
YDL029W	YMR216C	E8	+
YBL105C	YHR030C	E9	
YDR477W	YLR113W	E10	
YHR005C	YIL046W	E11	+
YGR092W	YHR030C	E12	
YDL029W	YPR054W	F1	a
YIL033C	YLR106C	F2	
YGR136W	YIL033C	F3	
YBL016W	YDR477W	F4	+
YDR523C	YOR157C	F5	+

Figure 3.4: Verification of predicted interactions by two-hybrid testing. **[a]** 65 pairs of yeast proteins were tested for physical interaction based on their co-occurrence within the same conserved cluster and the presence of orthologous interactions in worm and fly. Each protein pair is listed along with its position on the agar plates shown in [b] and [c] and the outcome of the two-hybrid test. **[b]** Raw test results are shown, with each protein pair tested in quadruplicate to ensure reproducibility. Protein 1 vs. 2 of each pair was used as prey vs. bait, respectively. **[c]** This negative control reveals activating baits, which can lead to positive tests without interaction. Protein 2 of each pair was used as bait with an empty pOAD vector as prey. Activating baits are denoted by "a" in the list of predictions shown in [a]. Positive tests with weak signal (e.g., A1) and control colonies with marginal activation are denoted by "?" in the list; colonies D4, E2 and E5 show evidence of possible contamination and are also marked by a "?". Discarding the activating baits, 31 out of 60 predictions tested positive overall. A more conservative tally, disregarding all results marked by a "?", yields 19 out of 48 positive predictions.

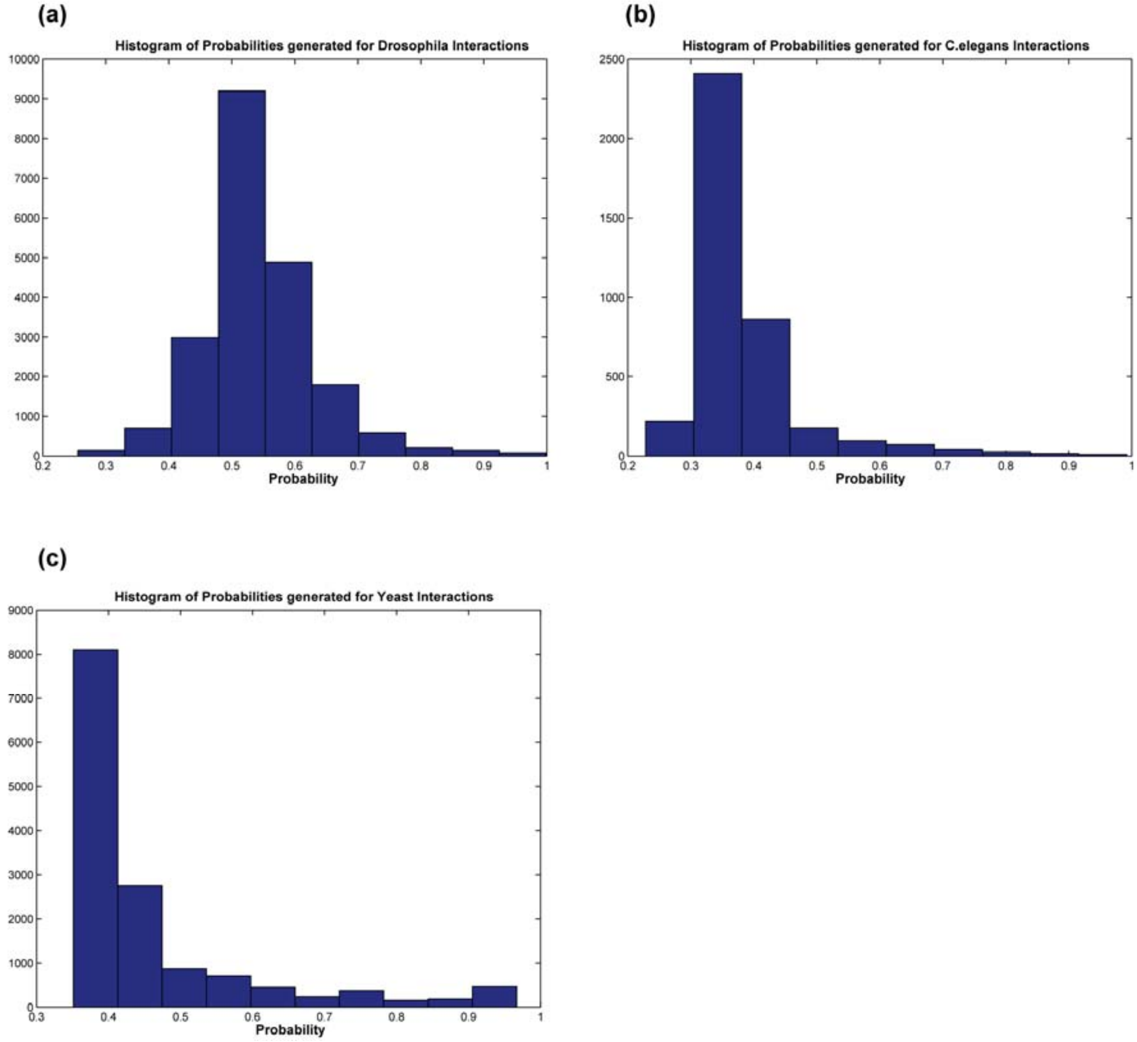


Figure 3.5: Histogram of probabilities assigned to experimentally observed interactions.
(a) fly, **(b)** worm and **(c)** yeast.

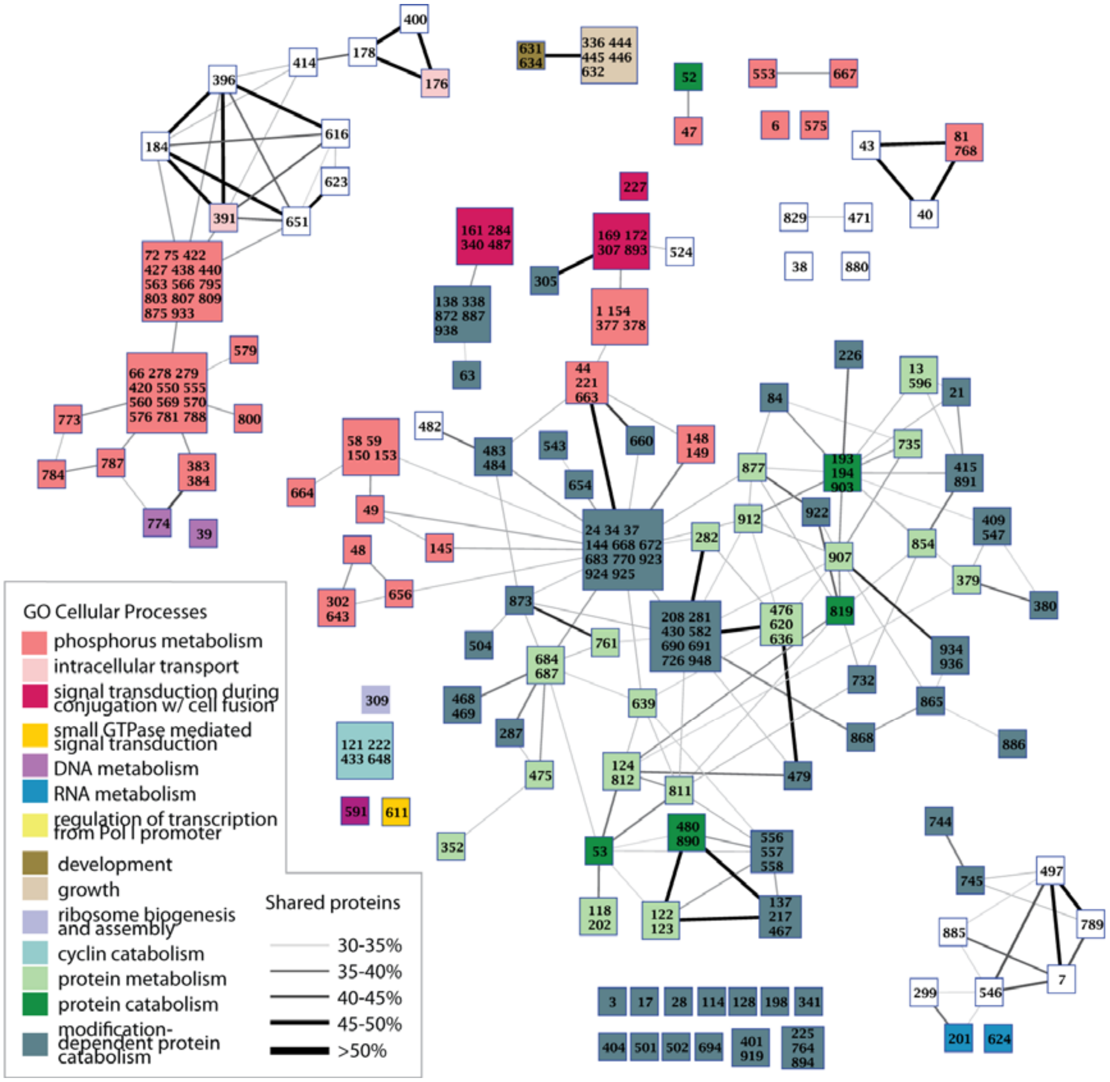


Figure 3.6: Modular structure of conserved complexes between yeast and worm.

The overview graph represents 220 conserved protein complexes. Each link indicates an overlap between complexes, where thickness is proportional to the percentage of shared proteins (Jaccard measure of intersection over union). Colors highlight complexes that are significantly enriched for proteins involved in the same GO cellular process (*p*<0.05, corrected for multiple testing). Complexes grouped into a single square share > 15% overlap with at least one other complex in the group, and are all of the same significant cellular process.

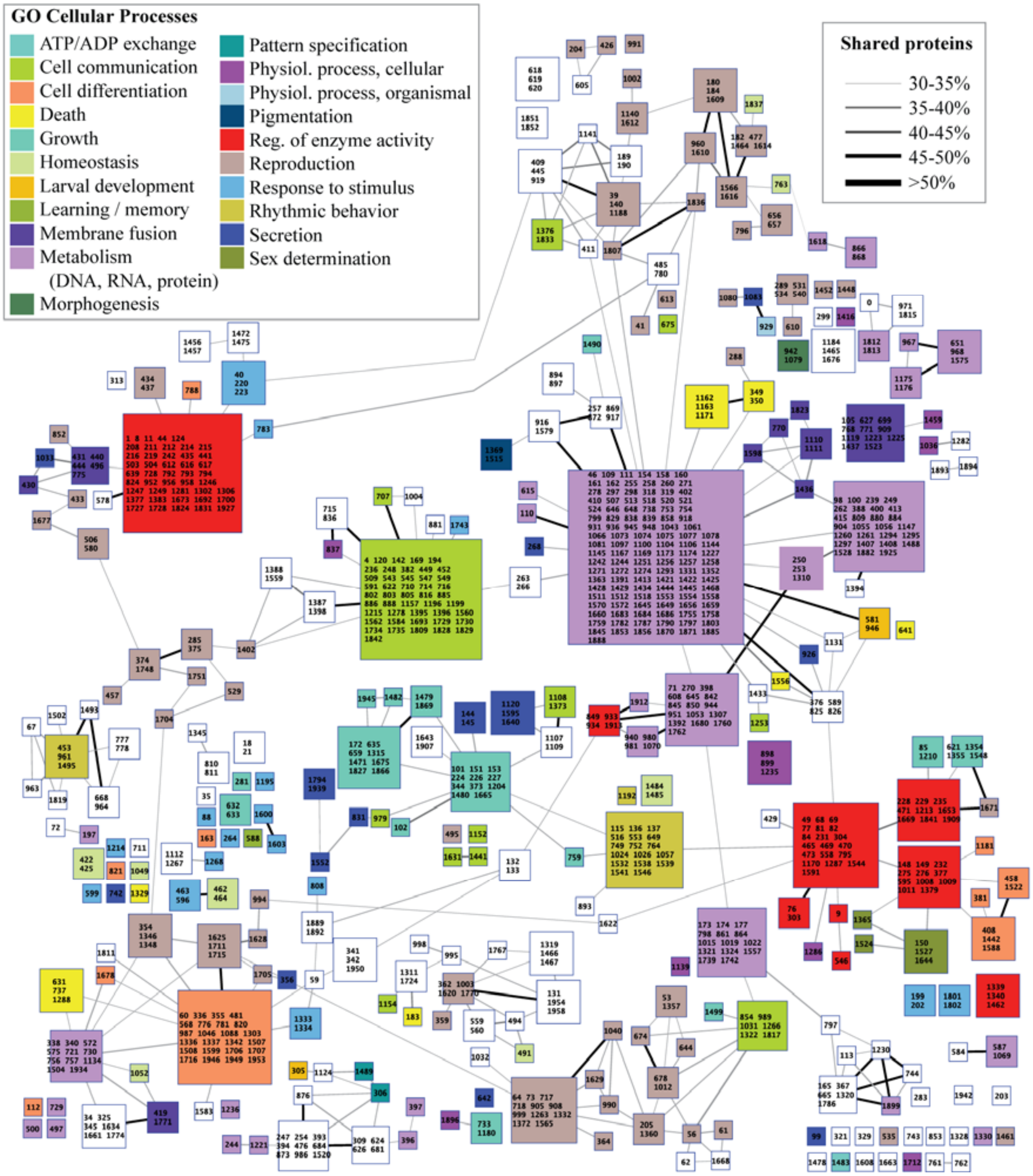


Figure 3.7: Modular structure of conserved protein complexes among yeast and fly.

The overview graph represents 835 conserved protein complexes.

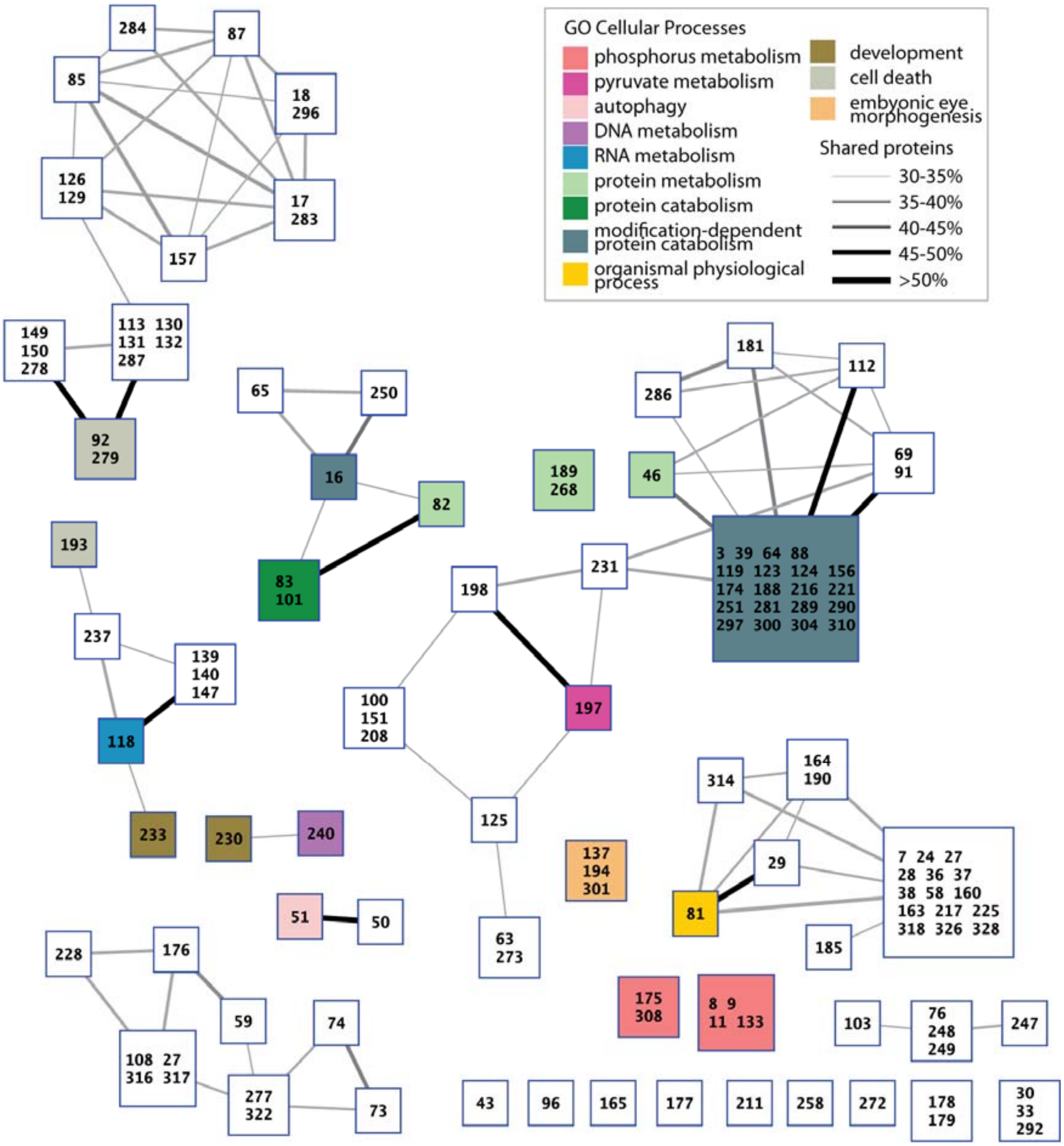


Figure 3.8: Modular structure of conserved protein complexes among worm and fly.

The overview graph represents 132 conserved protein complexes.

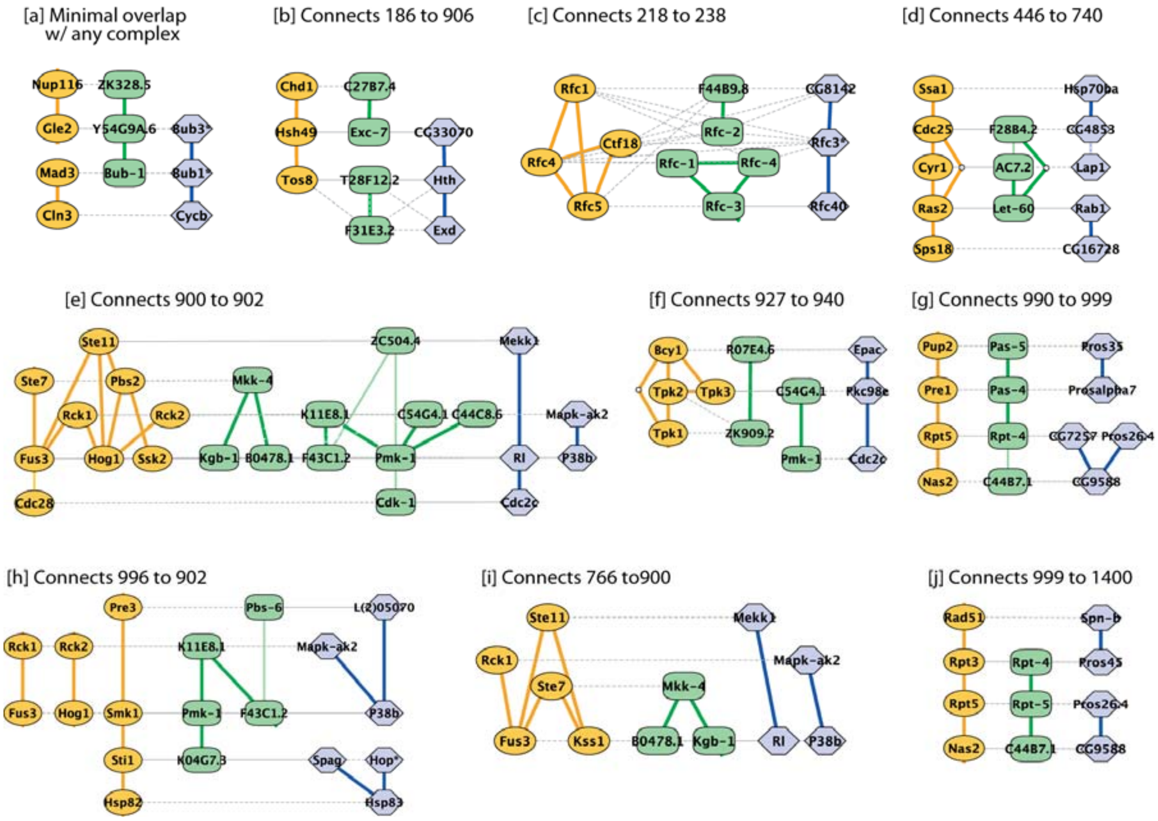


Figure 3.9: Conserved pathways among yeast, worm and fly.
Proteins from yeast (orange ovals), worm (green rectangles) or fly (blue hexagons) are connected by links representing protein-protein interactions. Dotted gray links indicate sequence similarity.

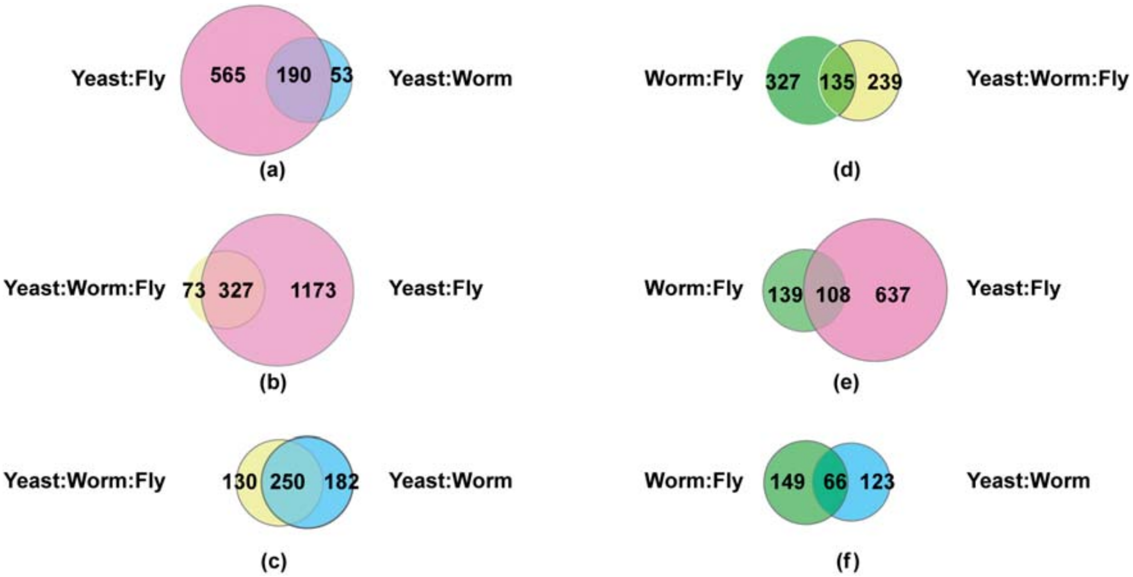


Figure 3.10: Comparison of 2-way anf 3-way complexes.
Shows are Venn diagrams depicting the relationships between the computed 2-way and 3-way complexes in terms of the number of distinct protein that are included in each set of complexes.

4 Comparative network analysis of the protein network of the malaria parasite *Plasmodium falciparum*

With the recent accumulation of protein interactions in public databases, cross-species comparisons are becoming critical for analyzing the large networks formed by these interactions to delineate protein function and evolution[88]. At a fundamental level, protein networks can be compared to identify "interologs", i.e. interactions that are conserved across species[60]. Beyond comparison of interactions individually, methods such as PathBLAST[28,54] create a global alignment between two protein networks to identify dense clusters of conserved interactions, suggestive of protein complexes. Such comparative approaches are important because they can tease conserved components of cellular machinery out of a highly connected network and increase overall confidence in the underlying interaction measurements.

Plasmodium falciparum is the pathogen responsible for over 90% of human deaths from malaria[89]. As such, it has been the focus of a major research initiative, involving complete DNA sequencing of the genome[90], large-scale expression analyses[91,92], and protein characterization of its lifecycle stages[93]. The *Plasmodium* genome sequence is relatively distant from those of most other eukaryotes, with more than 60% of the 5,334 encoded proteins lacking significant sequence similarity to other organisms[90]. To systematically elucidate functional relationships among these proteins, a large two-hybrid study has recently mapped a network of 2,847 interactions involving 1,312 proteins in *Plasmodium*[19]. This network adds to a growing collection of available interaction maps and raises questions about whether the divergence of *Plasmodium* at the sequence level is reflected in the configuration of its protein network. Here, we examine conserved structures between the *Plasmodium* protein network and those of model organisms and show that its patterns of interaction, like its genome sequence, set it apart from other species.

4.1 Cross-species comparison of the Plasmodium protein network

4.1.1 Data sources

We assembled protein-protein interaction networks for *Plasmodium falciparum*[19], the budding yeast *Saccharomyces cerevisiae*[37], the nematode worm *Caenorhabditis elegans*[40], the fruit fly *Drosophila melanogaster*[39], and the bacterial pathogen *Helicobacter pylori*[70] in the context of the Cytoscape network visualization and modeling environment[25]. Annotation and amino-acid sequence of each interacting protein in *Plasmodium* was obtained from PlasmoDB[94]. To obtain data for other species, we downloaded interactions from the Database of Interacting Proteins (DIP)[37] as of December 2004. The yeast interactions were attributed to a combination of two-hybrid studies[18,20], co-immunoprecipitation studies[12,17], and classical small-scale experiments. Interaction sets for worm, fly, and bacteria were each drawn from single two-hybrid studies[39,40,70]. Corresponding protein sequences were obtained from the *Saccharomyces* Genome Database[67], WormBase[95], FlyBase[81], or The Institute for Genomic Research (TIGR)[96], respectively.

4.1.2 Comparative analysis of the Plasmodium PPI network

We compared the protein-protein interaction network of *Plasmodium* reported by LaCount *et al*. to protein networks for the budding yeast *Saccharomyces cerevisiae*, the nematode worm *Caenorhabditis elegans*, the fruit fly *Drosophila melanogaster*, and the bacterial pathogen *Helicobacter pylori*. Surprisingly, pair-wise alignment of these networks using NetworkBLAST[54] revealed that *Plasmodium* had only three conserved complexes with yeast

(Figures 4.1a-c) and had none with any other species. However, yeast, fly, and worm shared substantial numbers of conserved complexes with each other (Figure 4.2a). For instance, yeast and fly had the highest degree of conservation with 61 conserved complexes.

The relatively low similarity between the *Plasmodium* network and those of other eukaryotes suggested that it encodes important functional differences worthy of further investigation. As an alternative explanation, it was possible that differences in the number of complexes were related to network size. Thus, in addition to searching for conserved complexes, we investigated whether the observed similarities and differences were reflected in the probability of conservation of each protein interaction individually (Figure 4.2b). For each pair of species, a protein-protein interaction was considered "conserved" if both proteins had homologs that interacted in the opposite species (BLAST E-value $\leq 1\times10^{-4}$, normalized for genome size). A global pair-wise similarity metric was then defined as the overall fraction of interactions that were conserved, restricted to proteins with at least one homolog in the opposite species.

Figure 4.2c expresses the pair-wise interaction similarities as a phylogenetic tree drawn using the method of Kitsch[97]. This tree was relatively robust to sampling errors as determined by bootstrap analysis: 86.2% of trials placed *Plasmodium* as an outgroup relative to yeast, worm, and fly. Among the three model eukaryotes, yeast and worm were closest based on interaction similarity (Figure 4.2b) while yeast and fly were closest based on conserved complexes (Figure 4.2a). This discrepancy was likely due to network size or coverage. Nonetheless, the particular phylogenetic placement of *Plasmodium* was consistent across both analyses and with the accepted taxonomical relationships among these species as established by morphological and sequence comparisons[90].

A second possibility for the low similarity of the *Plasmodium* network to other species was that its interaction network had been measured predominantly among proteins expressed in the asexual stages of the parasite (see LaCount *et al.*[19]). There are two ways in which this sampling could affect network similarity. First, it was possible that a high (low) level of mRNA expression increases (decreases) the number of interactions identified for the corresponding proteins and thus alters the topology of the *Plasmodium* network relative to other species. To investigate the potential relationship between protein interactions and mRNA expression, we plotted the number of interactions of each protein in the *Plasmodium* network as a function of its mRNA expression level. *Plasmodium* genome-wide expression data were obtained from Le Roch *et al.*[92] which includes experiments from the erythrocytic (asexual) stages as well as the mosquito salivary-gland sporozoite stage and the sexual gametocyte stage. As shown in Figure 4.5, we found no relation between the number of protein interactions and expression level (in any stage). Thus, the bias in protein sampling does not appear to affect the specific topology of the network. Second, it was possible that proteins from asexual stages tended to have lower similarity across species than did other stages of the *Plasmodium* life cycle. In fact, we found that the *Plasmodium* interaction set was enriched for proteins with homologs in other species and that all five protein interaction networks were enriched for yeast homologs in particular (Table 4.1 and 4.2). Such enrichment was observed even in worm, for which baits were explicitly selected to be non-homologous to yeast[40]. This effect is in need of further study, but might indicate a bias of the yeast two-hybrid system in measuring interactions among yeast homologs, since all two-hybrid constructs must be expressible in the yeast cell.

A final possibility was that the *Plasmodium* network might have a substantially higher proportion of false-positive interactions relative to the networks of yeast, fly, and worm. Lacking a "gold standard" set of true interactions, we characterized the relative quality of the *Plasmodium* network by examining: (1) its global topological properties, and (2) the signal-to-noise ratio of its protein complexes. Several common topological measures[98] were computed on each network, including the average number of interactions per protein (average degree), the average shortest path length between proteins, and the average clustering coefficient (Table 4.1). The number of interactions per protein in the *Plasmodium* interaction network followed a scale-free distribution, similar to other networks (Figure 4.3a). Moreover, the *Plasmodium* network was never the outlier in any of the various measurements, suggesting that its global organization was consistent with the others.

Next, we applied the PathBLAST procedure to identify dense interaction complexes within each species independently. A total of 29 single-species complexes were identified for *Plasmodium*, three of which are shown in Figures 4.1d-f. This number was the median of the range observed over the five species (Table 4.1). Single-species complexes were used to assess the overall quality of each network by computing their signal-to-noise ratio (SNR), a standard measure from information theory and signal processing[68]. SNR was computed by comparing the scores of complexes identified in the observed versus random interaction data for each species (see Methods). *Plasmodium*, worm, and fly had very similar SNR values (Figure 4.3b), while SNR of the yeast network was slightly higher and that of the *H. pylori* network slightly lower. The network distances of *Plasmodium* versus yeast, worm, or fly do not appear to depend on SNR.

4.1.3 *Analysis of conserved and distinct complexes in Plasmodium*

Analysis of the three conserved and 29 *Plasmodium*-specific protein complexes suggests new functional predictions for *Plasmodium* proteins. For instance, the conserved protein complex shown in Figure 4.1a predicts that the proteins PF10_0244 and MAL6P1.286 may play previously uncharacterized roles in endocytosis. The counterpart of PF10_0244 in the yeast network, Ede1, localizes to the cortical patch[99] of the cell membrane at sites of polarized growth and appears to be involved in endocytosis[100]. Myo5 and Myo3, yeast counterparts of MAL6P1.286, are class I myosins that also localize to actin cortical patches[101] where the calmodulin protein Cmd1 has been implicated in the uptake step of receptor-mediated endocytosis[102]. Taken together, this evidence suggests a role for this complex in calmodulin-mediated endocytosis. Calmodulin inhibitors have been shown to attenuate growth[103] and chloroquine extrusion (effecting drug resistance)[104] in malarial parasites, and endocytosis has recently been linked to the mechanism of anti-malarial drugs including chloroquine and artemisin[105]. The proximity of calmodulin to the formation of endocytic vacuoles in *Plasmodium* provides for a discrete hypothesis linking endocytosis, drug resistance, and drug mechanism of action.

The conserved complexes shown in Figures 4.1b,c contain yeast proteins involved in the unfolded protein response (UPR) pathway in the endoplasmic reticulum, which is linked to increased chaperone production, proteosomal degradation and specific gene expression changes[106]. The proteins Rpt1-5 comprise the regulatory subunit of the proteasome (Figure 4.1b). Rpt3 interacts with Lhs1, which is regulated by the UPR pathway[107]. These proteins are connected to a mesh of mitogen-activated (MAP) and serine/threonine kinases associated with maintenance of cell wall integrity; it is possible that these kinases also transmit signals to the mini-chromosome maintenance (MCM) complex as part of the UPR. Interestingly, the MCM complex links many of the same kinases as the UPR in both species. Whether these connections are coordinated with or independent of the unfolded protein response remains to be investigated.

Within the 29 *Plasmodium*-specific complexes, chromatin remodeling was a prominent function, as shown in Figure 4.1e. This complex involves the chromatin-remodeling protein ISWI (MAL6P1.183) interacting with a nucleosome assembly protein (PFI0930c)[108]. PF11_0429 has a PHD domain and PFL0130c has an HMG domain, both postulated to be involved in the remodeling process[109]. Together, these known functions suggest that other proteins in the complex, such as PF08_0060, PFB0765w, and PFL0625c, also participate in chromatin remodeling. For instance, although PFL0625c is annotated as a translation initiation factor, its yeast homolog has been found in complex with histone acetyltransferases[12]. Of the 29 complexes distinct to *Plasmodium,* three have the further distinction that the majority of their proteins have no homologs in human or yeast (at a BLAST E-value $< 1\times10^{-2}$). One such example is shown in Figure 4.1f. Six of the 13 proteins in this complex have predicted trans-membrane domains[56]. PF14_0678 is a 35 kDa exported protein located at the membrane of the parasitophorous vacuole of the infected erythrocyte[110]. The remaining proteins in this complex are unannotated due to lack of homology with other organisms. The complex in Figure 4.1d suggests a link between translation (several translation initiation factors and ribosomal subunits) and exported proteins involved in cell invasion. The latter proteins include MSP1, MSP9, several rhoptry proteins and antigen 332, associated with cytoadherence[111]. MSP9 (PFL1385c) is central to this complex.

An important question regarding the 29 *Plasmodium*-specific complexes is whether these complexes are truly unique to the pathogen or, alternatively, scored just below the significance threshold in other species despite having homologous proteins and protein interactions. To investigate this question, Table 4.5 lists the number of *Plasmodium* proteins covered by the *Plasmodium*-specific complexes that had homologs in yeast, fly and/or worm. Also listed are the number of interactions within each complex that are conserved across species (BLAST E-value $\leq 1\times10^{-4}$). From the table, it is apparent that although the complexes unique to Plasmodium have a number of proteins with homologs across species, these homologs have very few interactions conserved. Hence, we conclude that these complexes are not seen in yeast, worm or fly, at least in the interaction networks that are currently available. It is not the case that these complexes scored just below a threshold cutoff.

4.1.4 Distribution of GO Cellular Components across species.

Several cellular components that we expected to be present, such as the proteasome, were missing from the set of complexes conserved between

Plasmodium and other species. To investigate this issue, we plotted the distributions of known functional annotations (Gene Ontology Cellular Component Level Three)[56] among *Plasmodium* proteins, protein interactions, and conserved interactions (Figure 4.4 and 4.6; note that a protein or an interaction can participate in multiple categories). Considerable fractions of all three datasets were associated with intracellular organelles, membrane-bound organelles, or the cytoplasm (Figure 4.4a). Other cellular components, such as the membrane and extra-organismal space, were represented among proteins and interactions but to a lesser extent among conserved interactions (Figure 4.4b). Many membrane-associated components were also reported in the 29 *Plasmodium*-specific complexes and are suggestive of machinery unique to the organism. Finally, components such as the proteasome and cytoskeleton were represented among proteins but were absent from the interaction set and hence not found as conserved interactions or complexes (Figure 4.4c). Interactions among proteins in these components may have yet to be uncovered. These observations are reinforced by a complementary analysis of the functional distributions of yeast, worm, and fly (Figure 4.6). For instance, a set of interactions among membrane proteins is found in all networks, and this set is typically conserved across yeast/worm/fly, but the membrane interactions set in *Plasmodium* shows no homology to other species (compare blue to red bars). Extra-organismal proteins and protein interactions are much more abundant in *Plasmodium* than in other species (especially yeast and fly; a few are conserved between *Plasmodium* and worm). Accordingly, the functional categories listed in Figure 4.4b are of interest as potentially containing protein interactions that are unique to the pathogen, especially considering that many of the proteins known to participate in pathogenesis and cellular invasion come from these categories. Categories in Figure 4.4c were represented in the *Plasmodium* genome but generally absent from its interaction network, indicating possible false negatives. In Figure 4.6, we can see that some of these categories, such as proteasome and cytoskeleton, are in fact well represented in the protein networks of yeast, worm, and fly. On other hand, the 43S preinitiation complex is absent not only from the network of *Plasmodium*, but the other three networks also. This functional complex may have therefore been consistently missed by two-hybrid experiments.

4.2 Discussion

In summary, we have characterized conserved patterns of interaction between the network of *Plasmodium falciparum* and those of other species and reported the specific network regions that are conserved. All of the examined networks contain dense complex-like structures of interactions, some of which are shared by yeast, worm, and fly but not *Plasmodium*. These relationships are not clearly related to noise or bias in the *Plasmodium* interaction set. Some of the observed differences are almost certainly due to incomplete coverage in one or more networks: for instance, the present *Plasmodium* interaction set is focused on the asexual lifecycle stages. Nevertheless, our comparison reflects the relative degree of similarity between the different networks. These differences are observed even when considering only those genes that are homologous across species.

It is generally expected that conserved genes retain their functions and interactions. From this comparison, a different principle emerges: conservation of specific groups of related genes does not necessarily imply conservation of interaction among those genes. Further studies may distinguish the true differences from those related to network coverage and, ultimately, direct new pharmaceuticals to the protein complexes unique to this parasite.

4.3 Methods

4.3.1 Identification of conserved and species only complexes

Identification of protein complexes was performed using the PathBLAST family of network alignment tools, as previously described[54]. Briefly, these methods integrate protein interaction data from two species with protein sequence homology to generate an "aligned network", in which each node represents a pair of homologous proteins (one from each species) and each link represents a conserved interaction. The network alignment is searched to identify high-scoring subnetworks, for which the score is based on the density of interactions within the subnetwork as well as confidence estimates for each protein interaction (see below). The search is then repeated over 100 random trials, in which the interactions of both species are arbitrarily reassigned while maintaining the same number of interactions per protein, resulting in a distribution of random subnetwork scores pooled over all trials. Dense subnetworks that score in the top fifth percentile of this random score distribution are considered significant and reported as "conserved complexes." The search for "single-species complexes" is identical to the search for conserved complexes except that an individual protein network is searched instead of the network alignment. This process identifies dense subnetworks constrained by the interactions of one species rather than two.

4.3.2 Interaction confidence scores

We estimated the probability that each measured protein interaction is true using a logistic regression model based on mRNA expression correlation, the network cluster coefficient, and the number of times the interaction had been experimentally observed (see Chapter 2). For yeast, worm and fly, mRNA expression data was obtained from the Stanford Array Database[61] as of 5/01/2004. Expression correlation among *P. falciparum* genes was estimated from 48 arrays of mRNA expression collected across the different lifecycle stages by Bozdech *et al.*[91].

4.3.3 Phylogenetic tree construction

The Kitsch algorithm (provided by the PHYLIP package[97] assumes the presence of an evolutionary clock and is based on pairwise distances between species, which were computed as follows. For each pair of species, an interaction between proteins *a* and *b* was considered "conserved" if both proteins had sequence-similar counterparts a' and b' (BLAST E-value $< 1\times10^{-4}$) that interacted in the opposite species. A pairwise similarity between networks was computed as $s_{12} = (c_1+c_2) / (t_1+t_2)$, where c is the number of conserved interactions and t is the total number of interactions in species 1 or 2, respectively (with all interactions restricted to the set of proteins with homologs in the opposite species). Pairwise network distance was then defined as $1 - s_{12}$. The resulting phylogenetic tree reported in Figure 4.2c is the consensus over 10,000 bootstrap simulations. Values of c and t for each network are listed in Table 4.3.

4.3.4 SNR of protein complexes

Signal to noise ratio (SNR) is a standard measure used in information theory and signal processing to assess data quality. We compute SNR of the single-species complexes as follows. The search for dense interaction complexes is initiated from each node (protein) and the highest scoring complex from each is reported (see the PathBLAST section above). This yields a distribution of complex scores over all nodes in the network. A score distribution is also generated for 100 randomized networks which have identical degree distribution as the original network. SNR ratio is computed from these original and random score distributions (representing signal and noise, respectively) according to the standard formula[68] using the root-mean-square (rms):

$$\text{SNR} = \log_{10} \frac{\text{rms(original complex scores)}}{\text{rms(random complex scores)}}, \quad \text{with rms}(x) = \sqrt{\frac{1}{M}\sum_{i=1}^{M} x_i^2}$$

and where x_i is the score of a complex and M is the total number of complexes.

4.3.5 Simulating false-positives and negatives in protein networks

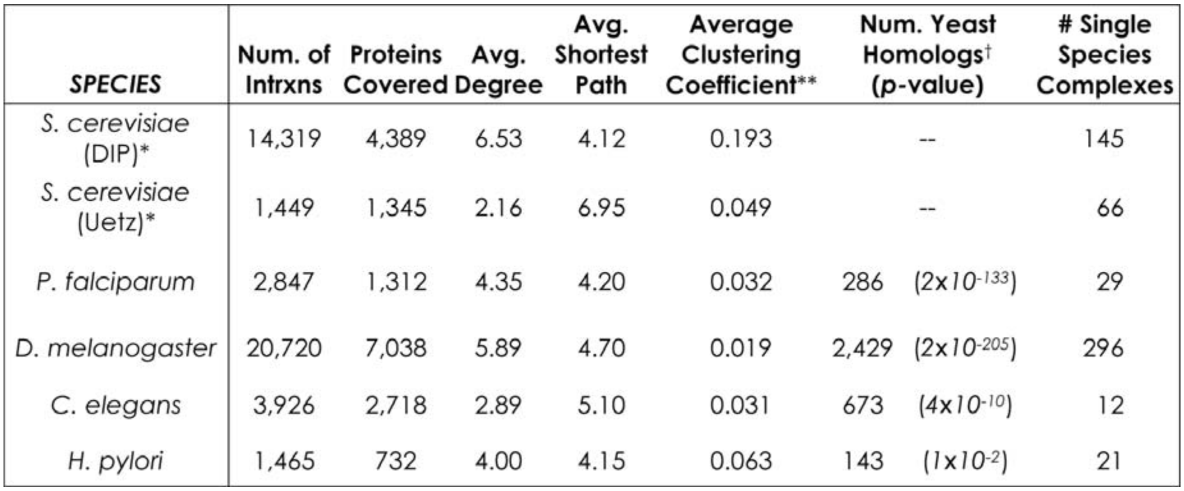
The percentage of false positives and false negatives in each interaction network was increased (Figure 4.3b x-axis) by randomizing the interactions in the network, keeping the degree distribution fixed. At each iteration, two interactions were selected (at random, say 'a-b' and 'A-B') and their targets exchanged, creating new interactions ('a-B' and 'A-b'). The shuffling was performed only if the newly created interactions did not already exist in the original network. When choosing an interaction partner at random (as above), there is a far greater chance that the resulting unobserved interaction does not occur *in vivo* (false positive) than vice versa (true positive). Therefore, each time a "true" edge is moved during randomization, the shuffling process replaces the true edge with a false negative and, the vast majority of the time, creates a false positive edge in its place. This process of network randomization simultaneously introduces both types of errors (conversely, reassigning a "false" edge has little effect on either measure).

Decay of certain global properties and signal to noise ratio (SNR)[68] was recorded during this process. We calculated the average clustering coefficient[98], and the overlap of the data set with previously established domain interactions[112]. Domain overlap was calculated as the fraction of interactions whose interacting proteins had domains that interact as defined in the Pfam database. These measurements are shown in Table 4.4.

Acknowledgements

This chapter contains the complete reprint of the work Suthram S, Sittler T, Ideker T. The Plasmodium protein network diverges from those of other eukaryotes. *Nature* 2005;438:108-12. I was the primary co-author of this paper.

Table 4.1: Topological properties of protein interaction networks.



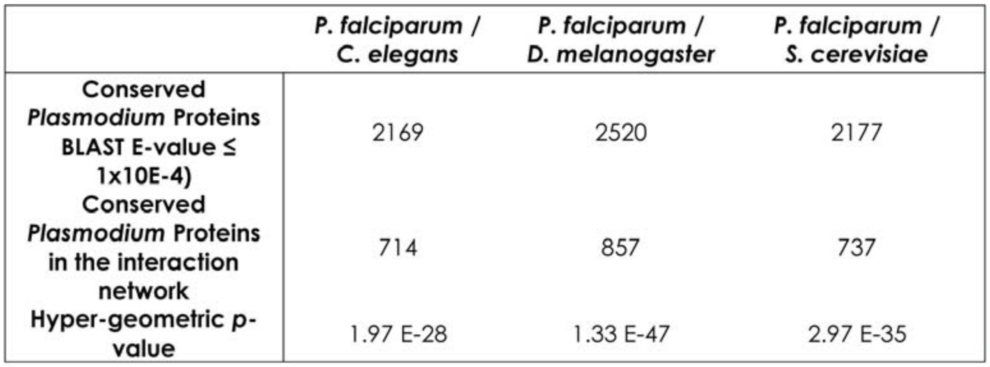
SPECIES	Num. of Intrxns	Proteins Covered	Avg. Degree	Avg. Shortest Path	Average Clustering Coefficient**	Num. Yeast Homologs† (*p*-value)	# Single Species Complexes
S. cerevisiae (DIP)*	14,319	4,389	6.53	4.12	0.193	--	145
S. cerevisiae (Uetz)*	1,449	1,345	2.16	6.95	0.049	--	66
P. falciparum	2,847	1,312	4.35	4.20	0.032	286 (2x10^{-133})	29
D. melanogaster	20,720	7,038	5.89	4.70	0.019	2,429 (2x10^{-205})	296
C. elegans	3,926	2,718	2.89	5.10	0.031	673 (4x10^{-10})	12
H. pylori	1,465	732	4.00	4.15	0.063	143 (1x10^{-2})	21

*Unlike other networks which are generated from single two-hybrid studies, the network of yeast interactions in DIP[37] consists of many experiments and experimental types. A separate analysis is included considering only data from a single two-hybrid screen by Uetz *et al*[20].

**The clustering coefficient measures local density of the network around a protein and is computed as previously described[98].

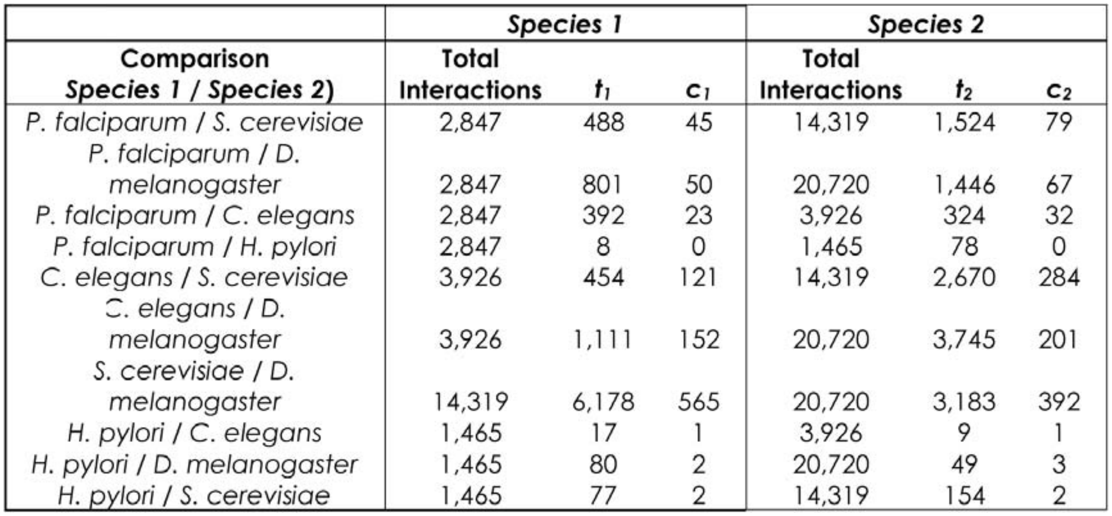
†Yeast homologs are determined using a conservative BLAST E-value threshold of ≤1E-10. The *p*-values score the significance of enrichment for yeast homologs within the set of proteins covered by each interaction network, using the hypergeometric test. These enrichments are significant over a broad range of E-value thresholds (data not shown).

Table 4.2: The Plasmodium network is enriched for proteins that have homologs in other species.



	P. falciparum / C. elegans	*P. falciparum / D. melanogaster*	*P. falciparum / S. cerevisiae*
Conserved *Plasmodium* Proteins BLAST E-value ≤ 1x10E-4)	2169	2520	2177
Conserved *Plasmodium* Proteins in the interaction network	714	857	737
Hyper-geometric *p*-value	1.97 E-28	1.33 E-47	2.97 E-35

Table 4.3: Comparison of number of conserved interactions across species.

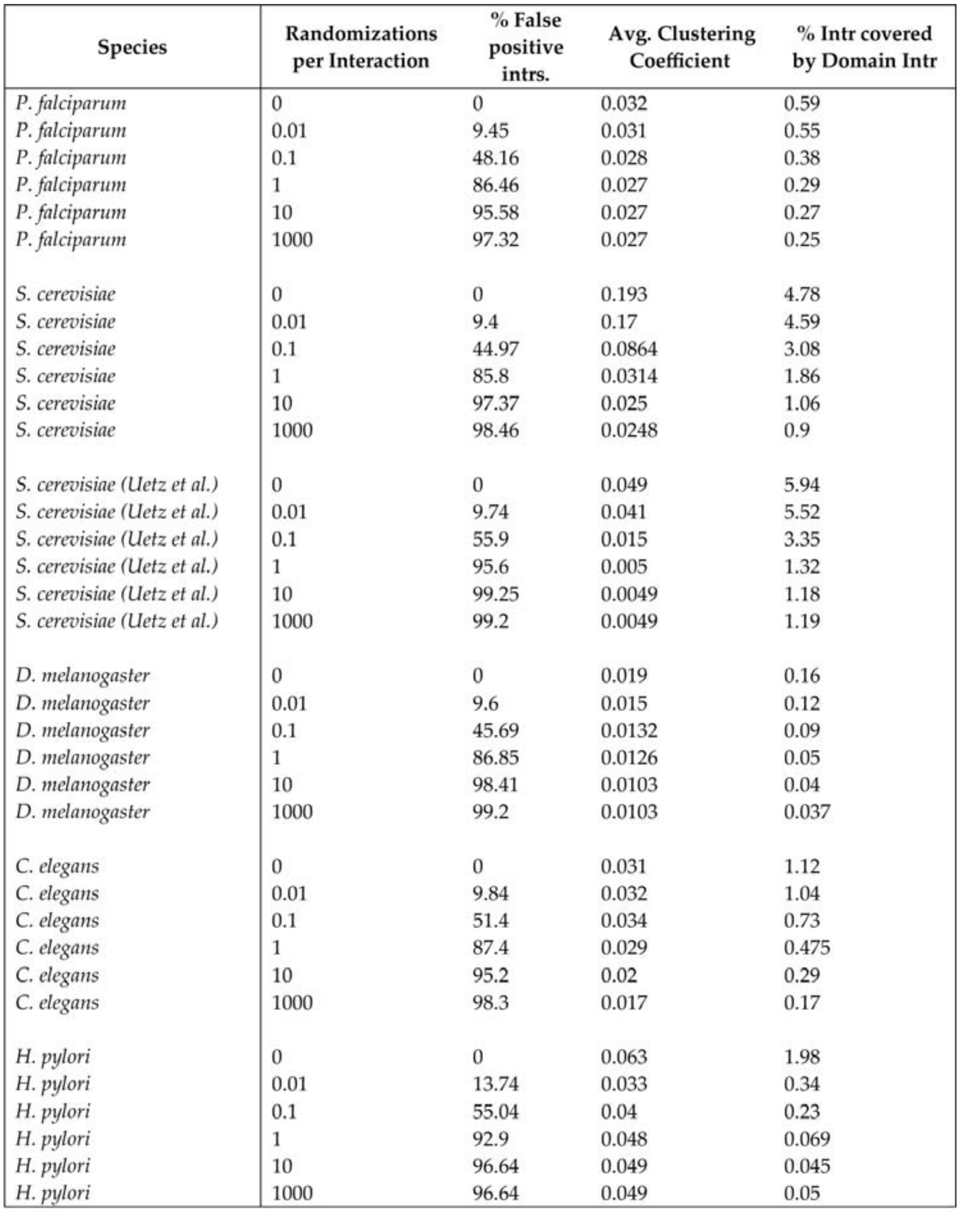


	Species 1			***Species 2***		
Comparison *Species 1 / Species 2)*	**Total Interactions**	t_1	c_1	**Total Interactions**	t_2	c_2
P. falciparum / S. cerevisiae	2,847	488	45	14,319	1,524	79
P. falciparum / D. melanogaster	2,847	801	50	20,720	1,446	67
P. falciparum / C. elegans	2,847	392	23	3,926	324	32
P. falciparum / H. pylori	2,847	8	0	1,465	78	0
C. elegans / S. cerevisiae	3,926	454	121	14,319	2,670	284
C. elegans / D. melanogaster	3,926	1,111	152	20,720	3,745	201
S. cerevisiae / D. melanogaster	14,319	6,178	565	20,720	3,183	392
H. pylori / C. elegans	1,465	17	1	3,926	9	1
H. pylori / D. melanogaster	1,465	80	2	20,720	49	3
H. pylori / S. cerevisiae	1,465	77	2	14,319	154	2

t_1 = *Species 1* interactions whose proteins have homologs in *species 2*. c_1 = The subset of interactions from t_1 that are conserved with interactions in *species 2*.

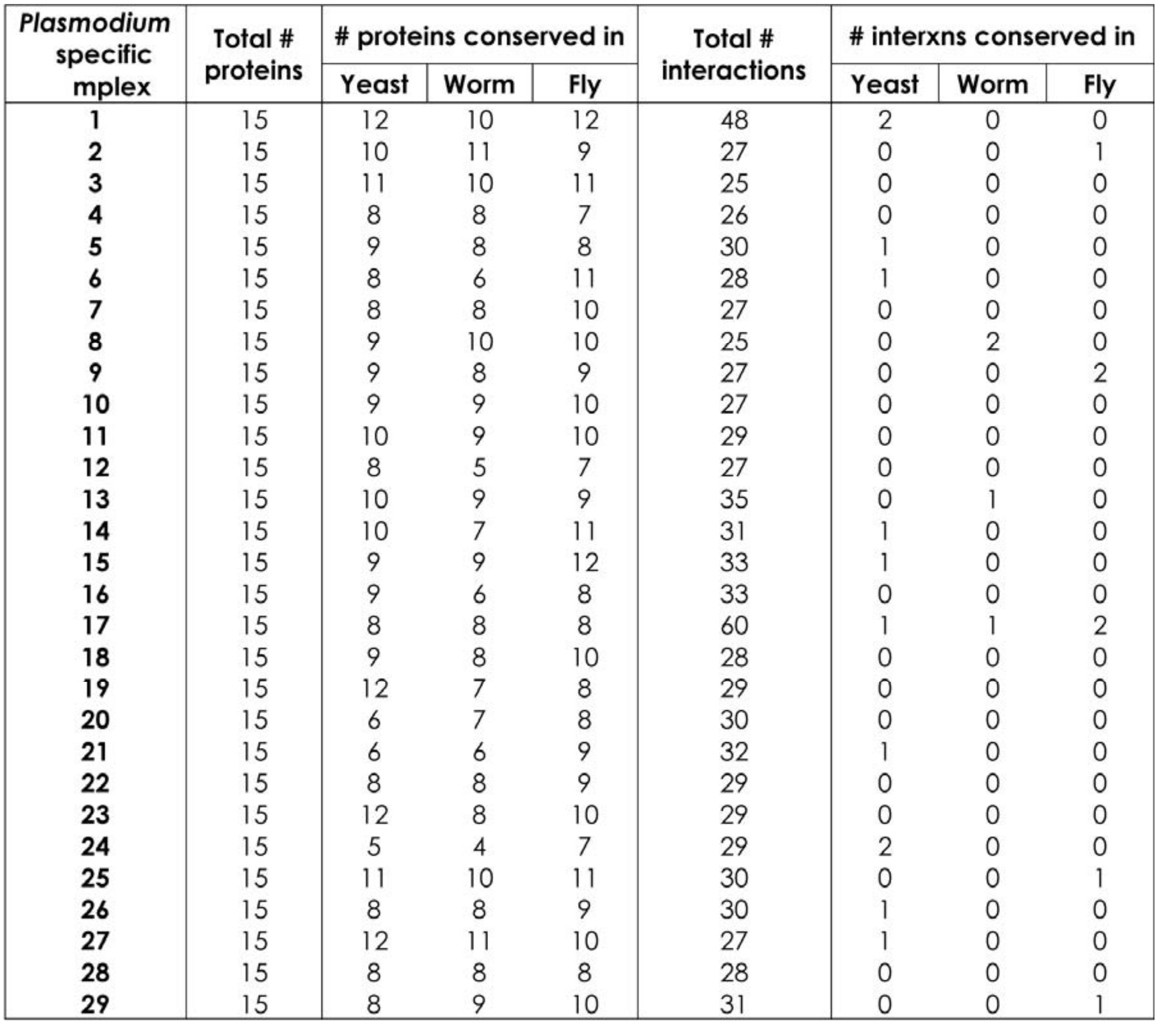
t_2 = *Species 2* interactions whose proteins have homologs in *species 1*. c_2 = The subset of interactions from t_2 that are conserved with interactions in *species 1*.

Table 4.4: Decay of global properties with increase in randomization.



Species	Randomizations per Interaction	% False positive intrs.	Avg. Clustering Coefficient	% Intr covered by Domain Intr
P. falciparum	0	0	0.032	0.59
P. falciparum	0.01	9.45	0.031	0.55
P. falciparum	0.1	48.16	0.028	0.38
P. falciparum	1	86.46	0.027	0.29
P. falciparum	10	95.58	0.027	0.27
P. falciparum	1000	97.32	0.027	0.25
S. cerevisiae	0	0	0.193	4.78
S. cerevisiae	0.01	9.4	0.17	4.59
S. cerevisiae	0.1	44.97	0.0864	3.08
S. cerevisiae	1	85.8	0.0314	1.86
S. cerevisiae	10	97.37	0.025	1.06
S. cerevisiae	1000	98.46	0.0248	0.9
S. cerevisiae (Uetz et al.)	0	0	0.049	5.94
S. cerevisiae (Uetz et al.)	0.01	9.74	0.041	5.52
S. cerevisiae (Uetz et al.)	0.1	55.9	0.015	3.35
S. cerevisiae (Uetz et al.)	1	95.6	0.005	1.32
S. cerevisiae (Uetz et al.)	10	99.25	0.0049	1.18
S. cerevisiae (Uetz et al.)	1000	99.2	0.0049	1.19
D. melanogaster	0	0	0.019	0.16
D. melanogaster	0.01	9.6	0.015	0.12
D. melanogaster	0.1	45.69	0.0132	0.09
D. melanogaster	1	86.85	0.0126	0.05
D. melanogaster	10	98.41	0.0103	0.04
D. melanogaster	1000	99.2	0.0103	0.037
C. elegans	0	0	0.031	1.12
C. elegans	0.01	9.84	0.032	1.04
C. elegans	0.1	51.4	0.034	0.73
C. elegans	1	87.4	0.029	0.475
C. elegans	10	95.2	0.02	0.29
C. elegans	1000	98.3	0.017	0.17
H. pylori	0	0	0.063	1.98
H. pylori	0.01	13.74	0.033	0.34
H. pylori	0.1	55.04	0.04	0.23
H. pylori	1	92.9	0.048	0.069
H. pylori	10	96.64	0.049	0.045
H. pylori	1000	96.64	0.049	0.05

Table 4.5: Conservation statistics of Plasmodium-specific complexes.



Plasmodium specific mplex	Total # proteins	# proteins conserved in			Total # interactions	# interxns conserved in		
		Yeast	Worm	Fly		Yeast	Worm	Fly
1	15	12	10	12	48	2	0	0
2	15	10	11	9	27	0	0	1
3	15	11	10	11	25	0	0	0
4	15	8	8	7	26	0	0	0
5	15	9	8	8	30	1	0	0
6	15	8	6	11	28	1	0	0
7	15	8	8	10	27	0	0	0
8	15	9	10	10	25	0	2	0
9	15	9	8	9	27	0	0	2
10	15	9	9	10	27	0	0	0
11	15	10	9	10	29	0	0	0
12	15	8	5	7	27	0	0	0
13	15	10	9	9	35	0	1	0
14	15	10	7	11	31	1	0	0
15	15	9	9	12	33	1	0	0
16	15	9	6	8	33	0	0	0
17	15	8	8	8	60	1	1	2
18	15	9	8	10	28	0	0	0
19	15	12	7	8	29	0	0	0
20	15	6	7	8	30	0	0	0
21	15	6	6	9	32	1	0	0
22	15	8	8	9	29	0	0	0
23	15	12	8	10	29	0	0	0
24	15	5	4	7	29	2	0	0
25	15	11	10	11	30	0	0	1
26	15	8	8	9	30	1	0	0
27	15	12	11	10	27	1	0	0
28	15	8	8	8	28	0	0	0
29	15	8	9	10	31	0	0	1

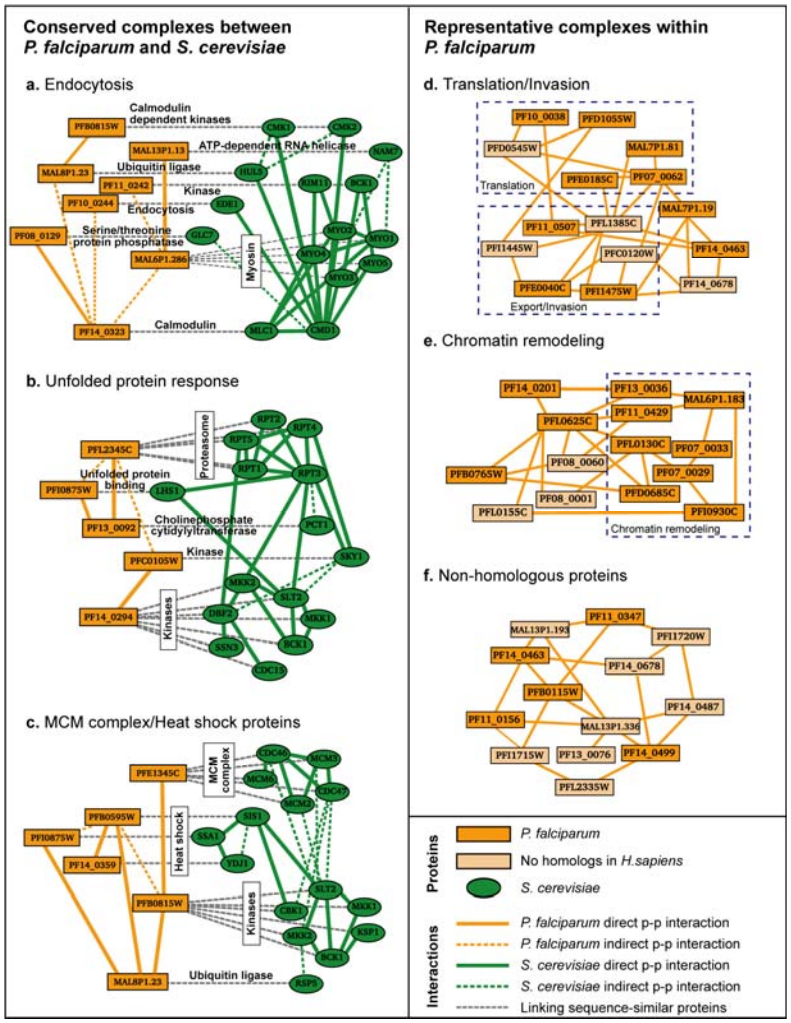
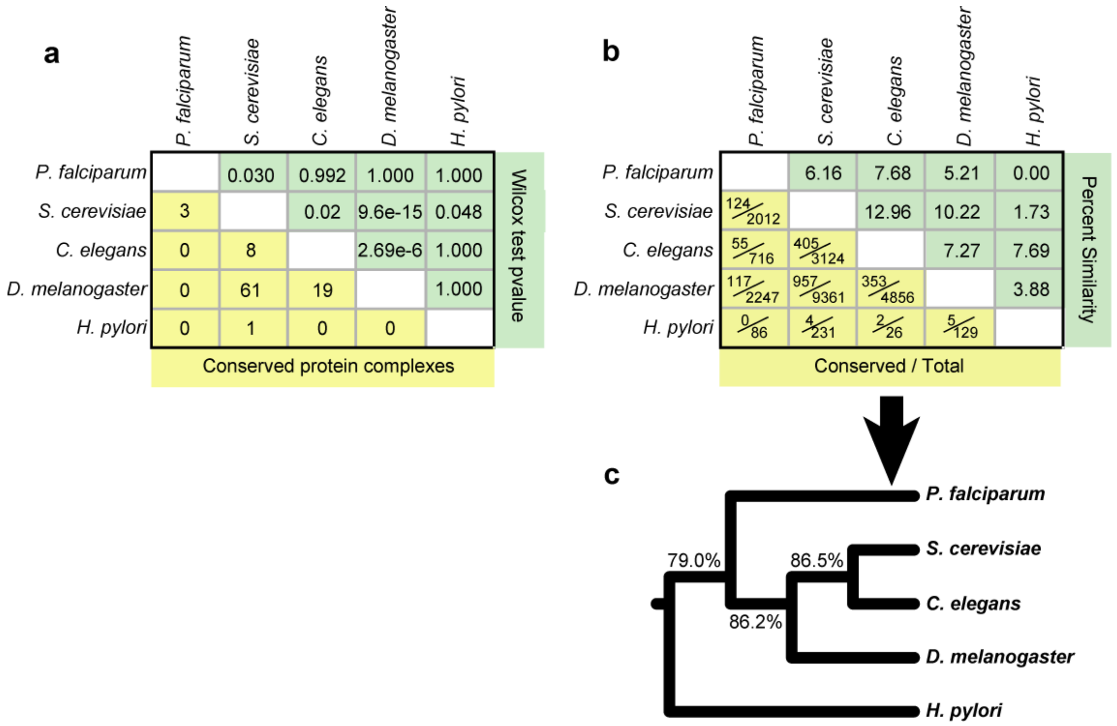


Figure 4.1: Conserved and distinct complexes within *P. falciparum*.
Panels (**a-c**) show the three conserved complexes identified between *P. falciparum* and *S. cerevisiae*. Orange vs. green nodes correspond to *P. falciparum* vs. *S. cerevisiae* proteins. Solid links represent direct interactions, while dashed (indirect) links represent interactions mediated by one other protein. Grey dashed lines connect sequence-similar proteins across the two species. Panels (**d-f**) show three representative complexes found within the *P. falciparum* network only. Cream-colored nodes denote *Plasmodium* proteins without human homologs (using a permissive BLAST E-value threshold ≤ $1\text{x}10^{-2}$ to allow for distant homologs).



	P. falciparum	*S. cerevisiae*	*C. elegans*	*D. melanogaster*	*H. pylori*
P. falciparum		0.030	0.992	1.000	1.000
S. cerevisiae	3		0.02	9.6e-15	0.048
C. elegans	0	8		2.69e-6	1.000
D. melanogaster	0	61	19		1.000
H. pylori	0	1	0	0	

	P. falciparum	*S. cerevisiae*	*C. elegans*	*D. melanogaster*	*H. pylori*
P. falciparum		6.16	7.68	5.21	0.00
S. cerevisiae	124/2012		12.96	10.22	1.73
C. elegans	55/716	405/3124		7.27	7.69
D. melanogaster	117/2247	957/9361	353/4856		3.88
H. pylori	0/86	4/231	2/26	5/129	

Figure 4.2: Network similarity across five species.

Panel **(a)** displays the results of all pair-wise PathBLAST comparisons. The number of conserved complexes is shown for each pair of species (yellow). The Wilcoxon rank-sum *p*-value (green) represents the significance of the distribution of all complex scores versus the distribution of complex scores found in equivalent random networks. In panel (**b)**, the interaction-by-interaction similarity between networks is reported as both fractional values (yellow) and percents (green). Panel (**c)** displays the phylogenetic tree constructed using these similarities. Percentages indicate the reproducibility of each branch during bootstrap analysis.

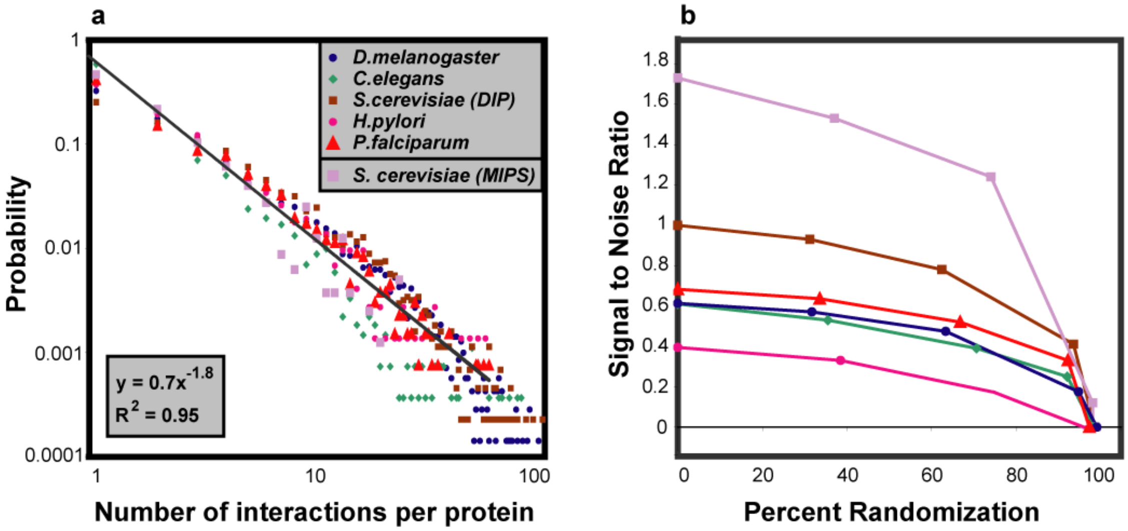


Figure 4.3: Properties of the protein interaction networks of five species. **(a)** Scale-free network behavior is shown in a manner similar to Barabasi *et al*.[98] The linear fit is for *Plasmodium* only. **(b)** Dependence of signal to noise ratio on error rate. Each network was modified by randomly shuffling from 0 to 100% of its protein interactions to simulate the addition of false positives and negatives. The subsequent decrease in SNR, converging to SNR=0 at 100% noise, validates that each network contains a substantial fraction of true positive interactions. In addition to the high-throughput networks in this study, a literature-curated network (*S. cerevisiae* physical interactions according to the MIPS[57]) is provided as a positive control. Similar trends are observed for the average clustering coefficient and the percent of interactions covered by established protein-domain interactions (Table 4).

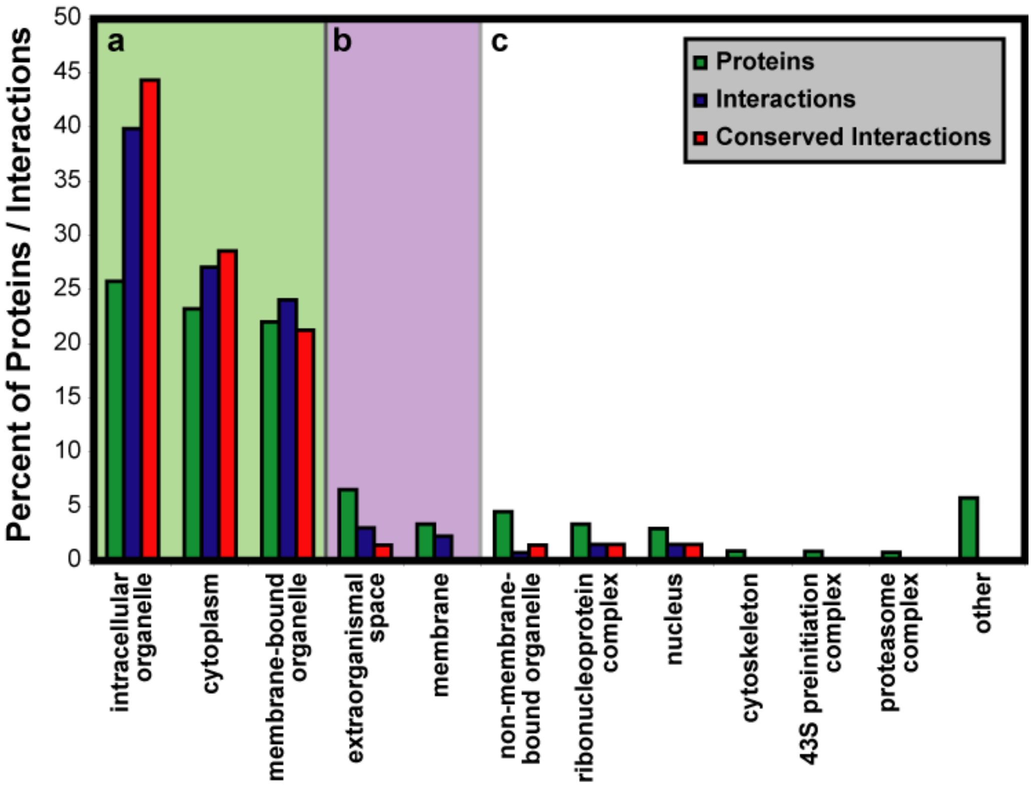


Figure 4.4: Functional roles among *Plasmodium* proteins (green), protein interactions (blue) and conserved interactions (red).

The histograms show the distribution of GO Cellular Component assignments over all annotated *Plasmodium* proteins or interactions. Interactions are considered "annotated" if the interacting proteins share the same GO category (these interactions are listed in Table 6). For conserved interactions, the percentages in each category are cumulative over the three pairwise comparisons of *Plasmodium* versus the other three eukaryotes yeast, fly, or worm. Note that a protein or interaction can participate in multiple categories.

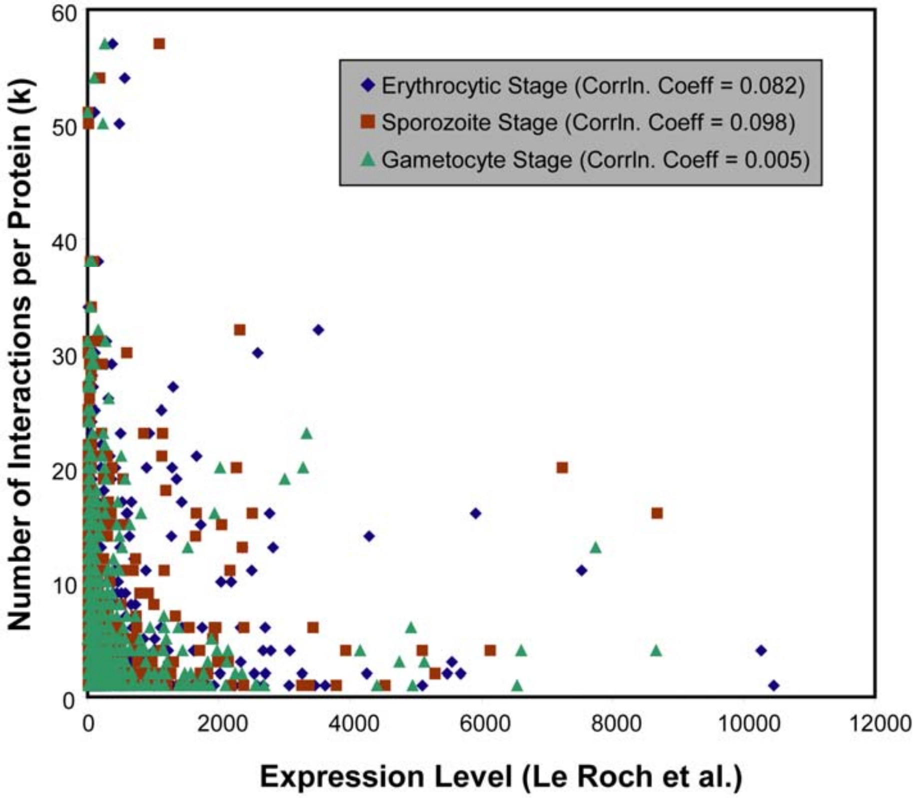


Figure 4.5: Number of interactions per protein versus mRNA expression level.

Absolute expression levels were obtained from Le Roch *et al*.[92], which includes experiments from the erythrocytic asexual stages (blue diamonds), the mosquito salivary-gland sporozoite stage (red squares) and the sexual gametocyte stage (green triangles). No significant correlation was observed for any stage.

a. Yeast

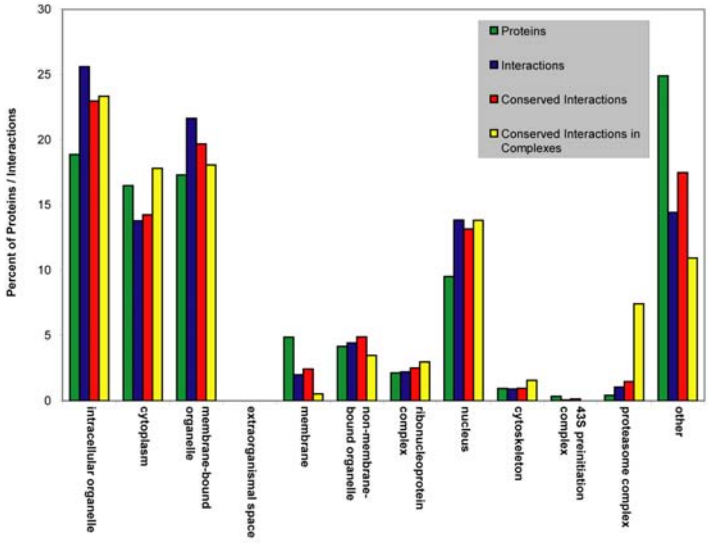
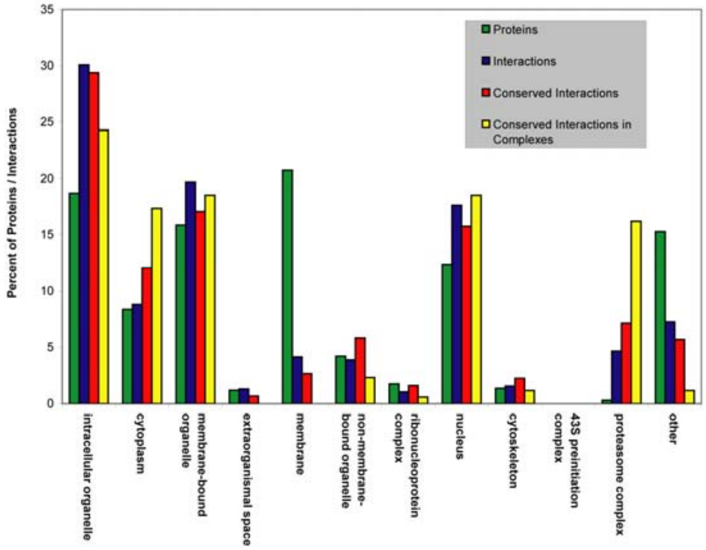


Figure 4.6: Gene Ontology (GO) enrichment among cellular components in yeast (a), worm (b) ,and fly (c).

These histograms are complementary to Figure 4.4 and indicate the representation of common cellular components in each species' genome (green), set of interactions (blue), conserved interactions (red), and conserved complexes (yellow). The percentages of conserved interactions are combined over the separate pairwise comparisons for each species (eg. in panel [a], data from yeast vs. worm, fly or *Plasmodium* comparison was used). Note that a protein or interaction can participate in multiple categories.

b. Worm



c. Fly

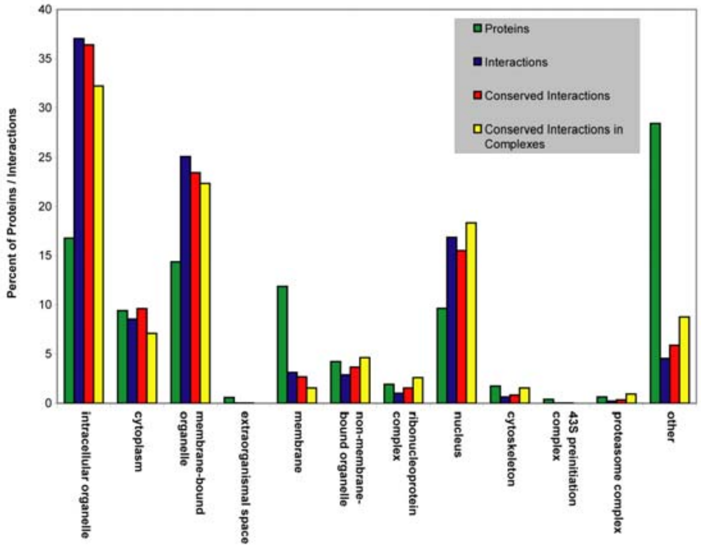


Figure 4.6 continued

5 Network Integration: An efficient method for interpreting eQTL associations using protein networks

The technique of expression Quantitative Trait Loci (eQTL) is becoming increasingly widespread for revealing the genetic loci in control of specific changes in gene expression[30,31]. eQTLs are a variant of the more basic concept of Quantitative Trait Loci, which measures the association between a quantitative phenotype (such as height and weight) and a panel of polymorphic genetic markers distributed across the genome[113]. For the special case of eQTL analysis, the phenotype of interest is a gene expression level measured with DNA microarrays[30]. Since a microarray monitors expression levels of all genes, separate statistical tests are performed to compute scores of association of each genetic marker with each gene expression level.

Two of the core challenges[32,35] in understanding and explaining eQTL associations are:

Fine Mapping: Due to the spacing of genetic markers and/or linkage disequilibrium, several genes can reside near each marker. Typically, no more than one of these genes is responsible for the observed expression phenotype.

Identifying the true causative gene requires additional data, since all genes at a locus are indistinguishable based on the eQTL measurements alone.

Lack of mechanistic explanation: A gene-phenotype association typically lends little insight into the underlying molecular mechanism for the association.

Several bioinformatic approaches have been proposed recently to address these two issues[31,34,36,114,115]. For the problem of "fine mapping", the main bioinformatic focus has been on predicting which genes within a given locus are the true regulators of expression of the target phenotype. For instance, Kulp and Jagalur[114] sought to infer the true causal genes using a Bayesian network model constructed from expression correlations detected within the eQTL profiles. Another powerful approach has been to complement eQTLs with data on physical molecular interactions. Tu *et al.*[36] modeled each eQTL association as a sequence of transcriptional and protein-protein interactions that transmits signals from the locus to the affected target. This method is promising since it prioritizes candidate genes by their network proximity to the affected target gene and also provides a model of the underlying regulatory pathways. In addition, assembly of protein interaction networks is a burgeoning area in genomics and the amount and quality of protein interaction data are rapidly improving. Therefore, integrating eQTL data with additional independent information may significantly reduce the noise and improve the statistical power of the analysis[116].

Here, we describe a new integrative approach (named '**eQ**TL **E**lectrical **D**iagrams' or **eQED**) which also combines eQTL data with protein interaction networks but predicts the true causal gene at each locus with substantially higher accuracy than the previous method. eQED models the flow of information from a locus to target genes as electric currents through the protein network. Currents can be simulated simultaneously for all loci influencing a target, allowing multiple loci to reinforce each other when they fall along a common regulatory pathway.

5.1 "Electric circuit" analysis of eQTLs

5.1.1 Definition of terms

In what follows, the genes near a polymorphic genetic marker are called *candidate genes*, and the genes with an associated change in expression are called *targets*. The particular candidate gene that is truly responsible for the downstream change in expression of a target is called the *true causal gene*. Collectively, the set of candidate genes near a marker define a *locus*. Finally, the proteins and their interactions in the protein network are referred to as *nodes* and *edges* respectively.

5.1.2 *Open problems motivated by previous method*

For a given locus and associated target, the Tu *et al.*[36] method works by executing a random walk through the protein network starting at the target. At every step of the walk, the next edge to be followed depends on its predefined weight (see Methods). The walk ends when it reaches one of the candidate genes in the locus. The random walk is repeated 10,000 times, and the candidate gene that is visited most often is predicted to be the true causal gene. Figure 5.1A shows a sample network, while Figure 5.1B shows a sample random walk on this network according to the Tu *et al.* approach. Gene L3 is visited most often and, hence, is reported as the causal gene.

Given that the protein network is large, many random walks must be executed in order for the predictions to be accurate. Moreover, a single random walk from the target to any candidate gene may require many steps. These two issues can lead to random walk simulations that last a prohibitively long time. In Tu *et al.*, the authors make a key approximation which allows them to achieve feasible simulation times: they constrain the path taken by each random walk to be acyclic (i.e., no genes can be revisited). As a consequence, many walks result in "dead ends" unable to reach any candidate gene, but all walks are at least relatively short. This "greedy" approximation may lead to different predictions from typical random walk models[117] which may affect their accuracy. In addition, since biological networks are scale-free[118], they contain a large number of dead ends (i.e., nodes with a single edge). The many dead ends greatly reduce the absolute number of visits to the candidate genes, thereby reducing the overall confidence in the final causal gene prediction for a given number (e.g. 10,000) of walks.

5.1.3 The eQED model

The eQED approach seeks to address the above open problems by replacing the random walk model with a framework based on electric circuits. There is considerable prior work establishing the equivalence between electric networks and random walks (see Methods). The eQTL associations and the corresponding protein network are abstracted as an analog electric circuit model grounded at a given target gene. The weights on the edges of the molecular network are modeled as conductances (1/resistance) in the electric circuit. The *p*-values of association between each genetic locus and expression of the target are modeled as independent sources of current. An electric circuit abstraction is constructed for every locus-target association (which we call the *single-locus* model, Figure 5.1C). Further details of the model are provided in the Methods section.

After solving the circuit for currents, the causal gene is predicted as the one with the highest current running through it. Analyzing the network as an electric circuit provides a deterministic "steady state" solution, in contrast to a stochastic random walk. Moreover, the number of dead-end nodes in the network does not affect the final result as the total current through them is always zero (Figure 5.1C).

5.1.4 Application to eQTL associations in yeast

As a proof of principle, we applied the eQED approach to analyze the results of a genome-wide eQTL study in yeast by Brem *et al.*[30]. This study reported associations between 2,956 genetic markers and 5,727 gene expression levels measured across 112 yeast strains (Methods). All locus / target pairs with a gene association *p*-value ≤ 0.05 were considered; within this set, we selected only those loci containing more than one candidate gene (i.e., for which the true causal gene was ambiguous). At the same time, we assembled a pooled interaction network consisting of 17,171 transcriptional and protein-protein interactions reported in previous large-scale studies (Methods). Given this network, the set of locus / target pairs was further filtered to include only those loci for which at least two of their candidate genes had at least one transcriptional or protein-protein interaction, yielding a total of 131,863 locus / target pairs. The single-locus model of eQED was applied to each locus / target pair, and a causal gene prediction was made in each case. This step-by-step procedure is diagrammed in Figure 5.2.

To estimate the accuracy of the predictions, we compiled a set of "gold standard" cause-effect pairs from two large gene knockout expression profiling studies in yeast, Hughes *et al.*[119] and Hu *et al.*[120], as well as from a gene over-expression study by Chua *et al.*[121]. In these studies, strains harboring a single gene knockout or over-expression construct (the "true causal gene") had been analyzed using whole-genome microarrays to identify a resulting set of differentially-expressed genes (the "targets"). We filtered these three data sets to include only those causal gene / target pairs that were present in the molecular network used by eQED and for which the causal gene was associated with the target gene at $p \leq 0.05$ in Brem *et al.* (see Methods). The resulting gold-standard set contained 548 causal gene / target pairs.

Table 5.1 reports the number of correct predictions of the causal gene for each method. The single locus model of eQED correctly predicted 392 of the 548 gold standards (72% accuracy). In comparison, the approach by Tu *et al.* achieved 50% accuracy. Both methods performed substantially better than random selection of a gene at a locus, which achieved 22% accuracy.

5.1.5 Combining multiple loci

In our model, given a target gene and a corresponding significant marker, there exists only one causal gene. However, in eQTL studies, the expression level of a target gene typically has significant associations with more than one marker (and thus more than one causal gene). If these causal genes fall along common regulatory pathways, considering multiple loci together in the same eQED model might increase our confidence in the causal gene predictions. Motivated by these considerations, we explored a second circuit model, called *multiple-loci* eQED, in which currents were included for all significant loci associated with a target (see Methods). For example (Figure 5.1E), let the target T associate significantly with two loci. In the single-locus model, we would investigate the two associations separately, but in the multiple loci model their information is processed as a single circuit. Figure

5.1F shows a schematic of the multiple-locus model from Figure 5.1E. For each locus considered, the causal gene is predicted as the one having the highest current flowing through it.

The accuracy of the multiple-loci eQED model was estimated using the same gold-standard data set used for the single-locus model. As shown in Table 5.1, the multiple-loci model boosted prediction accuracy substantially over the single-locus case (80% versus 72%). Combining information from all significant loci for a given target also reduces computation time, as all loci are processed in a single eQED simulation instead of multiple runs.

5.1.6 Predicting the direction of signaling along protein interactions

A direct consequence of the electric circuit model is that the currents on the wires of the network suggest a direction of information flow in the biological system. In the case of transcriptional interactions, the current is restricted to flow from the transcription factor to the regulated gene, and not *vice versa* (Methods). In contrast, the direction of information flow along protein-protein interactions is not predetermined, since the underlying biochemical measurements typically report only whether an interaction exists, not its functional consequences. Therefore, for protein-protein interactions in particular, eQED provides a means of predicting the direction of signal transmission.

Multiple-loci eQED induces a current on each protein-protein interaction in the network. Repeated application over all targets yields a distribution of current values for each interaction. This distribution can be analyzed to determine whether the current is predominantly positive or negative (prior to the analysis positive and negative directions of flow are defined arbitrarily for each interaction). We evaluated three simple methods for summarizing this distribution of currents, by using either (1) the most extreme current; (2) the sum of currents; or (3) the skewness of the current distribution. Each of these three methods yielded a single value per interaction whose signs were interpreted as the predicted directions and whose magnitudes could be used to rank the predictions in order of confidence.

To assess the performance of directionality prediction, we once again compiled a set of gold-standards, consisting of protein-protein interactions for which the signaling directions are known. A total of 408 gold-standard interactions were obtained, including 103 signaling interactions recorded in the Kyoto Encyclopedia of Genes and Genomes (KEGG)[122] or the Munich

Information center for Protein Sequences (MIPS)[57], as well as the 305 highest-confidence kinase-substrate interactions reported in a systematic analysis of phosphorylation by Ptacek *et al.*[123]. Figure 5.3A shows the accuracy of the three methods at recapitulating the known directions of signaling. Although the "sum of currents" method yielded very high accuracy (> 80%) for the 40 highest-ranking predictions, the "most extreme current" method retained moderate accuracy (generally > 75%) out through the best 80 predictions (corresponding to the largest area under the curve). In contrast to these first two methods, the third method based on "skewness" was not an accurate predictor of directionality.

Based on this analysis, we used the "most extreme current" method to predict directionality of information flow for all protein-protein interactions in the eQED network. We made a total of 368 predictions with absolute most extreme current >= 623, corresponding to the 75% accuracy mark above.

5.1.7 Prediction of regulatory pathways

The currents computed by eQED provide an estimate of the influence of each protein interaction on the regulation of the target gene. To reveal how individual high-current interactions might assemble into regulatory pathways, we sought to connect each causal gene to its target by finding an optimal path through the network, defined as the shortest route with the highest total sum of currents across its interactions. The union of all optimal paths leading from each predicted causal gene into a given target reveals its regulatory network. We also filter the regulatory network to include only those PPIs which have a predicted direction of influence (see previous section). Figure 5.3B-D shows the regulatory network obtained for three example target genes: HMG2, ARG5/6, and AAD15. Although the causal genes are often at the head of each path comprising the regulatory network, in some cases a path contains a chain of causal genes in series. For instance, both ARO80 and RLR1 associate significantly with the target ARG5/6 and share the same regulatory pathway. This is a direct consequence of integrating the information about all significant loci when running eQED (multiple-loci model). As a result, the casual genes not only reinforce each other but also increase the overall confidence of the underlying regulatory network.

5.1.8 Application to gene-association studies in human

During the past few years, a substantial body of eQTL data has been generated in higher eukaryotes, including a number of studies in mouse and

Arabidopsis thaliana (see www. genenetwork.org). Large eQTL studies are now also available for humans[124-126]. All of these datasets associate genetic loci with gene expression levels without explicitly identifying the causal genes at each locus, raising the important question of whether they could be identified using an integrative network-based approach such as eQED.

As for human, a network-based analysis of eQTLs will require a substantial database of protein-protein and transcriptional interactions. In terms of protein-protein interactions, several large networks have recently been mapped for humans[41,42,127]. The remaining hurdle is therefore the availability of large-scale measurements of transcriptional interactions. Although no systematic study has yet been published, several such efforts are underway using systematic chromatin immunoprecipitation experiments in human cell lines and in-vitro technologies such as the protein binding microarray (PBM)[6]. As these networks become available, the success of eQED in yeast suggests that it may also provide a powerful means for identifying human disease genes and their associated transcriptional regulatory pathways in higher eukaryotes.

5.2 Materials and Methods

5.2.1 *Electric circuits and random walks*

There is considerable literature establishing the analogy between random walks and electric networks[117,128,129]. In particular, Doyle and Snell[117] showed that there always exists a random walk equivalent of linear electrical circuits. Random walks on a network can be abstracted as a Markov chain and consequently, be represented using a transition state matrix. Consider an electric network *E* where the conductance on an edge (*x,y*) is represented by C_{xy}. A random walk can then be defined on *E*, which has the transition state probabilities:

$P_{xy} = \frac{C_{xy}}{C_x}$ where $C_x = \sum_{i \in N(x)} C_{xi}$ and *N(x)* is the set of neighbors of *x* in the network.

Since an electric network is a connected graph, it is possible to travel between any two states. A Markov chain with such a property is known as an *ergodic chain*. For an ergodic chain represented by the transition matrix *P*, there exists a fixed vector $w = (w_1, w_2, \ldots, w_n)^T$, such that $wP=w$, where w_j represents the steady-state proportion of times the walker remains in state *j*.

In the case of random walks derived from electric networks, it can be shown that:

$$w_j = {}^{C_j}\!\!\diagup_{C} \text{◎◎where◎} C = \sum_x C_x \text{ ◎◎◎}$$

An ergodic chain is called *time-reversible* if $w_x P_{xy} = w_y P_{yx}$. Thus, in the case of the random walk derived from an electric circuit,

$$w_x P_{xy} = \frac{C_x}{\sum_x C_x} . \frac{C_{xy}}{C_x} = \frac{C_{xy}}{\sum_x C_x} = \frac{C_{yx}}{\sum_y C_y} = \frac{C_y}{\sum_y C_y} . \frac{C_{yx}}{C_y} = w_y P_{yx}$$

As a result, the random walk *P* is also time-reversible. Finally, using the above properties we can show that when a unit current flows into an electric network at node "a" and leaves at node "b", then the amount of current through any intermediary node or edge is proportional to the expected number of times a random walker will pass through that node or edge [see Doyle *et al.*[117] for details].

We demonstrate this equivalence using the sample network of Figure 5.1A. Figure 5.1C is the electric network model of the sample network. Here, we add a new node L which is connected to all the candidate genes in the locus. The edges connecting L to L1, L2 and L3 have infinite conductance and for all purposes, L is no different from any of L1, L2 or L3. The conductance on the remaining edges is equal to their weight in the sample network. The target gene T is treated as "ground" for the electric network. There is an independent source of current sending 10,000 A of current into the network at L. We solve the network using Kirchhoff's law and Ohm's Law[130] to get the currents through each edge and node. Figure 5.1D shows the sample network represented as a random walk derived from the electric network of Figure 5.1C. The random walk is repeated 10,000 times. The number of times each edge and node was visited in the random walk converges to the amount of current through those edges and nodes in the electric network (Figures 5.1C and 5.1D).

5.2.2 eQTL Associations

Yeast eQTLs were obtained from Brem *et al.*[30], consisting of whole genome expression data for 112 yeast strains, which were genotyped across 2,956 genetic markers. Genetic similarity between strains, referred to as population substructure, can lead to false-positive relationships where the observed phenotype correlates well with the phylogenetic relationships between the strains and the markers do not predict phenotype beyond the phylogeny. We corrected for the population substructure problem using the method of Zhao *et al.*[131]. The resulting marker-gene associations were converted to gene-gene associations by assigning genes to their nearest marker (within 10 kb) on the genome. Finally, all genes assigned to the same marker were defined to belong to the same locus.

5.2.3 High-confidence physical interaction network

PPI interactions were obtained from a modified form of the STRING database (Search Tool for the Retrieval of Interacting Proteins, version 6.3)[48], extended to incorporate additional information on potential interactions. STRING reports a confidence score for each protein interaction based on numerous experimental and computational evidences. We implemented a naïve Bayes classifier which takes the STRING score as one line of evidence. As a second line of evidence, we incorporated quantitative genetic interactions from Collins *et al.*[132] who analyzed double-mutants to detect both aggravating and alleviating genetic interactions. Genetic interactions may also be used as indirect predictors of physical protein interactions[29,133]. As a third and final line of evidence, we used recently published protein interaction data[12,134] that were not included in the 6.3 version of STRING. For fitting the parameters of the model, a positive training set of 11,814 distinct interactions was created from pairs of proteins falling within known pathways recorded in the Kyoto Encyclopedia of Genes and Genomes (KEGG)[122] as well as from small-scale binary physical interactions and protein complexes from the Munich Information center for Protein Sequences (MIPS)[57]. The negative training set of 35,676 interactions was obtained by randomly pairing proteins. We filtered the top 22,428 interactions with log-likelihood scores (LLS) > 3.0.

Pooling with Transcriptional Interactions

A pooled molecular interaction network was constructed by merging the above protein-protein interactions (PPIs) with transcription factor-DNA interactions (TF-DNA) obtained from Beyer *et al.*[135]. This study combined several lines of evidence in a Bayesian framework to assign LLS to each TF-DNA link. The 11,513 TF-DNA interactions with LLS > 3.0 were included in the final set.

To ensure that all interactions in the network (PPI and TF-DNA) represented physical binding events (as opposed to functional linkages), we required that each included interaction has been reported in at least one experiment indicating direct physical interaction between the proteins. In addition, to enrich for interactions within regulatory pathways, the network was restricted to regulatory proteins. We included proteins that were assigned to the following MIPS categories: (1) regulation of glycolysis and gluconeogenesis, (2) regulation of electron transport and membrane-associated energy conservation, (3) regulation of respiration, (4) regulation of energy conversion / regeneration, (5) regulation of DNA processing, (6) mitotic cell cycle and cell cycle control, (7) transcriptional control, (8) regulation of splicing, (9) translational control, (10) protein fate (folding, modification, destination), (11) regulation of metabolism and protein function, (12) cellular communication and signal transduction mechanism (13) cell rescue, defense and virulence, (14) cellular sensing and response to external stimulus, (15) cell fate, (16) development (systemic), and (17) cell type differentiation. Altogether, the final pooled network consisted of 4,466 proteins and 17,171 non-redundant interactions.

Network filtering based on target-gene

In this study, we analyze transcriptional regulation of the target gene. For that purpose, we make the assumption that the expression of the target gene is modulated by the causal gene through a transcription factor of the target gene. Hence, for every target gene we filter the high confidence network such that the target gene is connected to the rest of the network through TF-DNA interactions only. As some of the target genes in the Brem study do not have TF-DNA edges, we cannot use our approach to analyze them. This reduced the number of target genes that were analyzed in this study to 3,711.

Tu *et al.*[36] weighted each node in the network using the mRNA expression correlation between the node and the target gene. We use the same idea; however, in our framework weights are most naturally placed on edges as opposed to nodes. The weight on each edge (*u,v*) is defined to be the average mRNA correlation of *u* and *v* with the target gene. Thus, our approach is meant to model as closely as possible the scheme of Tu *et al*. and differs only in that weights are placed on edges versus nodes.

5.2.4 eQED Model

The eQED model used in this paper utilizes the relationship between electric circuits and random walks (see previous sections). The exact equivalence between electric circuits and random walks follows only when the network under consideration is completely undirected. However, in our study we also use directed edges (TF-DNA interactions) and consequently, we employ a heuristic motivated by the undirected case. Specifically, let *S(N,E)* represent the molecular interaction network (*N* being the set of genes in the network and *E* being the set of interactions. Let *C(e)* denote the conductance on edge *e*, while *I(L)* represents the independent input current at locus *L*. Let d_e be a new variable associated with the directed edge *e* and let $D \subseteq E$ be the set of all known directed edges in the network.

$$\text{Objective:} \quad Min \sum_{(u,v)\in D} (d(u,v) - (V(u) - V(v))$$

$$\forall u,v \notin D : I(u,v) = C(u,v) \times (V(u) - V(v)) \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (1)$$

$$\forall u,v \in D : I(u,v) = C(u,v) \times d(u,v) \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (2)$$

$$\forall v \neq t : \sum_u I(u,v) = 0 \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (3)$$

$$\forall (u,v) \in D : d(u,v) \geq (V(u) - V(v)) \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (4)$$

$$\forall (u,v) \in D : d(u,v) \geq 0 \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (5)$$

where *u* and *v* are any nodes in the network and *t* is the target gene and, *V* is voltage on the nodes of the electric circuit. Here, equations (1) and (2) are derived from Ohm's law which states that the current flowing through any two points is directly proportional to the voltage difference and the conductance between them. Further, equation (3) corresponds to Kirchoff's current law in electric circuit theory which states that the total sum of current through any point in the circuit is zero[130]. The wires of a simple resistive circuit (as shown in Figure 5.1C) do not have explicit directionality, such that current can flow in either direction. However, the molecular network used in this study includes TF-DNA interactions that, by definition, transmit signal from the transcription factor to the DNA and not *vice versa*. Electrical circuits account for directed links by using diodes, which constrain current to flow in one direction only. Equations (4) and (5) are constraints to ensure that the current only flows in the correct direction on known directed edges. For instance, let *(u,v)* be a directed edge with the signal going from *u* to *v*. If *V(u)* > *V(v)*, then to minimize the objective function, *d(u,v)* will take the value *(V(u)-V(v))*. As a result, the equation becomes same as (1). However, if *V(u)* < *V(v)*, then due to (5), *d(u,v)* will be equal to 0, implying that there is no current on that edge. We implemented the above linear programming approach in Matlab[55] using the MOSEK package Version 5[136].

Acknowledgements

This chapter contains the complete reprint of the work Suthram S, Beyer A., Ideker T. eQED: An efficient method for interpreting eQTL associations using protein networks. *Proceedings of 3rd Annual RECOMB Satellite Conference on Systems Biology,* 2007. I was the primary author of this paper.

Table 5.1: Causal gene prediction accuracy.

Method	Number of Correct Predictions
Random	118
Tu *et al.*	262
Shortest Path**	351
eQED (single locus)	392
eQED (Multiple loci)	438

All predictions were against a gold standard data set of 548 causal gene-target pairs compiled from yeast gene-expression knockout studies by Hughes *et al.*[119] and Hu *et al.*[120] and a gene over-expression study by Chua *et al.*[121].

** A naïve method in which the causal gene is selected to be the gene in the locus that is connected by the shortest path to the target.

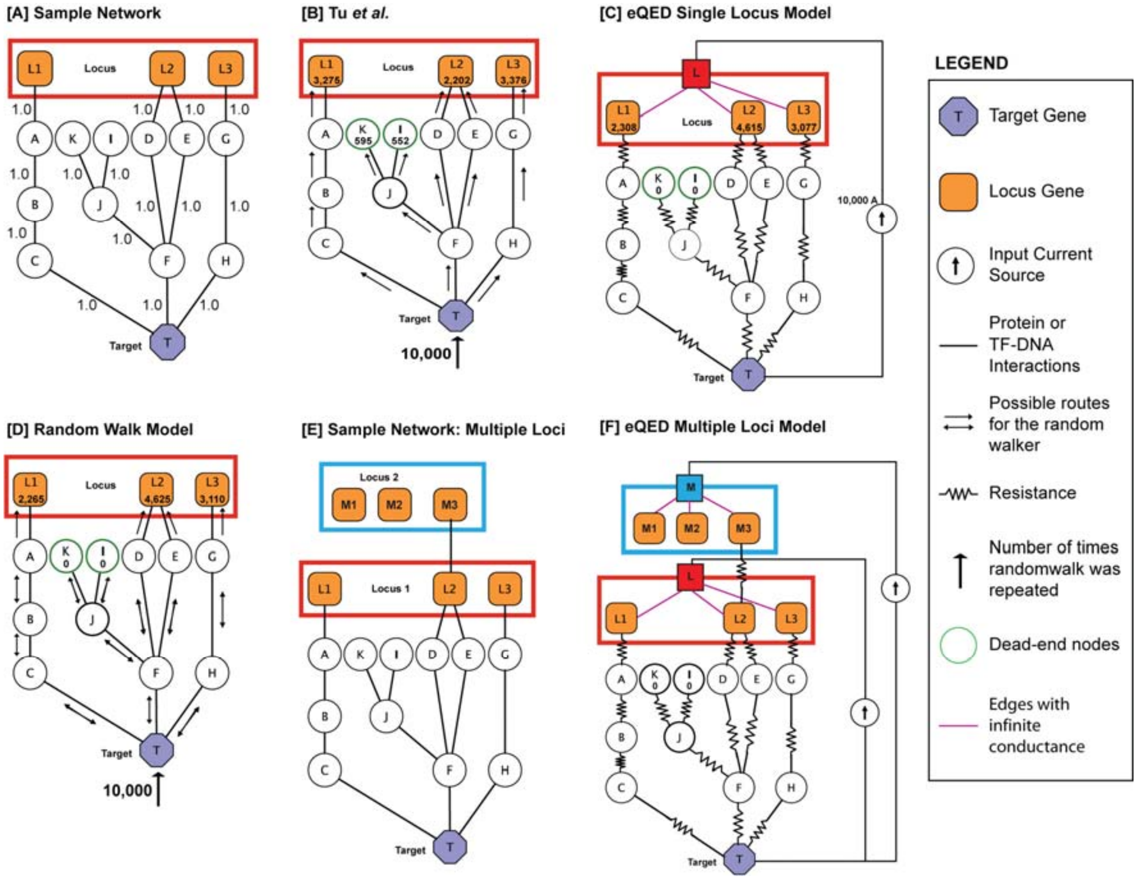


Figure 5.1: Examples of the electrical circuit approach and the eQED model. **[A]** Sample network. **[B]** The "greedy" random walk approach by Tu *et al*. **[C]** The single locus model of eQED. Gene T in the blue hexagon is the target gene. The locus marked by the red box, containing candidate genes L1, L2 and L3, associates significantly with the target T. The numbers next to the locus genes corresponds to the number of times the gene was visited in the random walk approaches or the amount of current through them in the electric circuit approach. **[D]** The random walk derived from [C]. **[E]** The sample network with two significant loci. **[F]** The multiple loci model of eQED.

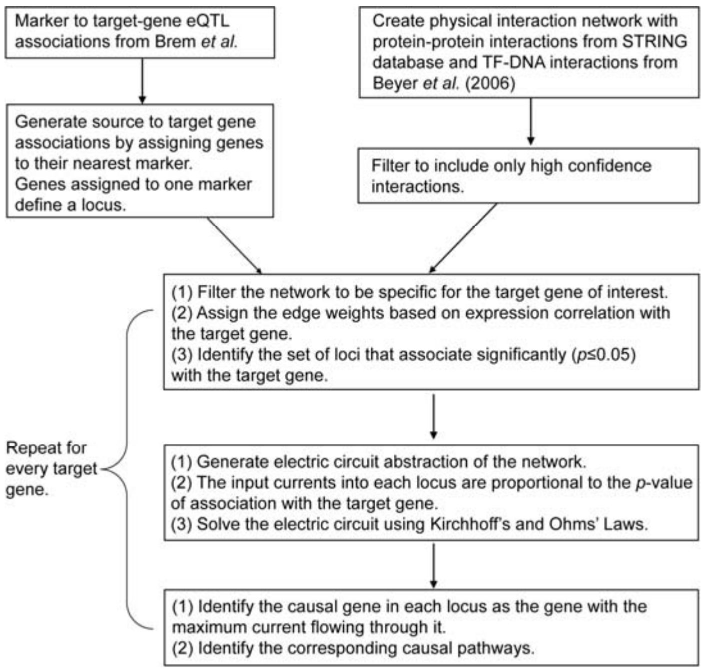

Marker to target-gene eQTL associations from Brem *et al.*
Create physical interaction network with protein-protein interactions from STRING database and TF-DNA interactions from Beyer *et al.* (2006)
Generate source to target gene associations by assigning genes to their nearest marker. Genes assigned to one marker define a locus.
Filter to include only high confidence interactions.
(1) Filter the network to be specific for the target gene of interest. (2) Assign the edge weights based on expression correlation with the target gene. (3) Identify the set of loci that associate significantly ($p \leq 0.05$) with the target gene.
Repeat for every target gene.
(1) Generate electric circuit abstraction of the network. (2) The input currents into each locus are proportional to the *p*-value of association with the target gene. (3) Solve the electric circuit using Kirchhoff's and Ohms' Laws.
(1) Identify the causal gene in each locus as the gene with the maximum current flowing through it. (2) Identify the corresponding causal pathways.

Figure 5.2: Flowchart of the eQED method.

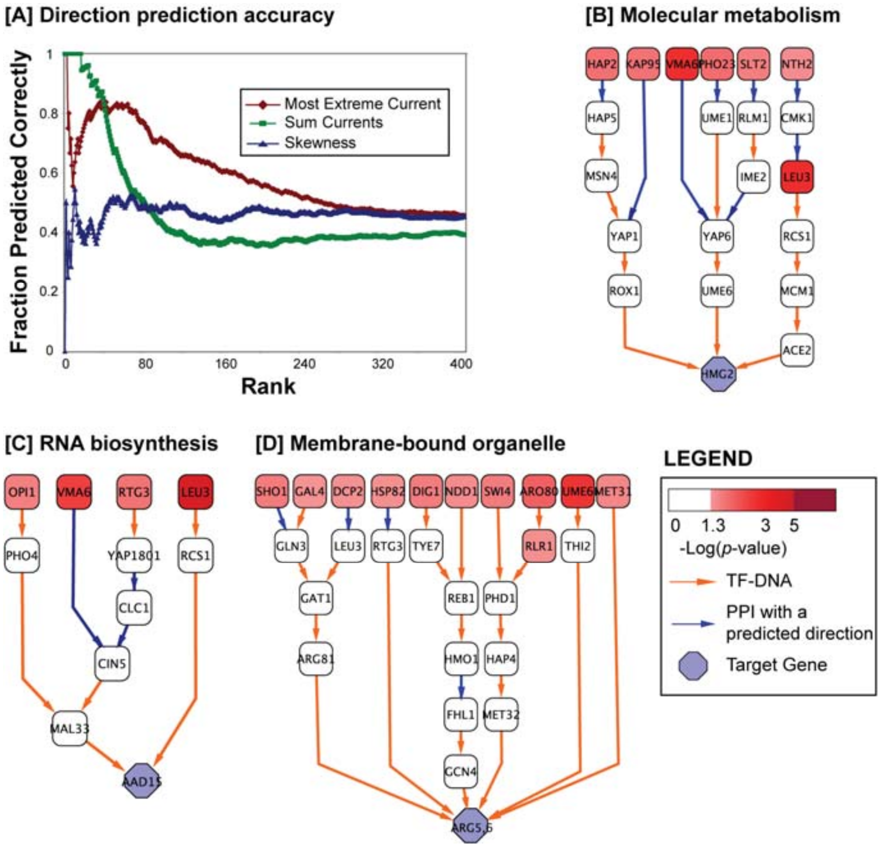


Figure 5.3: Inferred pathways and directionality prediction.
[A] The accuracy of the direction prediction methods. The "gold" standard set of protein interactions were ranked according to the different metrics and, the ranks in each approach was plotted as the x-axis. **[B-D]** The regulatory networks for three example target genes. The nodes colored in shades of red correspond to predicted causal genes. The intensity of color corresponds to their *p*-value of association with the target.

6 Conclusion

The amount data on protein interactions has been growing steadily. Figure 1.3 shows a graph of the vast increase in the number of protein interactions since 1999. In fact, the number of species for which large-scale protein network data is now available is also large. Simultaneously, high-throughput technology has resulted in other large-scale data such as generation of large scale generation of genetic interaction data[137], protein-DNA binding data[7] and genome wide association data[138]. The availability of these vast amounts of data, each addressing a different aspect of the cellular function in various species, calls for a comprehensive approach to analyzing them. As protein networks form the backbone of the all mechanisms within a cell, the focus of this dissertation was to get a better understanding of cellular function through the analysis of protein networks.

Most high-throughput methods contain a considerable number of false-positives. Therefore, I first addressed the issue of noise in the large-scale protein interaction networks in Chapter 2. Specifically, I generated a new method to assign confidence scores to protein-protein interactions and benchmarked our approach against existing methods.

As is often the case in biology, an approach based on cross-species comparisons provides a valuable framework for analyzing protein networks. I used this method to compare the protein networks of the three model eukaryotes and also the parasite *Plasmodium falciparum* in chapters 3 and 4. This approach was very fruitful as I was able to identify regulatory pathways and modules that were conserved between them. In addition, the conserved modules were also used to make new predictions of cellular function and protein interaction. These results are available online along with other modules generated by different algorithms at http://www.cellcircuits.org/[139].

In chapter 5, I used the strategy of network integration in order to understand the functioning of the yeast cell. I combined the knowledge of the protein network with that of eQTL data available for yeast to explain the observed gene expression associations.

6.1 Future directions

Cross-species sequence comparison has been the mainstay of genome analysis[13]. In contrast cross-species network analysis is still in its infancy. The method of NetworkBLAST suggested in chapter 3 compares protein networks across three species successfully. However, this method becomes computationally intensive when comparing more species simultaneously. Further, the scoring model used in NetworkBLAST is simple. I envision that better scoring models such as that of Koyuturk *et al.*[140] could be applied which take into account the evolution of protein networks similar to evolutionary models such as the Jukes-Cantor Model used in sequence analysis[141]. Currently, there are few models that describe how protein networks evolve and what principles guide the loss or gain of interactions[142,143]. In particular, Andreas Wagner and colleagues have developed a model of evolution of a single interaction network[143]. Their model includes duplication and divergence events affecting the yeast protein interaction network. Such models of network evolution will be important for identifying regions of the network that are more adaptable to change due to environmental pressure. They may also help in designing better therapeutics by targeting regions of the protein network that are conserved in general categories of pathogens but are absent from the human host.

High-throughput technologies generate large measurements for many different aspects of cellular function. Analyzing them comprehensively using network integration provides a clearer picture of the entire system. Integration of different sources of information has been primarily used for the inference of protein networks. For instance, Jansen *et al.*[50] and Lee *et al.*[144] used a probabilistic method to combine various data types (such as mRNA expression, phylogenetic profiles) to predict new protein interactions in yeast. In these methods, the data types being integrated are all shown to be correlated to true protein interactions. However, the main drawback of these methods is that they assume that each dataset is independent of the others. Therefore, new methods need to be developed that can predict new protein interactions by also accounting for the correlations between the different data types. Another exciting direction for such network integration techniques is to predict new protein interactions which can then be used to direct new large-scale experiments such as yeast two-hybrid.

Network integration of fewer data sources has been used previously to generate models of cellular machinery. For instance, Tan *et al.*[145] integrated protein network and transcription factor-DNA interactions to find protein complexes that were not only regulated by common transcription factors, but also conserved between yeast and fly. I also applied the idea of network integration to identify causal genes by combining yeast eQTL data with protein interactions. With the availability of new sources of data, these methods of integration have to be updated and improved. Finally, one of the main aims of systems biology and bioinformatics, in general, is to identify plausible drug targets for various diseases. The technique of network integration provides an exciting tool to tackle this issue.

7 References

1. Consortium IHG. Initial sequencing and analysis of the human genome. *Nature* 2001;409:860-921.

2. Team CHG. The sequence of the human genome. *Science* 2001;291:1304-51.

3. Hood L. Systems biology: integrating technology, biology, and computation. *Mech Ageing Dev* 2003;124:9-16.

4. Ideker T, Galitski T, Hood L. A new approach to decoding life: systems biology. *Annu Rev Genomics Hum Genet* 2001;2:343-72.

5. Watson A, Mazumder A, Stewart M, Balasubramanian S. Technology for microarray analysis of gene expression. *Curr Opin Biotechnol* 1998;9:609-14.

6. Berger MF, Philippakis AA, Qureshi AM, He FS, Estep PW, 3rd, Bulyk ML. Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. *Nat Biotechnol* 2006;24:1429-35.

7. Harbison CT, Gordon DB, Lee TI, Rinaldi NJ, Macisaac KD, Danford TW, Hannett NM, Tagne JB, Reynolds DB, Yoo J, Jennings EG, Zeitlinger J, Pokholok DK, Kellis M, Rolfe PA, Takusagawa KT, Lander ES, Gifford DK, Fraenkel E, Young RA. Transcriptional regulatory code of a eukaryotic genome. *Nature* 2004;431:99-104.

8. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL. GenBank. *Nucleic Acids Res* 2007;35:D21-5.

9. Chatton B, Bahr A, Acker J, Kedinger C. Eukaryotic GST fusion vector for the study of protein-protein associations in vivo: application to interaction of ATFa with Jun and Fos. *Biotechniques* 1995;18:142-5.

10. Matyus L. Fluorescence resonance energy transfer measurements on cell surfaces. A spectroscopic tool for determining protein interactions. *J Photochem Photobiol B* 1992;12:323-37.

11. Fields S, Song O. A novel genetic system to detect protein-protein interactions. *Nature* 1989;340:245-6.

12. Gavin AC, Bosche M, Krause R, Grandi P, Marzioch M, Bauer A, Schultz J, Rick JM, Michon AM, Cruciat CM, Remor M, Hofert C, Schelder M, Brajenovic M, Ruffner H, Merino A, Klein K, Hudak M, Dickson D, Rudi T, Gnau V, Bauch A, Bastuck S, Huhse B, Leutwein C, Heurtier MA, Copley RR, Edelmann A, Querfurth E, Rybin V, Drewes G, Raida M, Bouwmeester T, Bork P, Seraphin B, Kuster B, Neubauer G, Superti-Furga G. Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* 2002;415:141-7.

13. Sharan R, Ideker T. Modeling cellular machinery through biological network comparison. *Nat Biotechnol* 2006;24:427-33.

14. von Mering C, Krause R, Snel B, Cornell M, Oliver SG, Fields S, Bork P. Comparative assessment of large-scale data sets of protein-protein interactions. *Nature* 2002;417:399-403.

15. Samanta MP, Liang S. Predicting protein functions from redundancies in large-scale protein interaction networks. *Proc Natl Acad Sci U S A* 2003;100:12579-83.

16. Ideker T, Lauffenburger D. Building with a scaffold: emerging strategies for high- to low-level cellular modeling. *Trends Biotechnol* 2003;21:255-62.

17. Ho Y, Gruhler A, Heilbut A, Bader GD, Moore L, Adams SL, Millar A, Taylor P, Bennett K, Boutilier K, Yang L, Wolting C, Donaldson I, Schandorff S, Shewnarane J, Vo M, Taggart J, Goudreault M, Muskat B, Alfarano C, Dewar D, Lin Z, Michalickova K, Willems AR, Sassi H, Nielsen PA, Rasmussen KJ, Andersen JR, Johansen LE, Hansen LH, Jespersen H, Podtelejnikov A, Nielsen E, Crawford J, Poulsen V, Sorensen BD, Matthiesen J, Hendrickson RC, Gleeson F, Pawson T, Moran MF, Durocher D, Mann M, Hogue CW, Figeys D, Tyers M. Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry. *Nature* 2002;415:180-3.

18. Ito T, Chiba T, Ozawa R, Yoshida M, Hattori M, Sakaki Y. A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc Natl Acad Sci U S A* 2001;98:4569-74.

19. LaCount DJ, Vignali M, Chettier R, Phansalkar A, Bell R, Hesselberth JR, Schoenfeld LW, Ota I, Sahasrabudhe S, Kurschner C, Fields S, Hughes RE. A protein interaction network of the malaria parasite Plasmodium falciparum. *Nature* 2005;438:103-7.

20. Uetz P, Giot L, Cagney G, Mansfield TA, Judson RS, Knight JR, Lockshon D, Narayan V, Srinivasan M, Pochart P, Qureshi-Emili A, Li Y, Godwin B, Conover D, Kalbfleisch T, Vijayadamodar G, Yang M, Johnston M, Fields S, Rothberg JM. A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. *Nature* 2000;403:623-7.

21. Bang S, Choi J, Park J, Park SJ. A Hub-protein based Visualization of Large Protein-Protein Interaction Networks. *Conf Proc IEEE Eng Med Biol Soc* 2007;1:1217-20.

22. Hanisch D, Sohler F, Zimmer R. ToPNet--an application for interactive analysis of expression data and biological networks. *Bioinformatics* 2004;20:1470-1.

23. Ho E, Webber R, Wilkins MR. Interactive Three-Dimensional Visualization and Contextual Analysis of Protein Interaction Networks. *J Proteome Res* 2007.

24. Hu Z, Mellor J, Wu J, DeLisi C. VisANT: an online visualization and analysis tool for biological interaction data. *BMC Bioinformatics* 2004;5:17.

25. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* 2003;13:2498-504.

26. Grigoriev A. 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* 2001;29:3513-9.

27. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 1997;25:3389-402.

28. Kelley BP, Sharan R, Karp RM, Sittler T, Root DE, Stockwell BR, Ideker T. Conserved pathways within bacteria and yeast as revealed by global protein network alignment. *Proc Natl Acad Sci U S A* 2003;100:11394-9.

29. Kelley R, Ideker T. Systematic interpretation of genetic interactions using protein networks. *Nat Biotechnol* 2005;23:561-6.

30. Brem RB, Kruglyak L. The landscape of genetic complexity across 5,700 gene expression traits in yeast. *Proc Natl Acad Sci U S A* 2005;102:1572-7.

31. Schadt EE, Lamb J, Yang X, Zhu J, Edwards S, Guhathakurta D, Sieberts SK, Monks S, Reitman M, Zhang C, Lum PY, Leonardson A, Thieringer R, Metzger JM, Yang L, Castle J, Zhu H, Kash SF, Drake TA, Sachs A, Lusis AJ. An integrative genomics approach to infer causal associations between gene expression and disease. *Nat Genet* 2005;37:710-7.

32. Rockman MV, Kruglyak L. Genetics of global gene expression. *Nat Rev Genet* 2006;7:862-72.

33. Yvert G, Brem RB, Whittle J, Akey JM, Foss E, Smith EN, Mackelprang R, Kruglyak L. Trans-acting regulatory variation in Saccharomyces cerevisiae and the role of transcription factors. *Nat Genet* 2003;35:57-64.

34. Lee SI, Pe'er D, Dudley AM, Church GM, Koller D. Identifying regulatory mechanisms using individual variation reveals key role for chromatin modification. *Proc Natl Acad Sci U S A* 2006;103:14062-7.

35. Schadt EE, Lum PY. Thematic review series: systems biology approaches to metabolic and cardiovascular disorders. Reverse engineering gene networks to identify key drivers of complex disease phenotypes. *J Lipid Res* 2006;47:2601-13.

36. Tu Z, Wang L, Arbeitman MN, Chen T, Sun F. An integrative approach for causal gene identification and gene regulatory pathway inference. *Bioinformatics* 2006;22:e489-96.

37. Xenarios I, Salwinski L, Duan XJ, Higney P, Kim SM, Eisenberg D. DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions. *Nucleic Acids Res* 2002;30:303-5.

38. Mishra GR, Suresh M, Kumaran K, Kannabiran N, Suresh S, Bala P, Shivakumar K, Anuradha N, Reddy R, Raghavan TM, Menon S, Hanumanthu G, Gupta M, Upendran S, Gupta S, Mahesh M, Jacob B, Mathew P, Chatterjee P, Arun KS, Sharma S, Chandrika KN, Deshpande N, Palvankar K, Raghavnath R, Krishnakanth R, Karathia H, Rekha B, Nayak R, Vishnupriya G, Kumar HG, Nagini M, Kumar GS, Jose R, Deepthi P, Mohan SS, Gandhi TK, Harsha HC, Deshpande KS, Sarker M, Prasad TS, Pandey A. Human protein reference database--2006 update. *Nucleic Acids Res* 2006;34:D411-4.

39. Giot L, Bader JS, Brouwer C, Chaudhuri A, Kuang B, Li Y, Hao YL, Ooi CE, Godwin B, Vitols E, Vijayadamodar G, Pochart P, Machineni H, Welsh M, Kong Y, Zerhusen B, Malcolm R, Varrone Z, Collis A, Minto M, Burgess S, McDaniel L, Stimpson E, Spriggs F, Williams J, Neurath K, Ioime N, Agee M, Voss E, Furtak K, Renzulli R, Aanensen N, Carrolla S, Bickelhaupt E, Lazovatsky Y, DaSilva A, Zhong J, Stanyon CA, Finley RL, Jr., White KP, Braverman M, Jarvie T, Gold S, Leach M, Knight J, Shimkets RA, McKenna MP, Chant J, Rothberg JM. A protein interaction map of Drosophila melanogaster. *Science* 2003;302:1727-36.

40. Li S, Armstrong CM, Bertin N, Ge H, Milstein S, Boxem M, Vidalain PO, Han JD, Chesneau A, Hao T, Goldberg DS, Li N, Martinez M, Rual JF, Lamesch P, Xu L, Tewari M, Wong SL, Zhang LV, Berriz GF, Jacotot L, Vaglio P, Reboul J, Hirozane-Kishikawa T, Li Q, Gabel HW, Elewa A, Baumgartner B, Rose DJ, Yu H, Bosak S, Sequerra R, Fraser A, Mango SE, Saxton WM, Strome S, Van Den Heuvel S, Piano F, Vandenhaute J, Sardet C, Gerstein M, Doucette-Stamm L, Gunsalus KC, Harper JW, Cusick ME, Roth FP, Hill DE, Vidal M. A map of the interactome network of the metazoan C. elegans. *Science* 2004;303:540-3.

41. Rual JF, Venkatesan K, Hao T, Hirozane-Kishikawa T, Dricot A, Li N, Berriz GF, Gibbons FD, Dreze M, Ayivi-Guedehoussou N, Klitgord N, Simon C, Boxem M, Milstein S, Rosenberg J, Goldberg DS, Zhang LV, Wong SL, Franklin G, Li S, Albala JS, Lim J, Fraughton C, Llamosas E, Cevik S, Bex C, Lamesch P, Sikorski RS, Vandenhaute J, Zoghbi HY, Smolyar A, Bosak S, Sequerra R, Doucette-Stamm L, Cusick ME, Hill DE, Roth FP, Vidal M. Towards a proteome-scale map of the human protein-protein interaction network. *Nature* 2005;437:1173-8.

42. Stelzl U, Worm U, Lalowski M, Haenig C, Brembeck FH, Goehler H, Stroedicke M, Zenkner M, Schoenherr A, Koeppen S, Timm J, Mintzlaff S, Abraham C, Bock N, Kietzmann S, Goedde A, Toksoz E, Droege A, Krobitsch S, Korn B, Birchmeier W, Lehrach H, Wanker EE. A human protein-protein interaction network: a resource for annotating the proteome. *Cell* 2005;122:957-68.

43. Deane CM, Salwinski L, Xenarios I, Eisenberg D. Protein interactions: two methods for assessment of the reliability of high throughput observations. *Mol Cell Proteomics* 2002;1:349-56.

44. Deng M, Sun F, Chen T. Assessment of the reliability of protein-protein interactions and protein function prediction. *Pac Symp Biocomput* 2003:140-51.

45. Bader JS, Chaudhuri A, Rothberg JM, Chant J. Gaining confidence in high-throughput protein interaction networks. *Nat Biotechnol* 2004;22:78-85.

46. Goldberg DS, Roth FP. Assessing experimentally derived interactions in a small world. *Proc Natl Acad Sci U S A* 2003;100:4372-6.

47. Marcotte EM, Pellegrini M, Thompson MJ, Yeates TO, Eisenberg D. A combined algorithm for genome-wide prediction of protein function. *Nature* 1999;402:83-6.

48. von Mering C, Jensen LJ, Snel B, Hooper SD, Krupp M, Foglierini M, Jouffre N, Huynen MA, Bork P. STRING: known and predicted protein-protein associations, integrated and transferred across organisms. *Nucleic Acids Res* 2005;33:D433-7.

49. Myers CL, Robson D, Wible A, Hibbs MA, Chiriac C, Theesfeld CL, Dolinski K, Troyanskaya OG. Discovery of biological networks from diverse functional genomic data. *Genome Biol* 2005;6:R114.

50. Jansen R, Yu H, Greenbaum D, Kluger Y, Krogan NJ, Chung S, Emili A, Snyder M, Greenblatt JF, Gerstein M. A Bayesian networks approach for predicting protein-protein interactions from genomic data. *Science* 2003;302:449-53.

51. Qi Y, Klein-Seetharaman J, Bar-Joseph Z. Random forest similarity for protein-protein interaction prediction from multiple sources. *Pac Symp Biocomput* 2005:531-42.

52. Qi Y, Bar-Joseph Z, Klein-Seetharaman J. Evaluation of different biological data and computational classification methods for use in protein interaction prediction. *Proteins* 2006.

53. Suthram S, Sittler T, Ideker T. The Plasmodium protein network diverges from those of other eukaryotes. *Nature* 2005;438:108-12.

54. Sharan R, Suthram S, Kelley RM, Kuhn T, McCuine S, Uetz P, Sittler T, Karp RM, Ideker T. Conserved patterns of protein interaction in multiple species. *Proc Natl Acad Sci U S A* 2005;102:1974-9.

55. Matlab. MATLAB Software.

56. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* 2000;25:25-9.

57. Mewes HW, Albermann K, Heumann K, Liebl S, Pfeiffer F. MIPS: a database for protein sequences, homology data and yeast genome information. *Nucleic Acids Res* 1997;25:28-30.

58. http://*chianti.ucsd.edu/Suthram2006* SW.

59. Pagel P, Mewes HW, Frishman D. Conservation of protein-protein interactions - lessons from ascomycota. *Trends Genet* 2004;20:72-6.

60. Yu H, Luscombe NM, Lu HX, Zhu X, Xia Y, Han JD, Bertin N, Chung S, Vidal M, Gerstein M. Annotation transfer between genomes: protein-protein interologs and protein-DNA regulogs. *Genome Res* 2004;14:1107-18.

61. Ball CA, Awad IA, Demeter J, Gollub J, Hebert JM, Hernandez-Boussard T, Jin H, Matese JC, Nitzberg M, Wymore F, Zachariah ZK, Brown PO, Sherlock G. The Stanford Microarray Database accommodates additional microarray platforms and data formats. *Nucleic Acids Res* 2005;33:D580-2.

62. Bader GD, Hogue CW. An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinformatics* 2003;4:2.

63. Hu H, Yan X, Huang Y, Han J, Zhou XJ. Mining coherent dense subgraphs across massive biological networks for functional discovery. *Bioinformatics* 2005;21 Suppl 1:i213-i221.

64. Spirin V, Mirny LA. Protein complexes and functional modules in molecular networks. *Proc Natl Acad Sci U S A* 2003;100:12123-8.

65. Fraser HB, Hirsh AE, Steinmetz LM, Scharfe C, Feldman MW. Evolutionary rate in the protein interaction network. *Science* 2002;296:750-2.

66. Sindhwani V, Rakshit S, Deodhare D, Erdogmus D, Principe JC, Niyogi P. Feature selection in MLPs and SVMs based on maximum output information. *IEEE Trans Neural Netw* 2004;15:937-48.

67. Christie KR, Weng S, Balakrishnan R, Costanzo MC, Dolinski K, Dwight SS, Engel SR, Feierbach B, Fisk DG, Hirschman JE, Hong EL, Issel-Tarver L, Nash R, Sethuraman A, Starr B, Theesfeld CL, Andrada R, Binkley G, Dong Q, Lane C, Schroeder M, Botstein D, Cherry JM. Saccharomyces Genome Database (SGD) provides tools to identify and analyze sequences from Saccharomyces cerevisiae and related sequences from other organisms. *Nucleic Acids Res* 2004;32:D311-4.

68. Shanmugan KS. Digital and analog communication systems. New York: Wiley, 1979:xviii, 600.

69. Aebersold R, Mann M. Mass spectrometry-based proteomics. *Nature* 2003;422:198-207.

70. Rain JC, Selig L, De Reuse H, Battaglia V, Reverdy C, Simon S, Lenzen G, Petel F, Wojcik J, Schachter V, Chemama Y, Labigne A, Legrain P. The protein-protein interaction map of Helicobacter pylori. *Nature* 2001;409:211-5.

71. Matthews LR, Vaglio P, Reboul J, Ge H, Davis BP, Garrels J, Vincent S, Vidal M. Identification of potential interaction networks using sequence-based searches for conserved protein-protein interactions or "interologs". *Genome Res* 2001;11:2120-6.

72. Sharan R, Ideker T, Kelley B, Shamir R, Karp RM. Identification of protein complexes by comparative analysis of yeast and bacterial protein interaction data. *J Comput Biol* 2005;12:835-46.

73. Dandekar T, Schuster S, Snel B, Huynen M, Bork P. Pathway alignment: application to the comparative analysis of glycolytic enzymes. *Biochem J* 1999;343:115-24.

74. Forst CV, Schulten K. Evolution of metabolisms: a new method for the comparison of metabolic pathways using genomics information. *J Comput Biol* 1999;6:343-60.

75. Ogata H, Fujibuchi W, Goto S, Kanehisa M. A heuristic graph comparison algorithm and its application to detect functionally related enzyme clusters. *Nucleic Acids Res* 2000;28:4021-8.

76. Bergmann S, Ihmels J, Barkai N. Similarities and differences in genome-wide expression data of six organisms. *PLoS Biol* 2004;2:E9.

77. Enard W, Khaitovich P, Klose J, Zollner S, Heissig F, Giavalisco P, Nieselt-Struwe K, Muchmore E, Varki A, Ravid R, Doxiadis GM, Bontrop RE, Paabo S. Intra- and interspecific variation in primate gene expression patterns. *Science* 2002;296:340-3.

78. Stuart JM, Segal E, Koller D, Kim SK. A gene-coexpression network for global discovery of conserved genetic modules. *Science* 2003;302:249-55.

79. Tanay A, Sharan R, Kupiec M, Shamir R. Revealing modularity and organization in the yeast molecular network by integrated analysis of highly heterogeneous genomewide data. *Proc Natl Acad Sci U S A* 2004;101:2981-6.

80. Harris TW, Chen N, Cunningham F, Tello-Ruiz M, Antoshechkin I, Bastiani C, Bieri T, Blasiar D, Bradnam K, Chan J, Chen CK, Chen WJ, Davis P, Kenny E, Kishore R, Lawson D, Lee R, Muller HM, Nakamura C, Ozersky P, Petcherski A, Rogers A, Sabo A, Schwarz EM, Van Auken K, Wang Q, Durbin R, Spieth J, Sternberg PW, Stein LD. WormBase: a multi-species resource for nematode biology and genomics. *Nucleic Acids Res* 2004;32 Database issue:D411-7.

81. FlybaseConsortium. The FlyBase database of the Drosophila genome projects and community literature. *Nucleic Acids Res* 2003;31:172-5.

82. Xie H, Wasserman A, Levine Z, Novik A, Grebinskiy V, Shoshan A, Mintz L. Large-scale protein annotation through gene ontology. *Genome Res* 2002;12:785-94.

83. Gautier-Bert K, Murol B, Jarrousse AS, Ballut L, Badaoui S, Petit F, Schmid HP. Substrate affinity and substrate specificity of proteasomes with RNase activity. *Mol Biol Rep* 2003;30:1-7.

84. Schmid HP, Pouch MN, Petit F, Dadet MH, Badaoui S, Boissonnet G, Buri J, Norris V, Briand Y. Relationships between proteasomes and RNA. *Mol Biol Rep* 1995;21:43-7.

85. Yu TW, Bargmann CI. Dynamic regulation of axon guidance. *Nat Neurosci* 2001;4 Suppl:1169-76.

86. Spira ME, Oren R, Dormann A, Ilouz N, Lev S. Calcium, protease activation, and cytoskeleton remodeling underlie growth cone formation and neuronal regeneration. *Cell Mol Neurobiol* 2001;21:591-604.

87. T.Kamada, S.Kawai. An algorithm for drawing general undirected graphs. *Information Processing Letters* 1989;31:7-15.

88. Conant GC, Wagner A. Convergent evolution of gene circuits. *Nat Genet* 2003;34:264-6.

89. Miller LH, Baruch DI, Marsh K, Doumbo OK. The pathogenic basis of malaria. *Nature* 2002;415:673-9.

90. Gardner MJ, Hall N, Fung E, White O, Berriman M, Hyman RW, Carlton JM, Pain A, Nelson KE, Bowman S, Paulsen IT, James K, Eisen JA, Rutherford K, Salzberg SL, Craig A, Kyes S, Chan MS, Nene V, Shallom SJ, Suh B, Peterson J, Angiuoli S, Pertea M, Allen J, Selengut J, Haft D, Mather MW, Vaidya AB, Martin DM, Fairlamb AH, Fraunholz MJ, Roos DS, Ralph SA, McFadden GI, Cummings LM, Subramanian GM, Mungall C, Venter JC, Carucci DJ, Hoffman SL, Newbold C, Davis RW, Fraser CM, Barrell B. Genome sequence of the human malaria parasite Plasmodium falciparum. *Nature* 2002;419:498-511.

91. Bozdech Z, Llinas M, Pulliam BL, Wong ED, Zhu J, DeRisi JL. The transcriptome of the intraerythrocytic developmental cycle of Plasmodium falciparum. *PLoS Biol* 2003;1:E5.

92. Le Roch KG, Zhou Y, Blair PL, Grainger M, Moch JK, Haynes JD, De La Vega P, Holder AA, Batalov S, Carucci DJ, Winzeler EA. Discovery of gene function by expression profiling of the malaria parasite life cycle. *Science* 2003;301:1503-8.

93. Florens L, Washburn MP, Raine JD, Anthony RM, Grainger M, Haynes JD, Moch JK, Muster N, Sacci JB, Tabb DL, Witney AA, Wolters D, Wu Y, Gardner MJ, Holder AA, Sinden RE, Yates JR, Carucci DJ. A proteomic view of the Plasmodium falciparum life cycle. *Nature* 2002;419:520-6.

94. Fraunholz MJ, Roos DS. PlasmoDB: exploring genomics and post-genomics data of the malaria parasite, Plasmodium falciparum. *Redox Rep* 2003;8:317-20.

95. Chen N, Harris TW, Antoshechkin I, Bastiani C, Bieri T, Blasiar D, Bradnam K, Canaran P, Chan J, Chen CK, Chen WJ, Cunningham F, Davis P, Kenny E, Kishore R, Lawson D, Lee R, Muller HM, Nakamura C, Pai S, Ozersky P, Petcherski A, Rogers A, Sabo A, Schwarz EM, Van Auken K, Wang Q, Durbin R, Spieth J, Sternberg PW, Stein LD. WormBase: a comprehensive data resource for Caenorhabditis biology and genomics. *Nucleic Acids Res* 2005;33 Database Issue:D383-9.

96. (www.tigr.org) T.

97. Felsenstein J. PHYLIP - Phylogeny Inference Package (Version 3.2). *Cladistics* 1989;5:164-166.

98. Barabasi AL, Oltvai ZN. Network biology: understanding the cell's functional organization. *Nat Rev Genet* 2004;5:101-13.

99. Gagny B, Wiederkehr A, Dumoulin P, Winsor B, Riezman H, Haguenauer-Tsapis R. A novel EH domain protein of Saccharomyces cerevisiae, Ede1p, involved in endocytosis. *J Cell Sci* 2000;113 (Pt 18):3309-19.

100. Engqvist-Goldstein AE, Drubin DG. Actin assembly and endocytosis: from yeast to mammals. *Annu Rev Cell Dev Biol* 2003;19:287-332.

101. Goodson HV, Anderson BL, Warrick HM, Pon LA, Spudich JA. Synthetic lethality screen identifies a novel yeast myosin I gene (MYO5): myosin I proteins are required for polarization of the actin cytoskeleton. *J Cell Biol* 1996;133:1277-91.

102. Salisbury JL, Condeelis JS, Maihle NJ, Satir P. Calmodulin localization during capping and receptor-mediated endocytosis. *Nature* 1981;294:163-6.

103. Scheibel LW, Colombani PM, Hess AD, Aikawa M, Atkinson CT, Milhous WK. Calcium and calmodulin antagonists inhibit human malaria parasites

(Plasmodium falciparum): implications for drug design. *Proc Natl Acad Sci U S A* 1987;84:7310-4.

104. Sanchez CP, McLean JE, Stein W, Lanzer M. Evidence for a substrate specific and inhibitable drug efflux system in chloroquine resistant Plasmodium falciparum strains. *Biochemistry* 2004;43:16365-73.

105. Hoppe HC, van Schalkwyk DA, Wiehart UI, Meredith SA, Egan J, Weber BW. Antimalarial quinolines and artemisinin inhibit endocytosis in Plasmodium falciparum. *Antimicrob Agents Chemother* 2004;48:2370-8.

106. Mori K. Frame switch splicing and regulated intramembrane proteolysis: key words to understand the unfolded protein response. *Traffic* 2003;4:519-28.

107. Baxter BK, James P, Evans T, Craig EA. SSI1 encodes a novel Hsp70 of the Saccharomyces cerevisiae endoplasmic reticulum. *Mol Cell Biol* 1996;16:6444-56.

108. Langst G, Becker PB. Nucleosome mobilization and positioning by ISWI-containing chromatin-remodeling factors. *J Cell Sci* 2001;114:2561-8.

109. Jacobson S, Pillus L. Modifying chromatin and concepts of cancer. *Curr Opin Genet Dev* 1999;9:175-84.

110. Johnson D, Gunther K, Ansorge I, Benting J, Kent A, Bannister L, Ridley R, Lingelbach K. Characterization of membrane proteins exported from Plasmodium falciparum into the host erythrocyte. *Parasitology* 1994;109 (Pt 1):1-9.

111. Mattei D, Scherf A. The Pf332 gene of Plasmodium falciparum codes for a giant protein that is translocated from the parasite to the membrane of infected erythrocytes. *Gene* 1992;110:71-9.

112. Finn RD, Marshall M, Bateman A. iPfam: visualization of protein-protein interactions in PDB at domain and amino acid resolutions. *Bioinformatics* 2005;21:410-2.

113. Griffiths AJF. Modern genetic analysis : integrating genes and genomes. New York: W.H. Freeman and Co., 2002:xix, 736 p.

114. Kulp DC, Jagalur M. Causal inference of regulator-target pairs by gene mapping of expression phenotypes. *BMC Genomics* 2006;7:125.

115. Perez-Enciso M, Quevedo JR, Bahamonde A. Genetical genomics: use all data. *BMC Genomics* 2007;8:69.

116. Beyer A, Bandyopadhyay S, Ideker T. Integrating physical and genetic maps: from genomes to interaction networks. *Nat Rev Genet* 2007;8:699-710.

117. Doyle PG, Snell JL. Random walks and electric networks. [Washington, D.C.]: Mathematical Association of America, 1984:xiii, 159 p.

118. Albert-László Barabási RA. Emergence of Scaling in Random Networks. *Science* 1999;286:509 - 512.

119. Hughes TR, Marton MJ, Jones AR, Roberts CJ, Stoughton R, Armour CD, Bennett HA, Coffey E, Dai H, He YD, Kidd MJ, King AM, Meyer MR, Slade D, Lum PY, Stepaniants SB, Shoemaker DD, Gachotte D, Chakraburtty K, Simon J, Bard M, Friend SH. Functional discovery via a compendium of expression profiles. *Cell* 2000;102:109-26.

120. Hu Z, Killion PJ, Iyer VR. Genetic reconstruction of a functional transcriptional regulatory network. *Nat Genet* 2007;39:683-687.

121. Chua G, Morris QD, Sopko R, Robinson MD, Ryan O, Chan ET, Frey BJ, Andrews BJ, Boone C, Hughes TR. Identifying transcription factor functions and targets by phenotypic activation. *Proc Natl Acad Sci U S A* 2006;103:12045-50.

122. Kanehisa M. The KEGG database. *Novartis Found Symp* 2002;247:91-101; discussion 101-3, 119-28, 244-52.

123. Ptacek J, Devgan G, Michaud G, Zhu H, Zhu X, Fasolo J, Guo H, Jona G, Breitkreutz A, Sopko R, McCartney RR, Schmidt MC, Rachidi N, Lee SJ, Mah AS, Meng L, Stark MJ, Stern DF, De Virgilio C, Tyers M, Andrews B, Gerstein M, Schweitzer B, Predki PF, Snyder M. Global analysis of protein phosphorylation in yeast. *Nature* 2005;438:679-84.

124. Dixon AL, Liang L, Moffatt MF, Chen W, Heath S, Wong KC, Taylor J, Burnett E, Gut I, Farrall M, Lathrop GM, Abecasis GR, Cookson WO. A genome-wide association study of global gene expression. *Nat Genet* 2007;39:1202-7.

125. Goring HH, Curran JE, Johnson MP, Dyer TD, Charlesworth J, Cole SA, Jowett JB, Abraham LJ, Rainwater DL, Comuzzie AG, Mahaney MC, Almasy L, MacCluer JW, Kissebah AH, Collier GR, Moses EK, Blangero J. Discovery of expression QTLs using large-scale transcriptional profiling in human lymphocytes. *Nat Genet* 2007;39:1208-16.

126. Stranger BE, Nica AC, Forrest MS, Dimas A, Bird CP, Beazley C, Ingle CE, Dunning M, Flicek P, Koller D, Montgomery S, Tavare S, Deloukas P, Dermitzakis ET. Population genomics of human gene expression. *Nat Genet* 2007;39:1217-24.

127. Mathivanan S, Periaswamy B, Gandhi TK, Kandasamy K, Suresh S, Mohmood R, Ramachandra YL, Pandey A. An evaluation of human protein-protein interaction data in the public domain. *BMC Bioinformatics* 2006;7 Suppl 5:S19.

128. Christos Faloutsos KSM, Andrew Tomkins. Fast discovery of connection subgraphs. *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining* 2004:118 - 127

129. Newman MEJ. A measure of betweenness centrality based on random walks. *elsevier/Social Networks* 2005;27:39-54.

130. Irwin JD, Wu C-H. Basic engineering circuit analysis. Upper Saddle River, N.J.: Prentice-Hall, 1999:xiii, 976 p.

131. Zhao K, Aranzana MJ, Kim S, Lister C, Shindo C, Tang C, Toomajian C, Zheng H, Dean C, Marjoram P, Nordborg M. An Arabidopsis example of association mapping in structured samples. *PLoS Genet* 2007;3:e4.

132. Collins SR, Schuldiner M, Krogan NJ, Weissman JS. A strategy for extracting and analyzing large-scale quantitative epistatic interaction data. *Genome Biol* 2006;7:R63.

133. Ye P, Peyser BD, Spencer FA, Bader JS. Commensurate distances and similar motifs in genetic congruence and protein interaction networks in yeast. *BMC Bioinformatics* 2005;6:270.

134. Krogan NJ, Cagney G, Yu H, Zhong G, Guo X, Ignatchenko A, Li J, Pu S, Datta N, Tikuisis AP, Punna T, Peregrin-Alvarez JM, Shales M, Zhang X, Davey M, Robinson MD, Paccanaro A, Bray JE, Sheung A, Beattie B, Richards DP, Canadien V, Lalev A, Mena F, Wong P, Starostine A, Canete MM, Vlasblom J, Wu S, Orsi C, Collins SR, Chandran S, Haw R, Rilstone JJ, Gandi K, Thompson NJ, Musso G, St Onge P, Ghanny S, Lam MH, Butland G, Altaf-Ul AM, Kanaya S, Shilatifard A, O'Shea E, Weissman JS, Ingles CJ, Hughes TR, Parkinson J, Gerstein M, Wodak SJ, Emili A, Greenblatt JF. Global landscape of protein complexes in the yeast Saccharomyces cerevisiae. *Nature* 2006;440:637-43.

135. Beyer A, Workman C, Hollunder J, Radke D, Moller U, Wilhelm T, Ideker T. Integrated assessment and prediction of transcription factor binding. *PLoS Comput Biol* 2006;2:e70.

136. Mosek P. Mosek Package.

137. Tong AH, Lesage G, Bader GD, Ding H, Xu H, Xin X, Young J, Berriz GF, Brost RL, Chang M, Chen Y, Cheng X, Chua G, Friesen H, Goldberg DS, Haynes J, Humphries C, He G, Hussein S, Ke L, Krogan N, Li Z, Levinson JN, Lu H, Menard P, Munyana C, Parsons AB, Ryan O, Tonikian R, Roberts T, Sdicu AM, Shapiro J, Sheikh B, Suter B, Wong SL, Zhang LV, Zhu H, Burd CG, Munro S, Sander C, Rine J, Greenblatt J, Peter M, Bretscher A, Bell G, Roth FP, Brown GW, Andrews B, Bussey H, Boone C. Global mapping of the yeast genetic interaction network. *Science* 2004;303:808-13.

138. Scott LJ, Mohlke KL, Bonnycastle LL, Willer CJ, Li Y, Duren WL, Erdos MR, Stringham HM, Chines PS, Jackson AU, Prokunina-Olsson L, Ding CJ, Swift AJ, Narisu N, Hu T, Pruim R, Xiao R, Li XY, Conneely KN, Riebow NL, Sprau AG, Tong M, White PP, Hetrick KN, Barnhart MW, Bark CW, Goldstein JL, Watkins L, Xiang F, Saramies J, Buchanan TA, Watanabe RM, Valle TT, Kinnunen L, Abecasis GR, Pugh EW, Doheny KF, Bergman RN, Tuomilehto J, Collins FS, Boehnke M. A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants. *Science* 2007;316:1341-5.

139. Mak HC, Daly M, Gruebel B, Ideker T. CellCircuits: a database of protein network models. *Nucleic Acids Res* 2007;35:D538-45.

140. Koyuturk M, Kim Y, Topkara U, Subramaniam S, Szpankowski W, Grama A. Pairwise alignment of protein interaction networks. *J Comput Biol* 2006;13:182-99.

141. Jukes TH, Cantor CR. Evolution of protien molecules. *Mammalian protein metabolism* 1969:21-132.

142. Teichmann SA, Babu MM. Gene regulatory network growth by duplication. *Nat Genet* 2004;36:492-6.

143. Wagner A. How the global structure of protein interaction networks evolves. *Proc Biol Sci* 2003;270:457-66.

144. Lee I, Date SV, Adai AT, Marcotte EM. A probabilistic functional network of yeast genes. *Science* 2004;306:1555-8.

145. Tan K, Shlomi T, Feizi H, Ideker T, Sharan R. Transcriptional regulation of protein complexes within and across species. *Proc Natl Acad Sci U S A* 2007;104:1283-8.