Context-sensitive Methods for Learning from Genomic Data

Chad L. Myers

A DISSERTATION

PRESENTED TO THE FACULTY

OF PRINCETON UNIVERSITY

IN CANDIDACY FOR THE DEGREE

OF DOCTOR OF PHILOSOPHY

RECOMMENDED FOR ACCEPTANCE

BY THE DEPARTMENT OF

COMPUTER SCIENCE

January 2008

Abstract

Recent developments in biotechnology have enabled high-throughput measurement of several cellular phenomena including gene expression, protein-protein interactions, protein localization, and DNA sequences. The wealth of data generated by this technology promises to support computational prediction of network models, but so far, successful approaches that translate these data into accurate, experimentally testable hypotheses have been limited. This dissertation focuses on machine learning and signal processing approaches that utilize contextual clues often inherent in genomic data to extract useful information and make precise predictions.

First, we describe methods for using microarray technology to detect chromosomal aberrations. Amplification and deletion of portions of chromosomes often serve as a mechanism of rapid adaptation and have been associated with numerous cancers. Accurate and precise identification of when and where these changes occur will help us understand this important adaptive mechanism and is an important step towards effective cancer treatment.

Secondly, we address the more general problem of integrating diverse types of functional genomic data to understand gene function and predict biological networks. We demonstrate that Bayesian methods can leverage unique noise characteristics of genomic data to predict accurate network models. We illustrate the practical use of these methods in a web-based system that supports intelligent exploration of large repositories of noisy genomic data. We have used this system to generate specific hypotheses about previously uncharacterized genes, many of which have been confirmed through experimental validation.

Finally, this dissertation addresses the question of how to use machine learning methods to direct genome-scale experiments. Until now, most bioinformatics methods have been used exclusively downstream of data-generating experiments. Here, we discuss approaches for using computational predictions to actually direct further large-scale experiments. We demonstrate that such approaches can dramatically improve the efficiency with which we use high-throughput genomic technology and, ultimately, help us to discover more novel biology.

Table of Contents

Figures

Tables

Acknowledgements

The work presented in this dissertation was partially supported by funds from the Princeton Program in Integrative Information, Computer and Application Sciences (PICASso) funded by the NSF and a Quantitative and Computational Biology Training Program grant funded by the NIH (T32 HG003284).

There are a number of people who made my graduate school experience both possible and rewarding, all of whom deserve recognition. First, to my advisor, Olga, thank you for taking a chance on an Electrical Engineering graduate student who knew very little biology— I would certainly not be where I am had I not worked with you. Your guidance, constant encouragement, and never-ending advocacy for your students are an inspiration for the mentor I hope to be. To S.Y. Kung, my former advisor, thank you for your support throughout my time at Princeton and for introducing me to the field of bioinformatics. To my dissertation committee, David Botstein, Leonid Kruglyak, S.Y. Kung, Kai Li, and also Steve Kleinstein, thank you for your valuable advice over the past few years and for sending many recommendation letters on my behalf.

To the members of the Troyanskaya Lab, including Matt Hibbs (the other "original" member), Curtis Huttenhower, Patrick Bradley, Maria Chikina, Yuanfang Guan, Florian Markowetz, Edo Airoldi, and David Hess, thanks for making the lab a fun environment, which made work and life in graduate school much more bearable. I am also grateful to several talented undergraduates who worked in the Troyanskaya Lab for their help in implementing various software projects related to my dissertation including Drew Robson, Adam Wible, Xing Chen, Rachel Sealfon, Rajiv Ayyangar, and Jon Ullman. To several experimental biologists in the Lewis-Sigler Institute and at the University of Toronto, including Camelia Chiriac, David Hess, Amy Caudy, Michael Costanzo, Anastasia Baryshnikova, Maitreya Dunham, Kara Dolinski, and David Gresham, thank you for teaching me the biology side of bioinformatics. Many of you also undertook significant experimental work to validate predictions from the methods described in this dissertation for which I am grateful. To several administrative and technical staff in both

Computer Science and the Lewis-Sigler Institute, including Melissa Lawson, Ginny Hogan, Laurie Bellero, Jen Havens, Francine Taylor, Jen Brick, Faith Bahadurian, John Wiggins, and Mark Schroeder, thank you for making my life easier on many different occasions. Finally, I owe thanks to a number of friends from Princeton for good company and good times outside of work including Chris Sadler, Noel and Andrea Eisley, Matt Hibbs, Bryan Patel, Kat Wakabayashi, Phil Lenart, Greg Reeves, Scott McAllister, Melanie Webb, Chris Bristow, Tom O'Connor, and Brigitte Brunelle.

To my family, Mom, Dad, Carissa, Brandon, and the newest members, Ed and Lydia, thank you for being the loving and supportive family that anyone would hope for. You always gave me confidence and encouragement when I needed it the most, and I would not be writing this today without that. To my extended family of grandparents, aunts and uncles, cousins, and in-laws, especially Julie and Ben, Ben Jr., Ross, and Becky, thank you for supporting us through graduate school— spending time with you always puts things into perspective. We appreciate all of those times you acted interested while we carried on about our "exciting" research. Lastly, to Sasha, my wife and best friend, thank you for your unconditional love and support in everything I do. I cannot imagine a better person to have with me on this journey through graduate school or through life.

Chapter 1

Introduction

1.1 Background

Recent advances in biotechnology have enabled us to begin quantitatively characterizing several different aspects of cellular mechanisms and behavior. The impetus for many of these technologies was the introduction of fast sequencing techniques [10], which have led to the sequencing of hundreds of organism over the past ten to fifteen years, including the first human genome sequence in 2001 [14,24]. Whole-genome sequence information spurred the development of several new technologies including DNA microarrays, which can simultaneously read out the expression of all genes in an organism's genome given a sample of cells (e.g. a tumor specimen) (reviewed in [10]). Other technology has enabled high-throughput investigation protein-protein interactions, which are at the heart of the mechanisms behind most cellular processes. Still other approaches can precisely identify the location of proteins in different cellular compartments, giving us hints about their function. In all, current technologies are able to capture cellular information at a variety of complementary levels from the genetic identify of a cell to a cell's diverse responses to environmental stimuli.

These data promise to revolutionize our understanding of core cellular processes and gene function. However, harnessing the rich information present in these large, often noisy, datasets is challenging. This dissertation approaches this problem from a computational perspective. Specifically, we explore strategies for extracting accurate biological hypotheses from high-throughput data and for integrating several diverse data sources to reveal a holistic view of cellular mechanisms. The rest of this chapter is organized as follows.

First, we give a brief overview of the types of genomic data available, how these data are measured, and what we hope to learn from them. Second, we discuss a few of the key challenges in developing methods for analyzing and integrating these data, which motivate

recurring themes throughout this dissertation. Finally, we describe the overall flow of the dissertation, including a brief summary of each chapter.

1.1.1 Gene Expression

Perhaps the most revolutionizing and certainly most abundantly used genomic technology over the past ten years is the DNA microarray. Microarrays enable biologists to simultaneously measure the expression of tens of thousands of genes on a single chip [3]. Briefly, the way microarray technology works is that short probe sequences matching the mRNA of genes of interest are either spotted or printed onto a glass or silicon slide (Figure 1.1).

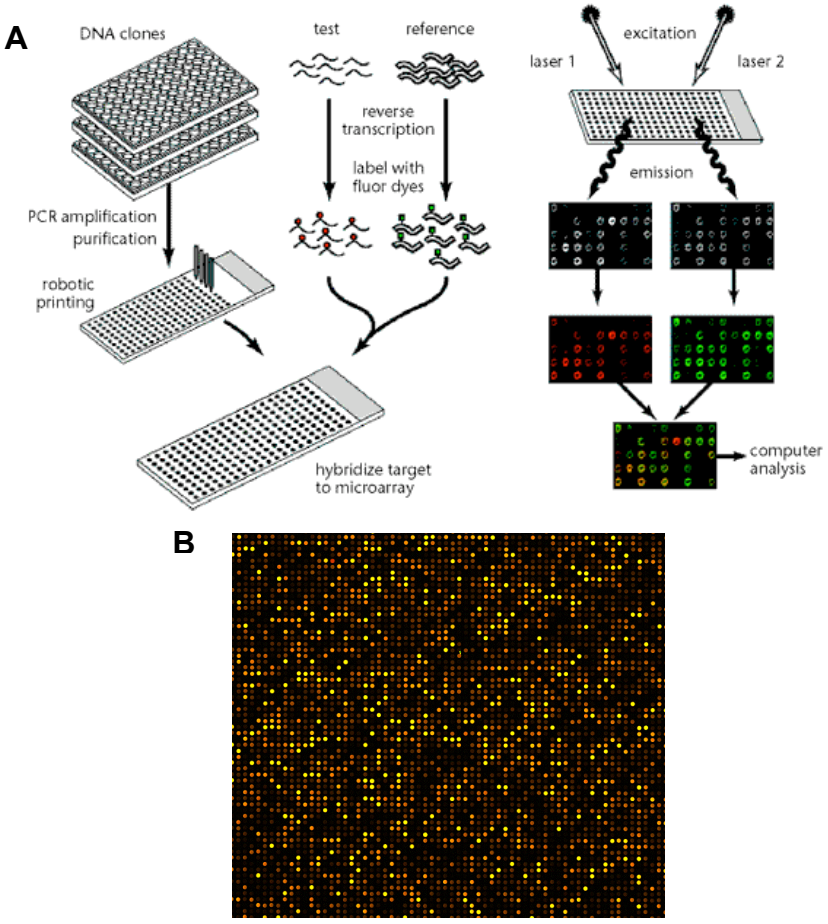


Figure 1.1. Schematic of DNA microarray technology. (A) Test and reference cDNA or cRNA samples are differentially labeled with Cy3 and Cy5 flourescent dye and hybrized to an array of target sequences [5]. Relative intensities at each set of probe sequences are then measured through excitation with a laser. (B) A raw image of microarray data from an Agilent *Saccharomyces cerevisiae* array of an *hsp82Δ* deletion mutant at 37 ^{o}C.

For two-color microarrays, cDNA or cRNA are prepared from a test and reference sample (e.g. diseased and normal tissue) and are differentially labeled with Cy3 and Cy5. The samples are mixed and hybridized to the array. The relative levels of mRNA are then inferred by measuring the relative intensities of Cy3 and Cy5 emission wavelengths upon excitation with a laser. Gene expression data is typically interpreted in the form of log-ratios of these intensities [20]. Another commonly used variation of microarrays is the single-channel array, which is based on the same principle of hybridization to target sequences. However, rather then measuring relative intensities, single-channel arrays attempt to directly measure absolute mRNA levels through the use of specific sequence variants and spike-in controls [15]. Microarray technology has rapidly evolved since its invention in 1995 [21], including improvements in both quality and the number of transcripts that can be measured at once. The technology has been rapidly adopted by the genomics community— as of October 2007, there were 6744 different microarray studies deposited in NCBI's Gene Expression Omnibus database [6]. Gene expression profiles measured over a variety of conditions capture the cell's programmed response to external stimuli, and thus, are a valuable source of functional information as discussed in detail in Chapters 4-6.

1.1.2 DNA Copy Number and Sequence Variation

Genomic sequence and copy number variation play a major role in susceptibility to disease and, in particular, several cancers. Variations of microarray technology have been used quite successfully to construct high-resolution maps of this variation. One specific technology discussed extensively in Chapters 2 and 3 is array comparative genomic hybridization (array CGH). The basis for the technology is the same as for expression microarrays, but instead of hybridizing representations of mRNA to an array of probe sequences, genomic DNA from a test and reference sample is prepared and hybridized [18]. Chromosomal aberrations relative to the reference sample are then inferred from stretches of adjacent probes with significantly higher or lower relative hybridization intensity. Using the latest versions of this technology, some studies have reported detecting even single copy number changes on the order of several hundred to

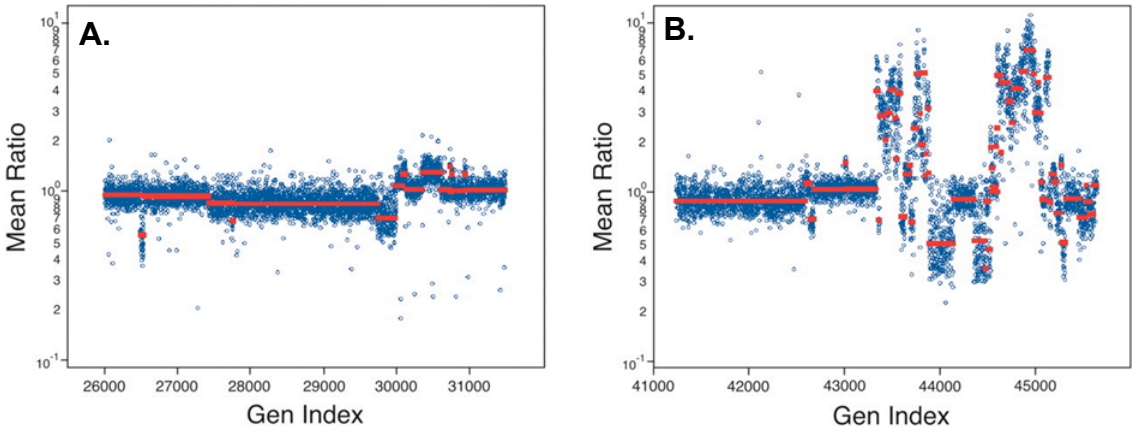


Figure 1.2. ROMA array CGH analysis for two abnormal chromosomes. ROMA oligonucleotide-based CGH analysis for chromosome 17 (A) and the X chromosome (B) of the tumor cell line SK-BR-3 [16]. The Y-axis plots the mean ratio of two hybridizations in log-scale and the X-axis is organized in genome order. Copy number polymorphisms appear as sustained spikes or troughs along each chromosome.

several thousand bases [18]. Figure 1.2 illustrates data from one existing high-density array CGH technique, Representational Oligonucleotide Microarray Analysis (ROMA), where two different chromosomes with various abnormalities have been identified [16].

Another promising application of microarray technology has been the detection of a smaller type of sequence variation, single nucleotide polymorphisms, or SNPs. Microarray technologies for detecting SNPs are often designed in a targeted fashion, where individual probes are specifically constructed to center around a suspected polymorphism [12]. Genomic DNA fragments with mismatches at the location of interest exhibit less hybridization efficiency and are thus detected as altered. Some groups have also demonstrated promising approaches for unbiased, global identification of SNPs based on whole-genome tiling microarrays, in which microarray probes are designed to cover the entire genome with short overlapping sequences [9].

1.1.3 Protein-protein Interactions

Another major focus of recent high-throughput genomic technology has been the identification of protein-protein interactions. Physical interactions between proteins are crucial for most cellular functions. For instance, protein-protein interactions are the means of signal transduction by which a cell receives information about its external environment. Furthermore, the very structure

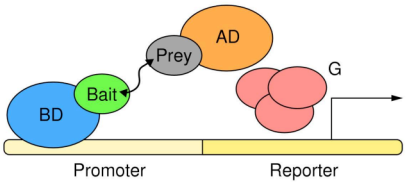


Figure 1.3. Yeast two-hybrid assay for detecting physical interactions. Interactions between a bait and prey protein are detected by fusing them to a DNA-binding domain and an activation domain, which when bound, activate transcription of a reporter gene [8].

of the cell and all of its machinery are formed through protein interactions [22]. We highlight two recent experimental techniques for capturing interactions that are responsible for the majority of available data.

Yeast two-hybrid

Yeast two-hybrid is a technology for specifically interrogating physical interactions between pairs of proteins. Two-hybrid relies on the two-domain structure of eukaryotic transcription factors to report an interaction. In eukaryotes, transcription factors bind short DNA sequences upstream, and recruit RNA polymerase to initiate transcription [1]. A two-hybrid positive interaction is obtained by fusing one protein to a DNA-binding domain (bait) while another protein is fused to an activation domain (prey) such that binding of the two proteins of interest in the nucleus "switches on" transcription of a reporter gene [17]. The power of the two-hybrid approach is that it can be used to efficiently query thousands of interactions in an unbiased manner. However, two-hybrid is notoriously known for its high rate of false positive interactions [22]. Dealing with this inherent noise is one of the challenges addressed by the methods described in Chapters 4-6.

Tandem affinity purification-mass spectrometry (TAP-MS)

Another promising technique for identifying protein-protein interactions in high-throughput is tandem affinity purification followed by mass spectrometry identification. The basis of the approach is to integrate a TAP tag into the open reading frame (ORF) of a target protein. The TAP tag consists of two IgG binding domains of *Staphylococcus aureus* protein A (ProtA) and a calmodulin binding peptide (CBP) separated by a TEV protease cleavage site [19] (Figure 1.4). The target protein and its interaction partners are then purified through two steps. First, the IgG domains bind with high affinity to a IgG matrix while contaminants are washed off. Then, the TEV

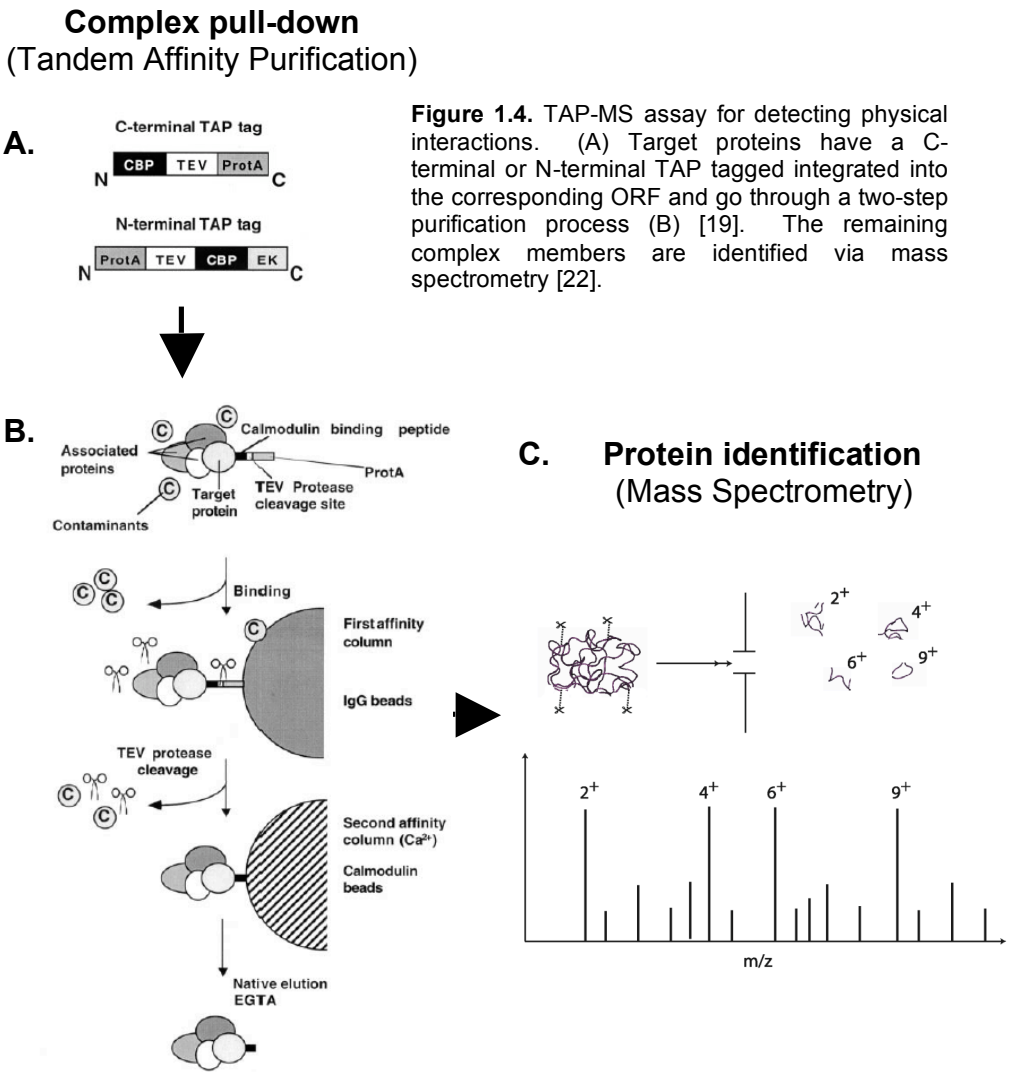


Figure 1.4. TAP-MS assay for detecting physical interactions. (A) Target proteins have a C-terminal or N-terminal TAP tagged integrated into the corresponding ORF and go through a two-step purification process (B) [19]. The remaining complex members are identified via mass spectrometry [22].

prot

ease is used to cut the the IgG domains free, and the eluate is incubated in calmodulin-coated beads in the presence of calcium [19]. The final result is the initial target protein and any interacting pairs, free from all contaminants.

Given a complex purified from the steps described above, the usual approach is to use mass spectrometry to identify the member proteins [22]. Briefly, mass spectrometry is a technique for precise identification of compounds and structure based on the mass-to-charge ratio of its constituent ions [22]. This combination of complex purification and precise identification of the components has been applied on a whole-proteome scale [7,13]. While TAP-MS also suffers from false positives like the two-hybrid system, it does have several advantages. First, it can detect not just protein-protein interactions, but whole protein complexes. Furthermore

these complexes can be detected in their natural environment, which is not true of the two-hybrid assay where interactions must occur within the nucleus to activate the reporter gene [22].

1.1.4 Genetic Interactions

Another highly informative type of genomic data that have been measured in high-throughput in recent years is genetic interaction data. A genetic interaction is said to occur between two genes whose simultaneous mutation results in a phenotype different from what is expected given the phenotypes of their individual mutations [4]. An extreme example of this phenomenon is synthetic lethality, where the simultaneous deletion of two genes causes cell death but a single deletion of either is healthy [4]. In general, genetic interactions indicate co-involvement in the same complex, pathway, or parallel pathways leading to the same essential function, and thus are highly informative of gene function [4]. One recent method, Synthetic Genetic Array (SGA) analysis, enables high-throughput investigation of double mutant combinations of yeast deletion strains [23]. SGA analysis is the focus of Chapters 7 and 8 and provides an ideal setting for applying genomic data integration technology to direct high-throughput experiments. Further background on genetic interactions is provided in Chapter 7.

1.1.5 Protein Localization

Clearly, an important indicator of protein function is cellular localization. Identifying where a protein is present in a cell can provide specific clues about the functional role it plays. Thus, several experimental groups have developed high-throughput microscopy assays to enable rapid localization of proteins (e.g. [11]). The typical strategy for these approaches is to construct a library of strains, each expressing one protein of interest fused with a green or red fluorescent protein (GFP or RFP). The library also consists of a set of “marker strains” with specific and known cellular localization patterns, which are used as a gold standard for each cellular compartment. This approach is illustrated in Figure 1.5 for two yeast query proteins (Utp13 and Cbf2) are matched against a known nucleolar marker (Sik1) [11].

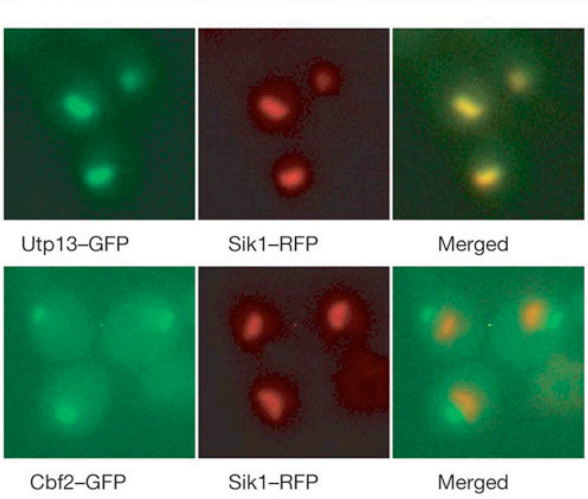


Figure 1.5. Protein localization in *Saccharomyces cerevisiae*. A library of GFP-tagged strains and a collection of genes with known localization is used for global analysis of protein localization in yeast. Here, Sik1 is a known nucleolar protein and is used to identify another nucleolar protein (Utp13) [11].

1.2 Dissertation Focus

This variety of relatively new high-throughput experimental technologies has fueled an explosion of data over the past ten years. Perhaps the best illustration of this is the rapid increase in the number of published microarray studies. Figure 1.6 plots the cumulative number of human microarray datasets deposited into NCBI's GEO database since January 2001 [6]. As of October 2007, there were 2208 published human datasets, and the total continues to grow quadratically [6] (Figure 1.6). Interestingly, by most metrics, the knowledge contributed by these studies has not experienced the same degree of growth. For instance, one reasonable metric for our current knowledge is the number of human genes with hand-curated Gene Ontology annotations [2]. Figure 1.7 plots the total number of such genes as curated by the European Bioinformatics Institute. As of October 2007, there were 8700 genes with annotations based on primary literature, and the trend appears to be relatively flat, linear growth. If we assume a total of 20,000 human genes, at this rate, it will take approximately 10 more years before we characterize a single function for each gene. Certainly most genes have more than one function, so that is only a small step towards a complete understanding of the cell. The situation is better in other

organisms, such as yeast, but even there, our ability to generate genomic data far surpasses our capacity for deriving specific, testable hypotheses from those data. This begs the question, why are we not more efficiently translating these data into *knowledge*?

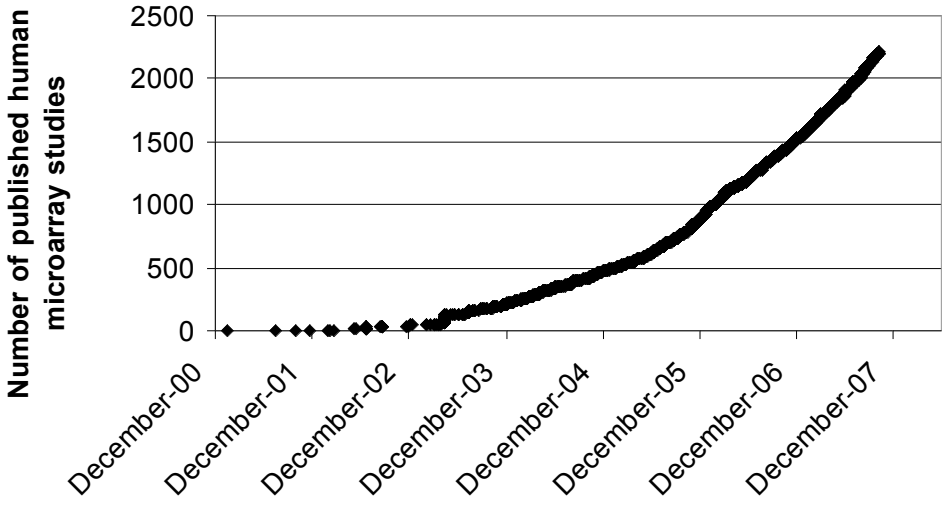


Figure 1.6. Number of published human microarray studies vs. date. All human microarray datasets were downloaded from NCBI's GEO database. Here we plot the cumulative number by date.

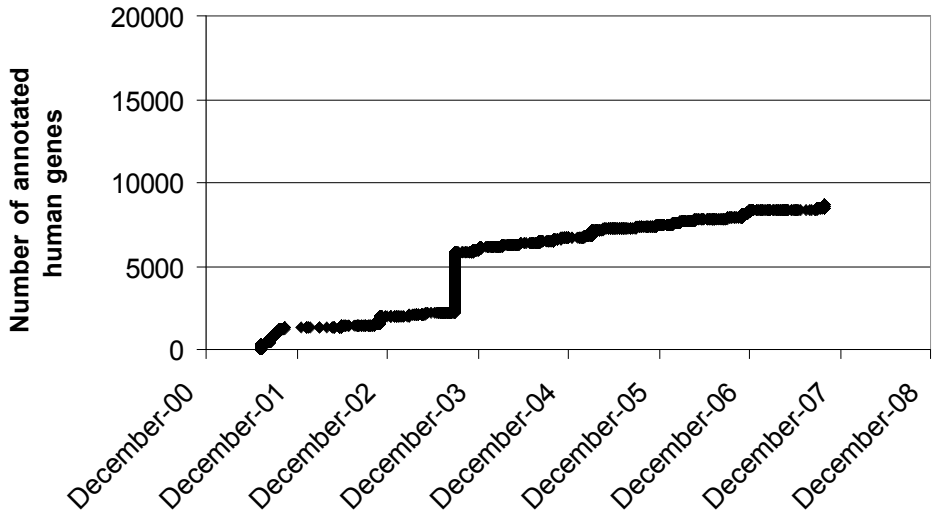


Figure 1.7. Number of human genes with annotations vs. date. Human GO annotations were downloaded from the European Bioinformatics Institute. We plot the cumulative number of genes annotated by year, excluding all IEA (Inferred from Electronic Annotation) and RCA (Reviewed Computational Analysis) annotations to capture how many have directly reference in the literature.

This dissertation focuses precisely on this question. I address a few key biological problems with variations of classic computational methods and demonstrate how they can turn vast repositories of relatively unreliable data into specific, testable hypotheses, many of which have been confirmed in our lab. While I describe solutions for a range of problems, there are three overarching themes that persist throughout my work.

First, taking full advantage of the rich information present in genomic data requires integrative methods. Many of the available experimental data are complementary in the cellular phenomena they capture and bringing them together illuminates patterns one might not otherwise see. Integration of diverse genomic data is challenging, however, because these data are heterogeneous both in structure and in their relevance to understanding gene function. If not done properly, integration of multiple noisy datasets can result in an even noisier combination. Much of this dissertation centers on robust methods for addressing this challenge.

The second recurring theme throughout the dissertation is the use of genomic data context for improving inferences drawn from noisy data. By context, we mean leveraging our prior knowledge, however incomplete, to focus and refine prediction methods. This does not mean the methods described here do not make predictions about novel biology— in fact, we describe several case examples where these methods have guided us directly to testable, correct hypotheses. Instead, one should consider our use of context as a way to “bootstrap” ourselves from relatively sparse knowledge to comprehensive, accurate predictions. Using clues about genomic context enables us to make more accurate predictions based on the same data.

Finally, every method described in this dissertation in some way incorporates genomic data visualization into the analysis pipeline. In my experience, the most effective means of translating raw genomic data into knowledge is not intelligent methods, but intelligent methods that are driven by biologists. All of the approaches described here are developed from the biologist user’s perspective and were ultimately created to enable hypothesis generation through intelligent exploration of genomic data.

1.3 Dissertation Organization

We begin in Chapter 2 with a discussion of methods for using either gene expression or array CGH microarray data to automatically and precisely identify chromosomal aberrations. Chapter 3 describes a software implementation of this approach, with an emphasis on aspects of visualization-based identification of chromosomal abnormalities from integrative analysis of gene expression and copy number microarray data. In Chapter 4, we transition into the more general problem of inferring biological networks from diverse genomic data including gene expression, protein-protein interactions, sequence, and localization data. Chapter 5 addresses the important issue of deriving gold standards for network inference from the Gene Ontology, which is a critical issue in applying machine learning approaches to the problem of genomic data integration. Chapter 6 describes further insights into the network inference problem, specifically, how biological context can be used to dramatically improve prediction performance. The last two chapters focus on using these data integration and network inference technologies to drive high-throughput genetic interaction screens. Chapter 7 presents a background on genetic interactions, a description of Synthetic Genetic Arrays (SGA), and our work on how the SGA platform can be used to quantitatively measure epistasis. Chapter 8 describes our iterative computational-experimental approach for mapping the global yeast genetic interaction network with the SGA platform. Chapter 9 concludes the dissertation with an outlook on possibilities for future research.

References

1. Alberts, B. (2002). Molecular biology of the cell. New York, Garland Science.
2. Ashburner, M., C. A. Ball, et al. (2000). "Gene ontology: tool for the unification of biology. The Gene Ontology Consortium." Nat Genet **25**(1): 25-9.
3. Brown, P. O. and D. Botstein (1999). "Exploring the new world of the genome with DNA microarrays." Nat Genet **21**(1 Suppl): 33-7.
4. Costanzo, M., G. Giaever, et al. (2006). "Experimental approaches to identify genetic networks." Curr Opin Biotechnol **17**(5): 472-80.

5. Duggan, D. J., M. Bittner, et al. (1999). "Expression profiling using cDNA microarrays." Nat Genet **21**(1 Suppl): 10-4.

6. Edgar, R., M. Domrachev, et al. (2002). "Gene Expression Omnibus: NCBI gene expression and hybridization array data repository." Nucleic Acids Res **30**(1): 207-10.

7. Gavin, A. C., P. Aloy, et al. (2006). "Proteome survey reveals modularity of the yeast cell machinery." Nature **440**(7084): 631-6.

8. Giorgini, F. and P. J. Muchowski (2005). "Connecting the dots in Huntington's disease with protein interaction networks." Genome Biol **6**(3): 210.

9. Gresham, D., D. M. Ruderfer, et al. (2006). "Genome-wide detection of polymorphisms at nucleotide resolution with a single DNA microarray." Science **311**(5769): 1932-6.

10. Hall, N. (2007). "Advanced sequencing technologies and their wider impact in microbiology." J Exp Biol **210**(Pt 9): 1518-25.

11. Huh, W. K., J. V. Falvo, et al. (2003). "Global analysis of protein localization in budding yeast." Nature **425**(6959): 686-91.

12. Kim, S. and A. Misra (2007). "SNP genotyping: technologies and biomedical applications." Annu Rev Biomed Eng **9**: 289-320.

13. Krogan, N. J., G. Cagney, et al. (2006). "Global landscape of protein complexes in the yeast Saccharomyces cerevisiae." Nature **440**(7084): 637-43.

14. Lander, E. S., L. M. Linton, et al. (2001). "Initial sequencing and analysis of the human genome." Nature **409**(6822): 860-921.

15. Lipshutz, R. J., S. P. Fodor, et al. (1999). "High density synthetic oligonucleotide arrays." Nat Genet **21**(1 Suppl): 20-4.

16. Lucito, R., J. Healy, et al. (2003). "Representational oligonucleotide microarray analysis: a high-resolution method to detect genome copy number variation." Genome Res **13**(10): 2291-305.

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

18. Pinkel, D. and D. G. Albertson (2005). "Array comparative genomic hybridization and its applications in cancer." Nat Genet **37 Suppl**: S11-7.

19. Puig, O., F. Caspary, et al. (2001). "The tandem affinity purification (TAP) method: a general procedure of protein complex purification." Methods **24**(3): 218-29.

20. Quackenbush, J. (2002). "Microarray data normalization and transformation." Nat Genet **32 Suppl**: 496-501.

21. Schena, M., D. Shalon, et al. (1995). "Quantitative monitoring of gene expression patterns with a complementary DNA microarray." Science **270**(5235): 467-70.

22. Shoemaker, B. A. and A. R. Panchenko (2007). "Deciphering protein-protein interactions. Part I. Experimental techniques and databases." PLoS Comput Biol **3**(3): e42.

23. Tong, A. H., G. Lesage, et al. (2004). "Global mapping of the yeast genetic interaction network." Science **303**(5659): 808-13.

24. Venter, J. C., M. D. Adams, et al. (2001). "The sequence of the human genome." Science **291**(5507): 1304-51.

Chapter 2

Using Genomic Context to Infer Chromosomal Aberrations from Gene Expression and Array CGH Data

2.1 Chapter Overview

We begin the dissertation with an illustration of how genomic context can be used to infer chromosomal aberrations from microarray data. Chromosomal copy number changes (aneuploidies) are common in cell populations that undergo multiple cell divisions including yeast strains, cell lines, and tumor cells. Identification of aneuploidies is critical in evolutionary studies, where changes in copy number serve an adaptive purpose, as well as in cancer studies, where amplifications and deletions of chromosomal regions have been identified as a major pathogenetic mechanism. Aneuploidies can be studied on whole-genome level using array CGH (a microarray-based method that measures DNA content), but their presence also affects gene expression. In gene expression microarray analysis, identification of copy number changes is especially important in preventing aberrant biological conclusions based on spurious gene expression correlation or masked phenotypes that arise due to aneuploidies. Previously suggested approaches for aneuploidy detection from microarray data mostly focus on array CGH, address only whole-chromosome or whole-arm copy number changes, and rely on thresholds or other heuristics, making them unsuitable for fully automated general application to gene expression data sets. There is a need for a general and robust method for identification of aneuploidies of any size from both array CGH and gene expression microarray data.

In this chapter, we present ChARM (Chromosomal Aberration Region Miner), a robust and accurate expectation-maximization based method for identification of segmental aneuploidies (partial chromosome changes) from gene expression and array CGH microarray data.

Systematic evaluation of the algorithm on synthetic and biological data shows that the method is robust to noise, aneuploidal segment size, and p-value cutoff. Using our approach, we identify known chromosomal changes and predict novel potential segmental aneuploidies in commonly used yeast deletion strains and in breast cancer. ChARM can be routinely used to identify aneuploidies in array CGH data sets and to screen gene expression data for aneuploidies or array biases. Our methodology is sensitive enough to detect statistically significant and biologically relevant aneuploidies even when expression or DNA content changes are subtle as in mixed populations of cells.

The work presented in this chapter is published in [28] and includes contributions from Maitreya Dunham and S.Y. Kung, and Olga Troyanskaya. Maitreya provided the yeast array CGH data for analysis and offered interpretation of the results, S.Y. gave feedback on the filtering and the expectation-maximization algorithm, and Olga supervised the project.

2.2 Background: Chromosomal Aberrations and Related Experimental Technology

Chromosomal amplifications, deletions, and rearrangements are thought to play important evolutionary roles in speciation [10] and adaptive mutation in yeast and microbial populations [9,16], and constitute a key mechanism in cancer progression [4,31]. Aneuploidies are especially common in cell populations that undergo multiple cell divisions such as laboratory strains or cell lines, and presence of amplifications or deletions of whole chromosomes or their parts (segmental aneuploidies) can have substantial effects on gene expression [18,11,14]. Thus, identification of aneuploidies is important in cancer pathogenesis and molecular evolution studies, as well as in every genome-scale gene expression microarray experiment because copy number changes can alter expression profiles and result in spurious correlations of functionally unrelated genes.

Recent developments in microarray technology have enabled genome-wide investigations of copy-number changes through array-based comparative genomic hybridization

(array CGH), where differentially labeled sample and reference DNA are hybridized to DNA microarrays [32,33]. This technology has proven effective in identifying aneuploidies in tumor cells [13,31,38,24], experimental evolution studies [9], and in yeast strains [18,30]. Routine application of array CGH to every strain or tissue used in gene expression studies is unfortunately not feasible. However, several studies have demonstrated that chromosomal abnormalities correlate with spatial biases in gene expression along chromosomes [18,31,37,11,14,27,34,24]. For example, Pollack et al. estimate that 62% of highly amplified genes in 37 breast cancer tumors demonstrate moderately or highly elevated expression. Thus, aneuploidies can be detected in gene expression or array CGH microarray data, and it is necessary to develop analysis methods that can accurately identify chromosomal abnormalities based on either.

Accurate identification of aneuploidies from thousands of array CGH or gene expression measurements requires robust computational methods. Most array CGH data analyses involve heuristics and threshold-based methods [18,9,34]. Recently, Autio *et al.* (2003) presented a dynamic-programming-based approach to identifying copy-number changes from array CGH data, which addressed the problem algorithmically for CGH data but lacked significance analysis [1]. Accurate identification of potential copy number changes based on gene expression data is even more challenging because of mRNA expression levels reflect transcriptional regulation as well as DNA copy number. Previous approaches for aneuploidy detection from gene expression data focus only on whole-chromosome or chromosomal-arm copy number changes, and most methods are based on heuristics or dataset-specific thresholds. In the most sophisticated method to date, Crawley *et al.* employ a sign test for detecting whole chromosome (or whole arm) expression biases [7]. Hughes and Roberts *et al.* use a simpler error-weighted mean approach for whole-chromosome aneuploidy detection and a heuristic scanning method that identifies adjacent occurrences of 4 over or under-expressed genes as potential segmental aneuploidies [18]. A visualization-based imbalance detection scheme for identifying biases common in cancer specimens as compared to normal samples is proposed by Kano *et al.* [21]. These methods address the problem of whole chromosome or chromosomal arm copy changes, but the issue of robust identification of segmental aneuploidies remains open.

Here we present ChARM, a robust and accurate statistical method for identification of segmental aneuploidies from gene expression or array CGH microarray data. Our technique provides three key improvements over previously suggested approaches. First, nearly all current aneuploidy detection schemes for expression data rely on thresholds for defining significant over- and under-expression levels (some requiring up to a 1.7-1.8 fold change). Recent studies suggest, however, that expression level changes do not always directly reflect copy change proportions, and thresholds determined for one data set often will not generalize to others [31]. Our method is statistical, and therefore generalizes to different datasets, microarray platforms, and organisms. Second, we focus on the problem of detecting segmental aneuploidy, which is generally more difficult than detecting whole-chromosome aneuploidy for which the methods developed by Hughes and Roberts *et al.* or Crawley *et al.* are effective. Third, our method is general and performs well with both gene expression and array CGH data.

ChARM employs an edge detection filter that identifies potentially aneuploid regions, an EM algorithm that finds maximum likelihood breakpoints based on a local search in these potential regions, and a statistical analysis that determines which predicted aneuploidies correspond to statistically significant biases as opposed to experimental noise. Our scheme can accurately identify known aneuploidies in biological gene expression or array CGH data [18], and rigorous performance analysis with synthetic data demonstrates that the method is robust to noise and aneuploidy size and thus can generalize to other microarray data sets. Applying ChARM to 300 gene expression profiles of laboratory yeast strains, we identify multiple previously unknown aneuploidies, most of which are supported by current biological knowledge of yeast chromosomal rearrangement mechanisms. Our analysis of breast cancer array CGH and gene expression microarray data identifies both known and novel areas of chromosomal instability and reveals two groups of immune system genes on different chromosomes that are overexpressed and often amplified in a subset of breast tumors. This novel result may, upon experimental verification, contribute to understanding of how cancers escape immune response.

2.3 ChARM: a Method for Detecting Segmental Aneuploidies

ChARM is composed of three sub-systems: an edge detection filter that identifies points on chromosomes where potential aneuploidies start or end, an EM-based edge-placement algorithm that statistically optimizes these start and end locations, and a window significance test that determines whether predicted amplifications and deletions are statistically significant or are artifacts of noise (Figure 2.1). The EM algorithm has a well-known tendency to find local rather than global maxima, but this three-stage structure is useful in setting initial conditions that ensure meaningful convergence. All three stages assume input in the form of array CGH or gene expression log ratios arranged in the order in which the corresponding genes appear along a single chromosome.

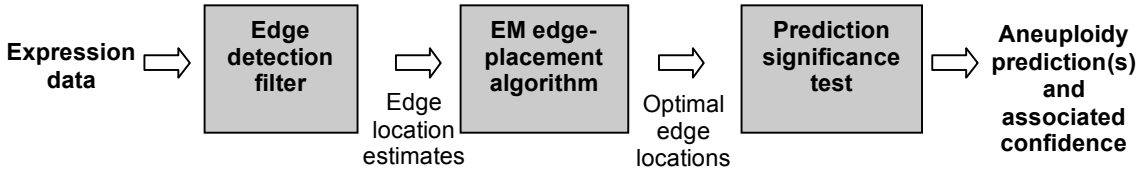


Figure 2.1. Three-stage segmental aneuploidy detection scheme. The edge detection filter estimates edge coordinates, which are then refined by the EM edge-placement algorithm. The resulting edges serve as input to the prediction significance test that analyzes statistical significance of spatial biases.

2.3.1 Edge Detection Filter

The edge detection filter estimates locations along the chromosome where abrupt changes in gene expression occur. This is accomplished by a simple cascade of a non-linear median filter, a linear smoothing filter, and a linear differentiator (Figure 2.2). The median filter functions as a high-level smoother, removing outliers, which are common in microarray data, and preserving only sustained changes in the input sequence. Finer smoothing, which is a necessary pre-processing step for the differentiator, is accomplished by a linear averaging filter with a smaller window size. The differentiator effectively computes the derivative over a short window flagging any substantial changes with large peaks. These peaks and the corresponding chromosomal locations serve as the input to the more precise EM algorithm.

2.3.2 Expectation-maximization Edge-placement Algorithm

The purpose of the EM edge-placement algorithm is to provide fine adjustments to the edge estimates from the previous filter. To facilitate convergence to statistically optimal gene indices, each edge is surrounded by a "radius of influence" (ROI), which includes an equal-length set of adjacent genes on either side that is allowed to affect the placement at a given iteration. Furthermore, each edge is associated with two distributions, one for each of the two distinct regions (left and right) it is potentially separating. Each iteration of the algorithm consists of two stages: a typical EM clustering stage for learning the maximum likelihood parameters of the two distributions for each ROI (see E-step, M-step 1 below) and an edge-placement stage which adjusts the edge position optimally given the learned parameters (see M-step 2 below). Before each edge adjustment, every pair of adjacent windows[1] is tested for similarity to ensure that the edge between these windows actually separates chromosomal regions of different copy number.

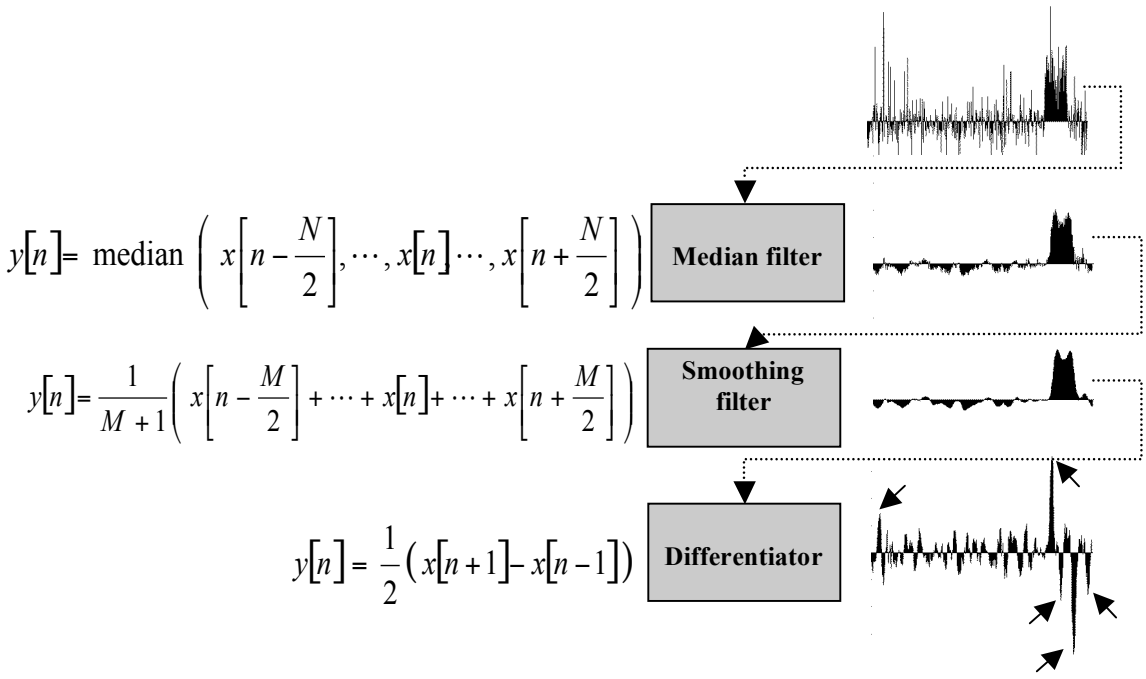


Figure 2.2. Preliminary edge detection filtering process illustrated on gene expression data positioned along the chromosome. Bars above the coordinate axis represent overexpression, bars below represent underexpression. The input-output relation for each of the filters is given on the left. $y[n]$ is the output as a function of $x[n]$ where n refers to gene index on the chromosome and *N* and *M* are the window sizes of each filter. Significant peaks are marked at the output of the differentiator.

[1] We refer to the regions between any two adjacent edges or between an edge and a chromosome end as "windows".

The algorithm converges when all edge positions are fixed for several iterations. Each of these steps is described in detail below.

Update membership (E-step)

Soft (fuzzy) memberships are computed for all genes in the radius of influence of an edge and are proportional to the probability of observing the gene given the left and right distributions associated with that edge. Let $G_i = [g_i, l_i]$ represent the log-transformed ratio (array CGH or expression) and location of gene i, $e_j^{(t)}$ denote edge j, and $\grave{e}_{j,1}^{(t)}$ and $\grave{e}_{j,2}^{(t)}$ the left and right edge distributions at iteration t of the EM algorithm. Also, let r_{inf} denote the radius of influence. Here, we assume that the set of genes in the ROI lie in two normal distributions, i.e. $\grave{e}_{j,k}^{(t)}$ is parameterized[2] by $\{\mu_{j,k}^{(t)}, \sigma_{j,k}^{(t)}\}$. Then, the conditional probability of observing gene i given the distribution $\grave{e}_{j,k}^{(t)}$ is:

$$P\left(G_i \middle| \grave{e}_{j,k}^{(t)}\right) = \begin{cases} N\left(g_i; \mu_{j,k}^{(t)}, \sigma_{j,k}^{(t)}\right) & \text{for } l_i \in \left[e_j^{(t)} - r_{\text{inf}}, e_j^{(t)} + r_{\text{inf}}\right] \\ 0 & \text{otherwise} \end{cases}$$

which allows us to compute the posterior probability of $\grave{e}_{j,k}^{(t)}$ given gene i as:

$$P\left(\grave{e}_{j,k}^{(t)} \middle| G_i\right) = \frac{P\left(G_i \middle| \grave{e}_{j,k}^{(t)}\right) P\left(\grave{e}_{j,k}^{(t-1)}\right)}{\sum_{m=1,2} P\left(G_i \middle| \grave{e}_{j,m}^{(t)}\right) P\left(\grave{e}_{j,m}^{(t-1)}\right)}$$

where $P\left(\grave{e}_{j,k}^{(t-1)}\right) = \frac{1}{n_g} \sum_{i=1}^{n_g} P\left(\grave{e}_{j,k}^{(t-1)} \middle| G_i\right)$ and n_g is the number of genes on the chromosome of interest.

Mean and variance computation (M-step 1)

Based on the membership $P\left(\grave{e}_{j,k}^{(t)} \middle| G_i\right)$ determined in the E-step, the maximum likelihood mean and variance parameters for the next iteration ($t+1$) are computed as follows:

---

[2] Note that in our implementation, we use normally distributed g_i's. Empirically, this has demonstrated adequate performance, but this approach can be generalized to other, more accurate models as well.

$$\mu_{j,k}^{(t+1)} = \frac{\sum_{i=1}^{n_g} P\left(\theta_{j,k}^{(t)} \middle| G_i\right) g_i}{\sum_{i=1}^{n_g} P\left(\theta_{j,k}^{(t)} \middle| G_i\right)} \qquad \sigma_{j,k}^{2\,(t+1)} = \frac{\sum_{i=1}^{n_g} \left(x_i - \mu_{j,k}^{(t+1)}\right)^2 P\left(\theta_{j,k}^{(t)} \middle| G_i\right)}{\sum_{i=1}^{n_g} P\left(\theta_{j,k}^{(t)} \middle| G_i\right)}$$

when $G_i \sim N(\mu_j^{(t)}, \sigma_j^{(t)})$ [8].

Edge adjustment (M-step 2)

For edge adjustment, we use the information theoretic notion of surprise (i.e. the amount of information learned from observing a probabilistic event). At each iteration, we restrict the possible edge locations to only the set of indices included in the current ROI. Each placement implies a different clustering of the genes around the edge into the left or right edge distributions. Each gene's placement in the implied cluster is treated as the observation of a random variable whose probability distribution is the gene's posterior probability of being associated with that cluster. For instance, if G_i falls in $\grave{e}_{j,1}^{(t)}$ for a particular placement of the edge $e_j^{(t)}$, the surprise of this event is $S(G_i) = -\log\left(P\left(\grave{e}_{j,k}^{(t)} \middle| G_i\right)\right)$. Then, the "minimum surprise" edge placement is given by:

$$e_j^{(t+1)} = \arg\min_{\mathrm{i}} -\left[\sum_{k=1}^{i-1} \log\left(P\left(\theta_{j,1}^{(t)} \middle| G_k\right)\right) + \sum_{k=i}^{2r_{\inf}+1} \log\left(P\left(\theta_{j,2}^{(t)} \middle| G_k\right)\right)\right]$$

where the indices $1 \ldots (2r_{\inf}+1)$ refer to those genes in the ROI. Upon adjusting the edge placement for each window, the window parameters are updated accordingly (i.e. $e_j^{(t)} \rightarrow e_j^{(t+1)}$, $\theta_{j,k}^{(t+1)} = \left[\mu_{j,k}^{(t+1)}, \sigma_{j,k}^{2\,(t+1)}\right]$).

Window similarity test

The window similarity test is needed at each iteration to ensure that edges about to be adjusted actually separate different windows with distinct chromosomal biases (separate aneuploidy predictions). The difference between left and right windows on either side of an edge must exceed a minimum signal-to-noise threshold or the edge is removed. As noted earlier, a window

that extends beyond the ROI includes all genes up to the next edge or chromosome end. We have evaluated several parametric and non-parametric statistical metrics for measuring the difference between two sets of samples including t-test, non-parametric t-tests, rank-sum test, Kolmogorov-Smirnov test. Empirically, the ratio of the difference in medians between two adjacent windows and the pooled absolute deviation from the median has demonstrated the best performance. Thus, we impose the following criterion on this modified signal-to-noise ratio (SNR) for removing an edge ($e_j^{(t)}$) at iteration *t*:

$$SNR_{i,j} = \frac{\left| med_{j,1} - med_{j,2} \right|}{\frac{1}{n_{j,1} + n_{j,2}} \left(\sum_{k \in w_{j,1}} \left| g_k - med_{j,1} \right| + \sum_{k \in w_{j,2}} \left| g_k - med_{j,2} \right| \right)} < SNR_{\text{thresh}}\left(\overline{\delta}_e\right)$$

for $med_{j,k} = median(w_{j,k})$ where $w_{j,1}$ and $w_{j,2}$ include all the genes in the adjacent windows with sizes $n_{j,1}$ and $n_{j,2}$ respectively. SNR_{thresh} is a threshold dependent on the current convergence behavior measured by $\overline{\delta}_e$, the average edge position change (in gene indices) from one iteration to the next. We raise the minimum SNR threshold as the edge positions begin to converge so that adjacent windows must be "more different" to remain separate as edges approach their final estimates.

2.3.3 Window Significance Analysis

Once the EM algorithm obtains precise window positions, the significance analysis scheme determines if each window represents a statistically significant spatial bias in DNA content or expression. We consider three statistical tests for assessing the significance of windows identified by the EM algorithm: a one-sample sign test, a mean permutation test, and a coefficient of variance permutation test, as well as combinations of the mean and sign tests and the variance and sign tests. The sign test is that reported by Crawley *et al.* in [7] with the modification that the threshold is chosen dynamically for each chromosome to allow for identification of biased regions exhibiting lower degrees of over or under-expression than the 1.7-1.8 fold threshold used by others. Both permutation tests require performing approximately 5,000 random permutations of

the genes on the chromosome and comparing the statistic (mean or variance) obtained on the actual arrangement with the most significant statistic for the same window size on each random permutation. We use the Bonferroni method to correct for multiple hypothesis tests on the same chromosome. Our permutation tests are designed specifically for the segmental aneuploidy problem, while other methods such as the sign test or the error-weighted mean approach proposed in [18] are more appropriate for chromosome-wide bias detection.

2.4 Evaluation

To systematically assess ChARM's accuracy and robustness, we evaluate it using a synthetic microarray measurement error model described below. Using this model, we assess which window significance test yields the best performance for aneuploidy detection and thoroughly evaluate the robustness of our scheme. We further evaluate our scheme on biological data (see Application to Biological Data).

2.4.1 Synthetic Data Model

We generate synthetic two-color microarray data according to the model proposed in [35]. Under this two-component model, reference (y_R) and test (y_T) intensity values are simulated as:

$$y_R = \alpha_R + \mu_R e^{\eta_S + \eta_R} + \varepsilon_S + \varepsilon_R \qquad y_T = \alpha_T + \mu_T e^{\eta_S + \eta_T} + \varepsilon_S + \varepsilon_T ,$$

where α is the mean background intensity, μ is the intensity contributed by the quantity of interest, and

$$\eta_S \sim N(0, \sigma_{\eta_S}) \quad \eta_R \sim N(0, \sigma_{\eta_R}) \quad \eta_T \sim N(0, \sigma_{\eta_T})$$
$$\varepsilon_S \sim N(0, \sigma_{\varepsilon_S}) \quad \varepsilon_R \sim N(0, \sigma_{\varepsilon_R}) \quad \varepsilon_T \sim N(0, \sigma_{\varepsilon_T})$$

This model was originally proposed for gene expression microarrays, but it is also appropriate for array CGH experiments with the modification that μ_R and μ_T are amounts of reference and test genomic DNA rather than mRNA. The parameters denoted by the subscript "*s*" are characteristics of the microarray spot and common to both reference and test samples. The

mean background intensities (α) are typically estimated by microarray image analysis software and used to compute estimates of test and reference signal intensities, x_T and x_R, as follows:

$$x_R = y_R - \hat{\alpha}_R \quad x_T = y_T - \hat{\alpha}_T .$$

We model the error in this background estimation, $\hat{\alpha}$, as an additional normally distributed error term, ε_{est}, so that the pre-log-ratio intensities are generated as:

$$x_R = \mu_R e^{\eta_S + \eta_R} + \varepsilon_S + \varepsilon_R + \varepsilon_{est} \quad x_T = \mu_T e^{\eta_S + \eta_T} + \varepsilon_S + \varepsilon_T + \varepsilon_{est}$$

Parameters for this model are estimated as suggested in [35] for biological array CGH and gene expression experiments (Table 2.1). Prior to noise addition, test and reference intensities across each synthetic chromosome for all simulations are drawn from a normal distribution with $\mu \sim N(3980, 800)$, and the mean background intensity is assumed to be 400 for test and reference samples with $\varepsilon_{est} \sim N(0, 40)$. Regions of aneuploidy are synthetically produced by setting all affected genes' test-to-reference ratio $\left(\frac{\mu_T}{\mu_R}\right)$ to 1.5[3] (prior to noise effects). Furthermore, to model expression scenarios realistically, 10% of the genes outside of aneuploidal regions are randomly set to over- or under-expressed with no spatial correlation.

Table 2.1. Estimated parameters for array CGH and expression human breast cancer data. Parameters were estimated as suggested by Rocke and Durbin (2001).

Parameter	Microarray type	
	Array CGH	Expression
$\hat{\alpha}_T . \hat{\alpha}_R$	59.2, 45.9	399, 238
$\hat{\mu}_T . \hat{\mu}_R$	111, 113	3980, 4130
$\hat{\sigma}_{\eta_S}, \hat{\sigma}_{\eta_T}, \hat{\sigma}_{\eta_R}$	.63, .059, .090	.53, .17, .13
$\hat{\sigma}_{\varepsilon_S}, \hat{\sigma}_{\varepsilon_T}, \hat{\sigma}_{\varepsilon_R}$	25, 11, 0	137, 54, 94

[3] As gene expression changes do not directly reflect DNA copy number, the test-to-reference ratio for a gene that has been duplicated will not necessarily be 2. We chose to set these ratios to 1.5 to provide a conservative evaluation of our method.

2.4.2 Choice and Performance of Window Significance Test

We first address the question of choosing the window significance test for our framework. We consider three window significance tests (sign test, mean test, coefficient of variance test) and evaluate their performance on simulated 50-gene aneuploidies under varying p-value cutoffs (Figure 2.3). Under all conditions tested, the mean and coefficient of variance permutation tests perform overwhelmingly better than the one-sample sign test, which is used by Crawley *et al.* [7] and Haddad *et al.* [14]. However, when an aneuploidy is located on the end of a chromosome, the mean test, which is generally very specific, can falsely report the region spanning the rest of the chromosome as significant based on the permutations. This shortcoming of the permutation-based approach can be overcome by combination with the simpler sign test. This combined mean permutation and sign test scheme performs best both in terms of specificity and sensitivity, and is thus used in the rest of evaluation experiments. A similar combination of the coefficient of variance test and the sign test is less effective because the variance-based test yields lower sensitivity due to the noisy characteristics of microarray data.

2.4.3 Robustness Evaluation

We also examine the performance of ChARM under varying noise conditions. The performance of the method is only minimally affected by additive noise (ε parameters) (data not shown). The effect of multiplicative error (η) in test and reference samples is shown in Figure 2.4. The sensitivity of the algorithm is robust ($\geq$ .9) to noise levels well above the biological range (Figure 2.4A, Table 2.1), and the specificity ranges from 1 to .94 for all noise parameters (data not shown). Our method provides accurate edge placement at biologically realistic noise levels (average edge coordinate error < 8%) (Figure 2.4B). Edge coordinate error is defined as $\Delta = \frac{\sum_i \left(\left| \hat{e}_{i,1} - e_{i,1} \right| + \left| \hat{e}_{i,2} - e_{i,2} \right| \right)}{\#\text{ of identified aneuploidies}}$, where parameters $\hat{e}_{i,1}$ and $\hat{e}_{i,2}$ refer to the edge estimates of the i^{th} prediction, and $e_{i,1}$ and $e_{i,2}$ are the known edge locations of the synthetic aneuploidy. Both

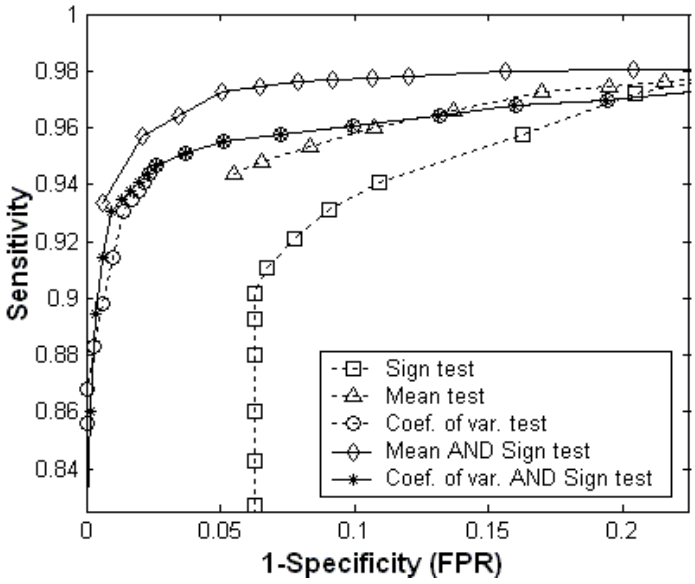


Figure 2.3. Receiver operating characteristic (ROC) curves for sign test, mean test, coefficient of variance, and combined tests with p-value cutoffs between 10^{-6} and .4. Performance was evaluated on synthetic data with simulated 50-gene aneuploidies and generated with $\sigma_{\eta_R}, \sigma_{\eta_T} = .25; \sigma_{\eta_S} = .15; \frac{\sigma_{\varepsilon_T}}{\alpha_T}, \frac{\sigma_{\varepsilon_S}}{\alpha_R}, \frac{\sigma_{\varepsilon_R}}{.5(\alpha_T + \alpha_R)} = .2.$ A combined mean and sign test shows the highest sensitivity at every false positive rate (FPR) tested.

s

ensitivity and edge placement error are more sensitive to multiplicative reference and test noise than to shared spot noise.

To test for bias in our method's performance toward particular aneuploidal segment sizes, we perform a similar noise analysis across a range of typical lengths (results not shown). At moderate biological noise levels (0.1), the algorithm identifies even small segments (< 10 genes) of copy-number change with very high specificity (> .95). Under severe noise conditions the sensitivity of the detection algorithm degrades quite noticeably for very small aneuploidies (much less than 100 genes in length). However, the algorithm is able to detect larger copy number changes (>100 genes) even under high noise conditions (σ_{η_T} 10 times greater than typical biological noise) with relatively high sensitivity. The edge coordinate errors behave similarly, although with less degradation. Both effects are due to the fact that separating signal from noise becomes more difficult as the length of spatial correlation decreases. Therefore our scheme is robust to noise and can accurately identify aneuploidy regions even under high noise conditions.

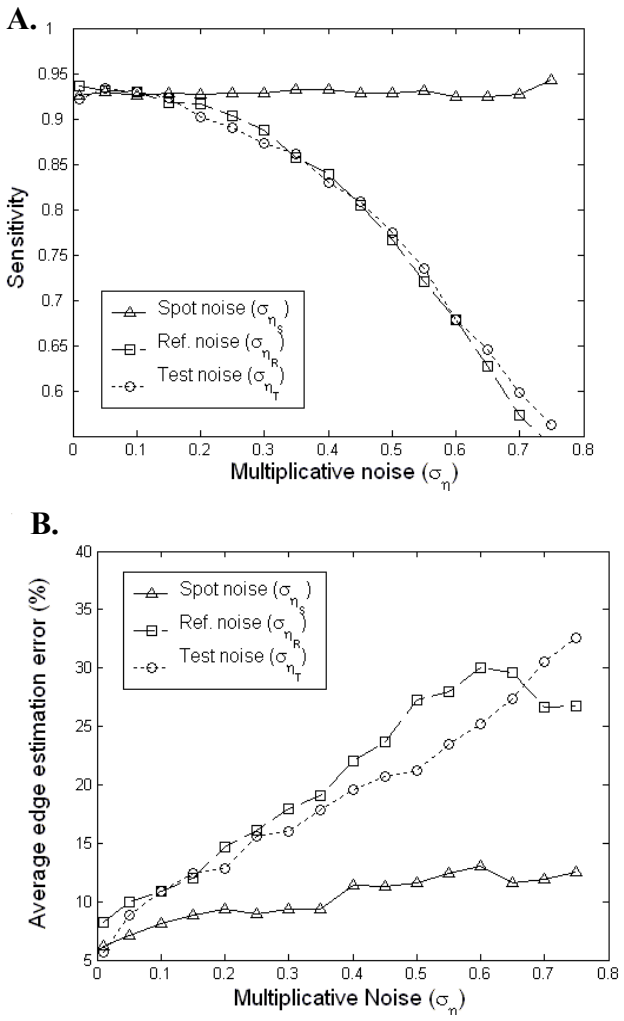


Figure 2.4. Effect of multiplicative noise on **A.** sensitivity and **B.** errors in edge coordinates (as % of total window size). Performance of the scheme in identifying a 50 gene aneuploidal segment was evaluated under varying degrees of noise. σ_{η_S} was varied while the remaining terms were fixed at .1. Similarly, σ_{η_T}, σ_{η_R} were varied with $\sigma_{\eta_S} = .5$. Biological noise is typically under 0.65 for $\acute{o}_{\varsigma_S}$ and under 0.2 for σ_{η_T}, σ_{η_R} (Table 1). P-value cutoffs were set at 10^{-3} and 10^{-2} for the sign and mean permutation tests respectively, and the tests were combined as previously described. The detection scheme with the combined mean and sign window significance test identifies most windows (>90%) with high accuracy in placement of edge coordinates (error < 0.1%) and is robust to high levels of spot, test, and reference noise (substantially higher than noise levels common in biological data shown in Table 2.1).

2.5 Biological Validation

We applied ChARM to the yeast deletion mutants' gene expression data set of Hughes and Marton *et al.* [17] and to gene expression and array CGH data for breast cancer patients from [34]. The results, presented below, demonstrate that our method can be successfully applied to

both gene expression and array CGH biological data for different organisms. We outline known amplifications and deletions that ChARM identifies and present some novel aneuploidies we find as well.

2.5.1 Segmental Aneuploidies in *S. cerevisiae* Deletion Mutants

We applied our method to the compendium of expression profiles of 300 *S. cerevisiae* deletion mutants and drug-treated strains developed and previously analyzed for aneuploidies by Hughes and Roberts *et al.* [18]. The analysis by Hughes *et al.* emphasizes whole-chromosome copy number changes, and they identify based on gene expression data and confirm by array CGH only two segmental aneuploidies[4]. Our method identifies these confirmed segmental aneuploidies (*rpl20aΔ/rpl20aΔ* and *rad27Δ/rad27Δ* strains) with high confidence (*rad27Δ/rad27Δ* sign test p-value of 10^{-5}, mean permutation test p-value of $<10^{-4}$; *rpl20aΔ/rpl20aΔ* sign test p-value of 10^{-7}, mean permutation test p-value of $<10^{-4}$). In addition to confirming the segmental aneuploidies identified by Hughes *et al.*, we identify a number of previously unknown potential aneuploidal regions[5], the top 100 (sign test p-values of $< 10^{-3}$ and mean permutation test p-values of $< 10^{-2}$) of which are pictured in Figure 2.5, and expression profiles of two are displayed in Figure 2.6. To assess the biological significance of these results, we use biological models of mechanisms of chromosomal breakage and aneuploidy formation in yeast. Chromosomal amplifications and deletions in yeast are thought to arise through ectopic recombination between homologous sequences, such as Ty transposons, transposon-related long terminal repeats (LTRs), or tRNA sequences (Infante *et al.*, 2003). Thus, presence of transposons, LTRs, or tRNA sequences near the edges of a predicted aneuploidy region can serve as biological evidence that the region in question truly contains an amplification or deletion.

[4] Hughes et al. identified one additional segmental aneuploidy (in *top3Δ* based on array CGH. This aneuploidy is not reflected in the gene expression data and thus cannot be identified by any gene expression analysis method.

[5] Predictions that represented two adjacent occurrences of Ty transposons or included centromeric regions were excluded from further analysis due to the potential of cross-hybridization artifacts.

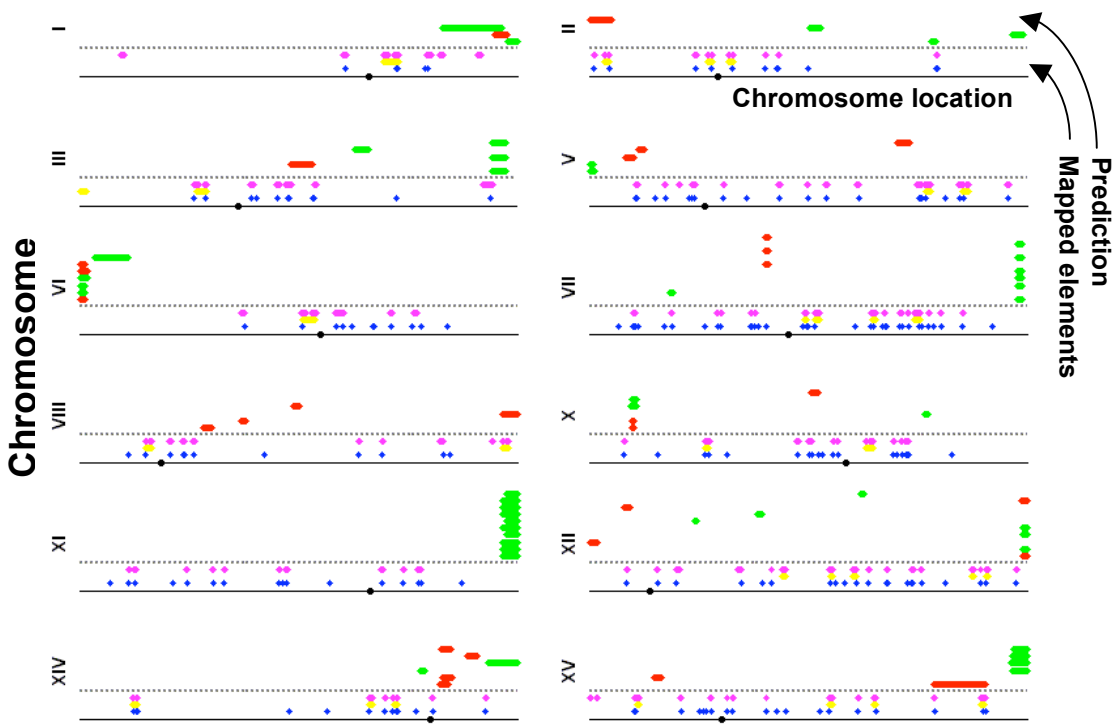


Figure 2.5. Chromosomal maps showing a subset of predicted aneuploidies (sign test p-values of $< 10^{-3}$ and mean permutation test p-values of $< 10^{-2}$) and biologically relevant mapped chromosomal elements. Aneuploidies are color-coded: red indicates amplification and green indicates deletion. Predictions shown in different rows on the same chromosome correspond to different yeast strains (e.g. Chr II), and multiple predications at the same chromosomal coordinate represent identical aneuploidies found in multiple strains (e.g. Chr XI). Proximity of predictions to LTR, transposon, and tRNA elements was evaluated through 10,000 random placements of same-sized regions on the chromosomal map and through finding the proportion of random regions with shorter distance (d_{rand}) to homologous elements than real predictions (d_{obs}) $\left[p = \frac{\text{count}(d_{rand} < d_{obs})}{\text{count}(rand\ placements)} \right]$.

In addition, increased chromosomal breakage may be observed in the conserved Y′ areas at the ends of the yeast chromosomes [5].

Our analysis reveals that 73% of predictions presented in Figure 2.5 are significantly (p-value < 0.1) closer to such homologous sequences than expected by chance or are located in the Y′ regions. These predictions likely correspond to novel segmental aneuploidies, while other predictions may represent array artifacts or aneuploidies that arose through an alternative molecular mechanism.

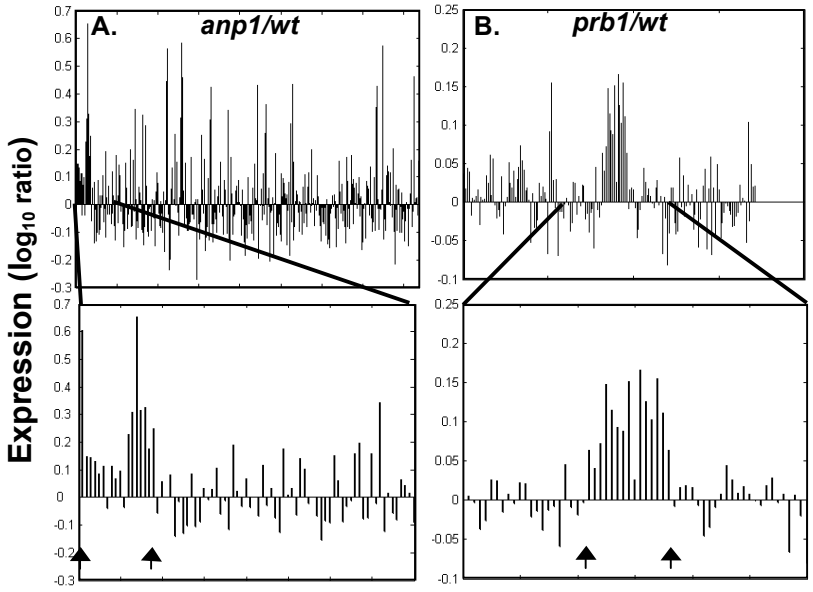


Figure 2.6. Gene expression levels plotted by chromosomal location in predicted aneuploidies: **A.** *anp1* (chromosome II, sign test p-value of $< 10^{-10}$, mean permutation test p-value of 10^{-3}) and **B.** *prb1* (chromosome III, sign test p-value of $< 10^{-10}$, mean permutation test p-value of $< 10^{-4}$) heterozygous deletion mutants. Aneuploidies predicted by our method are identified by arrows and correspond to spatial expression biases.

In yeast deletion mutant strains undergoing multiple divisions, an aneuploidy that compensates for or masks the deleted gene's phenotype could confer a selective advantage [9]. For example, growth defects caused by the deletion of *anp1* (Figure 2.6A), an endoplasmic reticulum (ER) protein with a role in retention of glycosyltransferases in the Golgi [20], may be alleviated by the amplification of the region on chromosome II that includes *SFT2*, a gene involved in ER-Golgi transport [6]. The *hdf1* deletion mutant also exhibits a compensatory mechanism. Hdf1 protein functions as a heterodimer with the Ku protein in maintaining normal telomere length and structure, but cells can maintain telomeres in the absence of telomerase through a recombination-dependent "survivor" pathway that replicates Y' regions of chromosomes [22]. Indeed, we identify amplifications in the Y' region of chromosomes II, VI, and XII in this *hdf1Δ/hdf1Δ* strain.

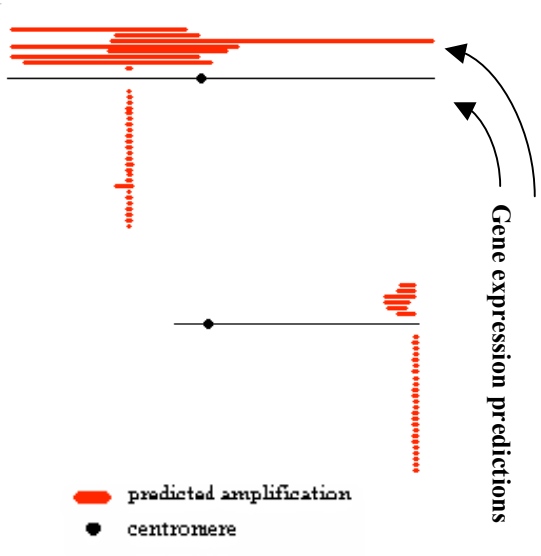
2.5.2 Identification of Aneuploidies in Breast Cancer Gene Expression and Array CGH Data

Genomic instability is thought to play a major role in oncogenesis, and breast tumors specifically are known to harbor multiple aneuploidies [34,12]. Using ChARM, we analyzed array CGH data from [34] for 44 breast tumors and the corresponding gene expression studies for 37 of these sample [36]. Our method identifies the known "hot spots" of amplifications and deletions in breast cancer [19,34], including multiple cases of deletions on 13q that include tumor suppressor protein Rb1 and on 17p that span tumor suppressor protein Tp53. Deletion of either Rb1 or Tp53 is known to cause chromosomal instability, and we do identify multiple additional aneuploidies in tumors with predicted Rb1 or Tp53 deletion [23]. We also identify a known 17q amplification that includes proto-oncogene ERBB2/HER2 [26].

One advantage of our method is the ability to make predictions based independently on array CGH or gene expression data. Overlaps in these independent predictions can be used to focus on potentially functionally relevant segmental aneuploidies. The two most striking overlap regions both include immune system proteins: genes that encode class II major histocompatability complex proteins (MHCII) on chromosome 6, and immunglobulin heavy chain genes on chromosome 14 (Figure 2.7). It is surprising to find such expression levels of these immune proteins in the tumor samples. One concern is that the data reflect the presence of a lymphocytic infiltrate in tumor tissue, however in such a case one would not expect correlated amplification data. Immune system effects on tumor progression are relatively poorly understood; a key question is why some tumors are recognized and destroyed by the immune system while others successfully proliferate.

Immunoglobulins, also known as antibodies, are secretable proteins produced by mature B lymphocytes. These molecules play an essential part in the adaptive immune system by binding and neutralizing foreign particles. As immunoglobulin gene expression typically occurs only in B lymphocytes after directed germline rearrangement, immunoglobulin heavy chain overexpression and amplification of the corresponding region is potentially an important finding, but requires further investigation into the functional status of the transcripts. MHCII is another key

component of adaptive immune response – it is a membrane protein whose primary role is the presentation of protein fragments for immune recognition. However, MHCII presentation of foreign proteins activates a response optimally in the presence of other costimulatory molecules, and MHCII overexpression outside of this immune context may lead to immune tolerance, a condition when tumors do not activate immune response [15,16]. One theory is that malignant tumors may induce tolerance with out-of-context immune stimuli, thereby evading immune response, which allows them to grow and proliferate [25]. No definitive evidence for this theory exists, but these effects have been observed in model systems [29,3] and MHCII overexpression has been associated with poor prognosis in melanomas [2]. Experimental verification of our findings may provide novel evidence of induction of immune tolerance in tumors.



2.6 Conclusions

We have demonstrated that segmental aneuploidies can be identified based on array CGH or gene expression microarray data and have presented a robust statistical method that can accurately locate aneuploidies in biological data. Evaluations on synthetic and biological data show that our method is robust to experimental noise and aneuploidy size and thus is appropriate for general and automated application to microarray data sets. ChARM allows routine screening

of gene expression data for aneuploidies and is sensitive enough to detect small statistically significant signal biases in mixed populations of cells. It is important to note that gene expression does not always reflect copy number and, furthermore, algorithms based on gene expression data alone cannot discriminate between spatial expression biases that arise from DNA abnormalities and biases that are the result of spatial coregulation or array artifacts. Our method can identify spatial expression biases due to either aneuploidies or technology artifacts and thus can be used as a general screening tool for gene expression microarray data. In cases when ChARM is used to screen for aneuploidies only, gene expression microarray data should be normalized for special artifacts prior to applying ChARM [39]. Applying ChARM to biological data, we have identified multiple previously unknown aneuploidies in public yeast gene expression data, several of which are supported by biological evidence, and potential amplification and overexpression of immune genes in breast cancer. These predictions should be further evaluated through targeted laboratory investigation.

References

1. Autio, R., S. Hautaniemi, et al. (2003). "CGH-Plotter: MATLAB toolbox for CGH-data analysis." Bioinformatics **19**(13): 1714-5.
2. Brocker, E. B., L. Suter, et al. (1985). "Phenotypic dynamics of tumor progression in human malignant melanoma." Int J Cancer **36**(1): 29-35.
3. Byrne, S. N. and G. M. Halliday (2003). "High levels of Fas ligand and MHC class II in the absence of CD80 or CD86 expression and a decreased CD4+ T cell Infiltration, enables murine skin tumours to progress." Cancer Immunol Immunother **52**(6): 396-402.
4. Cahill, D. P., K. W. Kinzler, et al. (1999). "Genetic instability and darwinian selection in tumours." Trends Cell Biol **9**(12): M57-60.
5. Chan, C. S. and B. K. Tye (1983). "Organization of DNA sequences and replication origins at yeast telomeres." Cell **33**(2): 563-73.
6. Conchon, S., X. Cao, et al. (1999). "Got1p and Sft2p: membrane proteins involved in traffic to the Golgi complex." Embo J **18**(14): 3934-46.

7. Crawley, J. J. and K. A. Furge (2002). "Identification of frequent cytogenetic aberrations in hepatocellular carcinoma using gene-expression microarray data." Genome Biol **3**(12): RESEARCH0075.

8. Dempster, A. P., N. M. Laird, et al. (1977). "Maximum Likelihood from Incomplete Data Via Em Algorithm." Journal of the Royal Statistical Society Series B-Methodological **39**(1): 1-38.

9. Dunham, M. J., H. Badrane, et al. (2002). "Characteristic genome rearrangements in experimental evolution of Saccharomyces cerevisiae." Proc Natl Acad Sci U S A **99**(25): 16144-9.

10. Fischer, G., S. A. James, et al. (2000). "Chromosomal evolution in Saccharomyces." Nature **405**(6785): 451-4.

11. Fritz, B., F. Schubert, et al. (2002). "Microarray-based copy number and expression profiling in dedifferentiated and pleomorphic liposarcoma." Cancer Res **62**(11): 2993-8.

12. Gollin, S. M. (2004). "Chromosomal instability." Curr Opin Oncol **16**(1): 25-31.

13. Gray, J. W. and C. Collins (2000). "Genome changes and gene expression in human solid tumors." Carcinogenesis **21**(3): 443-52.

14. Haddad, R., K. A. Furge, et al. (2002). "Genomic profiling and cDNA microarray analysis of human colon adenocarcinoma and associated intraperitoneal metastases reveals consistent cytogenetic and transcriptional aberrations associated with progression of multiple metastases." Applied Genomics and Proteomics **1**: 123-134.

15. Hardwick, K. G. (1998). "The spindle checkpoint." Trends Genet **14**(1): 1-4.

16. Hendrickson, H., E. S. Slechta, et al. (2002). "Amplification-mutagenesis: evidence that "directed" adaptive mutation and general hypermutability result from growth with a selected gene amplification." Proc Natl Acad Sci U S A **99**(4): 2164-9.

17. Hughes, T. R., M. J. Marton, et al. (2000). "Functional discovery via a compendium of expression profiles." Cell **102**(1): 109-26.

18. Hughes, T. R., C. J. Roberts, et al. (2000). "Widespread aneuploidy revealed by DNA microarray expression profiling." Nat Genet **25**(3): 333-337.

19. Hyman, E., P. Kauraniemi, et al. (2002). "Impact of DNA amplification on gene expression patterns in breast cancer." Cancer Res **62**(21): 6240-5.

20. Jungmann, J. and S. Munro (1998). "Multi-protein complexes in the cis Golgi of Saccharomyces cerevisiae with alpha-1,6-mannosyltransferase activity." Embo J **17**(2): 423-34.

21. Kano, M., K. Nishimura, et al. (2003). "Expression imbalance map: a new visualization method for detection of mRNA expression imbalance regions." Physiol Genomics **13**(1): 31-46.

22. Lendvay, T. S., D. K. Morris, et al. (1996). "Senescence mutants of Saccharomyces cerevisiae with a defect in telomere replication identify three additional EST genes." Genetics **144**(4): 1399-412.

23. Lentini, L., L. Pipitone, et al. (2002). "Functional inactivation of pRB results in aneuploid mammalian cells after release from a mitotic block." Neoplasia **4**(5): 380-7.

24. Linn, S. C., R. B. West, et al. (2003). "Gene expression patterns and gene copy number changes in dermatofibrosarcoma protuberans." Am J Pathol **163**(6): 2383-95.

25. Mapara, M. Y. and M. Sykes (2004). "Tolerance and cancer: mechanisms of tumor evasion and strategies for breaking tolerance." J Clin Oncol **22**(6): 1136-51.

26. Menard, S., E. Tagliabue, et al. (2000). "Role of HER2 gene overexpression in breast carcinoma." J Cell Physiol **182**(2): 150-62.

27. Mukasa, A., K. Ueki, et al. (2002). "Distinction in gene expression profiles of oligodendrogliomas with and without allelic loss of 1p." Oncogene **21**(25): 3961-8.

28. Myers, C. L., M. J. Dunham, et al. (2004). "Accurate detection of aneuploidies in array CGH and gene expression microarray data." Bioinformatics **20**(18): 3533-43.

29. Ostrand-Rosenberg, S., S. Baskar, et al. (1996). "Expression of MHC Class II and B7-1 and B7-2 costimulatory molecules accompanies tumor rejection and reduces the metastatic potential of tumor cells." Tissue Antigens **47**(5): 414-21.

30. Perez-Ortin, J. E., J. Garcia-Martinez, et al. (2002). "DNA chips for yeast biotechnology. The case of wine yeasts." J Biotechnol **98**(2-3): 227-41.

31. Phillips, J. L., S. W. Hayward, et al. (2001). "The consequences of chromosomal aneuploidy on gene expression profiles in a cell line model for prostate carcinogenesis." Cancer Res **61**(22): 8143-9.

32. Pinkel, D., R. Segraves, et al. (1998). "High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays." Nat Genet **20**(2): 207-211.

33. Pollack, J. R., C. M. Perou, et al. (1999). "Genome-wide analysis of DNA copy-number changes using cDNA microarrays." Nat Genet **23**(1): 41-46.

34. Pollack, J. R., T. Sorlie, et al. (2002). "Microarray analysis reveals a major direct role of DNA copy number alteration in the transcriptional program of human breast tumors." Proc Natl Acad Sci U S A **99**(20): 12963-8.

35. Rocke, D. M. and B. Durbin (2001). "A model for measurement error for gene expression arrays." J Comput Biol **8**(6): 557-69.

36. Sorlie, T., R. Tibshirani, et al. (2003). "Repeated observation of breast tumor subtypes in independent gene expression data sets." Proc Natl Acad Sci U S A **100**(14): 8418-23.

37. Virtaneva, K., F. A. Wright, et al. (2001). "Expression profiling reveals fundamental biological differences in acute myeloid leukemia with isolated trisomy 8 and normal cytogenetics." Proc Natl Acad Sci U S A **98**(3): 1124-9.

38. Wilhelm, M., J. A. Veltman, et al. (2002). "Array-based comparative genomic hybridization for the differential diagnosis of renal cell cancer." Cancer Res **62**(4): 957-60.

39. Yang, Y. H., S. Dudoit, et al. (2002). "Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation." Nucleic Acids Res **30**(4): e15.

Chapter 3

Visualization-Based Analysis of Chromosomal Aberrations

3.1 Chapter Overview

In Chapter 2, we described ChARM, a method for accurate detection of chromosomal amplifications and deletions. One powerful approach to analysis of copy number changes based on microarray data is a combination of visualization-based and automated computational analysis, such as that provided by ChARM. To address this need, we have developed ChARMView, a visualization and analysis system for guided discovery of chromosomal abnormalities from microarray data. Our system facilitates manual or automated discovery of aneuploidies through dynamic visualization and integrated statistical analysis. ChARMView is an effective and accurate visualization and analysis system for recognizing even small aneuploidies or subtle expression biases, identifying recurring aberrations in sets of experiments, and pinpointing functionally relevant copy number changes. In this chapter, we describe the functionality of ChARMView, and discuss several illustrative case examples where this system has been used to find biologically relevant chromosomal aberrations.

The work presented in this chapter is published in [18] and includes contributions from Xing Chen and Olga Troyanskaya. Xing developed a prototype of the software, and Olga supervised the project.

3.2 Background

As discussed in Chapters 1 and 2, aneuploidies (chromosomal copy number changes) constitute a key mechanism in cancer progression [3,20] and play important evolutionary roles in speciation [7] and adaptive mutation in yeast and microbial populations [6,13]. Array-based comparative

genomic hybridization (array CGH) has enabled fast genome-wide investigations of copy-number changes [21,22]. However, once microarray experiments have been performed, accurate identification of amplifications and deletions requires a combination of manual discovery through data visualization and sophisticated statistical analysis.

Computational methods can use additional data sources, such as gene expression, to facilitate the discovery and analysis of genomic aberrations. This is possible because the presence of amplifications or deletions of whole or partial chromosomes can have substantial effects on gene expression in the affected regions [14,10,12]. Gene expression microarray data can serve both as a second source of information for aneuploidy detection and perhaps as an indication of which genomic changes are most functionally relevant since mRNA transcript abundance more directly affects cellular phenotype than genomic DNA content. Therefore, an effective visualization and analysis system for aneuploidy detection should make use of both array CGH and gene expression data, and allow easy examination of overlaps in the corresponding data sets.

Existing visualization tools include Caryoscope [1], CGHAnalyzer [11], Java Treeview's Karyoscope [25], and SeeCGH [5]. All of these were developed specifically for the analysis of array CGH data and with the exception of CGHAnalyzer, none allow convenient visualization of multiple experiments. Additionally, while they all offer a number of useful approaches to visualization, none include automatic statistical prediction to complement manual discovery of amplifications and deletions (see Table 3.1 for a detailed comparison of features of our software as compared to those of existing applications). To facilitate discovery of genomic aberrations from microarray data, novel methodology is required that integrates visualization with sophisticated statistical analysis and enables visualization of multiple experiments and data types simultaneously.

Here we describe ChARMView – an integrated system that combines statistical analysis with effective visualization capabilities to enable interpretation of microarray data for aneuploidy discovery. Our system facilitates both manual and automated discovery of genomic aberrations

Table 3.1. ChARMView comparison with existing visualization and analysis software, including Caryoscope [1], CGHAnalyzer [11], Java Treeview's Karyoscope [25], SeeCGH [5], CGH-Explorer [17], CGH-PRO [4], and CGH-Miner [27].

Feature	ChARM View	Caryo-scope	Java TreeView	SeeGH	CGH Analyzer	CGH Explorer	CGH PRO	CGH Miner
Platform	most platforms (Java-based)	most platforms (Java-based)	most platforms (Java-based)	Windows	most platforms (Java-based)	most platforms (Java-based)	Linux, Windows	Windows, Unix, Excel add-in
Software availability	freely downloadable with registration	freely downloadable	freely downloadable	freely downloadable with registration	freely downloadable with registration	freely downloadable with registration	freely downloadable	freely downloadable with registration
Source-code license	GNU GPL	MIT license	GNU GPL	not available	freely downloadable	freely downloadable	GNU GPL	freely downloadable
External software dependencies	none	none	none	requires MySQL database	none	none	MySQL, R	R
Automatic statistical determination of single-array aberrations	Yes	No	No	No	No	Yes	Yes	Yes
Statistical analysis of manually selected regions	Yes	No	No	No	Yes	Yes	No	No
Simultaneous display of multiple experiments	Yes	No	No	No	Yes	Yes	Yes	Yes
Statistical analysis of aberrations occurring in multiple experiments	No	No	No	No	Yes	No	No	No
Aberration breakpoints/statistics export	Yes	No	No	No	Yes	Yes	Yes	Yes
Image export	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Command-line statistical analysis feature	Yes	No	No	No	No	No	No	No
Allows user-defined genomic feature annotation	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

from microarray data and can display multiple experiments and data types simultaneously. ChARM-View can be used to identify amplifications and deletions from array CGH or gene expression data independently or simultaneously, making it a powerful approach for identifying real and functionally relevant chromosomal changes.

3.3 Methodology: Statistical Analysis

ChARMView computational analysis automatically detects regions of non-random spatial bias and is appropriate for any genomic data associated with chromosomal coordinates. Statistical analysis is based on our algorithm ChARM (Chromosomal Aberration Region Miner), described in detail in [19]. ChARM identifies potential breakpoints by a differential filter followed by an accurate expectation-maximization approach. The statistical significance of each identified region is evaluated with a one-sample sign test and a permutation-based mean test. By their formulation, the significance tests are valid for any size segment, but do lose power with decreasing segment size. ChARM has been evaluated on gene expression and array CGH data: it is robust and accurate for regions as small as 4-5 probes, and sensitive enough to detect aneuploidies even in mixed populations of cells [19].

As a system for dynamic and real-time data analysis and visualization, ChARMView requires very fast statistical algorithms. However, the permutation-based test as originally described in Myers *et al.* [19] requires non-trivial computation since it involves performing several thousand permutations of the chromosome order. To speed up the mean permutation test for the software system, we have developed an accurate approximation that requires many fewer permutations. The original version of the test requires computing the mean of the region of interest and comparing this with the means of similar-sized segments in randomly permuted data. We have verified that means of typical chromosomal segments in array CGH and gene expression data can generally be reasonably approximated with a normal distribution. This is a generally well-accepted claim even for small groups (~10) unless the underlying population is extremely non-normal, which is typically not the case for log-transformed array CGH or gene expression data. The statistical significance of predicted aneuploidy region in

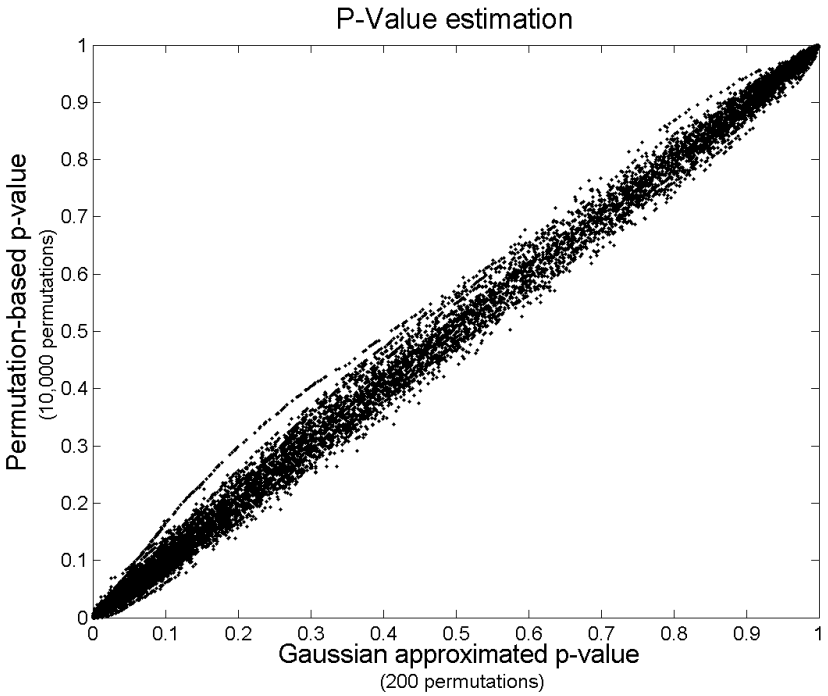


Figure 3.1. Gaussian approximation of the permutation-based p-value. ChARMView performs 200 permutations to assess statistical significance of predictions, then uses Gaussian approximation to estimate the p-value. The p-values based on 200 permutations and Gaussian approximation show .9991 correlation to the p-values based on 10,000 permutations as computed by ChARM.

ChARMView is obtained by computing means of 200 permutations of chromosome ordering of the actual data, estimating the parameters, and then integrating the tail of the underlying distribution beyond the observed value. Figure 3.1 illustrates the correlation between p-values generated from 10,000 random permutations and p-values obtained from a normal approximation whose parameters were estimated with only 200 permutations. This approximation yields the precision of several thousand permutations based on significantly less computation. Completing a fully automated statistical analysis on a typical gene expression dataset (6000 genes over 16 chromosomes, measured in 16 experiments) requires approximately 7 seconds/experiment for a total of less than 2 minutes on a Pentium 4 3.2 GHz desktop. ChARMView also allows users to manually select regions to test for statistical significance.

3.4 Methodology: Visualization-based Analysis

The most powerful aspect of ChARMView is integration of computational analysis with visualization. This combination of visualization and analysis enables users to view automated predictions of aneuploidies as well as analyze statistical significance of manually selected regions. Visualization is a critical complement to computational analysis as human perception can often identify subtle trends in the data that cannot be detected with purely computational methods. This is especially critical when comparing results of multiple experiments or experimental replicates, such as in cancer studies where researchers often search for recurring aneuploidies in a set of patients. ChARMView facilitates such discovery with visualization of multiple experimental replicates, experiments, and data types.

The most common way to increase confidence in results of an experiment is to produce replicate microarray experiments. Data from such replicate experiments is usually averaged for computational analysis. However, viewing such replicates simultaneously is an effective approach to analysis, as people are often perceptive of subtle but repeated trends that are difficult to capture with a statistical test. This visualization-based approach does not make any assumptions, such as independence assumption of the typically used Fisher meta-analysis test [8]. Thus, aligning corresponding chromosomal data from several replicates of the same experiment typically allows the user to spot trends that might otherwise go unnoticed. Figure 3.2 illustrates this phenomenon with two replicates of the same array CGH experiment.

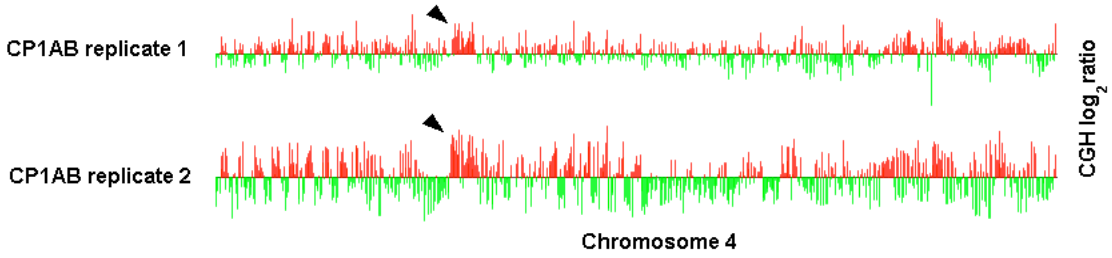


Figure 3.2. Simultaneous visualization of replicate aCGH experiments. A set of replicate array CGH experiments from Dunham *et al.* [6] displayed with ChARMView (chromosome 4 of CP1AB, replicates 1 and 2 shown). The region identified by the arrows is hard to distinguish from noise in either of the replicates when viewed separately, but is clearly a region of positive bias when the replicates are viewed together. This is confirmed by statistical analysis.

The simultaneous display feature of ChARMView is also useful for visual analysis of computational prediction results for multiple experiments. This is an effective method for identifying common genomic aberrations in otherwise uncorrelated experiments or a characteristic aberration in a set of samples with a common phenotype. For example, a set of breast cancer samples [23,26] can share the same bias in gene expression that corresponds to a predicted aneuploidy or a localized expression bias, as shown in Figure 3.3. Overlapping predictions serve as independent confirmations that the predicted aberration is real. Furthermore, results of such analysis of multiple samples can then be used to correlate specific chromosomal aberrations with phenotypic or clinical parameters.

As array CGH techniques become more widely applied, the generation of copy number data is rarely the end goal of biological studies. Instead, a key challenge is deciphering which parts of a karyotypic profile are responsible for particular phenotypes. While sophisticated statistical and computational methods will certainly be required to answer these questions, the most effective approaches will also need to harness the power of human visual perception. To address this issue, ChARMView can display and analyze both array CGH and gene expression microarray data and display these diverse data and predictions for corresponding chromosomes simultaneously. Simultaneous display of array CGH and gene expression data enables researchers to observe the effect that amplification or deletion of particular sequences of genomic DNA has on the abundance of mRNA transcripts (Figure 3.4). We have noted a number of cases where large amplifications or deletions result in no detectable change in gene expression. These regions may be less likely to cause a particular phenotype than aneuploidies that result in drastic changes in gene expression. ChARMView facilitates convenient discovery of these changes, focusing further experimental investigation.

A final unique characteristic of ChARMView is that its visualization and statistical tools are developed for general use, independent of data type and organism. Any dataset with features that can be associated with chromosomal position can be imported and analyzed with

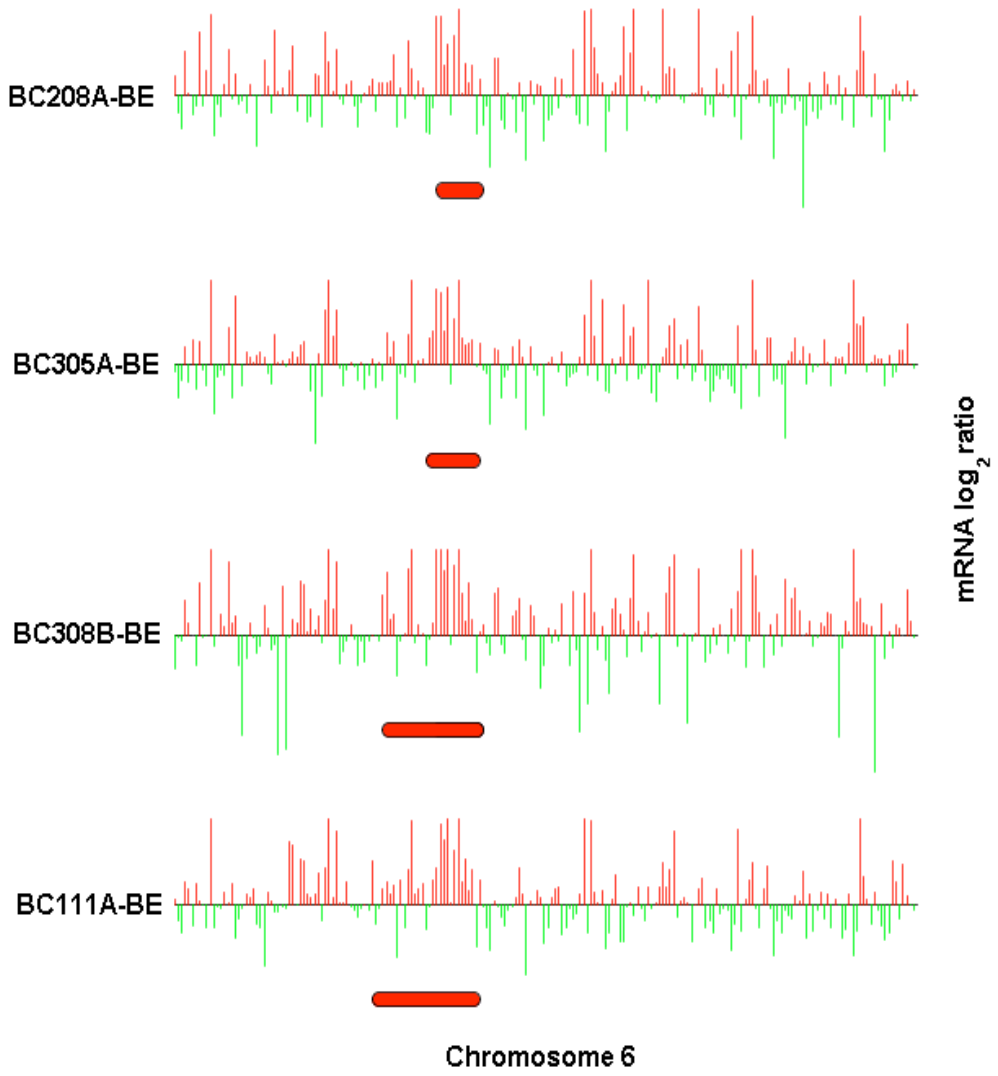


Figure 3.3. Simultaneous visualization of multiple independent expression microarrays. Simultaneous visualization of overlapping significant expression biases in a set of four independent breast tumor samples from Sorlie *et al.* [26] (chromosome 6 of breast tumor expression profiles BC208A-BE, BC305A-BE, BC308B-BE, and BC111A-BE shown). Each red bar below the data indicates a predicted aberration identified independently on the corresponding experiment.

ChARMView. For instance, the software has been particularly useful in identification of aneuploidies based on gene expression datasets although array CGH is the typical experimental approach for probing genomic amplifications or deletions. ChARMView has also been used to identify spatially-correlated biases in gene expression that are unrelated to altered chromosome structure. Generally, our tool can be used to identify any region of non-randomness with respect to position in genomic data with inherent ordering. In addition to its usefulness for a variety of data types, ChARMView can be applied to a variety of organisms. By default, the system

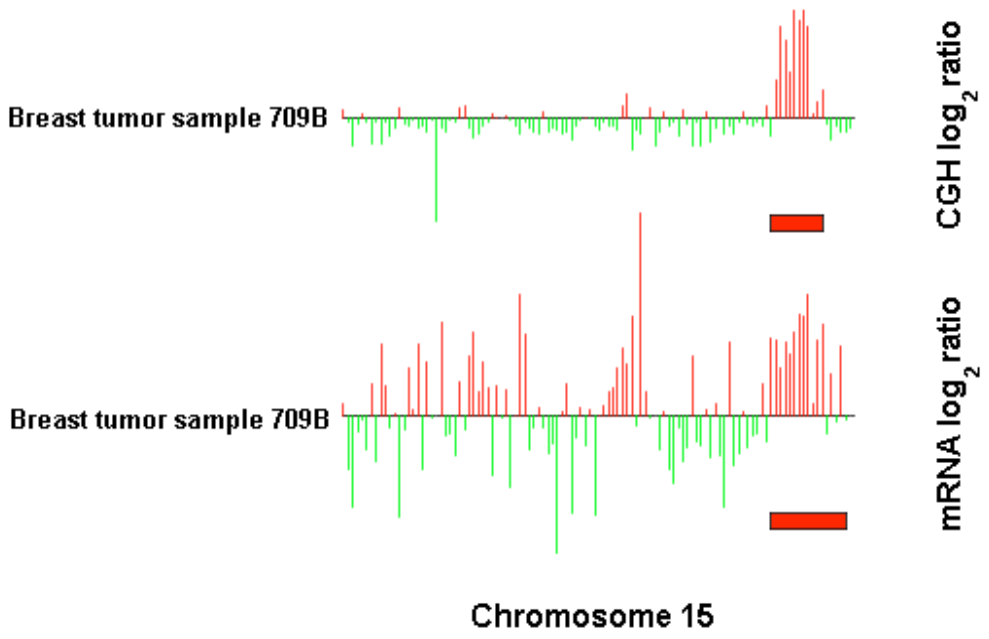


Figure 3.4. Identifying functionally relevant genomic aberrations. A small amplification evident in the array CGH data breast cancer data (top) [23], and its effect on mRNA expression (bottom) [26] (chromosome 15 of breast tumor sample 709B shown).

provides chromosomal coordinates for *Saccharomyces cerevisiae* data with ORF identifiers and human data with Unigene identifiers. However, any data that can be mapped to a set of linear chromosomes can be imported and analyzed by ChARMView.

3.5 Illustration of Application

We have applied ChARMView to a number of array CGH and gene expression datasets, including data derived from both *Saccharomyces cerevisiae* and human experiments. Here we present an example application of our software to array CGH data from experimental evolution experiments in which eight strains of budding yeast were analyzed for chromosomal copy number changes after 100–500 generations of growth in glucose-limited chemostats [6]. Dunham *et al.* confirmed aneuploidy regions identified by array CGH through pulsed-field gel electrophoresis, thus creating a standard for assessing our results. Our method identified all 12 of the confirmed aneuploidies and two additional regions of bias. The novel regions identified by our method correspond to biases smaller than the ones identified by

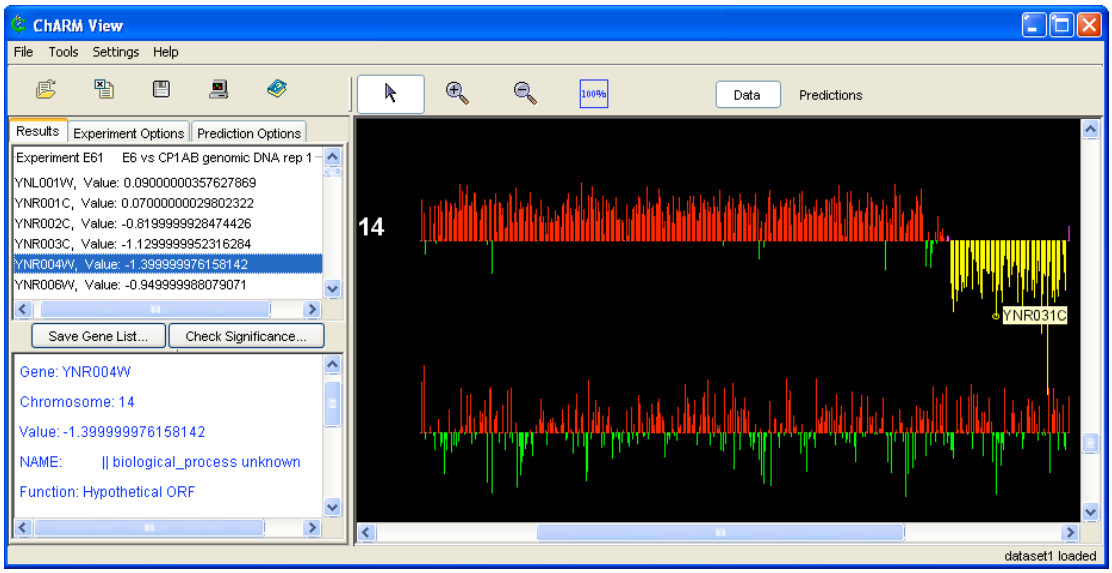


Figure 3.5. Screenshot of ChARMView applied to *S. cerevisiae* array CGH data. Screenshot of ChARMView analysis of *S. cerevisiae* molecular evolution experiments data from Dunham *et al* [6]. The right panel displays array CGH data arranged in the order of chromosomal position and amplification (red) and deletion (green) predictions. The left panel displays information for the selected region, including gene names and values and statistics for the selected amplification prediction.

Dunham *et al.* [6] and may reflect aneuploidy present in a subset of cells in the population or may be due to a hybridization artifact. Further laboratory experiments are required to further evaluate these predictions. Figure 3.5 shows a screenshot of our application upon finishing automated statistical analysis of one of these experiments.

We also present two specific instances from an array CGH breast cancer study where ChARMView can be used to visualize and accurately predict breakpoints of known amplifications. Figure 3.6A illustrates the results of ChARMView's automated statistical analysis on chromosome 1 array CGH profiles of three different breast tumor samples (110B, 112B, 122A) from [23]. The entire q arm of chromosome 1 is known to frequently amplified in breast cancer (typically observed in approximately 50-60% of tumors [9,24]). Thus, we expect the amplications here to begin at or near the centromeric end of the q arm. ChARMView predicts breakpoints 3, 1, and 0 probes from the centromeric end of the q arm for samples 110B, 112B, and 122A respectively.

ChARMView can also be used to accurately find much smaller regions of amplification or deletion and the associated breakpoints. Figure 3.6B illustrates this capability on chromosome 17 array CGH profiles of three breast tumor samples (123B, 309A, and BC-A) from

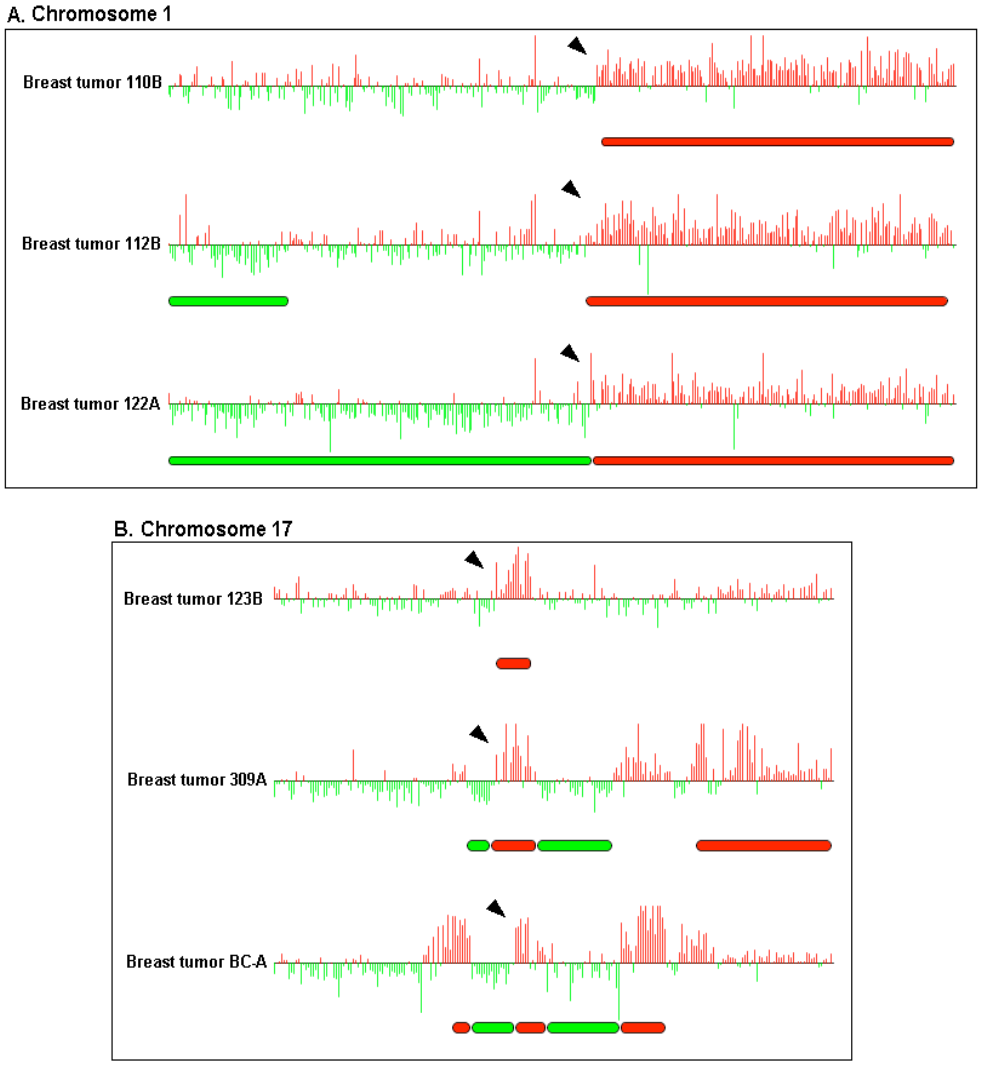


Figure 3.6. Predicted amplifications and deletions on breast cancer array CGH data. ChARMView automated predictions on three breast tumor array CGH profiles (110B, 112B, 122A) from chromosome 1 and three profiles (123B, 309A, and BC-A) from chromosome 17 of [23]. The predicted chromosome 1 breakpoints (identified by arrows in Figure 6A) are 3, 1, and 0 probes from the centromere. The predicted chromosome 17 amplification common to all three profiles (identified by arrows in Figure 6B) includes 7 genes known to be typically amplified with the *ERBB2* locus. All visible predictions have Bonferroni-corrected p-values less than .05 for both mean and sign significance tests. See Table 2 for a complete list of breakpoint predictions for each of the results pictured.

[23]. An amplicon frequently associated with breast tumors includes the *ERBB2* oncogene at 17q12. While breakpoints identified in individual tumors vary, recent studies have identified a group of 7 genes surrounding the *ERBB2* locus that are commonly amplified, including *NEUROD2*, *MLN64, PNMT, ERBB2*, *GRB7*, *ZNFN1A3,* and EST 48582 [15,16]. ChARMView's amplification predictions for the three tumor profiles shown include 15, 18, and 13 probes

respectively, all of which span the 7-gene region previously identified. All predictions shown in Figure 3.6 have Bonferroni-corrected p-values less than .05 for both mean and sign significance tests. Complete lists of predicted breakpoints for both chromosome 1 and chromosome 17 amplicons are included in Table 3.2.

Table 3.2. Examples of predicted breakpoints in breast tumor aCGH case study. Listing of Unigene IDs corresponding to predicted breakpoints for ChARMView results pictured in Figures 3.6A and 3.6B. The Unigene ID and gene name are the first and last markers included in the predicted amplification. All results listed have Bonferroni-corrected less than p-values of .05 for both mean and sign significance tests.

Tumor sample	Chrom.	Predicted start breakpoint	Adjacent gene (in amplicon)	Predicted end breakpoint	Adjacent gene (in amplicon)
110B	1	Hs.15871	ACP6	Hs.7395 (last marker)	TFB2M
112B	1	Hs.59889	HMGCS2	Hs.7395 (last marker)	TFB2M
122A	1	Hs.381235	SEC22L1	Hs.7395 (last marker)	TFB2M
123B	17	Hs.97477	LYZL6	Hs.276916	NR1D1
309A	17	Hs.73817	CCL18	Hs.267871	PTRF1
BC-A	17	Hs.635	CACNB1	Hs.2340	HAP1

3.6 Implementation and Usage

(logo design by Matt Hibbs)

ChARMView was implemented in Java using Swing set components to ensure cross-platform compatibility. Many of the visualization features were developed using the Open Source 2D graphics toolkit Piccolo developed at the University of Maryland [2]. ChARMView can be downloaded at http://function.princeton.edu/ChARMView and run on virtually any platform if the J2SE Java Runtime Environment version 1.4.2 or greater is present. A brief overview of the primary features of the software follows.

Loading data

ChARMView accepts all types of data from any organism provided that the features can be ordered on a set of linear chromosomes. Input files must be tab-delimited, specifically in the commonly-used .pcl format. Chromosome labels and position must be included in the input file unless the organism type is *Saccharomyces cerevisiae* or human with ORF or Unigene identifiers, which ChARMView is able to order without coordinates.

Viewing data

Figure 3.5 shows a typical ChARMView screenshot upon loading data and statistical analysis. The data display is zoomable and selectable with mouse-overs for identification of experiments and individual genes. Zoom features include standard single-click magnification, zoom to rectangle, and zoom reset (fit to screen) capabilities. When one or more gene or probe data points are selected, identifiers and associated annotation are displayed in the "Results" tab, which appears adjacent to the display panel. This allows users to select regions of interest on the display panel and retrieve lists of genes or probes within these regions. Additionally, any number of experiments may be viewed simultaneously by toggling the corresponding checkboxes in the "Experiment Options" tab, also adjacent to the display panel.

Analyzing data

ChARMView supports two different modes of analysis. The first employs the automated edge-finding algorithm discussed in Myers *et al.* [19] followed by statistical analysis. The second mode is for testing user-selected regions of data and only evaluates the statistical significance of the chosen region. Both methods of analysis rely on two tests of statistical significance: a mean-based permutation test, and a one-sample sign test. Details of both tests are discussed above and in Myers *et al.* [19]. P-values for these tests are reported for all regions found by the automated approach or selected by the user. Figure 3.5 displays a typical view of statistical results for a single experiment. Note that the red and green rectangles below the data correspond to regions of predicted aberration. The p-value cutoff at which results of the statistical

analysis appear in the display panel can be adjusted by applying p-value filters provided in the “Prediction Options” tab adjacent to the display panel.

A p-value filter consists of a logical combination of the mean permutation test and/or the one-sample sign test and real-valued cutoffs for each test. These combinations specify how the selected p-value cutoffs will be used to deem statistical significance. For instance, one possible p-value filter is "Sign AND Mean Tests" with Sign p-value cutoff of 0.001 and Mean p-value cutoff of 0.01, which will result in only predictions with both Bonferroni corrected sign p-values of less than .001 and mean p-values of .01 being displayed. The Bonferroni corrected p-value is obtained by multiplying the raw p-value from both significance tests by the number of regions tested for that chromosome. Another possibility is to apply "Sign OR Mean Tests", which results in a prediction being displayed if at least one of these criteria is met at the specified significance level. While we recommend the "Sign AND Mean Test" option for general use, other combinations may be useful under certain circumstances. Users can select any displayed prediction, which results in the genes or probes and associated annotation in that particular region to be displayed in the “Results” tab adjacent to the display panel (Figure 3.5).

Exporting results

Publication quality images can be exported in multiple formats at any stage of the visualization. This includes images of exclusively raw data, results of statistical analysis, or combinations of these. In addition, predictions resulting from automated or manual statistical analysis can be exported in tab-delimited text form with the associated gene or probe identifiers and corresponding p-values. A p-value filter similar to that described in “Analyzing data” can be applied to all exported results to allow full user control over which predictions are included. Finally, lists of genes or probes for any object selected on the display panel can also be exported to text files to facilitate immediate analysis of regions identified by manual inspection.

Command-line usage

ChARMView can also be used in command-line mode to make automated predictions of amplification or deletions. This command-line feature can be used by invoking ChARMView as follows:

```
java -Xmx300m -jar ChARM.jar  –inputFile <input-file>
                              -outputFile <output-file>
                              -organismType <organism-type>
                              -meanPvalCutoff <mean-pvalue-cutoff>
                              -signPvalCutoff <sign-pvalue-cutoff>
                              -sigTestType <significance-test-type>
```

The possible organism types, which determine reference chromosomal coordinates, are: 1, *Saccharomyces cerevisiae*; 2, human; 3, other (user-provided coordinates). Possible significance test options include: 1, mean AND sign tests; 2-mean OR sign tests; 3, mean test only; 4, sign test only. When run in command-line mode, ChARMView outputs all predicted regions of amplification and deletion meeting the specified significance level.

3.7 Conclusions

We have developed ChARMView, a statistical visualization system for analysis and discovery of genomic aberrations. Our system can analyze various types of genomic data, including gene expression and array CGH microarray data, for a variety of organisms, and has been developed to facilitate both manual discovery through powerful visualization as well as automated prediction through robust statistical analysis. ChARMView can identify and visualize even small copy number changes, and is sensitive enough to detect aneuploidies in mixed populations of cells. This combination makes ChARMView uniquely effective for identifying subtle trends, recurring aberrations in sets of experiments, and pinpointing functionally relevant copy number changes. Thus, this system is effective for identification of aneuploidies in cancer studies and molecular evolution experiments, as well as for routine analysis of microarray data for special biases.

References

1. Awad, I. A., C. A. Rees, et al. (2004). "Caryoscope: An Open Source Java application for viewing microarray data in a genomic context." BMC Bioinformatics **5**(1): 151.
2. Bederson, B. B., J. Grosjean, et al. (2004). "Toolkit design for interactive structured graphics." Ieee Transactions on Software Engineering **30**(8): 535-546.
3. Cahill, D. P., K. Kinzler, et al. (1999). "Genetic instability and darwinian selection in tumours." Trends Cell Biol **9**(12): M57-M60.
4. Chen, W., F. Erdogan, et al. (2005). "CGHPRO -- a comprehensive data analysis tool for array CGH." BMC Bioinformatics **6**(1): 85.
5. Chi, B., R. J. DeLeeuw, et al. (2004). "SeeGH--a software tool for visualization of whole genome array comparative genomic hybridization data." BMC Bioinformatics **5**(1): 13.
6. Dunham, M. J., H. Badrane, et al. (2002). "Characteristic genome rearrangements in experimental evolution of Saccharomyces cerevisiae." PNAS **99**(25): 16144-16149.
7. Fischer, G., S. A. James, et al. (2000). "Chromosomal evolution in Saccharomyces." Nature **405**(6785): 451-454.
8. Fisher, R. (1932). Statistical Methods for Research Workers. London, Oliver and Boyd.
9. Forozan, F., E. H. Mahlamaki, et al. (2000). "Comparative genomic hybridization analysis of 38 breast cancer cell lines: a basis for interpreting complementary DNA microarray data." Cancer Res **60**(16): 4519-25.
10. Fritz, B., F. Schubert, et al. (2002). "Microarray-based copy number and expression profiling in dedifferentiated pleomorphic liposarcoma." Cancer Res **62**(11): 2993-2998.
11. Greshock, J., T. L. Naylor, et al. (2004). "1-Mb resolution array-based comparative genomic hybridization using a BAC clone set optimized for cancer gene analysis." Genome Res **14**(1): 179-87.
12. Haddad, R., K. A. Furge, et al. (2002). "Genomic profiling and cDNA microarray analysis of human colon adenocarcinoma and associated intraperitoneal metastases reveals consistent cytogenetic and transcriptional aberrations associated with progression of multiple metastases." Applied Genomics and Proteomics **1**: 123-134.

13. Hendrickson, H., E. S. Slechta, et al. (2002). "Amplification-mutagenesis: Evidence that "directed" adaptive mutation and general hypermutability result from growth with a selected gene amplification." PNAS **99**(4): 2164-2169.

14. Hughes, T. R., C. J. Roberts, et al. (2000). "Widespread aneuploidy revealed by DNA microarray expression profiling." Nat Genet **25**(3): 333-337.

15. Kauraniemi, P., M. Barlund, et al. (2001). "New amplified and highly expressed genes discovered in the ERBB2 amplicon in breast cancer by cDNA microarrays." Cancer Res **61**(22): 8235-40.

16. Kauraniemi, P., T. Kuukasjarvi, et al. (2003). "Amplification of a 280-kilobase core region at the ERBB2 locus leads to activation of two hypothetical proteins in breast cancer." Am J Pathol **163**(5): 1979-84.

17. Lingjaerde, O. C., L. O. Baumbusch, et al. (2005). "CGH-Explorer: a program for analysis of array-CGH data." Bioinformatics **21**(6): 821-2.

18. Myers, C. L., X. Chen, et al. (2005). "Visualization-based discovery and analysis of genomic aberrations in microarray data." BMC Bioinformatics **6**(1): 146.

19. Myers, C. L., M. J. Dunham, et al. (2004). "Accurate detection of aneuploidies in array CGH and gene expression microarray data." Bioinformatics **20**(18): 3533-43.

20. Phillips, J., S. W. Hayward, et al. (2001). "The consequences of chromosomal aneuploidy on gene expression profiles in a cell line model for prostate carcinogenesis." Cancer Res **61**(22): 8143-8149.

21. Pinkel, D., R. Segraves, et al. (1998). "High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays." Nat Genet **20**(2): 207-211.

22. Pollack, J. R., C. M. Perou, et al. (1999). "Genome-wide analysis of DNA copy-number changes using cDNA microarrays." Nat Genet **23**(1): 41-46.

23. Pollack, J. R., T. Sorlie, et al. (2002). "Microarray analysis reveals a major direct role of DNA copy number alteration in the transcriptional program of human breast tumors." Proc Natl Acad Sci U S A **99**(20): 12963-8.

24. Rennstam, K., M. Ahlstedt-Soini, et al. (2003). "Patterns of chromosomal imbalances defines subgroups of breast cancer with distinct clinical features and prognosis. A study of 305 tumors by comparative genomic hybridization." Cancer Res **63**(24): 8861-8.

25. Saldanha, A. J. (2004). "Java Treeview." Bioinformatics **20**(17): 3246-3248.

26. Sorlie, T., R. Tibshirani, et al. (2003). "Repeated observation of breast tumor subtypes in independent gene expression data sets." Proc Natl Acad Sci U S A **100**(14): 8418-23.

27. Wang, P., Y. Kim, et al. (2005). "A method for calling gains and losses in array CGH data." Biostatistics **6**(1): 45-58.

Chapter 4

Inferring Biological Networks from Diverse Genomic Data

4.1 Chapter Overview

The previous two chapters have focused on using the chromosomal context of microarray data to make inferences about chromosomal aberrations. In this chapter, we transition to the more general problem of making inferences about biological networks based on large collections of diverse genomic data including gene expression, protein-protein interactions, genetic interactions, cellular localization and sequence information. Specifically, we describe, bioPIXIE, a general probabilistic system, we have developed for query-based discovery of pathway-specific networks. We illustrate both computational and experimental validation of this framework by accurately recovering known networks for 31 biological processes in *Saccharomyces cerevisiae* and experimentally verifying predictions for the process of chromosomal segregation and the protein chaperone Hsp90.

Much of the work presented in this chapter is published in [35] and includes contributions from Drew Robson, Adam Wible, Matt Hibbs, Camelia Chiriac, Chandra Theesfeld, Kara Dolinski, and Olga Troyanskaya. Drew and Adam developed the first prototype and web-interface for the system described here, and Matt provided suggestions on both the integration methods and interface design. Camelia contributed all experimental results described in this chapter, and Chandra and Kara provided biological interpretation of the Hsp90 findings and cross-talk analysis. Olga supervised the project.

4.2 Background: from Diverse Genomic Data to Networks

Understanding biological networks on a whole-genome scale is a key challenge in modern systems biology. Broad availability of diverse functional genomic data from protein-protein

interaction, gene expression, localization, and regulation studies should enable fast and accurate generation of network models through computational prediction and experimental validation. However, reliability of experimental results varies among data sets and technologies, and these data generally provide only pair-wise evidence for biological relationships between genes or proteins. Most cellular mechanisms, on the other hand, involve groups of genes or gene products that behave in a coordinated way to perform a specific biological process. We will refer to such groups of functionally related genes as process-specific networks. Although a wide variety of functional genomic data is available, and much has been learned from them, we are far from exploiting the full potential of these data for discovering such process-specific networks. There are several reasons for this: lack of accessibility to data and methods to analyze them, barriers to incorporating expert knowledge in the network discovery process, and noise and heterogeneity in high-throughput gene data.

The first problem is simply the lack of accessibility of both the data and analysis methods. Even when data are publicly available, results are often buried in large files, and computational methods developed to analyze them are often not available in forms that the typical biologist can use. Thus, experimental researchers are unable to identify interesting results from computational studies that are worth verifying. Instead, most biologists are limited to what the authors' of such studies deem important or interesting enough to highlight in the written publication. Our ability to effectively utilize genomic data for process-specific network discovery has thus been hampered by the lack of effective interfaces to both the data and the relevant analysis methods.

The second challenge is to allow biology researchers to integrate their biological knowledge in analysis. When biologists inquire about particular biological processes, they bring with them existing knowledge that can and should be used to generate the most sensitive and precise hypotheses possible. Such information is hard to extract automatically, and effectively incorporating biological expert knowledge is of course closely linked to the accessibility challenge noted above. Most previous methods for process-specific network prediction have not allowed biologists to use their previous knowledge in their area of interest to target the analysis process.

Biological research demands convenient and accessible systems that leverage existing knowledge to direct and facilitate discovery.

The third challenge in constructing accurate process-specific networks from diverse genomic data lies in the heterogeneity and high noise levels in large-scale data sets. High-throughput data by nature are often noisy and simple combinations of results from different types of experiments (e.g. conclusions of genome-scale two-hybrid experiments and microarray studies) are of limited effectiveness because they sacrifice either sensitivity or specificity.

Recent applications of probabilistic data integration to the related but simpler problem of predicting protein function from diverse genomic data have demonstrated that integrated analysis of heterogeneous sources provides a substantial increase in prediction accuracy. Much of the work in function prediction focuses on fusing information from multiple heterogeneous sources for pairs of proteins to make more reliable statements about pair-wise functional relationships. Bayesian networks [26,48] and variations of this approach [50,30,25] have been applied successfully to construct “functional linkage maps” whose connecting edges represent probabilistic support for a functional relationship between the adjacent proteins. Protein functions are then inferred through “guilt by association” with surrounding nodes of known function. Several studies have formalized this “guilt by association” approach by using Markov Random Field models to propagate known functional annotations through confidence-weighted edges [32,13,28].

Despite much investigation into heterogeneous data integration for the purpose of function prediction, there have been only limited attempts to use confidence-weighted linkage maps from integrated data to address the more biologically significant problem of how to group functionally related proteins together into process-specific networks. These network-level questions are distinctly different from function prediction problems and require new methodology for general data integration and network discovery. Previous work in identifying groups of genes involved in specific biological pathways from interaction networks has focused on mainly binary interactions, which are prone to false positives and inadequate coverage when only limited types of genomic evidence are used. For instance, two studies [5,45] describe approaches for finding

highly-connected subgraphs in binary interaction graphs from high-throughput experiments. They found that highly connected groups in these graphs often correspond to protein complexes or biological processes. Another study [18] introduced the notion of modular decomposition of protein-protein interaction networks to make inferences about pathways. While these approaches have demonstrated the promise of using protein-protein interaction networks for recognizing groups of proteins involved in specific processes, they are constrained by their reliance on limited types of interaction data and their use of binary, rather than probabilistic networks. A recent study extended these approaches to a weighted interaction network and used graph clustering analysis to detect coordinated functional modules [38]. A common theme among many of these studies is their unsupervised approach to network detection. However, incorporating expert knowledge in the search process can dramatically improve both the specificity and sensitivity of process-specific network discovery from protein-protein interaction data.

To our knowledge, the only existing work that leverages expert knowledge in constructing biological networks or protein complexes from integrated data is a network reliability approach to protein complex recovery [4] and a greedy search algorithm applied to a confidence-weighted protein-protein interaction network [6]. The former was specifically targeted towards protein complexes, while we focus on the more general problem of discovering not just physically interacting sets of proteins, but functional or process-specific networks. The latter algorithm [6] leveraged both physical and genetic interaction data with the goal of extracting more general protein networks. Distinctions between this work and our approach are that we integrate functional genomic data in a Bayesian framework that allows a probabilistic, rather than heuristic, graph search. This probabilistic search incorporates both direct and indirect protein-protein links while integrating a wider variety of data (e.g. microarray expression, co-localization). Furthermore, we are the first to our knowledge to develop an interactive, web-accessible system that both facilitates discovery of novel biological networks and allows exploratory analysis of the underlying genomic data that supports these predictions.

To address these challenges to discovering process-specific networks from functional genomic data, we have created a publicly available system called bioPIXIE (biological Process

Inference from eXperimental Interaction Evidence). The system allows users to enter a set of proteins and then uses a novel probabilistic graph search algorithm on a protein-protein linkage map derived from diverse genomic data to predict the surrounding process-specific network for the local neighborhood of interest. Most importantly, the system includes a convenient interface for dynamic visualization of the resulting predictions and provides analysis of their functional coherence. We have completed an extensive evaluation of our method against known pathways as well as experimentally verified a subset of predictions made by our system.

4.3 Methods for Inferring Networks from Diverse Data

Our method relies on four critical components: (1) Bayesian integration of heterogeneous data, (2) an expert-driven search paradigm, (3) a probabilistic graph search algorithm, and (4) an easily accessible interface for interpretation of the results (Figure 4.1). In simple terms, bioPIXIE integrates different types of data (gene expression, interaction data, high-throughput or single experiments, etc.) using a Bayesian framework that is learned from proteins (or genes) that are known to be functionally linked. This Bayesian data integration step reduces the heterogeneous input data to protein pairs with a score indicating the likelihood that they functionally interact, allowing different types of data to be combined with each other. Then, given a protein or group of proteins as a query set (the expert-driven search component), a novel probabilistic algorithm considers the integrated pair-wise relationships to build a local process-specific network around the query proteins.

4.3.1 Bayesian Integration of Heterogeneous Data

This component uses a Bayesian network to integrate diverse data to derive a probabilistic linkage map among proteins.

Functional genomic input data

We have collected a diverse set of evidence from over 950 publications from several databases, including complete physical and genetic interaction data from the GRID and BIND databases

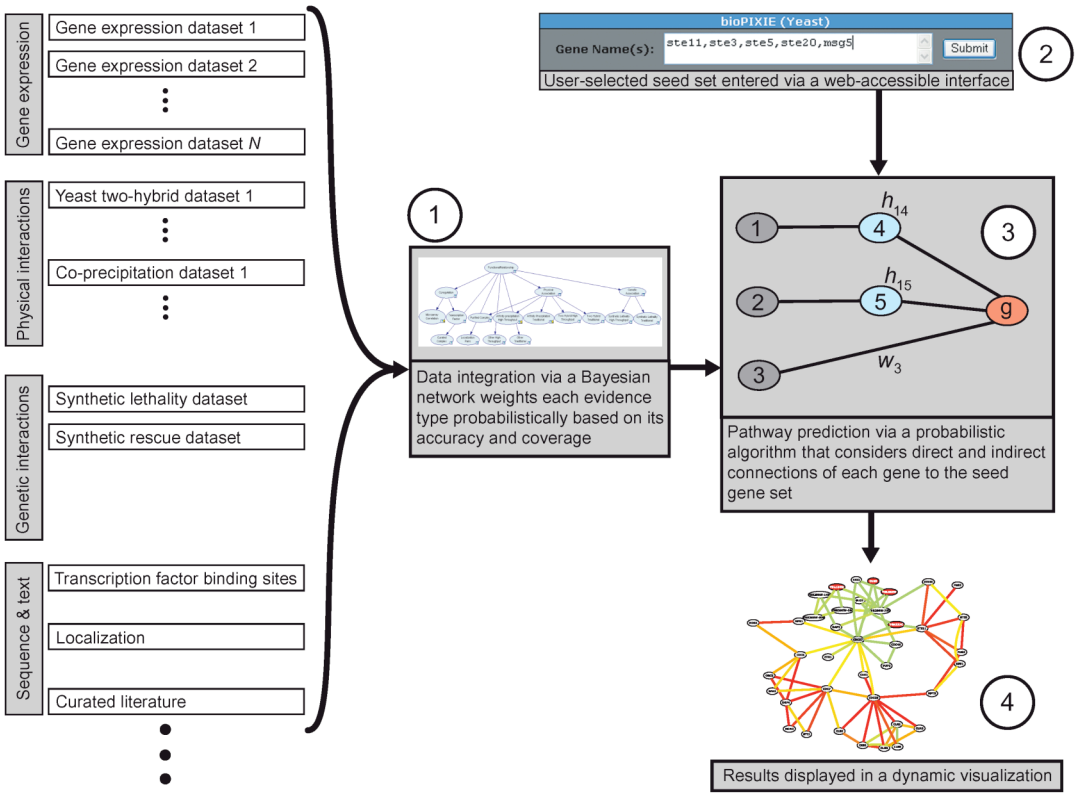


Figure 4.1. Overview of the bioPIXIE system. Diverse data sets are integrated with a Bayesian network, which weighs each evidence type probabilistically based on its accuracy (1). This Bayesian integration produces a graph with confidence-weighted relationships between each gene pair (characterized in supplemental Figure S1). Based on this integrated network graph and a user-defined query set of proteins of interest (2), the network prediction algorithm identifies novel network components by finding proteins with the maximum expected number of direct and indirect relationships with the query set (3). The resulting network is then displayed to the user using a spring model layout, such that the geometric proximity of genes reflects how related they are to each other, and the edge color reflects the confidence of pair-wise connections (4). Details of each component are presented in Methods.

(downloaded on 6/25/04), which contain both high-throughput interaction data sets and some interactions from individual experiments curated from the literature [8,10,3]. We also make use of cellular localization data [23], curated sequence data in the form of shared transcription factor binding sites from the *Saccharomyces cerevisiae* Promoter Database (SCPD) [56], and biological complex curated literature from the Saccharomyces Genome Database (SGD) [8]. Additionally, we have collected gene expression data from 10 different microarray studies, totaling more than 300 arrays and 29 distinct biological conditions [14,11,44,20,37,46,55,19,54,43]. Pearson correlation between genes across each set of related conditions is used as a measure of

similarity. Correlation coefficients in each dataset are converted to Z-scores and combined across datasets. References to all sources of genomic data are available as a Supplementary file.

Bayesian network structure and conditional probabilities

Given these diverse data, we can answer questions about pair-wise protein relationships using a Bayesian network that leverages our previous work [48]. A Bayesian network essentially weights each evidence type according to a measure of confidence in the source of that evidence and then estimates the posterior probability that a relationship exists between two proteins given all observed data [16]. The critical components of such a network are the structure, which determines relationships between evidence nodes, and the conditional probability tables (CPTs), which capture the reliability of each evidence type. The structure of the network used here is expert-based and derived from our previous work [48]. Unlike our previous work which also relied on experts for estimating the CPTs, here we generalize the framework and automatically learn the CPT for each evidence type using protein-protein relationships inferred by the GO biological process ontology.

Specifically, we obtained gold standard protein-protein relationships for learning the network CPTs by propagating each biological process annotation up to its ancestors and counting the number of unique annotations per GO term. Because the biological specificity of each term roughly corresponds to the number of total annotations, we chose two thresholds to define the set of positive (functionally related) and negative (not functionally related) protein pairs. Protein pairs whose most specific co-annotation occurs in GO terms of 300 total annotations or less were considered positives, while pairs whose most specific co-annotation occurs in GO terms of 1000 total annotations or *more* were considered negatives. The resulting set of positive and negative protein pairs is available as a supplementary file and can also be downloaded from the online supplement [51].

Given this set of gold standard pairs, we used the expectation-maximization algorithm [12] to compute the CPTs. As EM is guaranteed to identify a local, not global, maximum on the

likelihood surface, we computed a reasonable starting point for the algorithm based on independent counting of individual evidence sources. We used a discrete Bayesian network, and continuous-valued microarray expression correlation was discretized into 16 bins (see additional data file 1 for details). Both the structure and final learned conditional probabilities are available as a supplementary data file. The final probabilistic output of the Bayesian network for the whole yeast proteome is also available as supplementary file. We have performed cross-validation analysis by excluding all related GO relationships from the gold standard for each pathway we attempt to predict.

4.3.2 Expert-driven Search Paradigm

A critical aspect of our method is that we make use of existing expert biological knowledge to improve the accuracy of process-specific network prediction by allowing the biologist to drive the search process. Specifically, the user enters a list of proteins (of arbitrary size) he or she either expects to play a role in the same biological process, or wants to test for functional relationships. Our system then queries the surrounding confidence-weighted network derived from integrated data for additional related proteins. The resulting process-specific network is not a simple sub-section of the complete integrated protein-protein interaction graph; rather it is probabilistically biased by the graph search algorithm (described in detail below) toward the biological process represented in the set of query proteins. This paradigm is based on two important observations: (1) detailed knowledge of specific biological processes is typically learned in a directed fashion, not by taking a completely unsupervised view of high-throughput data, and (2) novel process-specific proteins can be predicted more precisely when we consider their relationship to groups of known proteins simultaneously. This query-driven process results in a view of the integrated genomic data in the context of the specific process being interrogated. Figure 4.3, discussed in detail in Results, illustrates this behavior for Rad23, a DNA repair protein.

4.3.3 Probabilistic Graph Search Algorithm

Given an initial set of query proteins defined by the user, we wish to find other proteins with significant connectivity back to the starting group. It is unrealistic to expect related proteins to have direct connections to all other proteins in the same biological process due to incomplete data. Furthermore, there are often protein pairs involved in the same process whose relationship

Table 4.1. Overview of graph search algorithm.

Start with user-defined query set of related proteins.

1. Find the n_1 *direct* neighbors with largest connections to the query set.
2. Find the n_2 *direct* or *indirect* neighbors with largest connections to the query set, requiring that all indirect paths pass through proteins from Step 1.
3. Return $n_1 + n_2$ proteins and associated links.

is not present in existing experimental data. Thus, we measure connectivity back to the original query set via both direct and indirect relationships. A brief overview of the algorithm is presented in Table 4.1. Because we used a Bayesian approach to data integration, weights of edges connecting pairs of proteins are precisely the posterior probability of a functional relationship between the proteins given all observed evidence for the pair, i.e. for each edge weight, e_{ij}, in the integrated network,

$$e_{ij} = P\left(\text{protein } i \text{ is functionally related to protein } j \mid \text{evidence}\right).$$

Given this formulation, the existence of any pairwise biological relationship can be treated as a Bernoulli random variable, X_{ij}, with probability of success e_{ij}. The number of direct relationships protein p_i shares with the original query set, Q, can then be found by summing over all p_i's connections to proteins in Q. Letting the random variable $S_Q(p_i)$ denote this sum, we obtain

$$S_Q(p_i) = \sum_{p_j \in Q} X_{ij} .$$

Then, the *expected* number of direct relationships to the query set for protein p_i is

$$E\left[S_Q(p_i)\right] = E\left[\sum_{p_j \in Q} X_{ij}\right] = \sum_{p_j \in Q} E\left[X_{ij}\right] = \sum_{p_j \in Q} e_{ij} .$$

Since not all proteins involved in a particular process will have high-probability direct relationships with other members of the same process, we also need to measure indirect connectivity to the query set. However, from a biological standpoint, not all indirect connections are actually meaningful. We expect there are a limited number of high-probability adjacent neighbors of the query set through which indirect connections are meaningful. Thus, our approach relies on a two-step search approach where a pre-defined number of direct neighbors are found (first neighborhood, referred to as N_1) after which the maximally connected indirect neighbors adjacent to the first neighborhood and the original query set are added (second neighborhood, referred to as N_2). Letting the random variable $S_{N_1 \to Q}(p_i)$ denote the number of 2-step indirect connections between protein p_i and the query set (Q) *through* first neighborhood proteins (N_1), we obtain

$$S_{N_1 \to Q}(p_i) = \sum_{p_k \in Q} \sum_{p_j \in N_1} X_{ij} X_{jk}$$

and the *expected* number of indirect connections through the first neighborhood is

$$E\left[S_{N_1 \to Q}(p_i)\right] = E\left[\sum_{p_k \in Q} \sum_{p_j \in N_1} X_{ij} X_{jk}\right] = \sum_{p_k \in Q} \sum_{p_j \in N_1} e_{ij} e_{jk} .$$

Here, we implicitly assume independence of X_{ij} and X_{jk}. This requires that the existence of a relationship between any proteins p_i and p_j be independent of the relationship between proteins p_j and p_k, which is a reasonable assumption. Also, we do not consider indirect connections beyond 2 steps from the query set. We have empirically evaluated the algorithm for more distant indirect relationships, but found the performance on 2-step relationships superior. The search algorithm is summarized in Table 4.2. We have empirically determined that a first

neighborhood of between 10 and 20 proteins (i.e. $10 \le n_1 \le 20$) provides the best precision and recall over a wide range of biological processes. This was determined by optimizing the

Table 4.2. Probabilistic graph search algorithm.

Allow user to determine query set, Q.

1. Find $\acute{I}_1 \equiv \left\{ n_1 \text{ proteins with largest } E\left[S_Q(p_i)\right] = \sum_{p_j \in Q} e_{ij} \right\}$
2. Find $\acute{I}_2 \equiv \left\{ n_2 \text{ proteins with largest } E\left[S_{N_1 \to Q}(p_i)\right] + E\left[S_Q(p_i)\right] = \sum_{p_k \in Q} \sum_{p_j \in N_1} e_{ij} e_{jk} + \sum_{p_k \in Q} e_{ik} \right\}$
3. Return $\{\acute{I}_1, \acute{I}_2\}$.

difference of recall and impurity (1-precision) with respect to the first neighborhood size (data not shown). The number of second neighborhood proteins returned (n_2) is determined by the density of the local network and the limits of the user interface. Thus, second neighborhood proteins are added to the graph until the total number of proteins reaches 40 or no neighbors with links exceeding the prior probability of interaction remain. From an information visualization perspective, a typical user is unable to draw useful information from interaction graphs of more than 40 proteins.

4.3.4 Publicly Available Interface

We provide public, web-based access to our integrated process-specific network analysis and visualization system [52]. This allows biologists to browse the integrated set of functional genomic data for proteins of interest, and explore our network predictions. Furthermore, users can directly query specific links leading to the reported predictions, an important part of the analysis pipeline.

4.3.5 Implementation

(logo design by Matt Hibbs)

The Bayesian network used in integrating genomic data was implemented using SMILE, a C++ library developed by the Decision Systems Laboratory at the University of Pittsburgh[53]. The user interface tool, GeNIe, useful for developing and analyzing Bayesian models was also used extensively during the development of bioPIXIE [53]. bioPIXIE's web interface is implemented in PHP and all genomic data is stored in a MySQL database. The graph server which performs probabilistic searches and renders results is implemented in C++ and renders graphs in SVG, which allows for user-friendly browsing and interactivity. AT&T's Graphviz is used for layout of all graphs.

4.4 Evaluation on Known Biological Networks

Our system achieves accurate network prediction by effectively integrating diverse data sets and probabilistically identifying new components of process-specific networks given only one or a few known members. We evaluated the ability of our approach to recover known process-specific networks given initial query sets by using a collection of well-annotated functional groups including KEGG pathways, sets of biological process GO terms, and MIPS protein complexes. We restricted our evaluation to groups of 15 to 250 total proteins in which at least half of the member proteins had one type of evidence linking them with another member protein. We identified 31 such groups from the set of KEGG pathways, MIPS protein complexes, and GO terms (see supplementary data file 2). We evaluated the performance of our method on each group by sampling 100 random query sets consisting of 10 proteins each from the pathway or complex of interest, applying our data integration and search algorithm, and analyzing the returned set of proteins for consistency with the remaining proteins in the group.

The advantage of using bioPIXIE to integrate multiple types of genomic data is illustrated in Figures 4.2 A, B, and C for three diverse KEGG pathways (graphs for all 31 processes are available in supplemental Figure S2 on the bioPIXIE website [51]). bioPIXIE dramatically and consistently improves the number of network components recovered over any of the individual types of evidence. For example, for KEGG cell cycle proteins (Figure 4.2A), given a random 10-protein query set, we identify an average of 42 of the remaining 77 proteins using integrated data,

whereas only 25 are identified by either physical or genetic evidence, and only 18 by microarray evidence alone. Different evidence types have varying degrees of relevance for different pathways—microarray correlation is very informative for ribosome proteins (Figure 4.2B) while physical interactions are more informative for proteins involve in ATP synthesis (Figure 4.2C).

This advantage of integrating diverse data types is confirmed in a more comprehensive evaluation of bioPIXIE's performance, where we averaged results over the entire set of 31 processes and complexes described above. Figure 4.2D compares the precision-recall characteristics of our network identification method using Bayesian integrated data versus using individual evidence types. Given only 10 query genes, the integrated version recovers 50% of the remaining members at a precision of 30% while the method applied to independent subsets achieves only 15% (physical association), 10% (genetic association), and 3% (microarray correlation) precision at the same recall (Figure 4.2D). Thus, combining data from multiple sources clearly improves network recovery.

One might expect that due to the relative sparsity of current functional genomic data, simple combinations of these sources followed by a straightforward search would be sufficient for precise network recovery. However, such combinations are substantially less effective than our approach as shown in Figure 4.2E, which plots the average precision-recall characteristics of two such approaches to integration and recovery. The first approach ("Binary recovery") uses all available evidence, but only as a binary "yes" or "no" depending on whether evidence of any type is present for a particular protein pair. Given a query, connected proteins are then added in an arbitrary order. The second approach ("Counting-based recovery") also uses all available evidence but counts observed evidence for each pair such that overlaps between multiple sources of evidence receive higher weights. Proteins are then added in order of weight for network recovery. Neither of these simpler approaches achieves accuracy similar to that of our method. In fact, the counting-based approach yields a 4-fold lower prediction precision than our approach and the binary approach results in a 10-fold lower prediction precision at 50% recall.

In addition to these two naive methods, we have also compared our system to two previously published methods for query-based protein complex discovery, SEEDY and

Complexpander [6,4]. bioPIXIE's performance is superior to both existing methods; it achieves an average of 30% precision at 50% recall while SEEDY yields 12% and Complexpander 7% at

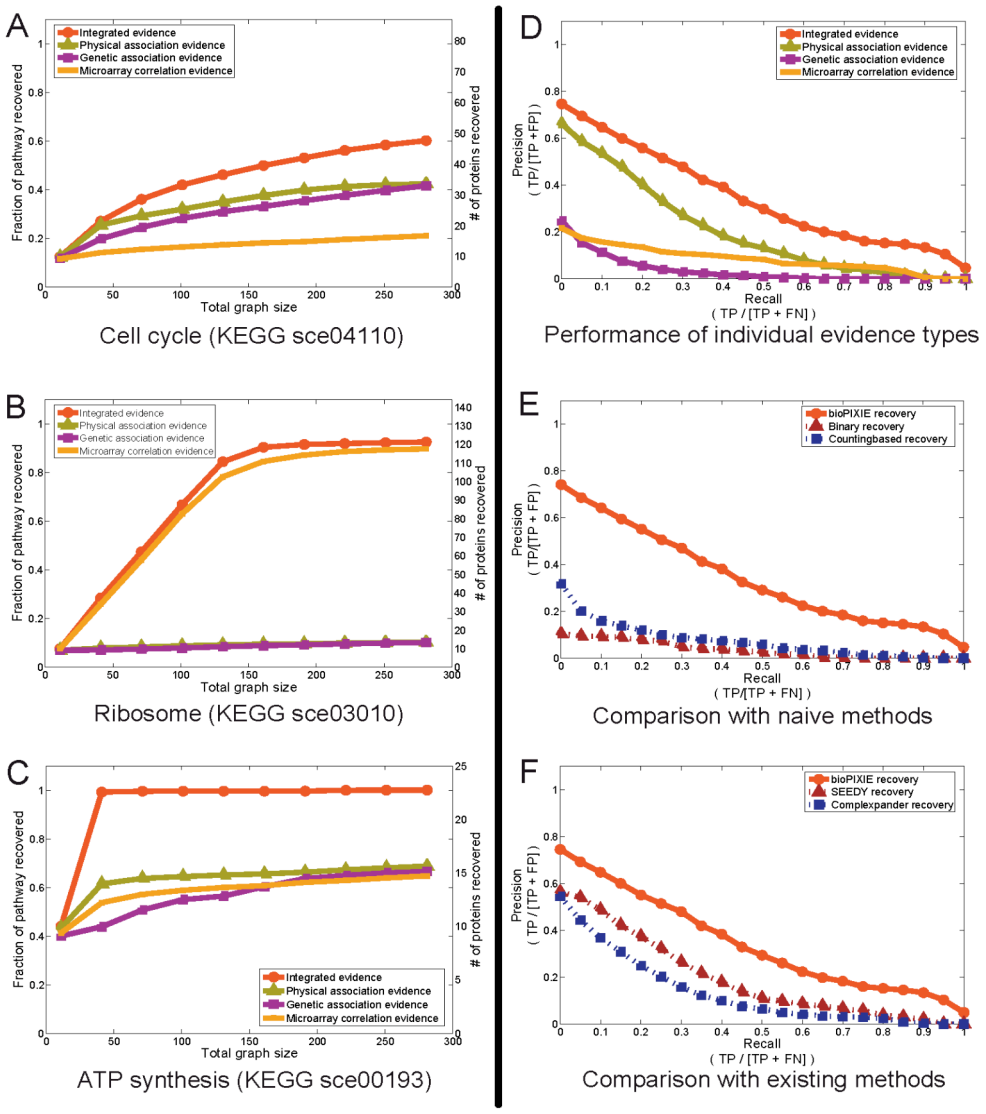


Figure 4.2. bioPIXIE network recovery evaluation. Figures 2A, B, and C illustrate typical network recovery performance for three KEGG pathways. For all pathways,10 proteins from the pathway were randomly picked as a query set and results shown are averages of 100 independent samplings. The fraction of the total known process components recovered is plotted versus the size of the graph grown from the query set. Figures 2D, E, and F represent an average over 31 KEGG pathways, GO biological processes, and MIPS complexes. Performance is measured and reported as the trade-off between precision (the proportion of correct pathway components returned to the total size of the returned network) and recall (the proportion of correct pathway components returned to the number of total non-query pathway proteins). Figure 2D illustrates the improvement gained by using our network prediction algorithm on a Bayesian integration of genomic evidence as compared to separate evidence types. bioPIXIE shows considerable improvement in both the number of known member proteins recovered and the precision of predicted members for the integrated evidence over any individual evidence type. Figure 2E illustrates the improved network recovery offered by the bioPIXIE algorithm versus more naïve approaches to integration and graph search. Specifically, we plot the performance of bioPIXIE on integrated data against a naïve binary approach for which information from all evidence types is used but only as a binary “yes” or “no” relationship, and a more sophisticated approach where overlapping evidence receives higher weights and connected proteins are recovered in order of confidence. Figure 2F compares the performance of bioPIXIE to two existing methods for query-based protein complex recovery [6,4].

50% recall (Figure 4.2F). Furthermore, calculating the average area under the precision-recall curve (AUC) for each pathway individually, we find that the average bioPIXIE AUC exceeds the average SEEDY AUC by more than one standard deviation for 22 of the 31 groups, while SEEDY outperforms bioPIXIE for only 1 of the 31 groups (see supplementary data file 3). Similarly bioPIXIE outperforms Complexpander for 26 of the 31 groups, while the converse never occurs (supplementary data file 3).

Another important characteristic of our method is its robustness to the quality and size of the query set. For each of the 31 groups of proteins described earlier, we evaluated the recovery performance for 20 query proteins of which between 1 and 19 were randomly chosen from the entire proteome and the rest were chosen from the appropriate process or complex. All 31 groups could tolerate 25% query set noise with less than a 10% reduction in the average AUC, 27 of those could tolerate 50% query set noise, and 14 of those could tolerate up to 75% random proteins in the query set (see Appendix A). Thus, our method is robust to imperfect query sets. We also evaluated the recovery performance over a range of query set sizes from 4 to 60 proteins to determine whether there was a noticeable decline in performance for very small query sets. We found that, in general, the quality of the network recovered from a pure query set of 4-5 proteins is comparable to the result of a much larger query (i.e. 40-50 proteins) on the same process, suggesting that relatively few proteins are required to obtain a signal (Appendix A). For instance, with only a 4-protein query set, bioPIXIE's maximum AUC score was within 10% of the maximum AUC score obtained on up to 60-protein query sets for 22 of the 31 processes (see Appendix A for supporting plot).

The query-driven nature of the search algorithm is a key factor in the accuracy of our method. The relationships between query proteins selected by the user affect which neighboring proteins are added to the final network. Thus, the network resulting from a query is not simply a sub-section of the complete integrated protein-protein interaction graph rooted at the query proteins; rather, it is probabilistically biased by the network search algorithm toward the specific biological context represented in the query set. Figure 4.3 illustrates this effect for the query protein Rad23. Rad23 is known to form a complex with Rad4 (NEF2) and participate in

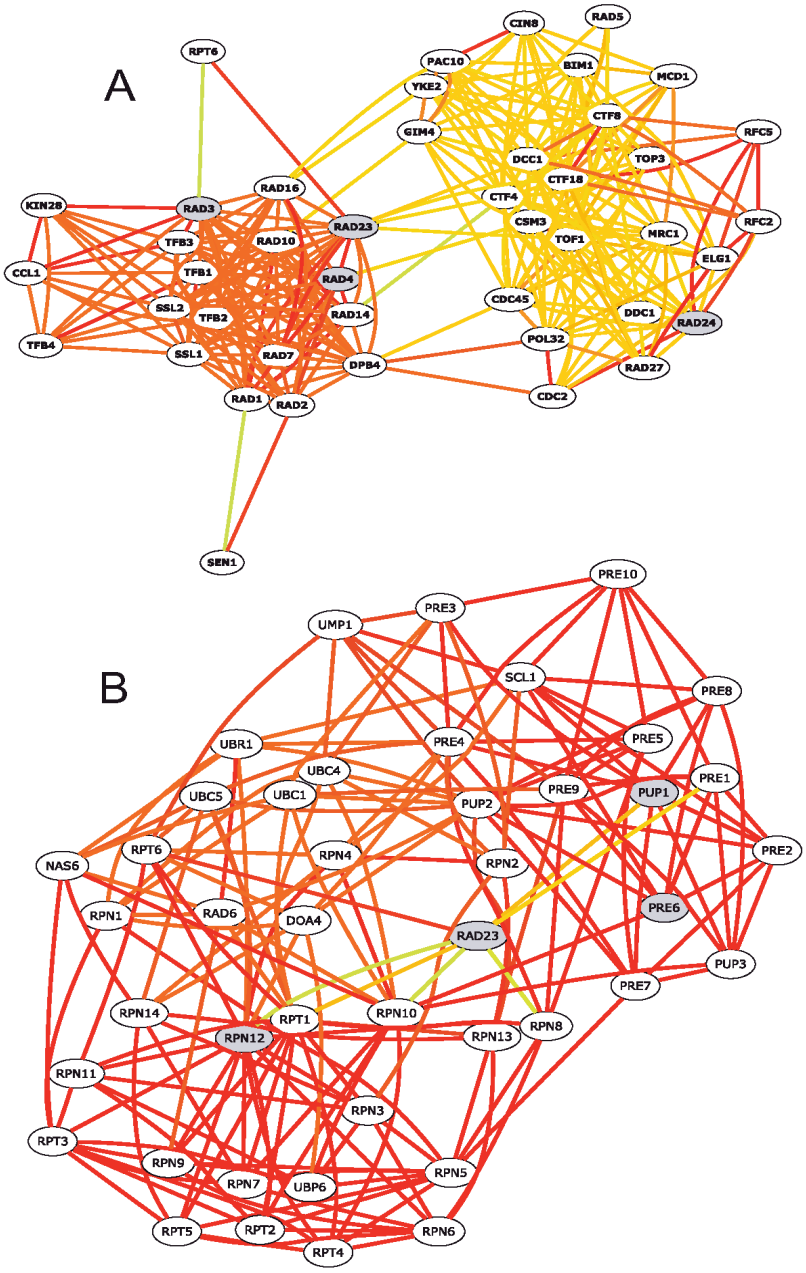


Figure 4.3. bioPIXIE query-driven context illustration. Nodes represent proteins, and edges represent functional links between them. Edge color indicates the confidence of the link (red edges are high confidence while green edges are low). Query proteins are indicated by gray nodes. Rad23 is known to form a complex with Rad4 (NEF2) and participate in nucleotide excision repair and has also been implicated in inhibiting the degradation of specific substrates in response to DNA damage. For Figure 3A, Rad23 was entered with Rad4, Rad3, and Rad24 and the resulting network is enriched (22 of 44, p-value $< 10^{-22}$) for DNA repair proteins (GO:0006281). For Figure 3B, Rad23 was entered with proteasome components Pup1, Pre6, Rpn12 and the recovered network is enriched (36 of 44, p-value $< 10^{-55}$) for ubiquitin-dependent catabolism proteins (GO:0006511) and only contains 2 DNA repair proteins (Rad6 and Rad23). Rad23 has high-confidence relationships with several proteins in both processes, but the network recovery algorithm is dependent on the context of the query, which results in two different views of Rad23 and its neighbors.

nucleotide excision repair [39]. Recent work has also suggested that Rad23 facilitates DNA repair by inhibiting the degradation of specific substrates in response to DNA damage [41,49]. Depending on which partners are included in a query with Rad23, the network recovered by our system can focus on Rad23's involvement in nucleotide excision repair or in ubiquitin-dependent protein catabolism. For instance, when the query includes DNA repair proteins Rad4, Rad3, and Rad24 in addition to Rad23, the recovered network of 44 total proteins (Figure 4.3A) is highly enriched for DNA repair (GO:0006281), with 22 of the 44 having direct or indirect annotations (p-value $< 10^{-22}$). However, when Rad23 is entered as a query with proteasome components Pup1, Pre6, Rpn12, the resulting network (Figure 4.3B) is instead enriched for ubiquitin-dependent catabolism (GO:0006511), with 36 of the 44 having direct or indirect annotations (p-value $< 10^{-55}$). Rad23 has high-confidence relationships with several proteins in both processes, but the recovered network returned by our system is dependent on the context implied by the query. This query-driven context facilitates accurate recovery of network components related to the biological process or pathway of interest.

4.5 Biological Validation of BioPIXIE

4.5.1 Experimental Validation of Novel Network Predictions

bioPIXIE does not simply recapitulate known biology, but it also predicts novel network components based on the diverse types of input data. In fact, the "false positives" identified by bioPIXIE in the evaluation above may be novel discoveries or known proteins that interact very closely with the biological process in question but are not annotated to it by the current standard. Thus, although the computational evaluation above is an accurate comparative evaluation of the methods, we wanted to experimentally confirm the quality of predictions made by our method. We have done so by using bioPIXIE to generate hypotheses about previously uncharacterized proteins in yeast and experimentally testing these hypotheses. Specifically, for several biological processes of interest, we entered member proteins as queries and identified uncharacterized proteins consistently returned in the predicted networks. One biological process with high-confidence uncharacterized proteins was the process of chromosomal segregation. When

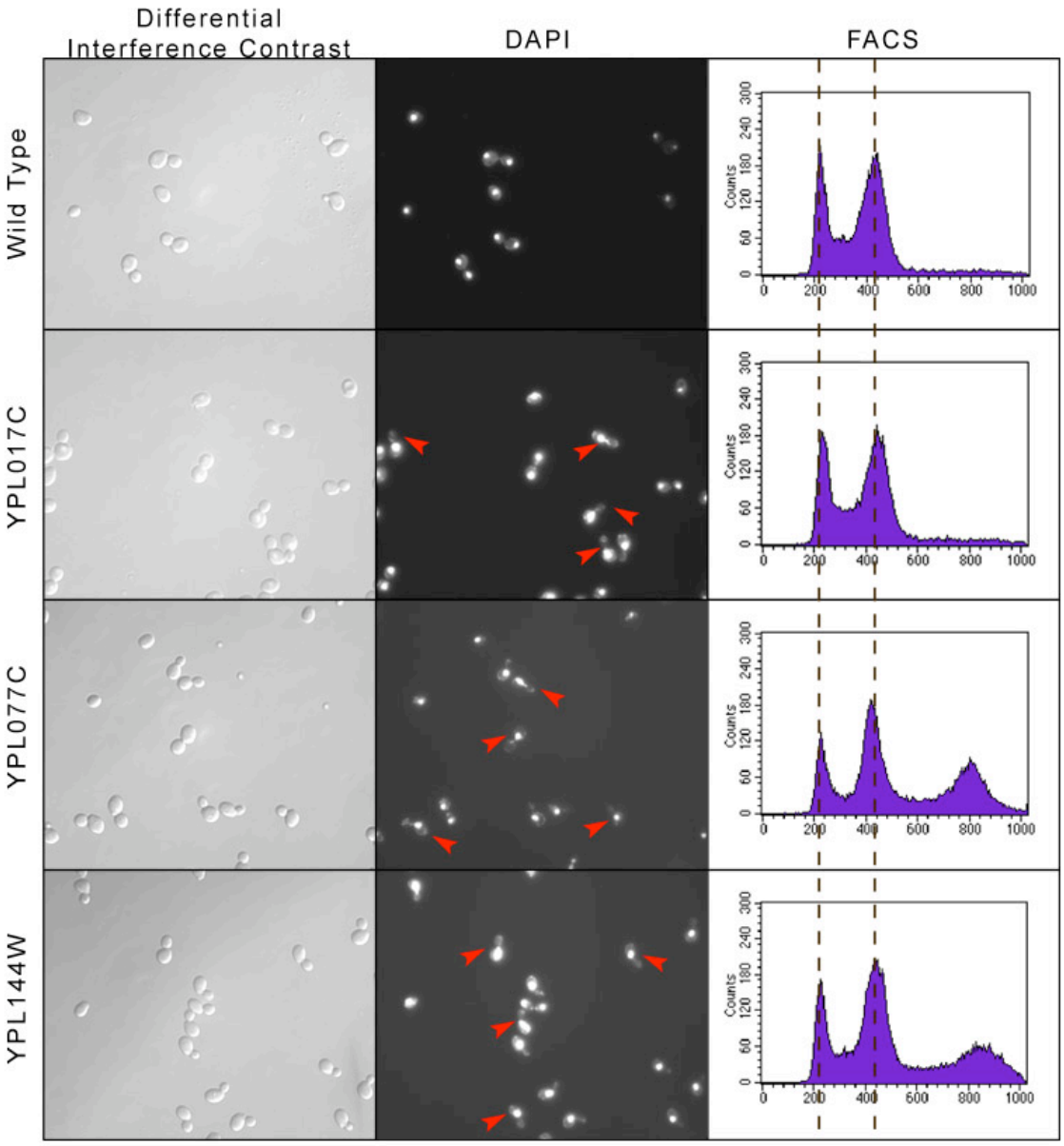


Figure 4.4. Experimental validation of bioPIXIE prediction for biological role of YPL017C, YPL077C and YPL144W. bioPIXIE was used to predict previously uncharacterized genes likely to participate in processes related to chromosomal segregation. Yeast cells were fixed, stained, and photographed using differential interference contrast imaging and DAPI staining. When compared with wild type cells, populations of cells lacking YPL017C, YPL077C or YPL144W have a higher proportion of large-budded cells with a single nucleus at the bud neck (75%, 55% and 62% as compared to 22% in wild type. Large budding cells are indicated by arrows. This morphology and failure of nuclear separation are analogous to that of ctf4Δ mutants [33], supporting the hypothesis that YPL017C, like CTF4, is involved in chromosome segregation.

compared with wild type cells populations of cells lacking the genes YPL017C, YPL077C or YPL144W all exhibit nuclear defects, containing a larger proportion of large budding cells with only one nucleus (indicated by arrows) as well as an increased fraction of clump morphologies (Figure 4.4). Nuclear defects, and sometimes ploidy problems, are both phenotypes associated with defects in chromosomal segregation. This difference is statistically significant in all three mutants, with the percentage of defective large budded cells at 22% in wild type, but at 75% in the YPL017C null mutant (p-value = 5.0x10^-9 , Fisher's exact test), 55% in the YPL077C null mutant (p-value = 1.98x10^-7 , Fisher's exact test), and 62% in the YPL144W null mutant (p-value = 1.17x10^-9 , Fisher's exact test). This morphology and failure of nuclear separation is consistent with the phenotype of mutants known to affect chromosome segregation such as ctf4Δ [33]. This example demonstrates that bioPIXIE facilitates experimental design by providing high-confidence predictions that can be readily tested experimentally using standard molecular biology techniques.

4.5.2 Example Use of the System: Prediction of Novel Targets for the Cdc37-Hsp90 Complex

We expect that bioPIXIE will be a convenient and effective tool for biologists to explore the growing sets of functional genomic data as well as direct further experimentation in their domains of interest. As an example of this type of exploratory analysis, we used bioPIXIE to examine the Cdc37-Hsp90 complex and found evidence for previously uncharacterized roles in important processes. Hsp90 is a molecular chaperone that participates in the folding of several proteins, including signaling kinases and hormone receptors, which are involved in growth and apoptotic pathways; it has thus been identified as a possible anticancer drug target. Hsp90 is a highly conserved protein found in organisms from bacteria to humans, and there are two Hsp90 homologs in yeast, HSC82 and HSP82 (reviewed in [21,7,9]).

Using bioPIXIE, we were able to identify known and novel targets of Hsp90 and its co-chaperones, in particular Cdc37. Cdc37 and other proteins associated with Hsp90 are thought both to function as chaperones themselves and potentially to determine Hsp90 target specificity.

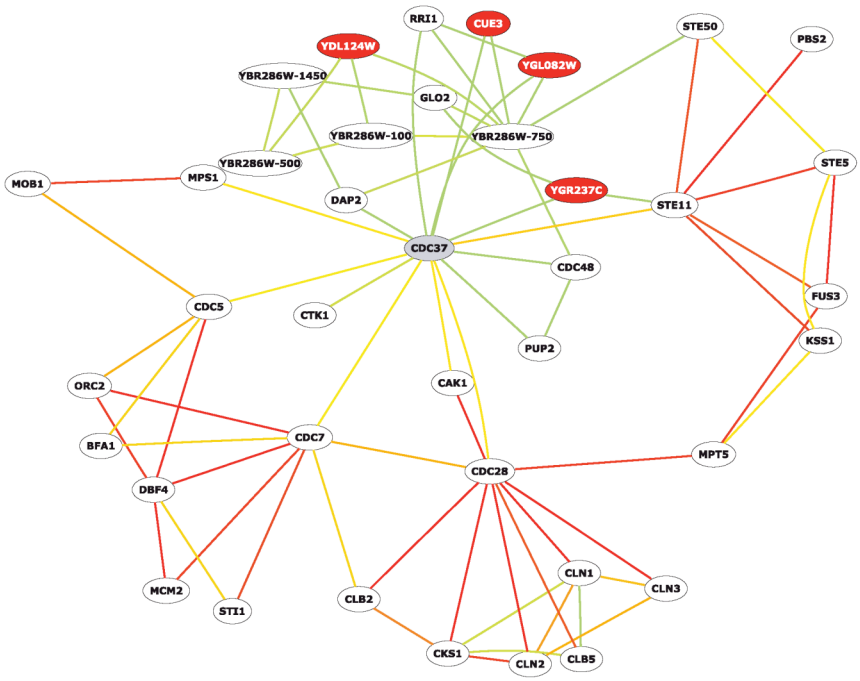


Figure 4.5. bioPIXIE output for Cdc37. Nodes represent genes, and edges represent functional links between them. Edge color indicates the confidence of the link (red edges are high confidence while green edges are low). In this example, *CDC37* was entered as input (gray node); other genes displayed (white nodes) were identified by the bioPIXIE prediction algorithm. Red nodes indicate that the gene is uncharacterized.

Cdc37 interacts with Hsp90 and is involved in the folding of protein kinases (CDKs, MAP kinases), and previous work has suggested that Cdc37 might be a general kinase chaperone [24]. When Cdc37 is entered as a seed protein into bioPIXIE, our algorithm detects associations between Cdc37 and several kinases that are known interaction partners (Cdc28 [21,17,34], Mps1 [42], Cak1 [17,34], Ste11 [1,31], Cdc5 [34]) (Figure 4.5). In addition, bioPIXIE predicts previously uncharacterized connections between Cdc37 and the protein kinase Ctk1, based on high-throughput affinity precipitation, thus providing further support for the hypothesis that Cdc37 may be a general kinase chaperone.

Furthermore, our algorithm predicts a potential novel role of the Cdc37-Hsp90 complex in DNA replication. Specifically, bioPIXIE identifies connections between components of this complex and Cdc7, a serine/threonine kinase involved in replication origin firing, which is regulated by Dbf4 in a manner analogous to the way that CDKs are regulated by cyclins [27]. Our system predicts this interaction (confidence of .49) based on a combination of two hybrid evidence and correlated expression data. Although this putative interaction was identified in a two

hybrid screen, it was not further characterized [34]. In further support of the DNA replication link, bioPIXIE also identifies previously uncharacterized interactions between Cdc7 and two other members of the Hsp90 complex, Sti1 and Cpr7. Sti1 is also functionally linked to Dbf4, a regulator of Cdc7, by the algorithm on the basis of a high-throughput genetic interaction [47] and correlated gene expression in a microarray experiment [20]. Because our system integrates diverse data sources, it highlights interesting interactions that may otherwise go unnoticed. Furthermore, bioPIXIE's network identification and interactive exploration features allow generation of novel, experimentally testable hypotheses, in this case that Cdc37- Hsp90 complexes may have a previously uncharacterized role in some aspect of DNA replication.

4.5.3 Experimental Evidence for an Hsp90 Role in DNA Replication

To investigate the hypothesis that Hsp90 plays a role in DNA replication in yeast, we experimentally characterized several mutants involving Cdc7-Dbf4 and Hsp90 and its co-chaperones. Indeed, we find there is strong evidence for this Hsp90-DNA replication link. Specifically, we identify several novel genetic interactions supporting this hypothesis including *cdc7-1—hsc82Δ*, *cdc7-1—cpr7Δ*, *cdc7-1—sti1Δ*, *cdc7-1—cdc37Δ*, and *dbf4-1—cpr7Δ*, and we confirm a previously known interaction between *dbf4-1—sti1Δ* (Figures 4.6, 4.7, and 4.8). We further investigated whether Hsp90 mutants exhibit defects in DNA replication by testing sensitivity to hydroxyurea (HU), which specifically inhibits DNA replication. We confirm hypersensitivity of the *hsc82Δ* (Hsp90 yeast homolog) single mutant under moderate exposure to HU, particularly at 37 ^{o}C (Figure 4.9). This sensitivity is consistent with the phenotypes of *cdc7-1* and *dbf4-1*, which are known to initiate DNA replication. This validation provides a compelling example where exploration through bioPIXIE was used to propose a specific, non-trivial hypothesis, which these experimental data suggest is correct.

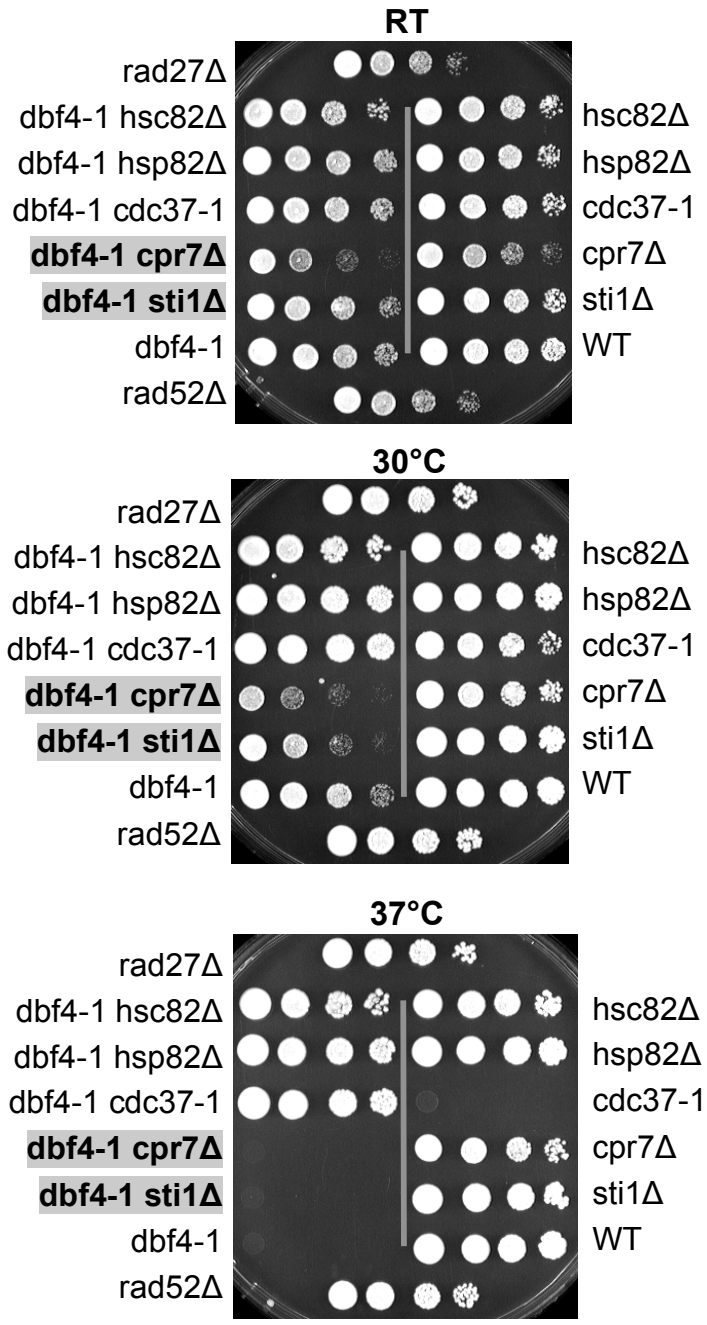


Figure 4.6. Single and double mutants between Hsp90 and co-chaperones and *dbf4-1*. All combinations of Hsp90 co-chaperone and Dbf4 haploid double mutants were formed and incubated at RT, 30 ^{o}C (semi-permissive) and 37 ^{o}C. Double mutants showing genetic interactions are highlighted.

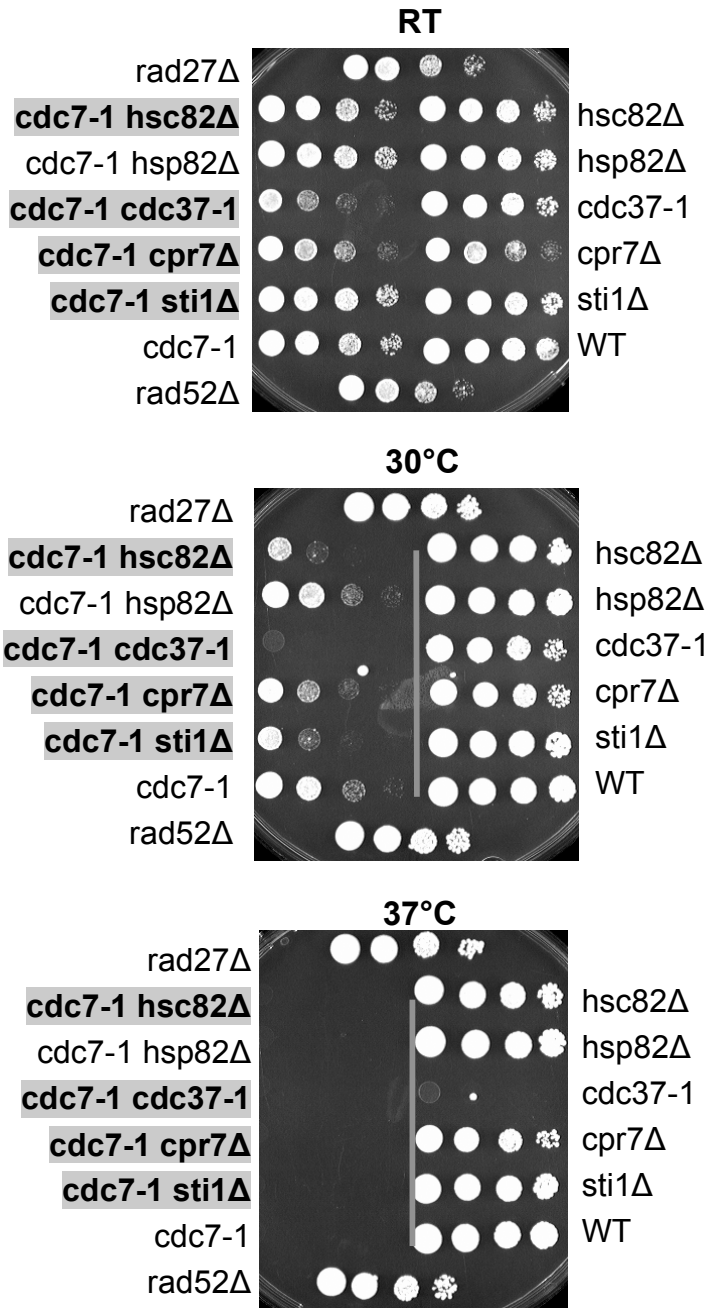


Figure 4.7. Single and double mutants between Hsp90 and co-chaperones and *cdc7-1*. All combinations of Hsp90 co-chaperone and Cdc7 haploid double mutants were formed and incubated at RT, 30 ^{o}C (semi-permissive) and 37 ^{o}C. Double mutants showing genetic interactions are highlighted.

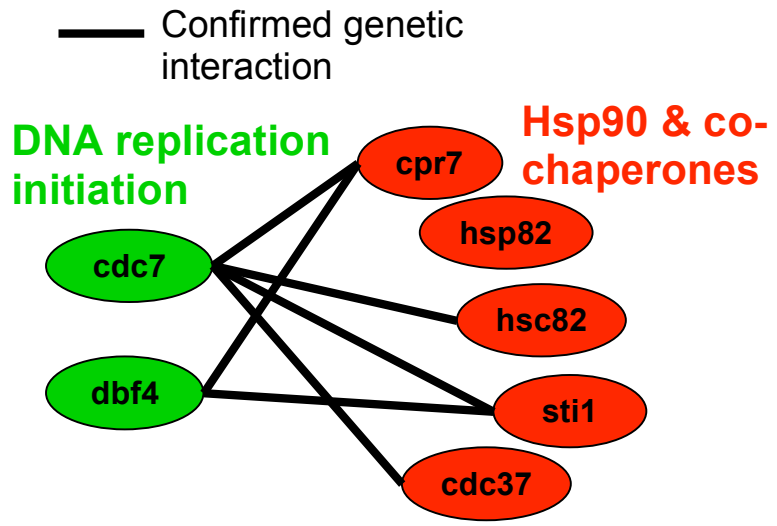


Figure 4.8. Experimentally confirmed genetic interactions between *cdc7/dbf4* and Hsp90 and co-chaperones. We tested all combinations of DNA replication *cdc7-1* and *dbf4-1* temperature sensitive mutants with Hsp90 and its co-chaperones for genetic interactions (see Figures 6 and 7). This figure summarizes all interactions confirmed among these double mutants.

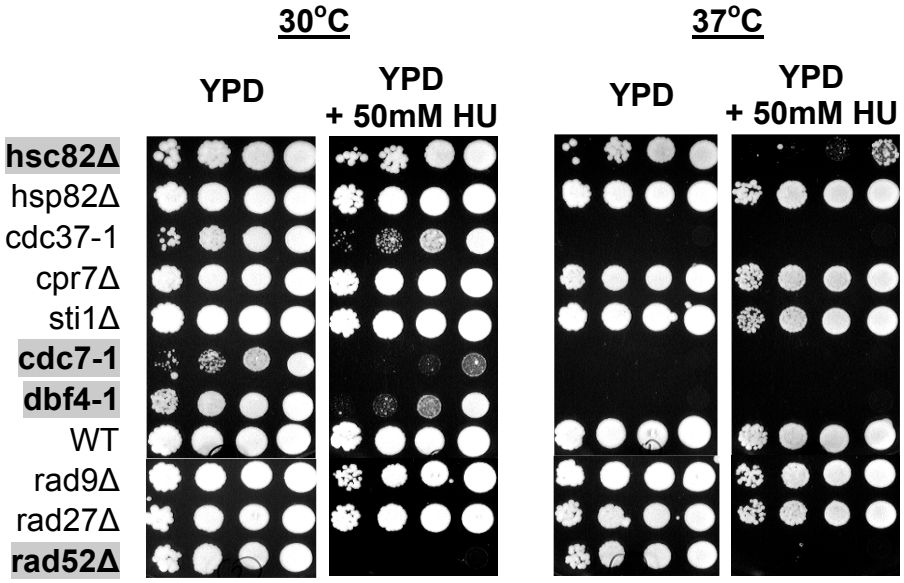


Figure 4.9. Hydroxyurea sensitivity of DNA replication and Hsp90 mutants. Single mutants related to DNA replication and Hsp90 were tested for sensitivity to hydroxyurea (HU), which inhibits DNA replication. Mutants showing hypersensitivity are highlighted. Both *cdc7-1* and *dbf4-1* mutants are sensitive to HU as well as the positive control, *rad52Δ*. Interestingly, the Hsp90 homolog also shows sensitivity to HU.

4.5.4 Experimental Methods

For confirmation of chromosome segregation mutants, haploid deletion mutant strains were obtained from YKO heterozygous diploid collection by sporulation (http://www-sequence.stanford.edu/group/yeast_deletion_project/spo.html), and confirmed by backcrosses. Haploid cells were grown to mid-log phase in YPD at 30°C. Cells were fixed in 4% formaldehyde, stained with DAPI and visualized under fluorescence microscope. Image acquisition was performed on an Zeiss AxioSkop equipped with a video CCD camera using SimplePCI software (Hamamatsu). For flow cytometry, mid-log phase haploid cells were fixed in ethanol and stained with sytox green.

For Hsp90 and Cdc7-Dbf4 experiments, the *cdc7-1, dbf4-1,* and the *cdc37-1* temperature sensitive mutants were obtained from Charlie Boone's lab (U. of Toronto) and backcrossed four times into the S288C background. Strains were mated and sporulated to obtain double mutant haploids, and confirmed by tetrad dissection. Single mutant and double mutant haploid cells were grown to saturation at RT overnight in liquid YPD media. 10x serial dilutions were spotted onto YPD plates and incubated at RT, 30°C and 37°C. For the hydroxyurea sensitivity analysis, 10x serial dilutions were spotted onto YPD plates +/- HU and incubated at RT, 30°C and 37°C.

4.6 Using the Predicted Functional Network for Understanding Links across Pathways

4.6.1 Cross-talk Analysis Method

To measure cross-talk between processes, we start with a single pathway as our query set, build the graph of interactions around this query using bioPIXIE, and analyze the resulting superset of proteins for statistical enrichment of *other* processes. More specifically, we first remove the original query set from the recovered set of proteins and obtain counts of proteins in the remaining set for every other possible interacting pathway. We then use a hypergeometric test to estimate the significance of the observed counts. For example, suppose we use a query pathway, Q, and with a graph of size X recover m proteins annotated to a different pathway,

R, of total size M. If there are N total known proteins in the organism of interest, the probability of observing a number this large or greater under the null assumption that the two pathways do not interact is:

$$\text{P-value} = 1 - \sum_{i=0}^{m-1} \frac{\binom{M}{i}\binom{N-M}{X-i}}{\binom{N}{X}}$$

We repeated this calculation for all pairwise combinations of pathways (see supplementary data file). We conservatively corrected for multiple hypothesis testing by Bonferroni correction and only report results with corrected p-values of $< 10^{-2}$.

4.6.2 Finding Functional Links between Processes

Our approach of combining data integration with a method for process-specific network discovery provides a convenient framework for addressing biological questions at a higher level. Thus, in addition to constructing specific and testable hypotheses about individual biological processes, we can use the system to discover novel interplay, or cross-talk, *among* biological networks. To investigate possible cross-talk among biological networks, we start with a single functional group as our query set, use bioPIXIE to predict additional network components, and analyze the resulting superset of proteins for statistical enrichment of *other* functional groups. By repeating this for each process of interest, we can construct a map of cross-talk that represents a variety of high-level biological relationships (see Materials and Methods for details of this analysis). We have applied this approach to map functional links among a set of 363 KEGG pathways, GO categories, and co-regulated transcription factor targets. By using this variety of classification systems, we can detect links across different biological relationships—from biological roles (GO process ontology) to cellular locations (GO component ontology) to metabolic pathways (KEGG). Upon mapping cross-talk among these groups, we clustered the results to reveal biologically significant groups of inter-related processes (Figure 4.6 and supplementary data files).

This analysis identifies a number of known or expected relationships between networks with related functions. For example, one would expect that the processes of actin cytoskeleton

organization, vesicle-mediated transport, and budding would be well connected with each other, and that proteins involved in these processes would share similar functional links to proteins localized to the sites of polarized growth or proteins that when mutated cause morphological defects. Indeed, these groups of genes are found in a tight cluster in our cross-talk analysis (Figure 4.10: top cluster).

In addition to such clusters that are expected based on current biological knowledge, we also identify novel relationships. For example, one such cluster contains four previously unrelated groups, namely genes that have Swi5 binding sites, genes with Ino2 binding sites, proteins with lyase activity, and genes that have Cbf1 binding sites. Swi5 activates genes expressed at the M/G1 boundary and during G1 phase of the cell cycle, and Ino2 regulates expression of phospholipid biosynthetic genes. Cbf1 is required for the function of centromeres and MET gene promoters, and recent work suggests a general role for Cbf1 in chromatin remodeling [29]. These four groups are found in the same cluster because they share significant links with ribosome biogenesis and assembly, nucleolus, RNA binding, and RNA metabolism. This suggests an explicit, functional link among the processes of cell cycle regulation, transcriptional regulation, inositol metabolism and protein synthesis.

While the cross-talk across all of these biological processes has not yet been well characterized, there is evidence in the literature that supports these predicted connections.

For instance, the expression pattern of CBF1, INO2, or SWI5 is well correlated with the expression of NOP7 (e.g. as cells undergo diauxic shift and during sporulation, CBF1 and NOP7 are co-expressed with a Pearson correlation of greater than 0.8 [8,14,11]). Du and Stillman found that Nop7/Yph1, a protein required for the biogenesis of 60S ribosomal subunits [22,2,36], associates with the Origin Recognition Complex (ORC), cell cycle-related proteins, and MCM proteins. As cells are depleted of Nop7p, they exhibit cell cycle arrest, and in wild-type cells, Nop7 levels vary in response to different carbon sources [15]. Taken together, these previous experimental results support our prediction linking metabolic pathways, the cell cycle, and ribosome assembly. It is important to note that while the characterization of Nop7 is consistent with this prediction, the individual experiments with Nop7 described above were not part of the

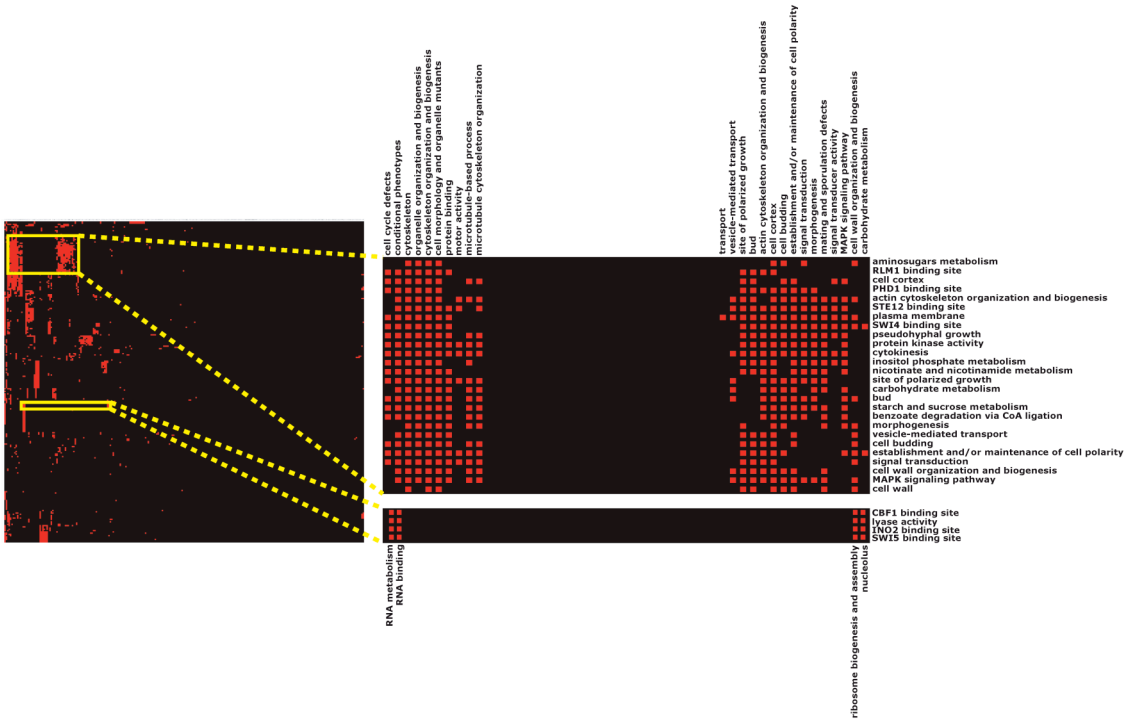


Figure 4.10. A map of cross-talk between 363 biological groups in *S. cerevisiae*. The combination of our Bayesian data integration system and our network discovery algorithm allows us to find biologically significant cross-talk among known biological groups. The interaction matrix was generated based on 363 KEGG pathways, GO categories, and co-regulated transcription factor targets. Rows of this matrix correspond to the query group and columns correspond to potential cross-talk partner processes; red boxes signify statistically significant links. The cross-talk matrix has been clustered [40] to reveal tightly-connected groups of interacting processes (i.e. clusters in this matrix correspond to sets of groups who interact with same partners). Highlighted clusters are discussed in the text. See supplemental Figure S10 [51] for a complete, labeled map.

input data to our system. Rather, our system was able to make the predicted links across these functional groups based on other heterogeneous, and mostly high throughout, data through bioPIXIE integration and network analysis. Thus, cross-talk analysis using bioPIXIE is effective in identifying novel interplay among pathways, biological processes, cellular locations, and regulatory modules.

4.7 Discussion and Future Directions

We have developed bioPIXIE, an analysis and visualization system for the discovery of biological process-specific networks. bioPIXIE's public interface allows researchers to use their knowledge to explore novel and previously known components of a variety of biological processes. The

system provides detailed information about experimental sources for each prediction, including links to original literature, and can be used to generate testable hypotheses. It is important to note that predictions made by bioPIXIE require further experimental validation; we hope that the public availability of our system and all results presented here will encourage such verification by yeast biology laboratories.

A key strength of our system is in addressing network-level behavior as opposed to focusing purely on pair-wise protein relationships. This is critical because many biologically significant questions involve the behavior of groups of proteins in networks or the interplay among networks with different functions. Furthermore, from a computational standpoint, the network-level approach to analysis and modeling of biological data is beneficial because subtle but coordinated group behavior can provide a more accurate picture of biological relationships than can be detected through pair-wise protein linkages. Although we focus on discovering networks, bioPIXIE can also be used for function prediction of individual proteins. Functions of uncharacterized proteins can be predicted either by analyzing uncharacterized components that are returned by the system given a known query set or by using an uncharacterized protein itself as the query, building the local interaction graph around it with our network-discovery algorithm, and analyzing the proteins in the final graph for statistical enrichment for particular functions. Another advantage of bioPIXIE is the probabilistic nature of the method that can easily adapt to new types of data. In the future, bioPIXIE will incorporate additional data sets from sources already modeled by the system as well as data from new approaches such as protein microarrays.

Another future direction for our method is to use process-specific neighborhoods generated by the system as a starting point for deciphering more precise details of biological relationships. Our notion of functional relationship is intentionally rather general so a wide variety of biological interactions can be detected. However, developing detailed models of how groups of functionally related proteins specifically relate with each other requires more precise definitions of relationships. We propose our method as a way to pinpoint groups of proteins acting together, after which other methods can be applied to investigate details of relationships between these

proteins. This narrowing process will undoubtedly improve downstream computational approaches.

Finally, our method may be applicable to higher eukaryotes. Additional challenges for such applications include handling multiple cell types, less comprehensive sets of functional genomics data, and incomplete genome annotation. Our method is general, and by extending the Bayesian network structure to organism-specific data sources and learning the corresponding integration weights from available annotation data, bioPIXIE can enable discovery and accurate modeling of previously uncharacterized process-specific networks in a diverse range of organisms. It is important to stress that the success of applying our method and other related approaches to higher eukaryotes depends on public availability of functional genomics data for these organisms and continued improvement of their annotation data, ideally through expert curation.

4.8 Conclusions

We have developed a novel probabilistic methodology for identification of biological process-specific networks based on diverse genomic data and have used this methodology to create a fully functional system for network analysis and visualization. bioPIXIE allows researchers to identify novel pathway components and to study specific interactions among them. Predictions made by our system are specific enough to be tested using common molecular biology techniques. Using this approach, we have accurately modeled multiple known processes in *S. cerevisiae*, characterized unknown components in these processes, and identified novel cross-talk relationships. We are making bioPIXIE publicly available through the web to ensure that analysis and interpretation of accurate network predictions we generate, as well as the underlying data, are conveniently accessible to biological researchers.

4.9 List of Supplemental Data Files

File name: bioPIXIE_bayesnet.dsl
File format: dsl, GeNIe recommended for viewing, available at http://www.sis.pitt.edu/~genie/downloads.html
Title: bioPIXIE Bayesian network for genomic data integration
Description:
This file contains the structure and final learned conditional probability tables used for integrating multiple heterogeneous sources of functional genomic data.

File name: bioPIXIE_evaluation_groups.txt
File format: txt, tab-delimited
Title: Evaluation pathways and protein complexes
Description:
This file contains a list of pathways and protein complexes that were used to evaluate the performance of bioPIXIE. The source of the group and the number of proteins in each is also included.

File name: comparison_AUCs.xls
File format: Microsoft Excel
Title: Results of comparison with existing methods
Description:
This file contains a comparison of the performance of bioPIXIE to existing methods for biological network recovery. The area under the precision-recall curve (AUC) is computed and plotted separately for each of the 31 evaluation pathways and complexes.

File name: bioPIXIE_data_sources.html
File format: HTML
Title: List of data sources
Description:
This file contains a list of references for all data incorporated into bioPIXIE.

File name: bioPIXIE_bayesnet_integration.zip
File format: zipped txt, tab-delimited
Title: bioPIXIE probabilistic network
Description:
This file contains the integrated, probabilistic functional network, a listing of pairwise probabilities between all genes.

File name: GO_gold_standard.zip
File format: zipped txt, tab-delimited
Title: bioPIXIE gold standard for learning
Description:
This file contains the pairwise gold standard used for learning the Bayesian network CPT's.

File name: bioPIXIE_pathwaycrosstalk.txt
File format: txt, tab-delimited
Title: Cross-talk between pathways as measured by bioPIXIE
Description:
This file contains a binary matrix of complexes, pathways, and processes where significant cross-talk between the pathways is indicated with a 1.

File name: bioPIXIE_crosstalk_clusters.txt
File format: txt, tab-delimited
Title: Clusters of pathway cross-talk
Description:
This file contains a list of pathway, process, and complexes clustered based on their cross-talk signature. It lists the complexes that are co-clustered and their common interacting partners.

File name: querysizedependence_AUCS.xls
File format: Microsoft Excel
Title: Results of query size sensitivity evaluation
Description:
This file contains the results of a query size sensitivity evaluation. The area under the precision-recall curve (AUC) is computed and plotted separately for each of the 31 evaluation pathways and complexes.

File name: querynoisedependence_AUCS.xls
File format: Microsoft Excel
Title: Results of query noise sensitivity evaluation
Description:
This file contains the results of a query noise sensitivity evaluation. The area under the precision-recall curve (AUC) is computed and plotted separately for each of the 31 evaluation pathways and complexes.

References

1. Abbas-Terki, T., O. Donze, et al. (2000). "The molecular chaperone Cdc37 is required for Ste11 function and pheromone-induced cell cycle arrest." FEBS Lett **467**(1): 111-6.

2. Adams, C. C., J. Jakovljevic, et al. (2002). "Saccharomyces cerevisiae nucleolar protein Nop7p is necessary for biogenesis of 60S ribosomal subunits." Rna **8**(2): 150-65.

3. Alfarano, C., C. E. Andrade, et al. (2005). "The Biomolecular Interaction Network Database and related tools 2005 update." Nucleic Acids Res **33 Database Issue**: D418-24.

4. Asthana, S., O. D. King, et al. (2004). "Predicting protein complex membership using probabilistic network reliability." Genome Res **14**(6): 1170-5.

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

6. Bader, J. S. (2003). "Greedily building protein networks with confidence." Bioinformatics **19**(15): 1869-74.

7. Bagatell, R. and L. Whitesell (2004). "Altered Hsp90 function in cancer: a unique therapeutic opportunity." Mol Cancer Ther **3**(8): 1021-30.

8. Balakrishnan, R., Christie, K. R., Costanzo, M. C., Dolinski, K., Dwight, S. S., Engel, S. R., Fisk, D. G., Hirschman, J. E., Hong, E. L., Nash, R., Oughtred, R., Skrzypek, M., Theesfeld, C. L., Binkley, G., Lane, C., Schroeder, M., Sethuraman, A., Dong, S., Weng, S., Miyasato, S., Andrada, R., Botstein, D., and Cherry, J. M. "Saccharomyces Genome Database." Retrieved 6/25/04, from ftp://ftp.yeastgenome.org/yeast/.

9. Beliakoff, J. and L. Whitesell (2004). "Hsp90: an emerging target for breast cancer therapy." Anticancer Drugs **15**(7): 651-62.

10. Breitkreutz, B. J., C. Stark, et al. (2003). "The GRID: the General Repository for Interaction Datasets." Genome Biol **4**(3): R23.

11. Chu, S., J. DeRisi, et al. (1998). "The transcriptional program of sporulation in budding yeast." Science **282**(5389): 699-705.

12. Dempster, A. P., N. M. Laird, et al. (1977). "Maximum Likelihood from Incomplete Data Via Em Algorithm." Journal of the Royal Statistical Society Series B-Methodological **39**(1): 1-38.

13. Deng, M., Z. Tu, et al. (2004). "Mapping Gene Ontology to proteins based on protein-protein interaction data." Bioinformatics **20**(6): 895-902.

14. DeRisi, J. L., V. R. Iyer, et al. (1997). "Exploring the metabolic and genetic control of gene expression on a genomic scale." Science **278**(5338): 680-6.

15. Du, Y. C. and B. Stillman (2002). "Yph1p, an ORC-interacting protein: potential links between cell proliferation control, DNA replication, and ribosome biogenesis." Cell **109**(7): 835-48.

16. Eddy, S. R. (2004). "What is Bayesian statistics?" Nat Biotechnol **22**(9): 1177-8.

17. Farrell, A. and D. O. Morgan (2000). "Cdc37 promotes the stability of protein kinases Cdc28 and Cak1." Mol Cell Biol **20**(3): 749-54.

18. Gagneur, J., R. Krause, et al. (2004). "Modular decomposition of protein-protein interaction networks." Genome Biol **5**(8): R57.

19. Gasch, A. P., M. Huang, et al. (2001). "Genomic expression responses to DNA-damaging agents and the regulatory role of the yeast ATR homolog Mec1p." Mol Biol Cell **12**(10): 2987-3003.

20. Gasch, A. P., P. T. Spellman, et al. (2000). "Genomic expression programs in the response of yeast cells to environmental changes." Mol Biol Cell **11**(12): 4241-57.

21. Gerber, M. R., A. Farrell, et al. (1995). "Cdc37 is required for association of the protein kinase Cdc28 with G1 and mitotic cyclins." Proc Natl Acad Sci U S A **92**(10): 4651-5.

22. Harnpicharnchai, P., J. Jakovljevic, et al. (2001). "Composition and functional characterization of yeast 66S ribosome assembly intermediates." Mol Cell **8**(3): 505-15.

23. Huh, W. K., J. V. Falvo, et al. (2003). "Global analysis of protein localization in budding yeast." Nature **425**(6959): 686-91.

24. Hunter, T. and R. Y. C. Poon (1997). "Cdc37: a protein kinase chaperone?" Trends in Cell Biology **7**(4): 157-161.

25. Jaimovich, A., G. Elidan, et al. (2005). "Towards an integrated protein-protein interaction network." Research in Computational Molecular Biology, Proceedings **3500**: 14-30.

26. Jansen, R., H. Yu, et al. (2003). "A Bayesian networks approach for predicting protein-protein interactions from genomic data." Science **302**(5644): 449-53.

27. Johnston, L. H., H. Masai, et al. (1999). "First the CDKs, now the DDKs." Trends Cell Biol **9**(7): 249-52.

28. Karaoz, U., T. M. Murali, et al. (2004). "Whole-genome annotation by using evidence integration in functional-linkage networks." Proc Natl Acad Sci U S A **101**(9): 2888-93.

29. Kent, N. A., S. M. Eibert, et al. (2004). "Cbf1p is required for chromatin remodeling at promoter-proximal CACGTG motifs in yeast." J Biol Chem **279**(26): 27116-23.

30. Lee, I., S. V. Date, et al. (2004). "A probabilistic functional network of yeast genes." Science **306**(5701): 1555-8.

31. Lee, P., A. Shabbir, et al. (2004). "Sti1 and Cdc37 can stabilize Hsp90 in chaperone complexes with a protein kinase." Mol Biol Cell **15**(4): 1785-92.

32. Letovsky, S. and S. Kasif (2003). "Predicting protein function from protein/protein interaction data: a probabilistic approach." Bioinformatics **19 Suppl 1**: i197-204.

33. Miles, J. and T. Formosa (1992). "Evidence that POB1, a Saccharomyces cerevisiae protein that binds to DNA polymerase alpha, acts in DNA metabolism in vivo." Mol Cell Biol **12**(12): 5724-35.

34. Mort-Bontemps-Soret, M., C. Facca, et al. (2002). "Physical interaction of Cdc28 with Cdc37 in Saccharomyces cerevisiae." Mol Genet Genomics **267**(4): 447-58.

35. Myers, C. L., D. Robson, et al. (2005). "Discovery of biological networks from diverse functional genomic data." Genome Biol **6**(13): R114.

36. Oeffinger, M., A. Leung, et al. (2002). "Yeast Pescadillo is required for multiple activities during 60S ribosomal subunit synthesis." Rna **8**(5): 626-36.

37. Ogawa, N., J. DeRisi, et al. (2000). "New components of a system for phosphate accumulation and polyphosphate metabolism in Saccharomyces cerevisiae revealed by genomic expression analysis." Mol Biol Cell **11**(12): 4309-21.

38. Pereira-Leal, J. B., A. J. Enright, et al. (2004). "Detection of functional modules from protein interaction networks." Proteins **54**(1): 49-57.

39. Prakash, S. and L. Prakash (2000). "Nucleotide excision repair in yeast." Mutat Res **451**(1-2): 13-24.

40. Saeed, A. I., V. Sharov, et al. (2003). "TM4: a free, open-source system for microarray data management and analysis." Biotechniques **34**(2): 374-8.

41. Schauber, C., L. Chen, et al. (1998). "Rad23 links DNA repair to the ubiquitin/proteasome pathway." Nature **391**(6668): 715-8.

42. Schutz, A. R., T. H. Giddings, Jr., et al. (1997). "The yeast CDC37 gene interacts with MPS1 and is required for proper execution of spindle pole body duplication." J Cell Biol **136**(5): 969-82.

43. Shakoury-Elizeh, M., J. Tiedeman, et al. (2004). "Transcriptional remodeling in response to iron deprivation in Saccharomyces cerevisiae." Mol Biol Cell **15**(3): 1233-43.

44. Spellman, P. T., G. Sherlock, et al. (1998). "Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization." Mol Biol Cell **9**(12): 3273-97.

45. Spirin, V. and L. A. Mirny (2003). "Protein complexes and functional modules in molecular networks." Proc Natl Acad Sci U S A **100**(21): 12123-8.

46. Sudarsanam, P., V. R. Iyer, et al. (2000). "Whole-genome expression analysis of snf/swi mutants of Saccharomyces cerevisiae." Proc Natl Acad Sci U S A **97**(7): 3364-9.

47. Tong, A. H., G. Lesage, et al. (2004). "Global mapping of the yeast genetic interaction network." Science **303**(5659): 808-13.

48. Troyanskaya, O. G., K. Dolinski, et al. (2003). "A Bayesian framework for combining heterogeneous data sources for gene function prediction (in Saccharomyces cerevisiae)." Proc Natl Acad Sci U S A **100**(14): 8348-53.

49. van Laar, T., A. J. van der Eb, et al. (2002). "A role for Rad23 proteins in 26S proteasome-dependent protein degradation?" Mutat Res **499**(1): 53-61.

50. von Mering, C., M. Huynen, et al. (2003). "STRING: a database of predicted functional associations between proteins." Nucleic Acids Res **31**(1): 258-61.

51. Website. "bioPIXIE Online Supplement." from http://pixie.princeton.edu/supplement.

52. Website. "bioPIXIE Public Home Page." from http://pixie.princeton.edu.

53. Website. "Decision Systems Laboratory Home Page." from http://www.sis.pitt.edu/~genie/.

54. Yoshimoto, H., K. Saltsman, et al. (2002). "Genome-wide analysis of gene expression regulated by the calcineurin/Crz1p signaling pathway in Saccharomyces cerevisiae." J Biol Chem **277**(34): 31079-88.

55. Zhu, G., P. T. Spellman, et al. (2000). "Two yeast forkhead genes regulate the cell cycle and pseudohyphal growth." Nature **406**(6791): 90-4.

56. Zhu, J. and M. Q. Zhang (1999). "SCPD: a promoter database of the yeast Saccharomyces cerevisiae." Bioinformatics **15**(7-8): 607-11.

Chapter 5

Gold Standards and Evaluation Methods for Functional Genomic Data

5.1 Chapter Overview

In Chapter 4, we discussed a general system for integration of diverse data and prediction of biological networks. One critical aspect of such a system is that we require gold standards of established functional information for both evaluation and training. In general, accurate evaluation of the quality of genomic or proteomic data and computational methods is vital to our ability to use them for formulating novel biological hypotheses and directing further experiments. There is currently no standard approach to evaluation in functional genomics. Our analysis of existing approaches shows that they are inconsistent and contain substantial functional biases that render the resulting evaluations misleading both quantitatively and qualitatively. These problems make it essentially impossible to compare computational methods or large-scale experimental datasets and also result in conclusions that generalize poorly in most biological applications.

In this chapter, we reveal issues with current evaluation methods and suggest new approaches to evaluation that facilitate accurate and representative characterization of genomic methods and data. Specifically, we describe a functional genomics gold standard based on curation by expert biologists and demonstrate its use as an effective means of evaluation of genomic approaches. Our evaluation framework and gold standard are freely available to the community through our website. Proper methods for evaluating genomic data and computational approaches will determine how much we, as a community, are able to learn from the wealth of available data. In this chapter, we describe our insight into this problem and propose several guidelines for reasonable gold standards for genomic data analysis and integration.

The work presented in this chapter was published in [26] and includes contributions from Daniel Barrett, Matthew Hibbs, Curtis Huttenhower, and Olga Troyanskaya. Daniel was instrumental in implementing the web-based evaluation system resulting from this study, Matt and Curtis provided ideas for both analysis and interface design, and Olga supervised the project. Also, Matt Brauer, Kara Dolinski, Maitreya Dunham, Rose Oughtred, and Charlotte Paquin contributed to the Gene Ontology-based evaluation standard.

5.2 Background: Genomic Data Evaluation

Recent advances in experimental methods have enabled the development of functional genomics, a genome-wide approach to understanding the inner workings of a cell. While such large-scale approaches will undoubtedly be instrumental in extending our knowledge of molecular and cellular biology, they produce enormous amounts of heterogeneous data of varying relevance and reliability. A key challenge in interpreting these data is separating accurate, functionally relevant information from noise.

Here we focus on using noisy genomic datasets to associate uncharacterized genes or proteins with biological processes. Recent literature on protein function prediction focuses on integrating multiple sources of evidence (e.g. physical interactions, genetic interaction, gene expression data) to assign proteins to processes [28,9,20,4] or to predict functional associations or interactions between related proteins [18,33,21,23,6,37]. Individual high-throughput datasets are typically noisy, but effective integration can yield precise predictions without sacrificing valuable information in the data. All of these methods require a gold standard, which is a trusted representation of the functional information one might hope to discover. Such a standard, coupled with an effective means of evaluation, can be used to assess the performance of a method and serves as a basis for comparison with existing approaches. Beyond methods for predicting protein function or interactions, evaluation against gold standards can be used to directly measure the quality of a single genomic dataset, a necessary step in developing and validating new experimental technology.

We have undertaken a study of proposed standards and approaches to evaluation of functional genomic data and highlight a number of important issues. We find that current approaches are inconsistent, making reported results incomparable, and often biased in such a way that the resulting evaluation cannot be trusted even in a qualitative sense. One specific problem we identify is substantial functional biases in typical gold standard datasets. We demonstrate this problem by evaluating several functional genomic datasets using the Kyoto Encyclopedia of Genes and Genomes (KEGG) [19] as a gold standard (Figure 5.1), as is commonly employed in the literature (e.g. [21,39]). A naïve evaluation in this manner identifies co-expression data as by far the most sensitive and specific genome-scale functional genomic data type (Figure 5.1a). However, this apparent superior performance is due to characteristics of a single pathway; when the ribosome (1 out of 99 total KEGG pathways) is removed from the gold standard, co-expression becomes one of the least informative datasets (Figure 5.1b). In addition to such substantial functional biases, we find that commonly used gold standards are highly inconsistent even for comparative evaluations and that most current evaluation methodologies yield misleading estimates of accuracy.

We have identified and describe these problems with current evaluation standards with the hope of instigating a community dialog on proper approaches to comparing genomic data and methods. As noted above, there are two typical approaches to using genomic data for analyzing protein function: methods that directly associate proteins with particular processes or functional classes, and methods that focus on predicting functional associations or interactions between pairs of proteins. We focus our attention toward standards for the latter, evaluating pairwise associations between genes produced by either experimental or computational techniques. Many of the problems we describe, however, apply to both approaches, and we suggest an alternative standard for evaluation that is appropriate in both settings. We provide both a trusted set of functional associations between proteins as well as a specific set of biological processes that maps proteins to well-defined functional classes. Both standards are based on curation by a panel of biological experts. Furthermore, we propose several guidelines for using these standards to perform accurate evaluation of methods and data. The resulting evaluation

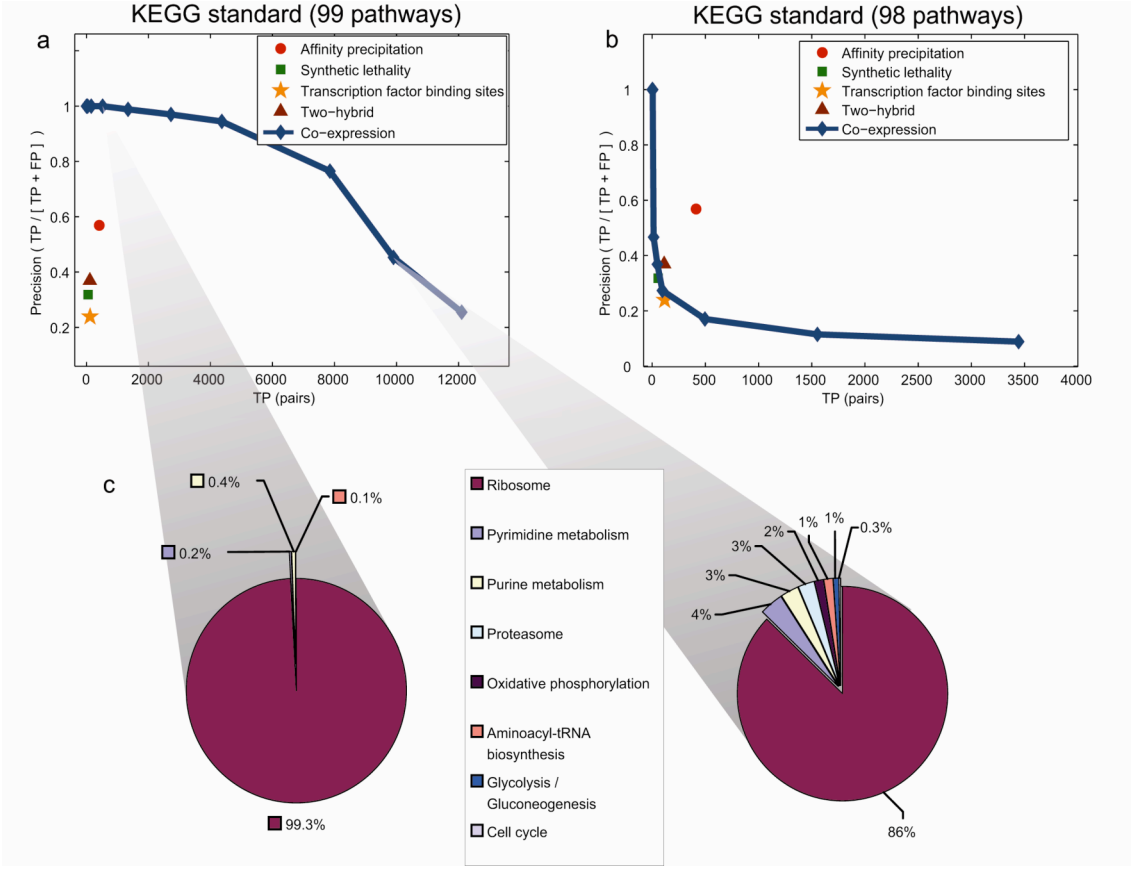


Figure 5.1. Inconsistencies in evaluation due to process-specific variation in performance. (a and b) Comparative functional evaluation of several high-throughput datasets based on a KEGG-derived gold standard. The evaluation pictured in (b) is identical to that in (a) except that one of ninety-nine KEGG pathways was excluded from the analysis ("Ribosome," sce03010). Gold standard positives were obtained by considering all protein pairs sharing a KEGG pathway annotation as functional pairs, while gold standard negatives were taken to be pairs of proteins occurring in at least one KEGG pathway but with no co-annotation. Performance is measured as the trade-off between precision (the proportion of true positives to total positive predictions) and true positive pairs. For the evaluation in (b), both precision and sensitivity drop dramatically for co-expression data. (c) Composition of correctly predicted positive protein-protein relationships at two different choices of precision-recall. Of the 0.1% most co-expressed pairs, 99.3% of the true positive pairs (842 of 848) are due to co-annotation to the ribosome pathway (left pie chart). This bias is less pronounced at lower precision but still present. Of the 1% most co-expressed pairs, 86% of the true positive pairs (8500 of 9900) are due to co-annotation to the ribosome pathway (right pie chart).

framework can be used to directly measure and compare the functionally relevant information present in raw high-throughput datasets as well as to evaluate or train computational genomics methods.

Our gold standard and evaluation methodology have been implemented in a web-based system [36] to facilitate community use for comparison among published datasets or methods. We demonstrate the use of our approach on genomic data from *Saccharomyces cerevisiae*.

Accurate evaluation methods are particularly critical for this model organism, because yeast is widely used as a platform for the development of both high-throughput experimental techniques and computational methods. However, the weaknesses we identify in existing evaluation methodologies as well as the solution we propose are applicable to data from other model organisms and humans.

5.3 Challenges in Effective Functional Evaluation

We first discuss commonly used gold standards and several fundamental issues with current approaches to evaluation of functional genomic data and methods. To address these problems, we propose a new gold standard based on expert curation and recommend appropriate uses of the standard that ensure accurate evaluation. Finally, we describe a web-based implementation of our evaluation framework, which is available for public use by computational and experimental biologists.

5.3.1 Existing Gold Standards

A number of different gold standards for evaluating yeast functional genomic data or methods have been proposed in the literature. Each standard generally consists of sets of gene or protein pairs grouped as either “positive” or “negative” examples. This is due in large part to the fact that some high throughput data takes the form of associations between genes or gene products (e.g. physical or genetic interactions). Furthermore, a pairwise approach to analysis is a natural way to view biological systems, which are composed of networks, or groups of interactions between gene products. Although this is a commonly adopted approach, others have trained classifiers for specific functional classes where individual proteins or genes are directly associated with functional classes or processes [28,4]. While we focus on data and methods for pairwise associations between proteins here, many of the issues described are equally problematic for such non-pairwise approaches, and we propose an alternative gold standard appropriate for both settings (see details in “Defining a new gold standard” in Methods).

Most functional genomics evaluations derive gold standard positives from functional classification schemes that capture associations of genes or proteins with specific biological processes as reported in the literature [34,13,35,31,21,22,39,37]. Such classifications are available from multiple sources including the Gene Ontology (GO) [2] (and associated annotation repositories such as the *Saccharomyces* Genome Database)[3], KEGG [19], the Munich Information Center for Protein Sequences (MIPS) [25], and the Yeast Protein Database (YPD) [10]. A common source of gold standard negatives is cellular localization data [18,17,21,27]. Most of these methods utilize a localization study in which 75% of the yeast proteome was GFP-tagged and classified into 22 different cellular compartments [15] and they assume that two proteins localizing to distinct compartments do not interact. Random pairs of proteins sampled from the proteome provide another common gold-standard negative, relying on the assumption that the expected number of functionally related or interacting pairs is much less than the total number of possible pairwise protein-protein combinations [11,6,29].

5.3.2 Inconsistencies among and within Different Standards

Perhaps the most apparent issue with functional genomic evaluation arises from the diversity of possible standards and lack of agreement among them. It has been noted that gold standard positive pairs derived from KEGG, MIPS, and GO biological process ontology show little overlap [7]. We find even less agreement among gold standards for physical interactions predictions, which are usually based on small interaction datasets obtained from protein-protein interaction databases such as the Database of Interacting Proteins (DIP) [38], the General Repository for Interaction Datasets (GRID) [8], or the Biomolecular Interaction Network Database (BIND) [1]. However, the more alarming problem is that even the *relative* performance of methods or datasets evaluated against these standards is not consistent. For example, using both the biological process GO and the KEGG pathways gold standard to evaluate the relative performance of commonly used data sets produces strikingly different results (Figure 5.2). This difference is likely due to the nature of the biological relationships each standard is trying to capture or simply variation in which specific proteins are present in the classification scheme or

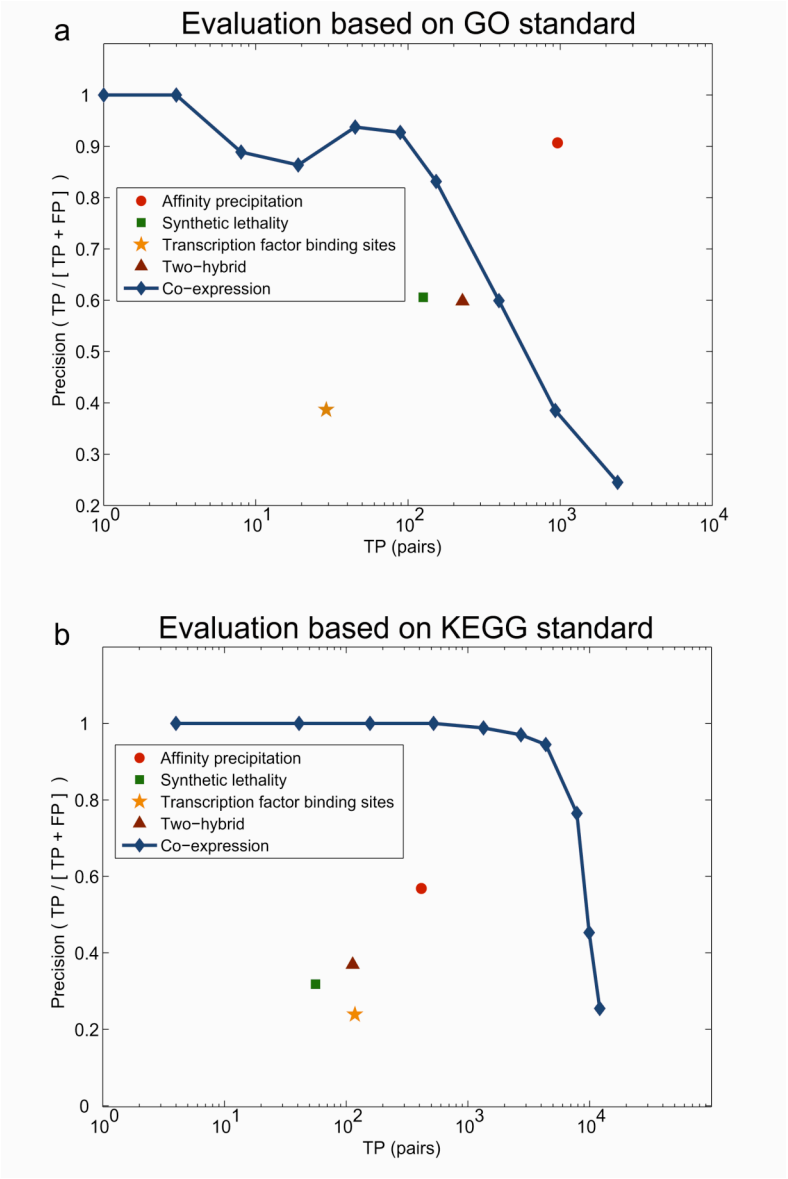


Figure 5.2. Comparison of functional genomic data evaluation on GO and KEGG gold standards. (a) Comparative functional evaluation of several high-throughput evidence types based on a typical Gene Ontology (GO) gold standard. Positive pairs were obtained by finding all protein pairs with co-annotations to terms at depth 8 or lower in the biological process ontology. Negative pairs were generated from protein pairs whose most specific co-annotation occurred in terms with more than 1000 total annotations. (b) Evaluation of the same data against a KEGG-based gold standard. Gold standard positives were obtained by considering all protein pairs sharing a KEGG pathway annotation as functional pairs, while gold standard negatives were taken to be pairs of proteins occurring in at least one KEGG pathway but with no co-annotation. There are several serious inconsistencies between the two evaluations. In addition to vastly different estimates of the reliability of co-expression data, other evidence types change relative positions. For instance, transcription factor binding site predictions appear competitive with both two-hybrid and synthetic lethality in the KEGG evaluation, but are substantially out-performed in the GO evaluation. These inconsistencies between the two gold standards demonstrate the need for a common, representative evaluation framework.

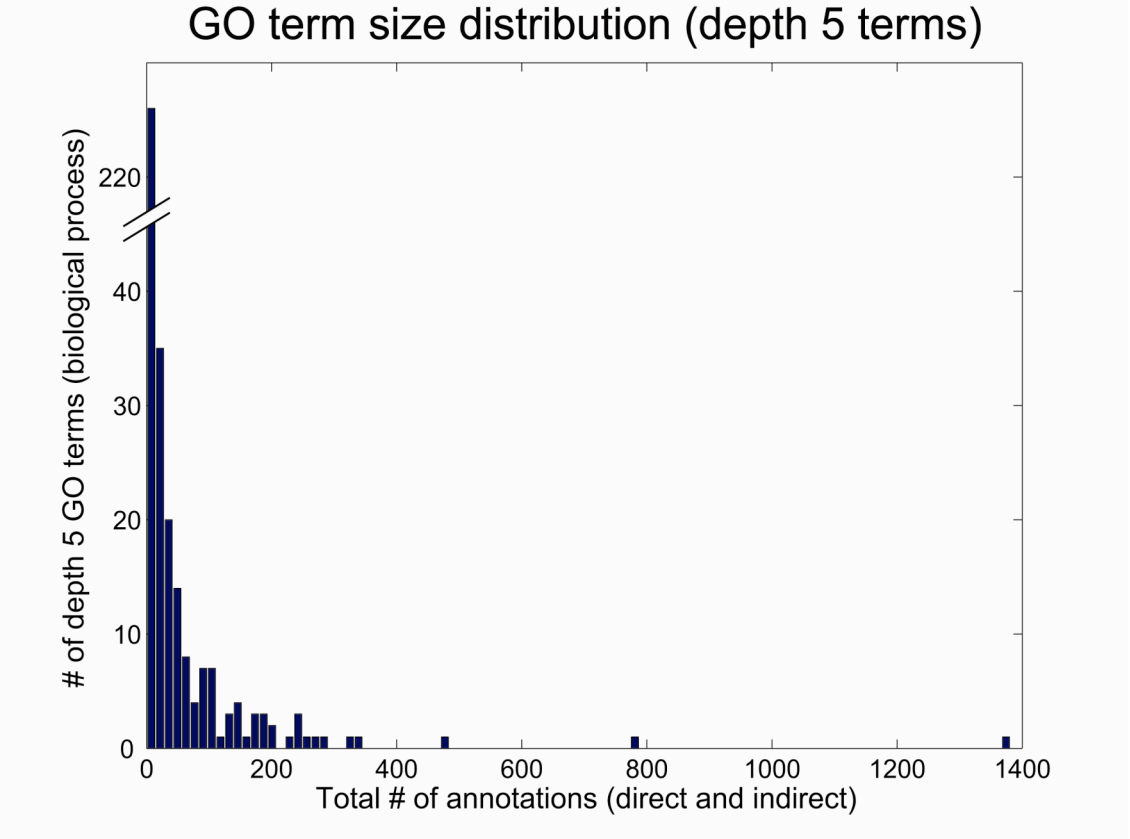


Figure 5.3. Size distribution of depth 5 biological process GO terms (*S. cerevisiae*). Depth and size are commonly used metrics for assessing the biological specificity of GO terms, a necessary step in creating a functional gold standard from the ontology. Here, the number of direct and indirect annotations was counted for each depth 5 GO term and counts were binned to obtain a histogram of sizes for depth 5 GO terms. This reveals a wide range of sizes for terms at the same depth (from 0 annotations to 1381 annotations), suggesting size and depth are not capturing the same notion of specificity, and that likely neither is an appropriate measure for true biological specificity. A sampling of the largest and smallest depth 5 GO terms is shown in Table 5.1.

interaction dataset. Although each standard is correctly evaluating some aspect of the data, without a common, representative evaluation framework, the community cannot assess the relative performance of novel methods or high-throughput techniques.

In addition to substantial inconsistencies among existing gold standards, variation in biological specificity within each standard has also impaired previous evaluation methods. Standards based on biological ontologies (e.g. GO or the MIPS Functional Catalogue) classify proteins at a broad range of resolutions (e.g. metabolism vs. carbohydrate metabolism). Although these ontologies can provide a powerful framework for defining a gold standard, there are a few caveats. A typical approach for using GO has been to pick a particular depth in the hierarchy below which term co-annotations imply gold standard positives. However, terms at the

same level can vary dramatically in biological specificity [24] (Fig 5.3 and Table 5.1). For example, at a depth of 5 in the biological process GO, the term "regulation of sister chromatid cohesion" (GO:0007063) with a single indirect gene product annotation appears alongside a much more general term "cellular protein metabolism" (GO:0044267), which has 1381 annotations. Widely varying degrees of specificity in a gold standard not only complicate evaluation methods but can also appear as inconsistencies in the data when training machine learning algorithms, which can result in poor performance.

Table 5.1. Example depth five biological process GO terms. GO term depth is a commonly used metric for biological specificity in the Gene Ontology. 5 of the smallest depth 5 GO terms and 4 of the largest depth 5 GO terms are listed above. The processes described range from very specific behaviors (e.g. contractile ring contraction) to less informative groupings (e.g. cellular protein metabolism), suggesting depth is a poor measure of specificity. The size distribution for all depth 5 GO terms is plotted in Fig. 3.

GO term	Term depth	Total annotations
lipoic acid metabolism (GO:0000273)	5	1
cytokinesis, contractile ring contraction	5	1
DNA ligation (GO:0006266)	5	1
lysosomal transport (GO:0007041)	5	1
regulation of sister chromatid cohesion	5	1
cytoskeleton organization and biogenesis	5	285
transcription (GO:0006350)	5	474
protein biosynthesis (GO:0006412)	5	775
cellular protein metabolism (GO:0044267)	5	1381

5.3.3 Functional Biases in Prediction Performance

The majority of current evaluation approaches are performed without regard to which biological processes are represented in the set of true positives (correctly predicted examples), and thus they are often unknowingly skewed toward particular processes. We illustrate this bias with an example using the KEGG pathways gold standard to evaluate genomic data (Figure 5.1). In this evaluation, the estimated reliability of microarray co-expression drops dramatically when a single pathway ("Ribosome" or sce3010) is excluded from the analysis. The substantial drop in precision suggests that a large fraction of the true positives predicted by co-expression are exclusively ribosome relationships. In fact, of the positive examples in the 1% most co-expressed

pairs, 86% (~8500 of 9900) are due to co-annotation to the ribosome pathway. This bias becomes even more pronounced at higher co-expression level cutoffs: of the 0.1% most co-expressed positive pairs, 99% (842 of 848) are from the ribosome pathway. We find a similar bias in evaluations using the GO and MIPS gold standards.

Thus, the traditional approach of using a general ROC curve (or related measure) without regard to which processes are represented can be misleading (see Methods for a discussion of ROC curves). This is particularly true when the data or computational predictions have process-dependent reliability as is often the case with genomic or proteomic data. The problem is magnified when the gold standard examples themselves are heavily skewed towards specific functional categories. While the general precision-recall characteristics such as those portrayed in Figure 5.1 are technically correct, they generalize poorly to non-ribosomal protein relationships. Thus, such an evaluation would be misleading for a scientist hoping to use these data to generate new hypotheses about proteins unrelated to the ribosome. We address this problem in our process-specific evaluation framework.

5.3.4 Gold Standard Negatives

Another shortcoming of current standards for gene/protein function prediction is the nature of the gold standard negative examples. In yeast, one proposed source of gold standard negatives is based on protein localization data [15,17] because pairs of proteins localizing to different cellular compartments are highly enriched for non-interacting proteins. However, localization data is likely not *representative* of "typical" unrelated protein pairs. For instance, Ben-Hur and Noble found the performance of SVM classifiers trained with localization negatives artificially inflated because this negative set is composed entirely of high-confidence pairs [6,5]. Using such a non-representative "easy" set of negatives will overestimate prediction accuracy, and the resulting classifier will generalize poorly to real biological problems.

Thus, although protein localization data is a strong negative indicator of functional relationships or interactions, we caution against its use as a general negative gold standard. This is particularly problematic for higher-level questions such as function prediction, because proteins

co-involved in some biological processes span cellular compartments. Perhaps a safer role for localization data is as the input to computational approaches. We suggest an alternative negative standard based on the biological process Gene Ontology that can provide representative negative examples (see "Suggestions for representative functional evaluation of data and methods").

5.3.5 Relative Size of Gold Standard Positive/Negative Sets

A final issue common among many evaluation standards in the literature is the relative size of the positive and negative example sets. The expected number of proteins involved in any particular biological process is a small percentage of the proteome, which should be reflected in evaluation standards. This imbalance is particularly problematic in methods based on pairwise associations between proteins, where the expected number of protein pairs sharing functional relationships is an even smaller fraction of all possible protein combinations. For instance, of the 18 million possible protein pairs in yeast, it is expected that less than 1 million are functionally related. This large difference makes the typical reporting of sensitivity and specificity misleading. For instance, a recently published method for predicting protein-protein interactions from several genomic features showed seemingly impressive 90% sensitivity and 63% specificity in evaluations [27], but would make correct predictions only 1 out of every 9 times when applied on a whole-genome scale, rendering the method impractical in many experimental contexts (details in additional file 4: Supplementary discussion).

Given this imbalance, an appropriate measure of functional relevance of genomic data or predictions is the precision or positive predictive value (PPV) $\left(\frac{TP}{TP+FP}\right)$ [17]. This measure rewards methods that generate firm positive predictions, without regard to the accuracy of negative predictions, which are less helpful in guiding laboratory experiments. Direct application of precision may be misleading, though, because this measure is only correct under the assumption that the ratio of positive to negative examples in the gold standard matches that in the application domain. If the ratio of positive to negatives in the gold standard is much larger

than in whole-genome data, as is often the case in published evaluations, then the number of false positive predictions will be small and will artificially inflate the precision statistic. For instance, the 90%-63% sensitivity-specificity example above used an approximately equal number of positive and negative examples (1500 and 2000 respectively), leading to 65% precision. However, application of this method on a whole-genome scale, where the ratio of positive to negative examples is roughly 20 times smaller, would lead to an expected precision of just 11% (details in additional file 4: Supplementary discussion).

To avoid such misleading evaluations, the balance of positives and negatives in the gold standard should match that of the application domain as closely as possible. Precision, or PPV, then becomes a direct, representative measure of how well one could expect a dataset or method to perform on whole-genome tasks. Of course, precision alone does not convey all of the important information, only the *quality* of the predictions made by a dataset or method. It must be reported in tandem with some measure of the *quantity* of true predictions made. A standard measure for this is the recall, or sensitivity $\left(\frac{TP}{TP + FN}\right)$, which is what is used in our evaluation framework (for more details, see Methods).

5.4 Suggestions for Representative Functional Evaluation of Data and Methods

In light of these problems with current gold standards and approaches to evaluation, we have compiled a new functional genomics gold standard and suggest several strategies for accurate comparative evaluation of genomic datasets and methods.

5.4.1 Defining a New Gold Standard

As discussed previously, a major issue with the current state of the community is inconsistency among the variety of standards used. Evaluations based on different standards (e.g. derived from KEGG versus GO) are often not comparable, even in a qualitative sense. Deriving a standard from these hierarchies is further complicated due to varying levels of biological

specificity of curated biological knowledge. Furthermore, each of the sources of curated information has inherent functional biases that can lead to incorrect estimates of accuracy.

To develop a unified standard for general application in functional genomics, several key criteria must be met. The standard must be cross-organismal to ensure relevance to a broad audience. Secondly, the standard should cover a wide variety of biological functions or processes to facilitate comprehensive evaluations. Finally, the standard should adapt quickly as biological knowledge expands. Although there are several sources of annotation that satisfy these criteria to varying extents (eg. KEGG, MIPS, and GO), GO is arguably the best option to serve as a foundation for the standard, as it is well-curated and was designed for complete coverage.

Although GO can serve as a good basis for a functional gold standard, effective mapping from organism-specific annotations to a set of positive and negative examples is critical. In particular, we have addressed the problem of varying levels of resolution in the GO hierarchy by selecting the gold standard set of terms through curation by six expert biologists. Through this formal curation process, the experts selected terms that are specific enough to be confirmed or refuted through laboratory experiments while also general enough to reasonably expect high-throughput assays to provide relevant information (see details in Methods and additional file 4: Supplementary discussion). The result of this process is a set of specific functional classes (GO terms) which can be used to generate an accurate set of positively related gene pairs or to directly evaluate or train computational approaches that explicitly associate proteins with particular biological processes. This standard created using expert knowledge is quite different from GO standards commonly used in the literature (Figure 5.4). It can serve as a *single*, common standard that addresses the specific concerns of functional genomics.

This curation can also be used to obtain a negative standard which addresses some issues with currently used methods. Specifically, our standard includes a set of negatives more broadly representative than sources such as localization while excluding likely positive examples (a shortcoming of approaches that use random sampling). Further, the standard approximates

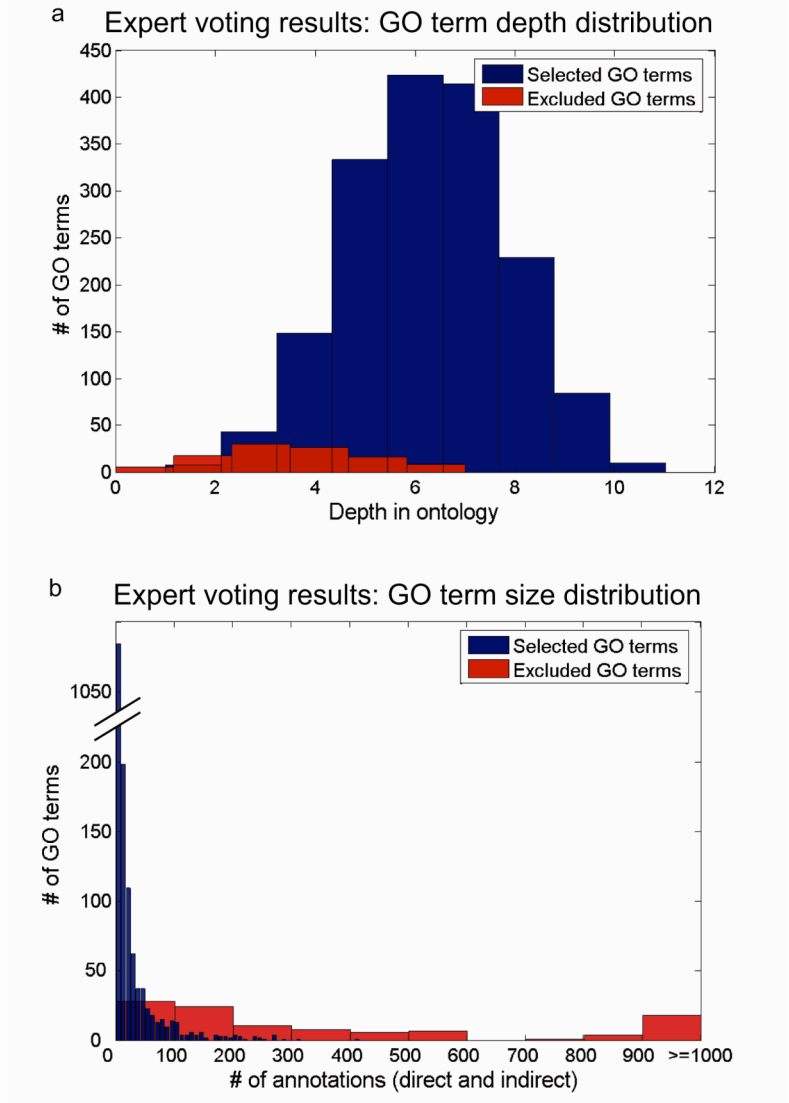


Figure 5.4. Depth and size properties of GO terms selected or excluded from the evaluation gold standard based on expert curation. The functional gold standard based on voting from an expert panel cannot be approximated by either a size or a depth measure of specificity. (a) Distribution of GO term depths for expert-selected terms (4-6 votes) and expert-excluded terms (1-3 votes). The selected set of terms cannot be separated from the "too general" excluded terms on the basis of depth. For instance, 53 of the 107 general GO terms appear at depth 4 or lower and 51 of 1692 specific GO terms appear at depth 3 or higher. (b) Distribution of GO term sizes (direct and indirect annotations) for the selected and excluded terms based on the expert voting analysis. As with term depth, size cannot effectively distinguish specific terms from those deemed too general by experts. For example, 28 of 107 GO terms deemed too general for inclusion in the standard have fewer than 100 annotations.

the correct relative balance of positive and negative sets enabling biologically relevant evaluations (see Methods for details).

5.4.2 Evaluating Genomic Methods and Data

In addition to defining a unifying standard, it is critical to use the standard in a manner that accurately reflects the biological reliability of datasets or methods. To expressly address the process-specific variability in accuracy, we developed an evaluation framework that facilitates identification of functional biases in current general evaluations. To accomplish this, we propose that two complementary modes of analysis accompany any evaluation of functional genomic data: (1) a genome-wide evaluation that estimates general reliability but also reports the functional composition of the results and (2) a process-specific evaluation in which the data or method is independently evaluated against a set of expert-selected processes.

Genome-wide evaluation

To provide a genome-wide analysis that also features information on the constituent biological processes, we have developed a hybrid evaluation framework that combines traditional measures of the precision-recall tradeoff with an analysis of the biological processes accurately represented in the data. In addition to the usual estimation of precision-recall characteristics, we compute the distribution of biological processes represented in the set of correctly classified positives (true positives) at every point along the precision-recall tradeoff curve (Figure 5.5). This distribution allows one to identify and measure any biases in the set of positive results toward a specific biological process and interpret evaluation results accordingly. Furthermore, all of this information is summarized and presented in a dynamic and interactive visualization framework that facilitates quick but complete understanding of the underlying biological information.

Figure 5.5 illustrates an example of a genome-wide evaluation of several high-throughput datasets using our framework. At first glance, a general evaluation indicates that the Gasch *et al.* microarray data is the second most reliable source for functional data (Figure 5.5a). However, an analysis of the processes represented in the set of correctly classified pairs reveals that approximately 60% of the correct predictions by the co-expression data are related to the process of ribosome formation (Figure 5.5a, bottom chart). This type of analysis is included for any

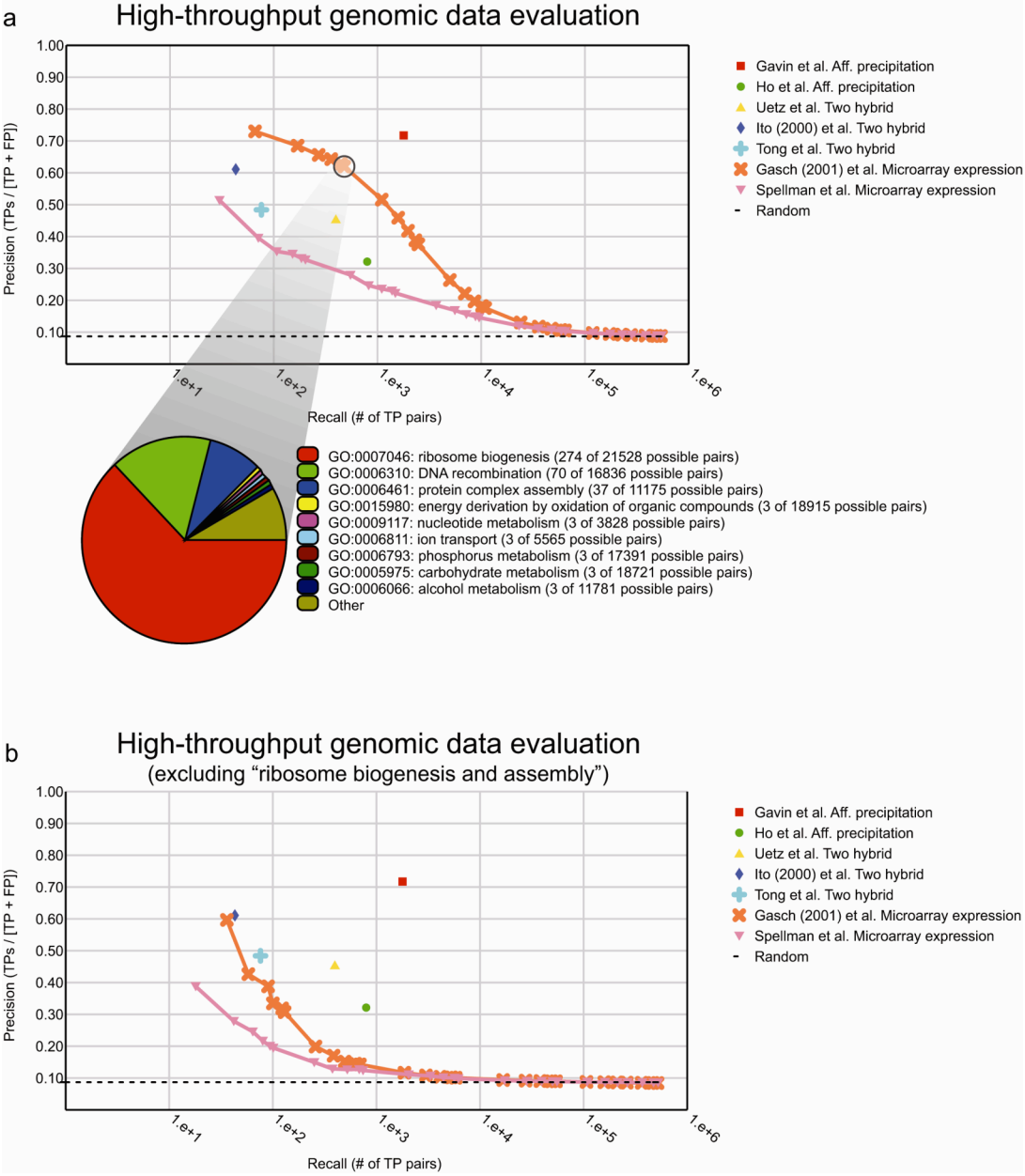


Figure 5.5. General (whole-genome) evaluation example. (a) Example of a genome-wide evaluation of several different high-throughput datasets using our framework. These datasets include five protein-protein interaction datasets, including yeast 2-hybrid [34,16,32] and affinity precipitation data [13,14], and two gene expression microarray studies [30,12]. Pearson correlation was used as a similarity metric for the gene expression data. The functional composition of the correctly classified set can be investigated at any point along the precision-recall trade-off, as is illustrated for the Gasch *et al.* co-expression data. This analysis reveals that a large fraction of the true positive predictions (> 60%) made by this dataset are associations of proteins involved in ribosome biogenesis. Of the 500 true positive pairs identified at this threshold, 298 are pairs between proteins involved in ribosome biogenesis, suggesting that the apparent superior reliability may not be general across a wider range of processes. (b) The same form of evaluation as in (a), but with a single GO term ("ribosome biogenesis and assembly," GO:0042254) excluded from the analysis, a standard option in our evaluation framework. With this process excluded, the evaluation shows that neither of the co-expression datasets is as generally reliable as the physical binding datasets. Additional functional biases can be interrogated through this analysis and corrected if necessary.

evaluation done with our system and interactive visualization allows for quick and accurate detection of any biases that might be present.

In addition to identifying biases in genome-wide evaluations of datasets or methods, our evaluation framework provides a way to normalize these biases out of the analysis. A user can choose to exclude all positive examples related to one or more biological processes. Figure 5.5b illustrates an example of this functionality for the evaluation discussed above. Based on the bias we observed, we excluded all proteins involved in ribosome biogenesis and assembly (GO term GO:0042254) and re-evaluated the same set of datasets. While none of the interaction datasets change significantly with this process excluded, both gene expression datasets show substantial decay in their precision-recall characteristics, suggesting they are generally less reliable at predicting functional relationships over a broad range of processes. This result is quite different from what we might have concluded had we not been able to discover and correct this process-specific bias.

Process-specific evaluation

Many biological laboratories focus on specific processes or domains of interest, even when using high throughput data/methods. In such situations, a targeted, process-specific evaluation is often more appropriate than a genome-wide evaluation. Our framework facilitates convenient and representative process-specific evaluations by performing independent precision-recall analysis for each process of interest.

For effective presentation of process-specific evaluation results, we have developed an interactive matrix-based view that facilitates comparative evaluation of multiple datasets across several targeted biological processes (Figure 5.6). This method allows for easy and dynamic inter-process and inter-dataset comparisons. In addition, precision-recall characteristics for any process are readily accessible, allowing for a more detailed view of the results. Thus, our framework combines general and specific evaluations, enabling accurate interpretation of functional genomics data and computational methods. This community standard can facilitate the

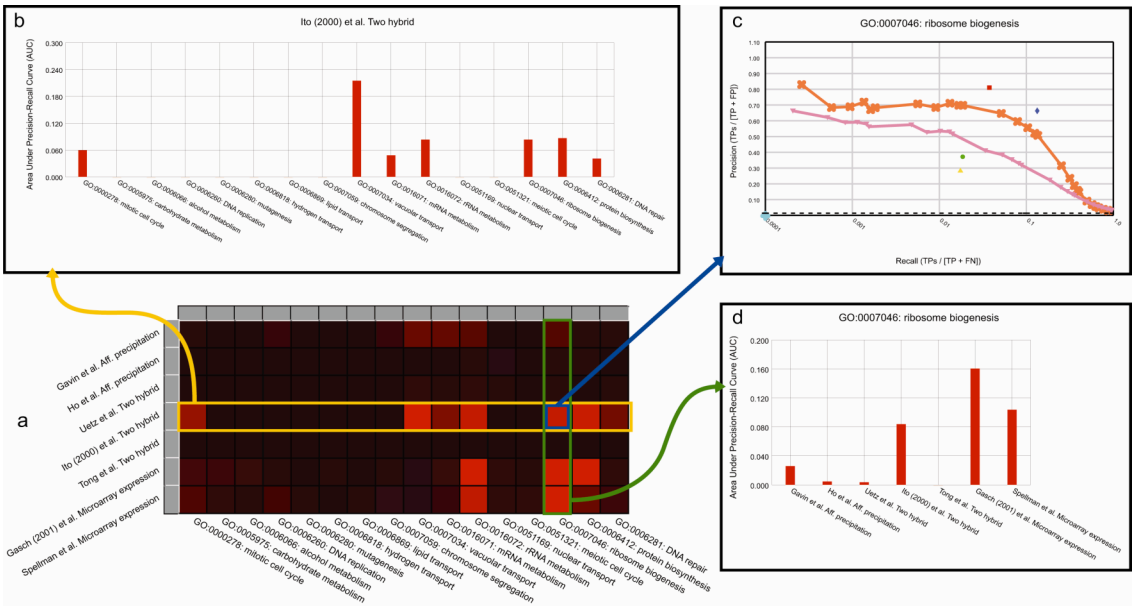


Figure 5.6. Process-specific evaluation example. A detailed understanding of which specific biological signals are present in a particular dataset is important for robust evaluation. Our evaluation framework allows users to query specific processes of interest. (a) Example of an evaluation of 7 high-throughput datasets over a set of 16 user-specified processes (GO terms). The precision-recall characteristics of each dataset-process combination were computed independently and the intensity of the corresponding square in the matrix is scaled according to the area under the precision-recall curve (AUPRC). (b) Detailed comparison of results for a single dataset, which can be accessed directly from the summary matrix. The AUPRC statistic of a particular dataset (e.g. Ito *et al.* two-hybrid) for each process is plotted to allow for comparison across a single dataset. (c) The actual precision-recall curve (from which the AUPRC was computed) is also easily accessible from our evaluation framework. Users can view underlying details of the AUPRC summary statistic which appears in the other three result views. (d) The AUPRC results for a single biological process across all datasets can also be obtained from an evaluation result. This allows for direct measure of which datasets are most informative for a process of interest.

comparisons necessary for formulating relevant biological hypotheses and determining the most appropriate dataset or method for directing further experiments.

5.5 Supporting Methods

5.5.1 GO-based Functional Gold Standard

With the Gene Ontology and corresponding annotations in hand, the main issue in generating a standard for evaluation is deciding which terms are specific enough to imply functional associations between gene products. As noted in Results and discussion, the typical approach to this problem has been to select a particular depth in the ontology, below which all co-annotated genes are taken to be positive examples. This has obvious problems in that biological specificity

varies dramatically at any given depth in the ontology (see Figure 5.3 and Table 5.1 for details). Another approach reported in the literature is to use term size (i.e. the number of gene product annotations) as a proxy for biological specificity. Using this approach, gene products co-annotated to terms smaller than a certain threshold are considered positive examples. The number of annotation genes, however, is not only a function of how specific a particular term is, but often how well-studied the area is. Thus size is not always an accurate indicator of specificity, and this problem only becomes worse in organisms that are less well-studied.

To address the issue of biological specificity of positive examples, we chose the less automated but more direct and biologically consistent approach of expert curation. For this task, we chose six biological experts with doctorate degrees in yeast genomics. This group contains a cumulative total of more than 40 years of post-doctoral experience working with yeast in a research setting. Instead of using characteristics of the GO term (e.g. depth in the hierarchy, number of annotations) to determine specificity, we instructed our expert panel to formally assess which GO terms are specific enough to imply a meaningful biological relationship between two annotated proteins. More precisely, we instructed the experts to select terms with enough specificity that predictions based on them could be used to formulate detailed biological hypotheses, which could be confirmed or refuted by laboratory experiments. This curation was performed for all GO terms from the biological process branch of the ontology without information of their hierarchical relationships, and each set of resulting responses was corrected for hierarchical inconsistencies. Responses for all experts were then merged by counting the number of votes for each GO term and terms that received more than three votes were selected for the positive evaluation standard. The final counts for all GO terms can be obtained from additional file 1: Biological expert voting results.

Given this set of specific GO terms, we can generate a positive pairwise gold standard by considering all proteins co-annotated to each term as positives. This set of specific functional classes can also be used to directly evaluate or train computational approaches that explicitly associate proteins with particular biological processes as well. For this, we start with the set of specific terms and obtain a non-redundant set by removing any terms whose ancestors are also

in the set. This set of terms can be obtained from additional file 3: Non-redundant set of specific GO terms.

We can also use the results of this voting procedure to define a representative set of negative examples. We expect that GO terms receiving 1 or fewer votes are too general to imply meaningful functional relationships between co-annotated proteins. Furthermore, GO terms with a very large number of direct and indirect annotations (i.e. a substantial fraction of the genome) are most certainly too general to imply meaningful functional relationships between co-annotated members. Thus, we obtain a set of gold standard negatives by finding pairs of proteins in which both members have annotations (other than "biological process unknown") but whose most specific co-annotation occurs in terms with more than 1000 total annotations (~25% of the annotated genome) and with one or fewer votes from our panel of six experts. The resulting negative set is more accurate than random pairs of proteins but is still large enough to reflect our understanding of the relative size of functionally related to unrelated pairs in the genome. Furthermore, this set of negative examples is more representative of the presumed distribution of biological negatives than alternate sources of negative evidence such as co-localization. The final gold standard based on this analysis can obtained from additional file 2: GO-based yeast functional gold standard.

The resulting set of gold standard positive and negative examples is quite different from previously used GO standards based on size or depth as a measure of biological specificity. Figure 5.4 illustrates this, plotting a histogram of GO term depth and size for both the excluded and included GO term sets based on the biological expert voting procedure described above. Because our gold standard is based on direct re-evaluation of the gene ontology with respect to functional genomics, there are a number of non-specific GO terms excluded based on the voting results that appear relatively deep in the ontology, and conversely, a number of relevant GO terms included that appear near the root (Figure 5.4). A similar trend is true of the GO term sizes of the selected and excluded set: many of the GO terms excluded on the basis of expert voting have relatively few annotations. This confirms our earlier observation that neither size nor depth in the ontology serve as good measures of biological specificity. Basing the criteria for generating

a GO-based gold standard instead on expert knowledge ensures that the standard is consistent in terms of the biological specificity of the relationships it is capturing and can therefore provide a meaningful basis for evaluation.

Other efforts have previously aimed to derive summary terms from the GO hierarchy, most notably the Saccharomyces Genome Database's (SGD) GO Slim set [2]. This set, however, is not generally appropriate for the purposes of functional evaluation as it was constructed to be a set of "broad biological categories" meant to span the entire range of processes [2]. The functional relationships captured by such broad terms are often too general to provide a meaningful basis for data evaluation. For example, protein biosynthesis (GO:0006412) is one such term included in the GO Slim set, which has approximately 800 annotated genes. A prediction of an uncharacterized protein's involvement in "protein biosynthesis" would not be specific enough to warrant further experimental investigation in most cases. Furthermore, from the perspective of defining an accurate pairwise evaluation standard, clearly not every pair of genes within this set (over 300,000 possible pairwise combinations) has a specific functional relationship.

5.5.2 Metrics for Evaluation: ROC and Precision-recall Curves

Sensitivity-specificity and precision-recall analysis are two approaches to measuring the predictive accuracy of data from two classes given the class labels (referred to here as positive and negative). Sensitivity and specificity are typically computed over a range of thresholds (for multi-valued data) and plotted with respect to one another. Such an analysis is known as a Receiver Operating Characteristic (ROC) curve and portrays the trade-off between sensitivity and specificity. Each threshold yields one point on the curve by considering protein pairs whose association in the data exceeds the threshold value to be positive predictions and other pairs to be negative. Precision-recall analysis is done in the same way, but with precision (or PPV) replacing specificity. Each of these quantities is calculated as described in Table 5.2.

ROC and precision-recall curves can be summarized with a single statistic: the area under the curve. For ROC curves, we refer to this statistic as the AUC, which is equivalent to the

Table 5.2. Definition of quantities relevant for dataset evaluation.

Quantity	*Definition*
True positives (TP)	protein pairs associated by data and annotated as positives in gold standard
False positives (FP)	protein pairs associated by data and annotated as negatives in gold standard
True negatives (TN)	protein pairs not associated by data and annotated as negatives in gold standard
False negatives (FN)	protein pairs not associated by data and annotated as positives in gold standard
Precision	$\frac{TP}{TP+FP}$
Recall	$\frac{TP}{TP+FN}$
Specificity	$\frac{TN}{TN+FP}$
Sensitivity	$\frac{TP}{TP+FN}$

Wilcoxon rank-sum (Mann-Whitney) statistic. Precision-recall characteristics can be summarized with a similar measure which we refer to as the AUPRC. For all plots shown here, we have used AUPRC because precision is more informative than specificity for the typical sizes of positive and negative example sets as discussed in the "Relative size of gold standard positive/negative sets" section of Results and discussion.

5.5.3 Implementation of Web-based Evaluation Framework

To facilitate community use of the standard, we have implemented our evaluation framework in a public, web-based system available at [36]. All evaluations are based on the standard described in "Defining a new gold standard", which is also available for download as additional file 2: GO-based yeast functional gold standard and additional file 3: Non-redundant set of specific GO terms. The website allows users to upload genomic datasets for evaluation and includes several widely used high throughput datasets (including those described

(logo design by Matt Hibbs)

here) for comparative evaluation. The methods for presenting evaluation results, including all graphs and interactive components, were implemented in SVG (Scalable Vector Graphics), which can be viewed on most browsers with freely available plugins (see Help at [36] for details). The web interface was implemented in PHP, with a back-end MySQL database and C++ evaluation server.

5.6 Conclusions

We have identified a number of serious issues with current evaluation practices in functional genomics. These problems make it practically impossible to compare computational methods or large-scale datasets and also result in conclusions or methods that generalize poorly in most biological applications. We have developed an expert-curated functional genomics standard and a methodological framework that address the problems we have identified. We hope these can serve as an alternative to current evaluation methods and will facilitate accurate and representative evaluation. Furthermore, we hope our analysis will initiate a broader community discussion about appropriate evaluation techniques and practices.

In recent years, the computational community has played an influential role in the field of genomics by contributing many valuable computational methods that facilitate discovery of biological information from high-throughput data. However, without an accurate understanding of how well the computational methods perform, the role of bioinformatics in directing experimental biology will remain limited. Lack of accurate assessment of the experimental methods themselves hinders both interpretation of the results and further development of genomic techniques. Thus, representative evaluation of computational approaches and high throughput experimental technologies is imperative to our ability as a community to harness the full potential of biological data in the post-genome era.

5.7 Supplemental Data Files

1. Biological expert voting results
File name: GO_curated_gold_standard_votingresults.txt
File format: tab-delimited text
Title: Biological expert voting results
Description:
This file contains the results of the voting procedure used to generate a functional gold standard based on the Gene Ontology (described in detail in Methods). Experts selected terms that are specific enough to direct laboratory experiments, but are also general enough to reasonably expect high-throughput assays to provide relevant information.

2. GO-based yeast functional gold standard
File name: GO_curated_gold_standard.txt.gz
File format: gzipped text
Title: GO-based yeast functional gold standard
Description:
This file contains the final pairwise gold standard set of positive and negatives resulting from our expert curation. Yeast protein pairs classified as positives are labeled with a "1" and pairs classified as negative in the standard are indicated with a -1.

3. Non-redundant set of specific GO terms
File name: GO_curated_nonredundant_terms.txt
File format: tab-delimited text
Title: Non-redundant set of specific GO terms
Description:
This file contains a non-redundant set of GO terms receiving more than 3 votes (of 6) from experts. The non-redundant set was obtained by removing any term whose ancestor in the hierarchy is also in the set.

4. Supplementary discussion
File name: supplementary_discussion.pdf
File format: pdf
Title: Supplementary discussion
Description:
This file contains a more detailed discussion of the relative size of gold standard positive and negative example sets and associated issues.

References

1. Alfarano, C., C. E. Andrade, et al. (2005). "The Biomolecular Interaction Network Database and related tools 2005 update." Nucleic Acids Res **33 Database Issue**: D418-24.

2. Ashburner, M., C. A. Ball, et al. (2000). "Gene ontology: tool for the unification of biology. The Gene Ontology Consortium." Nat Genet **25**(1): 25-9.

3. Ball, C. A., K. Dolinski, et al. (2000). "Integrating functional genomic information into the Saccharomyces genome database." Nucleic Acids Res **28**(1): 77-80.

4. Barutcuoglu, Z., R. E. Schapire, et al. (2006). "Hierarchical multi-label prediction of gene function." Bioinformatics.

5. Ben-Hur, A. and W. S. Noble (2005). "Choosing negative examples for the prediction of protein-protein interactions." BMC Bioinformatics **(in press)**.

6. Ben-Hur, A. and W. S. Noble (2005). "Kernel methods for predicting protein-protein interactions." Bioinformatics **21 Suppl 1**: i38-i46.

7. Bork, P., L. J. Jensen, et al. (2004). "Protein interaction networks from yeast to human." Curr Opin Struct Biol **14**(3): 292-9.

8. Breitkreutz, B. J., C. Stark, et al. (2003). "The GRID: the General Repository for Interaction Datasets." Genome Biol **4**(3): R23.

9. Clare, A. and R. D. King (2003). "Predicting gene function in Saccharomyces cerevisiae." Bioinformatics **19 Suppl 2**: II42-II49.

10. Costanzo, M. C., M. E. Crawford, et al. (2001). "YPD, PombePD and WormPD: model organism volumes of the BioKnowledge library, an integrated resource for protein information." Nucleic Acids Res **29**(1): 75-9.

11. Deane, C. M., L. Salwinski, et al. (2002). "Protein interactions: two methods for assessment of the reliability of high throughput observations." Mol Cell Proteomics **1**(5): 349-56.

12. Gasch, A. P., M. Huang, et al. (2001). "Genomic expression responses to DNA-damaging agents and the regulatory role of the yeast ATR homolog Mec1p." Mol Biol Cell **12**(10): 2987-3003.

13. Gavin, A. C., M. Bosche, et al. (2002). "Functional organization of the yeast proteome by systematic analysis of protein complexes." Nature **415**(6868): 141-7.

14. Ho, Y., A. Gruhler, et al. (2002). "Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry." Nature **415**(6868): 180-3.

15. Huh, W. K., J. V. Falvo, et al. (2003). "Global analysis of protein localization in budding yeast." Nature **425**(6959): 686-91.

16. Ito, T., T. Chiba, et al. (2001). "A comprehensive two-hybrid analysis to explore the yeast protein interactome." Proc Natl Acad Sci U S A **98**(8): 4569-74.

17. Jansen, R. and M. Gerstein (2004). "Analyzing protein function on a genomic scale: the importance of gold-standard positives and negatives for network prediction." Curr Opin Microbiol **7**(5): 535-45.

18. Jansen, R., H. Yu, et al. (2003). "A Bayesian networks approach for predicting protein-protein interactions from genomic data." Science **302**(5644): 449-53.

19. Kanehisa, M. and S. Goto (2000). "KEGG: kyoto encyclopedia of genes and genomes." Nucleic Acids Res **28**(1): 27-30.

20. Lanckriet, G. R., M. Deng, et al. (2004). "Kernel-based data fusion and its application to protein function prediction in yeast." Pac Symp Biocomput: 300-11.

21. Lee, I., S. V. Date, et al. (2004). "A probabilistic functional network of yeast genes." Science **306**(5701): 1555-8.

22. Lee, S. G., J. U. Hur, et al. (2004). "A graph-theoretic modeling on GO space for biological interpretation of gene clusters." Bioinformatics **20**(3): 381-8.

23. Lin, N., B. Wu, et al. (2004). "Information assessment on predicting protein-protein interactions." BMC Bioinformatics **5**(1): 154.

24. Lord, P. W., R. D. Stevens, et al. (2003). "Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation." Bioinformatics **19**(10): 1275-83.

25. Mewes, H. W., D. Frishman, et al. (2002). "MIPS: a database for genomes and protein sequences." Nucleic Acids Res **30**(1): 31-4.

26. Myers, C. L., D. R. Barrett, et al. (2006). "Finding function: evaluation methods for functional genomic data." BMC Genomics **7**: 187.

27. Patil, A. and H. Nakamura (2005). "Filtering high-throughput protein-protein interaction data using a combination of genomic features." BMC Bioinformatics **6**(1): 100.

28. Pavlidis, P., J. Weston, et al. (2002). "Learning gene functional classifications from multiple data types." J Comput Biol **9**(2): 401-11.

29. Qi, Y., J. Klein-Seetharaman, et al. (2005). "Random forest similarity for protein-protein interaction prediction from multiple sources." Pac Symp Biocomput: 531-42.

30. Spellman, P. T., G. Sherlock, et al. (1998). "Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization." Mol Biol Cell **9**(12): 3273-97.

31. Sprinzak, E., S. Sattath, et al. (2003). "How reliable are experimental protein-protein interaction data?" J Mol Biol **327**(5): 919-23.

32. Tong, A. H., B. Drees, et al. (2002). "A combined experimental and computational strategy to define protein interaction networks for peptide recognition modules." Science **295**(5553): 321-4.

33. Troyanskaya, O. G., K. Dolinski, et al. (2003). "A Bayesian framework for combining heterogeneous data sources for gene function prediction (in Saccharomyces cerevisiae)." Proc Natl Acad Sci U S A **100**(14): 8348-53.

34. Uetz, P., L. Giot, et al. (2000). "A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae." Nature **403**(6770): 623-7.

35. von Mering, C., R. Krause, et al. (2002). "Comparative assessment of large-scale data sets of protein-protein interactions." Nature **417**(6887): 399-403.

36. Website. "GRIFn Home Page." from http://function.princeton.edu/GRIFn.

37. Wong, S. L., L. V. Zhang, et al. (2005). "Discovering functional relationships: biochemistry versus genetics." Trends Genet **21**(8): 424-7.

38. Xenarios, I., D. W. Rice, et al. (2000). "DIP: the database of interacting proteins." Nucleic Acids Res **28**(1): 289-91.

39. Yamanishi, Y., J. P. Vert, et al. (2004). "Protein network inference from multiple genomic data: a supervised approach." Bioinformatics **20 Suppl 1**: I363-I370.

Chapter 6

Context-sensitive Data Integration and Prediction of Biological Networks

6.1 Chapter Overview

In Chapter 4, we discussed a general strategy for integrating genomic data and predicting biological networks that accounts for reliability of the input data. However, as discussed in Chapter 5, experimental technologies capture different biological processes with varying degrees of success, and thus, each source of genomic data can vary in relevance depending on the biological process one is interested in predicting. Accounting for this variation can significantly improve network prediction, but no previous approaches have explicitly leveraged this critical information about biological context.

In this chapter, we confirm the presence context-dependent variation in functional genomic data and propose a Bayesian approach for context-sensitive integration and query-based recovery of biological process-specific networks. By applying this method to *Saccharomyces cerevisiae*, we demonstrate that leveraging contextual information can significantly improve the precision of network predictions, including assignment for uncharacterized genes. We expect that this general context-sensitive approach can be applied to other organisms and prediction scenarios.

The work presented in this chapter was published in [21] and includes contributions from David Hess, Amy Caudy, and Olga Troyanskaya. David and Amy were complete all experimental work in confirming mitochondria-related mutants, and Olga supervised the project.

6.2 Background

Recent developments in biological technology have fueled the generation of numerous large genomic and proteomic datasets for several organisms. These data capture a wide range of biological phenomena including gene expression, genetic interactions, physical interactions between proteins, and sequence content. Many recent studies have shown that high-throughput data are often quite noisy and have varying degrees of reliability or relevance for understanding biological networks [9,24,5]. To address this heterogeneity and harness the wealth of information present in the data, several groups have designed methods for data integration to combine information from multiple sources of genomic or proteomic evidence in order to arrive at accurate and holistic network and gene predictions. For instance, Troyanskaya *et al.* used expert-based

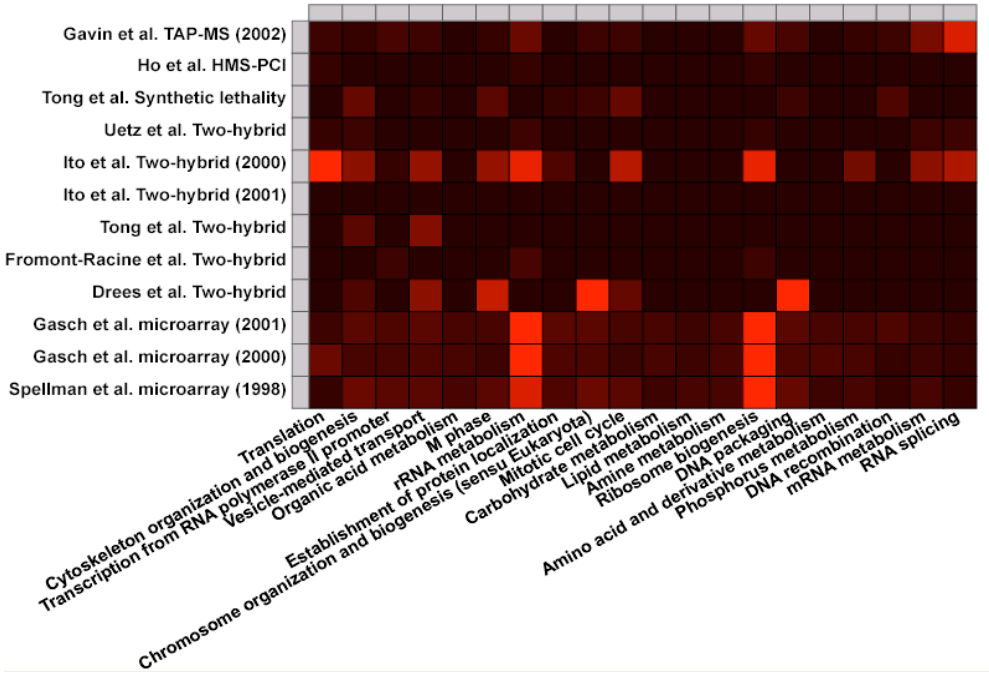


Figure 6.1. Dataset relevance across different biological contexts. We measured the relevance of several *Saccharomyces cerevisiae* genomic datasets for predicting function in a range of biological contexts (GO terms) using our previously published evaluation framework [19]. A selection of the datasets used in our integration appear on the rows, and contexts appear on the columns. The intensity of each square reflects the area under a precision-recall curve (AUPRC) for each dataset in the corresponding context. The relevance of each dataset varies substantially both in terms of precision and sensitivity across biological processes, and thus the relative weighting of data during integration depends critically on the context. For example, if one were interested in predicting proteins involved in ribosome biogenesis, any of the three gene expression datasets would be informative. If one were interested in chromosome organization, these data might offer little reliable information as compared to one of the two-hybrid datasets (e.g. Drees *et al.*).

Bayesian networks for inferring functional interactions between pairs of proteins given observed experimental data supporting those interactions [26]. Other studies have extended this idea by applying more sophisticated Bayesian approaches and other methods, most of which automatically learn reliability characteristics from the data given a trusted gold standard [15,27,17,14,23]. In general, all of these methods assess the reliability of input high-throughput genomic data and use these characteristics for more robust integration, which typically offers significant improvement in terms of both sensitivity and specificity in predicting protein-protein interactions or functional relationships.

While these earlier approaches to data integration address the heterogeneity in reliability among different datasets, they all fail to utilize one important source of variation: biological context. Most experiments are designed with a particular process or pathway in mind. For instance, a researcher studying meiosis in yeast might profile gene expression under specific conditions (e.g. in sporulation media) that result in a clear meiotic signal in the data but very little reliable information about the mitotic cell cycle. Furthermore, most experimental technologies target specific biological processes simply because of how they physically measure biological phenomena. Yeast two-hybrid technology for identifying interacting proteins, for example, relies on the two-domain structure of eukaryotic transcription factors to report an interaction. A two-hybrid positive interaction is obtained by fusing one protein to a DNA-binding domain (bait) while another protein is fused to an activation domain such that binding of the two proteins of interest "switches on" transcription of a reporter gene [22]. Thus, while two-hybrid results are generally informative for proteins which can be targeted to the nucleus, we should expect very little reliable information about membrane proteins or proteins with domains that prevent them from entering the nucleus. In fact, if an interaction including such a protein is reported, we should confidently reject it as a false positive.

We have explicitly measured context-dependent variation for a wide variety of public, genomic data for *Saccharomyces cerevisiae* (baker's yeast), including a large number of microarray datasets, protein-protein interaction data, and sequence data. Specifically, for each source of functional genomic data we measured precision-recall characteristics for a set of

experimentally relevant Gene Ontology (GO) terms covering a broad range of biological processes [19] (Figure 6.1). This analysis demonstrates that most datasets have a broad range of precision-recall characteristics depending on which processes they are compared against. More importantly, we find that the relative ordering of genomic data in terms of quality varies dramatically from process to process, suggesting the degree to which we should trust any dataset depends on the process we are interested in predicting.

While this context-dependent variation is not surprising given the inherent bias of different experimental techniques toward particular processes and different goals and conditions under which the data was measured, to our knowledge, no previous computational approaches for heterogeneous data integration or network prediction have explicitly leveraged this information. We demonstrate here that incorporating information about biological context in the integration and prediction process can significantly boost precision and sensitivity. We develop a system for predicting process-specific networks from diverse genomic data that uses biological context information to improve the recovery of known networks from integrated experimental data. We compare our contextual approach to our earlier work, which uses prior knowledge of gene function as a gold standard, but does not specifically leverage biological context [20], and demonstrate that considering context can yield a dramatic benefit. While we illustrate the effect of biological context for a specific method for network prediction here, we demonstrate that such context-specificity has a dramatic effect on dataset reliability and thus we expect that the general idea can be used to improve predictions in a variety of settings and for many organisms.

6.3 Methods

The objective of our approach is, given a diverse set of genomic data, to recover a process-specific network starting from a small related set of query proteins. Such algorithms have proven to be practical approaches for expert-driven search of genomic data, largely because they harness information from all available evidence in a robust way while also providing an intelligent interface for discovering functional modules and extracting the relevant portion of the interaction network [20]. This general approach of incorporating expert direction in the prediction process is

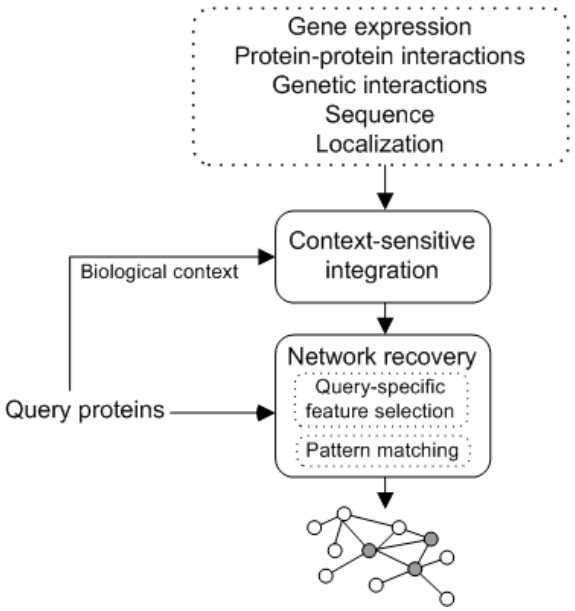


Figure 6.2. Overview of method for context-sensitive integration and prediction. Our approach is developed for the scenario where a user enters a query set of proteins and wishes to obtain a relevant network prediction based on a diverse set of experimental evidence. The method consists of two stages, the first a Bayesian network for data integration and the second a network recovery algorithm which uses the probabilistic network from the first stage to recover the network surrounding the entered query. The biological context of a prediction is inferred from the entered query set, and this information is fed into both stages to improve prediction precision.

p

articularly attractive because it offers a convenient method of learning the biological context and leveraging this information to arrive at more precise predictions. Our solution based on this premise can be divided into two distinct components: a data integration phase that forms a probabilistic protein-protein network as supported by experimental data, and a network search algorithm that, given the probabilistic network, recovers additional relevant proteins starting from a query set (Figure 6.2). Both phases of the network prediction process utilize information about biological context, which is inferred from the starting query set.

6.3.1 Bayesian Context-specific Integration

The integration phase consists of a Bayesian network, which captures the context-dependent reliability variation to integrate the diverse input data. The result of this phase is a probabilistic protein-protein interaction network reflecting the reliability of the supporting data in a given biological context. The input data used here and the details of the Bayesian network are described below.

Genomic input data

We have collected genomic data for *Saccharomyces cerevisiae* from over 6500 publications, including gene expression, literature-curated and high-throughput protein-protein and genetic interactions [1,25], protein localization data [13], transcription factor binding site data [29,12], and sequence data [28]. See Appendix A for a detailed description of how each data type was processed. The processed input data was separated first by experimental method responsible for producing the data, then by publication. To ensure that each input dataset had a reasonable number of observations for learning, publications with fewer than 50 observations were merged with other publications reporting results from the same experimental method. This process resulted in 174 different input data types for Bayesian integration.

Bayesian network

The goal of our integration scheme is to harness the information from the diverse data while not sacrificing precision. Furthermore, the integration is designed such that it can model and exploit the context-dependent relevance variation discussed earlier. Because many of the input data types represent functional interactions (either physical or other) between pairs of genes or proteins, we have adopted the approach of predicting functional associations. This approach has been used in several earlier studies [15,27,17,14], and the final integrated protein-protein linkage network is convenient for understanding and predicting network structure, which is our goal here. Several methods for associating proteins directly with processes or functional classes (function prediction) have also been applied successfully [18,16,6], but are less appropriate for the goal of network analysis and prediction.

Starting with the goal of predicting functional associations between genes, there are several choices of machine learning methods that might be appropriate. Here, we employ a Bayesian network because it is robust to diverse forms of input data, and it yields a generative model that is useful in terms of drawing relevant biological conclusions about the properties of the input data. Furthermore, a Bayesian framework is a convenient setting for incorporating contextual information as is illustrated below.

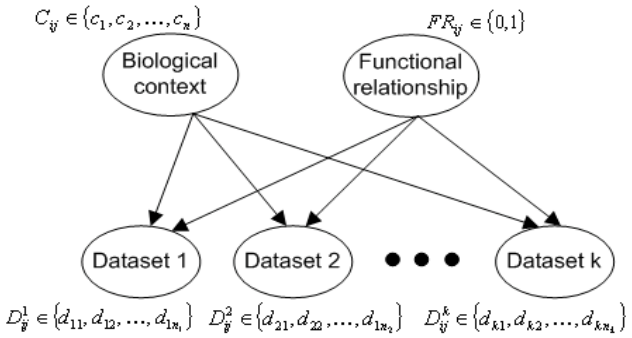


Figure 6.3. Bayesian network for context-sensitive integration. The data integration stage of our context-sensitive approach consists of a Bayesian network, which is used to integrate pairwise protein-protein association data to arrive at a single, probabilistic network. Biological context information is incorporated into the integration process by conditioning the probability distributions of each type of observed genomic data on both the presence or absence of a functional relationship between the pair of proteins in question and the biological context of interest. This structure captures both the inherent dataset quality as well as the relevance variation from one biological process to another. Evidence nodes are assumed to be discrete, and conditional probability tables (CPT's) are automatically learned from the data using a gold standard based on the biological process branch of the Gene Ontology (GO).

The simplest Bayesian approach for integration is to assume independence between all of the input datasets given knowledge of a functional relationship between any pair of proteins. In practice, this approach is quite powerful for genomic data and is competitive with more sophisticated alternatives, including methods where dependence among datasets is modeled (e.g. tree-augmented Bayesian networks [10], see Appendix B for a comparison). We begin with the naive approach and extend it to include contextual information as illustrated in Figure 6.3. Each input dataset is modeled with a discrete probability distribution conditioned on the presence or absence of a functional relationship *and* the biological context. Given a gold standard which associates observed data with known functional relationships and biological context (described in detail in the following section), we estimate the conditional distribution for each input dataset by simple counting. With these learned parameters, given a new protein-protein pair with observed data and a corresponding context (derived from the query as described below), we can then infer the probability of functional relationship between the two proteins, i.e.

$$P\left(FR_{ij} \middle| D_{ij}^1, D_{ij}^2, \ldots, D_{ij}^k, C_{ij}\right) = \alpha\, P\left(FR_{ij} \middle| C_{ij}\right) \prod_{n=1}^{k} P\left(D_{ij}^n \middle| FR_{ij}, C_{ij}\right)$$

where

$$P\left(D_{ij}^{n} = d \middle| FR_{ij} = f, C_{ij} = c\right) = \frac{\#\left(D_{ij}^{n} = d \wedge FR_{ij} = f \wedge C_{ij} = c\right)}{\#\left(FR_{ij} = f \wedge C_{ij} = c\right)}.$$

Here FR_{ij} refers to the presence or absence of a functional relationship between proteins *i* and *j*, D_{ij}^{n} refers to the observed association in dataset *n* between the proteins *i* and *j*, C_{ij} is the biological context of the pair, and α is a normalization constant.

Gold Standard for Bayesian Integration

The gold standard used in estimating the parameters for the Bayes net is a critical part of the prediction process. The gold standard used here is based on the biological process branch of the Gene Ontology [2] as proposed in [19]. For the global (non-context-sensitive) approach described here, we directly used the protein-protein pairwise standard for functional relationships published in Myers et al. as our global (non-context-sensitive) standard for functional relationship. For the context-sensitive approach, we require a gold standard that associates positive and negative examples of functionally related pairs of proteins to a set of biological contexts. For this, we used the non-redundant set of specific GO terms published in [19], which is a set of terms spanning the entire process ontology at a specificity sufficient for inferring useful functional information as curated by biology researchers. Specifically, we chose the 101 largest of these terms (those with more than 20 annotations), as the space of all possible contexts (c_1, ..., c_n). Positive examples for each context were derived by forming all possible pairs of proteins annotated to the corresponding term. Negatives were sampled from the negative gold standard described in [19]. Negative gold standard pairs are obtained by sampling from the set of negatives used for the global context until the ratio of positives to negatives matches the global prior. 20% of these negatives are sampled from protein pairs annotated as negative in the gold standard but for which one of the proteins is in the current context. The remaining 80% are sampled from the entire set of gold standard negatives. The reason for this distinction is that these two sets of negatives can be, qualitatively, quite different— the negatives touching the context of interest are generally more difficult to classify. A context-sensitive gold standard consisting of a mix of these two types of negatives provides the best performance based on

empirical evidence. During the inference process, context is inferred from the entered query proteins by mapping to the term in this comprehensive set containing the maximum number of proteins in the query.

6.3.2 Context-sensitive Network Recovery Algorithm

The problem of recovering a network from a starting query set given a probabilistic interaction graph of proteins has been addressed in previous work [4,3,8,20]. Approaches to this problem range from random walks on the probabilistic network [8], to methods based in network reliability theory [3], to variations of maximum adjacency [4,20]. We find that the performance of such methods often depends on the sparsity of the starting network, and it is difficult to find one that always provides superior performance. We describe an approach here that performs favorably on our probabilistic network, but emphasize that the larger point of incorporating biological context is independent of the specific network recovery algorithm used. Our network recovery algorithm consists of two steps: (1) a feature selection step that, given a query set of genes, determines a "characteristic" interaction profile for that group, and (2) a pattern matching step that finds additional proteins matching the characteristic profile.

Feature selection

Let Q be the query set of proteins of size N_Q chosen out of the entire proteome consisting of N_T proteins, and let $p_{ij} = P\left(FR_{ij} \middle| D_{ij}^1, D_{ij}^2, \ldots, D_{ij}^k, C_{ij}\right)$ be the probability of functional relationship between proteins i and j in the current biological context. Our goal is to select a set of features which are predictive of proteins related to the query set. Here, we treat each protein's interaction probabilities as a set of features, and thus feature selection is equivalent to finding a set of interaction partners which are common and discriminative of the query set. For each possible feature, k, we compute:

$$N_{Q,k}(t) = \left| \left\{ i \in Q: \; p_{ij} > t \right\} \right|$$
$$N_{T,k}(t) = \left| \left\{ i: \; p_{ij} > t \right\} \right|$$

where t is a threshold on the interaction probabilities. We can then assign a p-value measuring the significance of the association between feature *k* and the query set using the hypergeometric distribution, i.e.

$$f_k(t) = 1 - \sum_{n=0}^{N_{Q,i}(t)} \frac{\binom{N_Q}{n}\binom{N_T - N_Q}{N_{T,i}(t) - n}}{\binom{N_T}{N_{T,i}(t)}}$$

For each feature, we compute this p-value over a range of interaction probability thresholds and select the minimum. The selected features are then given by $F = \left\{ k \in \{1,2,\ldots,N_T\} : \min_t f_k(t) < 0.05 \right\}$.

Pattern matching

During the pattern matching phase, we identify remaining genes whose interaction profiles match the characteristic profile determined during the feature selection phase. Given the query set, Q, and selected features, we add proteins to the predicted network based on their similarity to the query proteins over the set of relevant features, F. Specifically, each candidate protein, i, is ranked according to the following adjacency score:

$$S_i = \sum_{j \in Q} \sum_{k \in F} p_{ik} p_{jk}$$

This metric ensures that only relevant features are used in predicting the final network, and each relevant feature (protein interaction) is weighted by our confidence in that particular interaction. Intuitively, this two-step approach of graph feature selection and pattern matching identifies a set of informative neighbors in the interaction network and ranks candidate proteins by measuring adjacency to the query set on paths *through* these informative neighbors.

6.4 Results

We demonstrate the importance of considering biological context for predicting biological networks by comparing our contextual approach with a simpler version that does not use information about biological context. Specifically, we replaced the context-sensitive Bayesian

network illustrated in Figure 6.3 with a simple, naive structure with no context node. For all experiments described here, both approaches start with a query set of proteins and use the same network recovery procedure, such that the only difference between the two is the presence or absence of contextual information during data integration.

We compared the simpler version of our method (with no contextual information) to existing approaches for network recovery [4,3] in our previous publication [20]. In summary, the non-contextual version of our method outperforms existing approaches for network recovery in terms of both precision and recall on a wide range of biological processes, complexes, and pathways. The details of this comparison are summarized in the Supplementary information. Evaluation results presented here illustrate further improvement offered by incorporating context information during integration and network recovery.

6.4.1 Contextual Network Recovery Evaluation

Perhaps the most important question to address with evaluation experiments is: does incorporating biological context information improve network prediction? To answer this question, we performed cross-validation experiments on *Saccharomyces cerevisiae* data for both our context-sensitive approach and the simpler non-contextual Bayesian integration and search algorithm. Specifically, for each of the GO terms in the evaluation gold standard [19], we withheld one-half of the annotated proteins for network recovery evaluation. The other half was used in training both Bayesian network configurations (with and without context nodes). Positive and negative examples (protein pairs) for the non-contextual configuration were derived as described in [19]. For the context-specific case, we obtained positive protein pairs for each context by considering all pairs between proteins annotated to the corresponding GO terms, except those selected in the corresponding cross-validation fold, as positive examples. To maintain the same ratio of positives to negatives, negative examples were sampled from the negatives described in [19]. Details on the training example selection are discussed in the Supplementary data.

On the proteins held out in each cross-validation fold, query sets of 10 proteins each were randomly sampled from each GO term, and we attempted to recover the remaining proteins

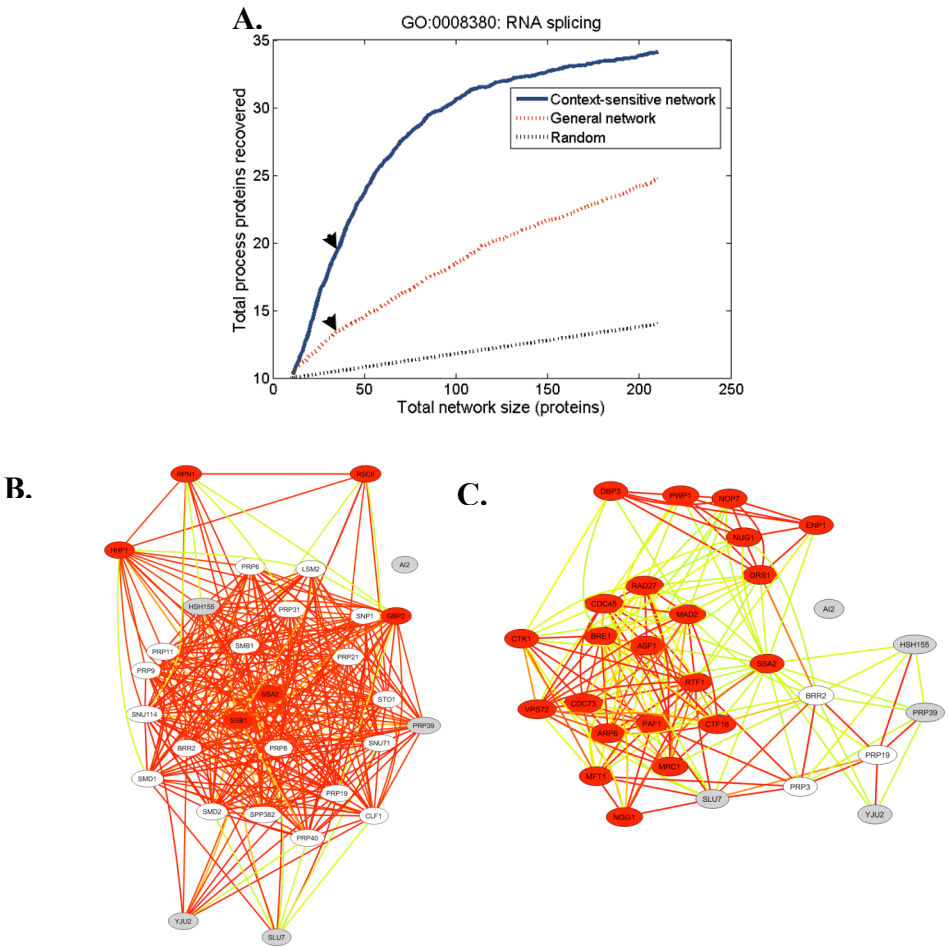


Figure 6.4. RNA splicing network recovery example. We compared the ability of the context-sensitive and global approaches to recover known networks of proteins using cross-validation experiments. Specifically, we started with a set of GO terms covering a wide range of biological processes [19], and measured each method's ability to recover held-out proteins given 10-protein queries from the same process. As proteins are added to the predicted network, we plot the number of true positive proteins present for each method, averaged over 20 query samplings (Figure 4a). On average, the context-sensitive approach recovers more held-out true positive proteins at better precision than the global approach. Specific examples of predicted networks from the context-sensitive and global approaches are pictured in Figures 4b and 4c respectively (sampled from the recovery curve at the point indicated in Figure 4a). Query proteins are colored gray, true positives are white, and false positives are red. For this particular query, the context-sensitive approach makes 24 of 30 correct predictions (80% precision) while the global approach only makes 8 of 30 correct predictions (27% precision).

w

ith both the context-sensitive and general approaches. All results presented here are averaged over 20 random query set samplings and two folds of cross-validation. We start by considering network recovery results for the RNA splicing context. Our context-sensitive integration and recovery dramatically improves both the precision and sensitivity of network recovery for RNA

splicing proteins (Figure 6.4). For example, starting with 10 randomly chosen RNA splicing proteins, the context-sensitive approach recovers an average of 25 proteins correctly in the first 50 predictions, while the global approach only recovers 15 proteins. Figures 6.4b and 6.4c illustrate the results of the same 5-protein query for both methods at the indicated point on the recovery performance curves. For this particular query, the context-sensitive prediction reports only 6 false positives resulting in 80% precision while the global network reports 22 false positives resulting in 27% precision. Both approaches are substantially better than random in terms of predictive power, but the contextual information clearly offers an improvement.

This improvement gained by using contextual information is consistent over a broad range of biological processes. We performed a similar evaluation to that described above for RNA splicing for 101 total GO terms from the evaluation set [19]. The results of this evaluation for a range of predicted network sizes are summarized in Table 6.1. As each approach added proteins to the predicted network, we measured the number of predicted, held-out true positives and averaged these estimates over several randomly sampled query sets. At each network size increment, we compared the average number of recovered true positive proteins for the context-sensitive versus global approaches and summarized the improvement over the set of evaluation GO terms for which both methods recovered at least 2 true positives (53 out of the 101 evaluation terms). For example, for networks of 40 recovered proteins (from a query of 10 proteins), the context-sensitive approach improved 51% of the GO terms by more than 2 standard deviations (estimated from random query samplings).

Conversely, the context-sensitive approach resulted in a deterioration of the performance by more than 2 standard deviations on only 8% of the GO terms. The average improvement in the number of true positives recovered across all terms for size 40 networks is 46%. This comparison is summarized in Figure 6.5. The improvement offered by context-sensitive integration and prediction is consistent across a range of network sizes (see Table 6.1 for a complete performance comparison).

Network size (proteins)	Fract. of processes where Context-sensitive > Global network *	Fract. of processes where Global > Context-sensitive network *	Fract. of processes where no significant performance difference	Average improvement (%)
15	0.49	0.13	0.38	50%
20	0.43	0.08	0.49	40%
25	0.52	0.11	0.38	53%
30	0.51	0.10	0.39	42%
35	0.53	0.09	0.38	44%
40¶	0.51	0.08	0.41	46%
50	0.54	0.07	0.39	42%
60	0.55	0.07	0.38	44%

* Networks are only counted as significantly different if the difference in number of true positive proteins recovered is more than two standard deviations over random samplings.
¶ Pictured in Figure 5.

Table 6.1. Comparison between context-sensitive and global network inference approaches. To compare the context-sensitive and global approaches to network prediction, we performed cross-validation experiments as described in the manuscript. On the proteins held out in each cross-validation fold, query sets of 10 proteins each were randomly sampled from each GO term, and we attempted to recover the remaining proteins with both the context-specific and global approaches. All results presented here are averaged over 20 query set samplings and two folds of cross-validation. This table compares the ability of the two approaches to recover held-out true positive proteins given a 10 protein query over a range of network sizes. Here, we restrict our evaluation to terms where both methods recovered at least 2 true positives (53 of the 101 total terms). For each network size, the fraction of the evaluation contexts for which the context-sensitive approach improves over the global approach by at least 2 standard deviations is highlighted in gray (std. dev. estimated from the 20 samplings). The fraction of processes for which the converse occurs is also reported as is the fraction with no significant difference between the two approaches. The average improvement across *all* processes is reported in the final column. Figure 5 corresponds to a network size of 40.

Correlation between Context-sensitive Improvement and Context specificity

Interestingly, the performance of the context-sensitive integration correlates with the specificity of the context. Table 6.2 illustrates this effect for the evaluation described above, considering all 101 evaluation GO terms. For contexts with between 20 and 49 annotations (proteins), the context-sensitive approach improves the result significantly for 8 terms and results in a performance deterioration for 7 terms. However, as the context size grows, the number of terms improved by context-sensitive integration increases (e.g. 50-79: 10 improved, 1 deteriorated; 80-110: 11 improved, 2 deteriorated). This is likely due to the fact that the number of training examples for each context scales quadratically with the number of proteins annotated. Smaller contexts may suffer from too few examples, which results in mediocre performance of our

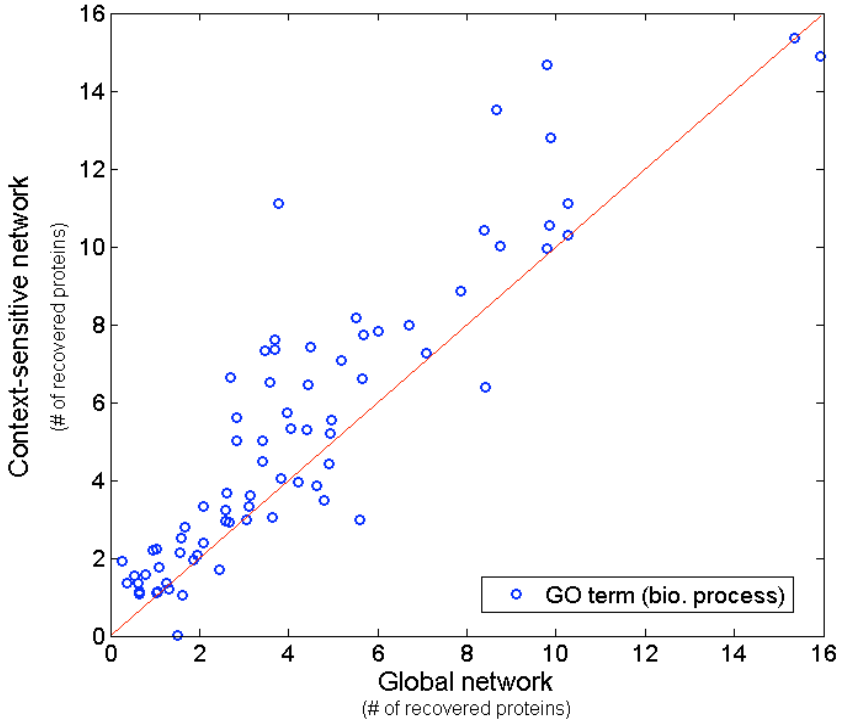


Figure 6.5. Network recovery evaluation summary. We compared the ability of the context-sensitive and global approaches to recover known networks of proteins using cross-validation experiments. Specifically, we started with a set of GO terms covering a wide range of biological processes [19], and measured each method's ability to recover held-out member proteins given 10-protein queries from the same process. As proteins were added to each process-specific network, we measured the number of true positives recovered. Figure 5 compares the number of true positives recovered for the two different methods for networks of 40 proteins on 101 different biological processes. The context-sensitive approach improves recovery by more than 2 std. dev. (estimated from query samplings) for 51% of the terms evaluated and only causes deterioration by more than 2 std. dev. on 8% of the terms. This improvement is consistent across network sizes (see Table 6.1 for a complete comparison).

approach. Another interesting result of this evaluation is that the two most severe performance deteriorations resulting from context-sensitive integration occur on the GO terms translation (GO:0043037) which has 424 annotations and cellular respiration (GO:0045333) which has 90 annotations. Both of these contexts clearly have a sufficient number of examples for learning, but represent fairly general processes compared to many of the other contexts (e.g. RNA splicing, sulfur metabolism). A possible explanation for this is that such contexts are so general, that context-specific learning is unable to identify a consistent signal between cross-validation folds, resulting in poor performance. In these cases, a global integration reflecting overall dataset reliability appears to be a safer alternative.

Table 6.2. Correlation between improvement due to context-sensitivity and the specificity of the context. This table list the total number of context networks where prediction improved and deteriorated for a range of context sizes.

# of proteins associated with context	20-49	50-79	80-110	111-140	>140	Total
# of terms improved by context-sensitive integration	8	10	11	6	11	46
# of terms deteriorated by context-sensitive integration	7	1	2	0	3	13

6.4.2 Comparing Dataset Relevance across Contexts

After confirming superior performance of the context-sensitive approach for a variety of biological processes, we investigated reasons for this improvement. The most informative aspect of our results is the learned parameters of the context-sensitive Bayesian network, which is designed to capture the relevance variation that motivated our approach. If our original observation of context-dependent relevance variation is correct, we expect to observe differences in the learned conditional probability distributions. To measure these differences, we computed $P(FR|D_i,C_j)$, the posterior probability of a functional relationship given an observation from a single dataset, D_i, across a range of biological contexts, C_j. To obtain a single measure reflecting the relevance of each dataset in each given, we then found the maximum posterior over all possible quantized observations for a given dataset. Comparing this posterior for several contexts to the same posterior inferred by the non-contextual Bayesian network yields insight into how dataset relevance variation is captured across different contexts. Figure 6.6 illustrates this comparison for 13 of the total 174 input datasets and two biological contexts: RNA splicing (GO:0008380) and Phosphorus metabolism (GO:0006793). The global network reports dataset relevance (posterior probability of FR) as inferred by the simpler Bayesian network (with no contextual information). As is demonstrated in the figure, there are several datasets for which the posterior from the global network is much larger than both contexts (e.g. ER-Golgi co-localization, Martin *et al.* microarray) suggesting these datasets are generally quite reliable but contain little information about either RNA splicing or phosphorus metabolism. Conversely, there are some datasets that appear relatively unreliable on the global scale, but are actually quite precise when examined in a

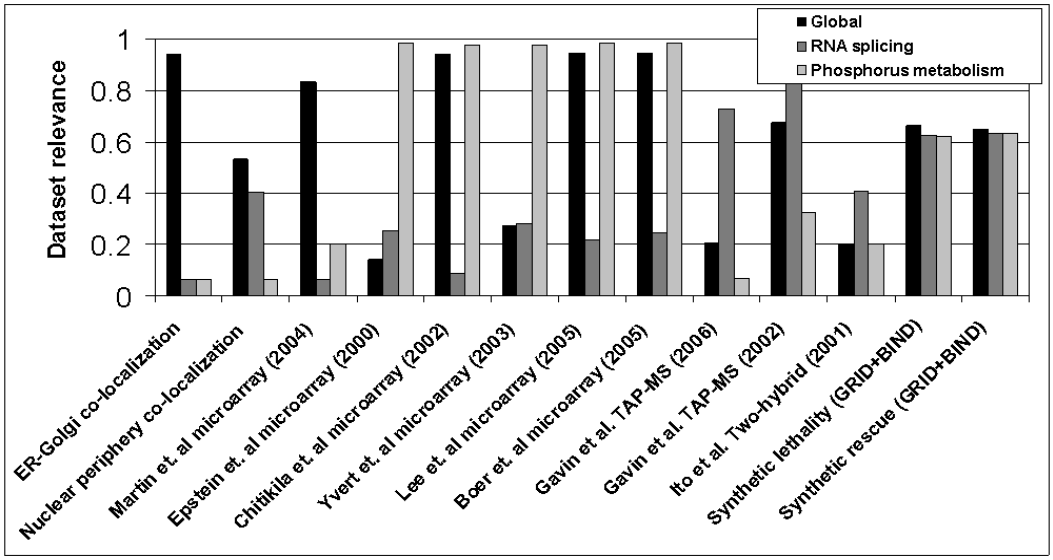


Figure 6.6. Bayes net learned dataset relevance. We analyzed the learned parameters of the context-sensitive Bayesian network to understand the improvement achieved by our method. Dataset relevance was measured by computing the maximum posterior probability of functional relationship for each dataset in each context. Figure 6a compares these relevance estimates for the global integration approach to the context-specific approach for RNA splicing and phosphorus metabolism contexts on a sampling of 13 datasets integrated by our approach. Datasets that one might expect to be relevant for predicting RNA splicing proteins are up-weighted relative to the global approach in the RNA splicing context (e.g. Gavin et al. TAP-MS data), and likewise, datasets that are likely relevant for understanding metabolism are up-weighted in the phosphorus metabolism context (e.g. Epstein et al., which profiled mitochondrial perturbations). Figures 6b and 6c compare these context-specific dataset relevance measures for the whole collection of 174 datasets to the global Bayesian network for RNA splicing and phosphorus metabolism, respectively. The most striking trend is that there are number of datasets which contain information globally but are uninformative (or contain no data) for these specific contexts. Modeling this variation during data integration helps to exclude false positives from irrelevant datasets that might otherwise result in poor network prediction.

spe

cific context. For instance, all three protein-protein interaction datasets pictured are up-weighted in the RNA splicing context, particularly the Gavin *et al.* TAP-MS (2006) interaction data, which measures a maximum posterior of .72 for the RNA splicing context compared to a .22 posterior in the simpler Bayesian network. From a biological standpoint, perhaps this is not too surprising since a large portion of the RNA splicing term is composed of the spliceosome complex, which would be readily detectable with physical binding assays. These protein-protein interaction datasets have no extra relevance for the phosphorus metabolism, but all of the microarray datasets included in Figure 6.6 are up-weighted in the phosphorus metabolism context, particularly the Epstein *et al.* dataset, which profiled several mitochondrial perturbations.

These differences between the global and context-specific posteriors are not limited to these 13 datasets, but occur in many of the datasets included in our integration (data not shown). Interestingly, there are a large number of datasets that have reasonably high posteriors in the global setting with near zero posteriors in the specific contexts. This suggests that many datasets either contain little or very unreliable information for these contexts. This knowledge is actually quite useful for improving predictions for a specific context, because it means we can confidently exclude a number of observations from the corresponding datasets as false positives. Generally, the chances of making a false positive prediction are high simply because of there are many more negative examples (proteins) than positive for network prediction problems. Thus, any reliable means of excluding false positives is an effective strategy for improving prediction performance.

6.4.3 Learning New Biology Using Contextual Information

We have shown through cross-validation experiments that using contextual information can generally improve the quality of network prediction, but these results are based on held-out, known annotations for genes or proteins. An interesting (and perhaps more biologically relevant) question is, does such an approach help us learn new biology with greater precision? While the true answer to this question requires experimental confirmation of novel predictions, we can derive some hints from our network recovery evaluation.

To compare the ability of the context-sensitive and global approaches to confidently associate previously uncharacterized proteins in *Saccharomyces cerevisiae* with portions of characterized networks, we performed a similar cross-validation experiment to that described previously. More specifically, on the proteins held out in each cross-validation fold, query sets of were randomly sampled from each GO term, and we used both methods to recover the remaining network. For each protein added to the network, we estimated the precision of that particular prediction based on known, held-out proteins for the corresponding cross-validation fold. Precision estimates were smoothed across each ranked list (order in which proteins were recovered for each network), and an uncharacterized gene appearing in any prediction was

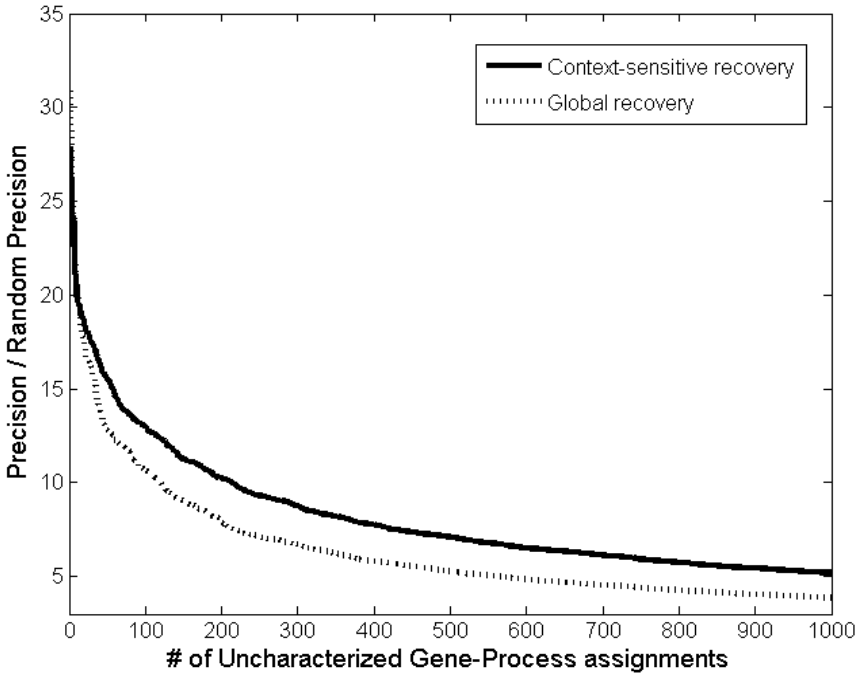


Figure 6.7. Precision of network prediction for uncharacterized genes. To assess the potential of context-sensitive prediction for learning new biology in *Saccharomyces cerevisiae*, we compared the ability of the context-sensitive and global approaches to predict precise networks involving uncharacterized genes. We performed cross-validation analysis as described in Section 3.1, and used held-out known proteins to assess the precision at which uncharacterized genes were predicted in networks across a range of biological processes. Figure 7 plots a range of precision measures (relative to random predictions) versus the number of uncharacterized genes recovered at that precision or higher. The context-sensitive approach tends to predict the involvement of more uncharacterized genes at higher precision than the global approach. For instance, at 10 times the precision expected by chance, the global scheme is able to predict networks for 118 previously uncharacterized proteins while the context-sensitive approach makes predictions for 214 uncharacterized proteins (81% improvement).

a

ssigned the corresponding precision. Uncharacterized genes were assumed to be genes annotated to the "biological process unknown" GO term (GO:0000004) as of 5/1/2006 [28]. Figure 6.7 illustrates the results of this analysis for the two methods by plotting the measured precision (relative to random) versus the number of uncharacterized proteins assigned with *at least* that precision.

The context-sensitive network prediction approach is generally able to make more network predictions at higher confidence. For instance, at 10 times the precision expected by chance, the global scheme is able to predict networks for 118 previously uncharacterized proteins while the context-sensitive approach makes predictions for 214 uncharacterized proteins (81% improvement). Interestingly, the difference between the two approaches is smaller for very high-

precision predictions (e.g. > 20 fold over random), suggesting there a limited number of uncharacterized proteins whose participation in certain networks is relatively easy to detect and varies little between the two methods. As we relax the precision criteria, however, the context-sensitive approach shows a clear and consistent improvement in precisely predicting uncharacterized genes in networks recovered from known sets of related proteins.

6.5 Biological Validation: Predicting Novel Mitochondria-related Genes[6]

We further validated our context-sensitive prediction approach by experimentally testing several predictions related to mitochondrial function. Specifically, we trained a Bayesian network based on the mitochondrion organization and biogenesis context (GO:0007005) and tested 30 novel predictions using two different assays for mitochondrial function. These 30 proteins included 17 completely uncharacterized proteins as well as 13 characterized proteins that had no previously reported association with mitochondria. We summarize the results of these experiments in the following sections.

6.5.1 Summary of Experimental Findings

We tested predictions of novel mitochondria-related proteins with two different assays. First, we checked for severe respiration defects in deletion mutants of all predicted genes, and secondly, we checked for altered frequency of petite colony formation, which indicates an absence of functional mitochondria (see the Experimental Methods section below for more details). In addition to the 30 novel predictions, 47 positive controls (genes known to be involved in proper mitochondrial functioning) and 48 genes chosen at random were tested with the same protocols to establish a baseline discovery rate for mitochondrial proteins. We confirmed mitochondrial defects (either respiratory deficiency or increased petite frequency) for 15 of the 30 novel predictions (50%) (Figure 6.8). 9 of these 15 novel phenotypes were for previously

[6] All experimental work presented in the section was completed by David Hess and Amy Caudy.

uncharacterized genes and 6 were previously characterized genes with no reported mitochondria association. We observed mitochondria defects for 34 of the 51 positive controls (67%) and for only 7 of the 48 genes selected at random (15%) (Figure 6.8). Thus, using our prediction approach we have increased the discovery rate of genes with mitochondria-related phenotypes from 15% to 50%, and have discovered a novel mitochondrial role for 15 new genes.

A striking characteristic of the distribution of phenotypes for our novel predictions is that they are not only highly enriched for real mitochondria phenotypes, but that they result largely in *subtle* phenotypes. Our positive controls, which were drawn from the set of known mitochondria proteins, exhibit 55% respiratory deficiency, a severe phenotype whose detection is relatively straightforward. We only confirm 10% respiratory deficiency in our novel predictions. However, 40% of our predictions have significantly increased petite colony frequency compared to only 12% with the same phenotype among the positive controls. This suggests an interesting characteristic of the yeast biology that is remains undiscovered: many of these uncharacterized genes result in mild phenotypes when knocked out and thus will not likely be readily detected by

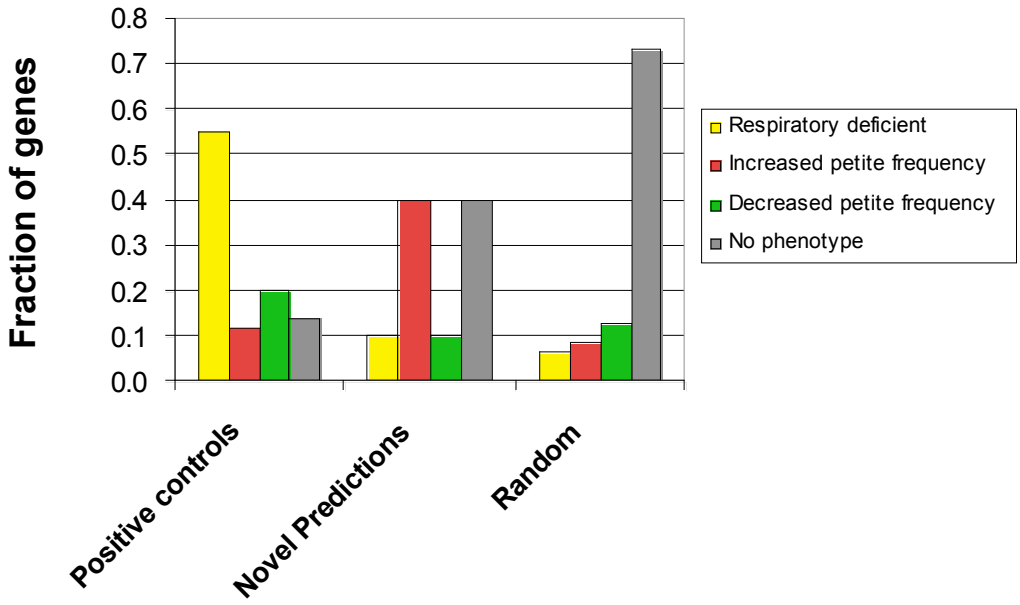


Figure 6.8. Summary of confirmed phenotypes for novel predictions and controls.
We tested our novel mitochondria predictions using a petite frequency assay, which indicates a lack of functional mitochondria. Here, we plot the percentage of phenotypes resulting from either the positive control group, the novel predictions, and randomly selected genes. The set of novel predictions is highly enriched for mutants exhibiting real phenotypes compared to genes selected at random. See section 6.5.2 for a detailed description of the experimental assay.

low-sensitivity whole-genome screens. Computational predictions, such as those made here, could play a key role in directing us towards putative functions for these proteins.

Interestingly, a few of our confirmed predictions showed physical interactions with actin-related proteins, suggesting their possible involvement in mitochondrial movement within the cell. The mechanism of mitochondrial movement in yeast cells is largely uncharacterized [7], so we pursued further experiments to elucidate the role of these proteins. Using a version of GFP (green fluorescent protein) tagged with a mitochondrial localization signal, we can clearly visualize the mitochondria in live cells (Figure 6.9A). When yeast cells prepare to divide, the mitochondria are localized to the bud neck. Just prior to cytokinesis, a portion of the mitochondria move into the daughter cell (anterograde movement) and a portion of the mitochondria move back into the mother cell (retrograde movement) [7]. This coordinated relocalization of the mitochondria is necessary to ensure that both the mother and daughter cells receive functioning mitochondria.

Two mutants in genes predicted by our approach demonstrate significant mitochondrial movement defects similar to those exhibited by our positive control, a *puf3* deletion (Figure 6.9B) [11]. Additionally, when the frequency of cells exhibiting anterograde or retrograde movement is calculated for these mutants, one mutant (*yir003wΔ)* has a strong defect in retrograde movement (Fig 6.9C). These data demonstrate a clear role for Yir003w, a completely uncharacterized protein, in mitochondrial localization. Furthermore, this presents a compelling case for the utility of our context-sensitive prediction framework— we have correctly predicted novel mitochondrial functions for 15 proteins, including what appears to be a key player in mitochondrial inheritance.

6.5.2 Experimental Methods

Respiration and petite frequency assay

First, several replicates of each deletion mutant are grown for 48 hours using glycerol as a carbon source. Strains severely deficient in their ability to maintain functional mitochondrial are unable to grow on glycerol and are classified as respiration deficient. Strains able to grow on glycerol are

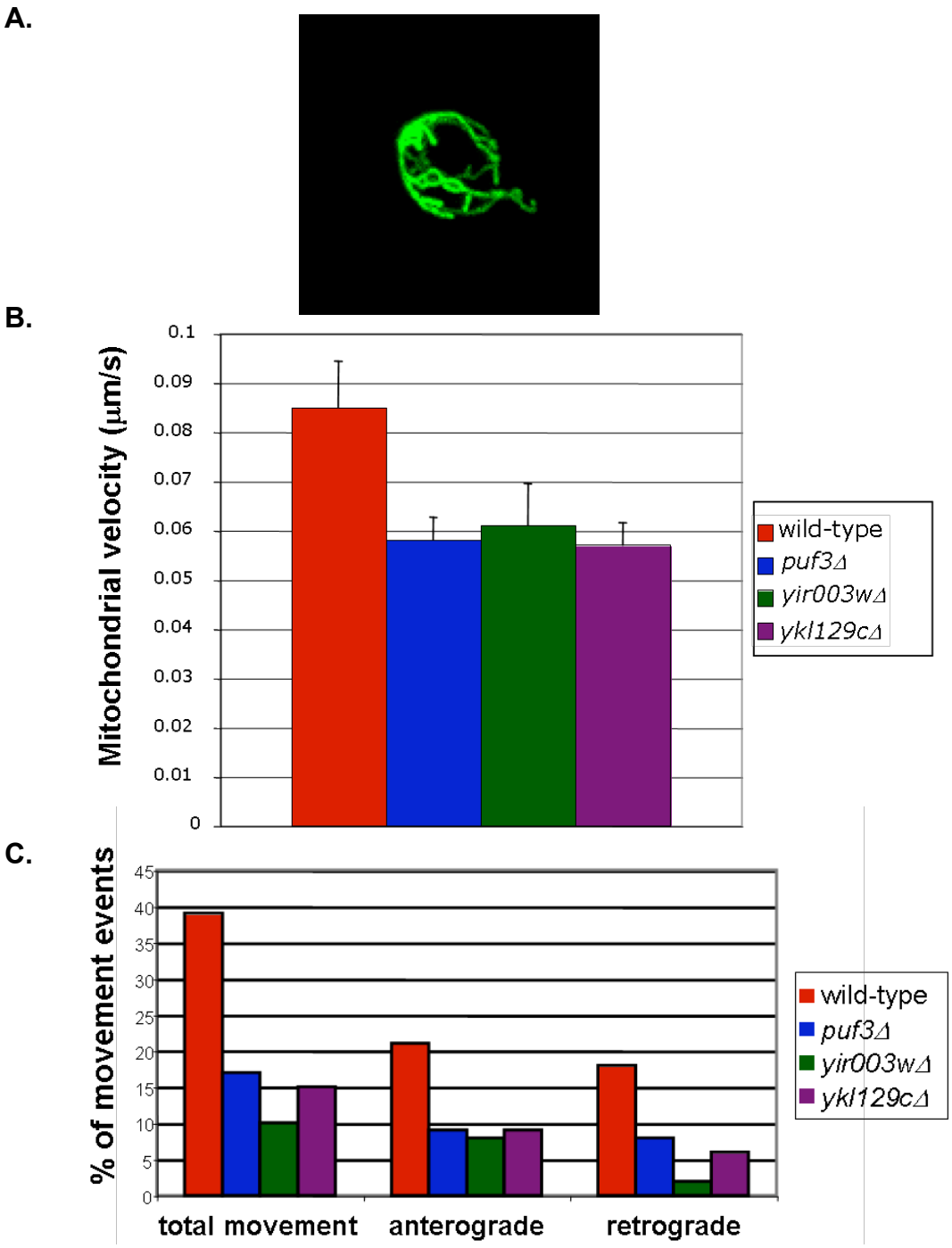


Figure 6.9. Experimental results for mitochondrial movement assay on two novel mitochondria-related proteins. (A) Example still frame of a yeast strain expressing the mitochondrial-localized GFP used in the mitochondrial movement assay. (B) Average mitochondrial velocity (μm/s) measured for the indicated strains. Measurements based on tracking >50 independent cells with motile mitochondrial over 5 seconds. (C) Direction of mitochondrial movement. For each strain >200 cells were tracked over a period of 3 seconds. The first set of bars displays the percent of cells with moving mitochondria over that time period. When movement was observed, it was characterized as anterograde (towards the daughter cell) or retrograde (towards the mother cell). The percentage of cells with anterograde and retrograde motion are displayed by the second and third set of bars, respectively.

diluted and plated for single colonies on rich media, which releases the requirement for functional mitochondria. Thus, as the colony forms, cells without functional mitochondria are generated. When the colony is fully formed, it is a mixture of cells with functional mitochondria and cells without functional mitochondria. This ratio is measured by resuspending the colony and plating a dilution such that 100-300 colonies are formed on a plate. By overlaying with soft agar containing tetrazolium, colonies with functional mitochondria are stained red while cells without functional mitochondria remain white. The ratio of white colonies to total colonies gives the petite frequency (a petite cell is a cell without functional mitochondria). Eight independent petite frequencies were measured for each strain tested. The distribution of these frequencies is compared to frequency of petite generation in wild-type yeast using the Mann-Whitney U test. Strains with a p-value of less than .01 are classified as a confirmed petite phenotype.

Mitochondrial movement assay

We track mitochondrial movement in the cell using a version of GFP (green fluorescent protein) tagged with a mitochondrial localization signal. Images are filmed over 2 minute time courses with 1 second resolution, which allows us to measure the rate and direction of mitochondrial movement. Movements are recorded by hand using the ImageJ software.

6.6 Discussion and Conclusions

In summary, incorporating contextual information in the data integration and prediction process can significantly improve prediction quality and provide important information about relevance of individual datasets in different contexts. As noted above, there are a very limited number of cases where the context-sensitive approach results in a loss of performance. This is typically due to the size of the GO terms corresponding to these contexts, and for such cases, global (non-context-sensitive) integration should be used (see Supplementary information for a detailed analysis). Incorporating context into the Bayesian integration phase requires context-specific examples, which can be very few in number for smaller contexts (GO terms). Interestingly, this

suggests a trade-off between the number of examples and the specificity of examples, which hints at why contextual information for network prediction is important. Put simply, the more specific we can be about the learning task, the better performance we can expect. This only holds true, however, if we can maintain a statistically representative example set, which requires a minimum number of examples. In general, this problem seems to affect a small minority of contexts evaluated here, and can be avoided by defining contexts more broadly.

We should emphasize that although we have implemented our approach using a Bayesian integration scheme and a particular search algorithm, the overall message of using contextual information is general and could be used to improve a variety of approaches to network prediction. We expect this concept to be particularly true as we begin to develop methods for integration and prediction in higher organisms, where there is not only variation in dataset relevance across biological process, but also across other aspects such as tissues or stages of development. An important consideration, however, is that to take advantage of this information, methods must be formulated in such a way that cross-context variation can actually by incorporated into the process. For instance, in our discussion here, we have assumed a query-based scheme, which inherently provides a straightforward approach to inferring the context of the prediction. Methods like this that allow expert direction are particularly well-suited to leveraging contextual information to improve prediction.

In conclusion, we have demonstrated evidence for context-dependent dataset reliability and illustrated a Bayesian integration and network recovery approach that makes use of this variation. Our approach achieves significant improvement in terms of both precision and sensitivity over a broad range of biological processes, and we have shown that it improves the estimated precision on predicting networks for previously uncharacterized genes. We further confirm several of these novel predictions by experimentally validating their role in mitochondrion organization and biogenesis. Biological context is an important consideration for any network prediction approach, and can be an effective means for managing data heterogeneity, particularly as we move toward developing computational methods for understanding networks in more complex organisms.

6.7 Supplemental Data Files

File name: GO_<GO ID>_ps_network_posteriors.txt
File format: Tab-delimited
Title: Dataset relevance measures for all 101 contexts
Description:
These files contain dataset relevance measure for all contexts and datasets. For each context, we report the posterior probability of functional relationship given evidence from that dataset in the corresponding context.

References

1. Alfarano, C., C. E. Andrade, et al. (2005). "The Biomolecular Interaction Network Database and related tools 2005 update." Nucleic Acids Res **33 Database Issue**: D418-24.
2. Ashburner, M., C. A. Ball, et al. (2000). "Gene ontology: tool for the unification of biology. The Gene Ontology Consortium." Nat Genet **25**(1): 25-9.
3. Asthana, S., O. D. King, et al. (2004). "Predicting protein complex membership using probabilistic network reliability." Genome Res **14**(6): 1170-5.
4. Bader, J. S. (2003). "Greedily building protein networks with confidence." Bioinformatics **19**(15): 1869-74.
5. Bader, J. S., A. Chaudhuri, et al. (2004). "Gaining confidence in high-throughput protein interaction networks." Nat Biotechnol **22**(1): 78-85.
6. Barutcuoglu, Z., R. E. Schapire, et al. (2006). "Hierarchical multi-label prediction of gene function." Bioinformatics **22**(7): 830-6.
7. Boldogh, I. R. and L. A. Pon (2007). "Mitochondria on the move." Trends Cell Biol.
8. Can, T., O. Camoglu, et al. (2005). Analysis of protein-protein interaction networks using random walks. Conference on Knowledge Discovery in Data, Chicago, IL.
9. Deng, M., F. Sun, et al. (2003). "Assessment of the reliability of protein-protein interactions and protein function prediction." Pac Symp Biocomput: 140-51.
10. Friedman, N., D. Geiger, et al. (1997). "Bayesian network classifiers." Machine Learning **29**(2-3): 131-163.

11. Garcia-Rodriguez, L. J., A. C. Gay, et al. (2007). "Puf3p, a Pumilio family RNA binding protein, localizes to mitochondria and regulates mitochondrial biogenesis and motility in budding yeast." J Cell Biol **176**(2): 197-207.

12. Harbison, C. T., D. B. Gordon, et al. (2004). "Transcriptional regulatory code of a eukaryotic genome." Nature **431**(7004): 99-104.

13. Huh, W. K., J. V. Falvo, et al. (2003). "Global analysis of protein localization in budding yeast." Nature **425**(6959): 686-91.

14. Jaimovich, A., G. Elidan, et al. (2005). "Towards an integrated protein-protein interaction network." Research in Computational Molecular Biology, Proceedings **3500**: 14-30.

15. Jansen, R., H. Yu, et al. (2003). "A Bayesian networks approach for predicting protein-protein interactions from genomic data." Science **302**(5644): 449-53.

16. Lanckriet, G. R., M. Deng, et al. (2004). "Kernel-based data fusion and its application to protein function prediction in yeast." Pac Symp Biocomput: 300-11.

17. Lee, I., S. V. Date, et al. (2004). "A probabilistic functional network of yeast genes." Science **306**(5701): 1555-8.

18. Letovsky, S. and S. Kasif (2003). "Predicting protein function from protein/protein interaction data: a probabilistic approach." Bioinformatics **19 Suppl 1**: i197-204.

19. Myers, C., D. Barrett, et al. (2006). "Finding function: evaluation methods for functional genomic data." BMC Genomics **7**(1): 187.

20. Myers, C. L., D. Robson, et al. (2005). "Discovery of biological networks from diverse functional genomic data." Genome Biol **6**(13): R114.

21. Myers, C. L. and O. G. Troyanskaya (2007). "Context-sensitive data integration and prediction of biological networks." Bioinformatics **23**(17): 2322-30.

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

23. Qi, Y., J. Klein-Seetharaman, et al. (2005). "Random forest similarity for protein-protein interaction prediction from multiple sources." Pac Symp Biocomput: 531-42.

24. Sprinzak, E., S. Sattath, et al. (2003). "How reliable are experimental protein-protein interaction data?" J Mol Biol **327**(5): 919-23.

25. Stark, C., B. J. Breitkreutz, et al. (2006). "BioGRID: a general repository for interaction datasets." Nucleic Acids Res **34**(Database issue): D535-9.

26. Troyanskaya, O. G., K. Dolinski, et al. (2003). "A Bayesian framework for combining heterogeneous data sources for gene function prediction (in Saccharomyces cerevisiae)." Proc Natl Acad Sci U S A **100**(14): 8348-53.

27. von Mering, C., M. Huynen, et al. (2003). "STRING: a database of predicted functional associations between proteins." Nucleic Acids Res **31**(1): 258-61.

28. Website. "Saccharomyces Genome Database." Retrieved 5/1/06, 2006, from ftp://ftp.yeastgenome.org/yeast/.

29. Zhu, J. and M. Q. Zhang (1999). "SCPD: a promoter database of the yeast Saccharomyces cerevisiae." Bioinformatics **15**(7-8): 607-11.

Chapter 7

Deriving Quantitative Epistasis Measures from Yeast Mutant Colony Growth

7.1 Chapter Overview

The earlier chapters of this dissertation have focused mainly on general methods for characterizing and integrating genomic data to understand gene function. We transition now to analysis of a specific type of genomic data, genetic interactions. Our motivation for a detailed study of genetic interactions is two-fold. First, genetic interaction data are among the most informative high-throughput sources of functional information. Figure 7.1 illustrates this with a precision-recall analysis of several different genomic datasets, comparing their ability to predict functional associations between genes [13]. Only one other dataset in this group provides more precise information, the Gavin et *al.* affinity precipitation data, and genetic interaction profiles are able to provide substantially more recall if the precision threshold is relaxed only slightly.

The second motivation for our focus on genetic interactions, is that recent methods have been developed that can rapidly generate enormous amounts of data. Specifically, one technique, Synthetic Genetic Array (SGA) analysis, uses robots to rapidly construct yeast deletion mutants and assay for growth defects [19]. Given how valuable genetic interactions in understanding gene function, ideally we would apply this approach to construct all possible double deletion mutants (about 18 million total), but even with this technology, that is estimated to take between five and ten years [19]. However, if we intelligently pick which double mutants to screen, we can potentially find the most interesting interactions much more quickly. This presents a perfect opportunity to apply the genomic integration technology we have described throughout this dissertation. Over the next two chapters, we demonstrate how such methods can

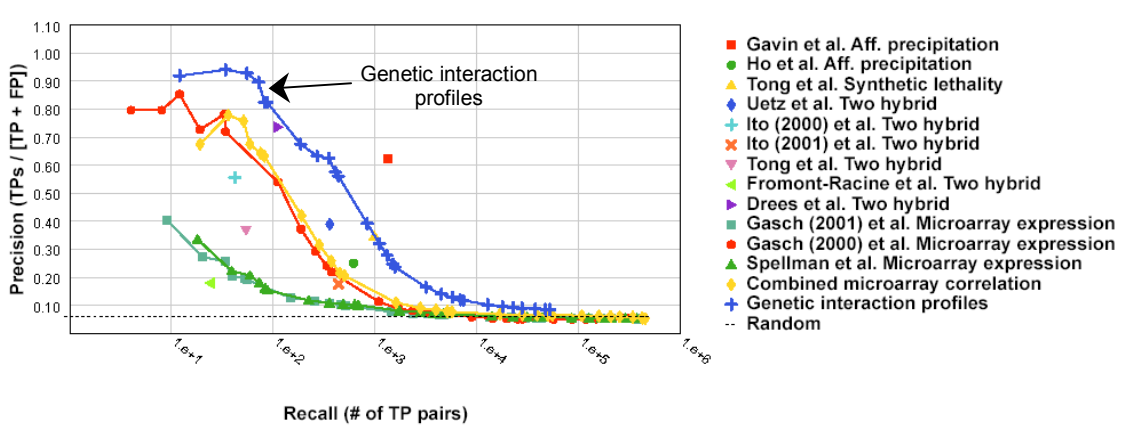


Figure 7.1. Evaluation of genetic interaction profiles at predicting functionally related pairs of genes. We computed correlation coefficients between genetic interactions profiles and measured their ability to prediction functional associations between genes [13]. We find that genetic interactions are among the most informative genomic data.

be used to dramatically improve the efficiency with which we apply high-throughput technology and give several examples of biological insight we gain in the process.

In this chapter, we give a detailed introduction to genetic interactions, or epistasis, and how they relate to a gene's role in the cell. We also describe SGA technology and our contribution to the SGA project, which is how to accurately derive quantitative epistasis measures from double mutant colony size data. Given this background and a precise measure of epistasis, we move on in Chapter 8 to discuss our iterative computational-experimental framework for efficiently mapping the global yeast genetic interaction network. The work presented in this chapter includes contributions from Michael Costanzo, Anastasia Baryshnikova, David Hess, and Olga Troyanskaya. Michael and Anastasia helped me in understanding the SGA technology and in developing the SGA epistasis score. David Hess provided several insights in measuring epistasis from colony data, and Olga supervised the project.

7.2 Background

Most phenotypic properties of an organism result from the collaborative interactions of genes and their products. Genetic interactions are broadly defined as pairs or groups of genes whose simultaneous perturbation results in a phenotypes different from what is expected given their individual phenotypes [3]. Genetic interactions are of particular interest because this general

phenomenon is believed to be the root of many common diseases, where multiple gene variants interact to alter the normal functioning of cellular processes [11]. This is in contrast to relatively rare Mendelian disorders which result from a mutation in a single gene. Beyond their relevance to our understanding human disease, genetic interactions reflect fundamental properties of the underlying genetic network. For instance, in yeast, it has recently been confirmed that only 20% of all genes are essential for cell viability [20,7]. The other 80% can be completely disabled in haploid cells growing under normal laboratory conditions with no severe fitness consequences, suggesting the cell has a great deal of built-in redundancy. Studying genetic interactions can reveal the underlying network that leads to this robust behavior.

Because of their value in dissecting the structure of the genetic network and their relevance to understanding human disease, geneticists have long been using combinatorial perturbations of genes for characterizing biological systems. Recently, genetic interactions have also been the focus of efforts to develop high-throughput screening technology in yeast [19,14] and higher eukaryotes such as worm [10]. One of these approaches designed for high-throughput screening in yeast (Synthetic Genetic Arrays) is the focus of this chapter.

7.2.1 A Quantitative Definition of Genetic Interaction

Statistical geneticists have also been studying the phenomena of genetic interactions for years, albeit from a different perspective than the classical geneticist. Global trends of genetic interaction have important implications on broad questions of interest to the statistical genetics community such as the evolutionary basis for recombination and sexual reproduction [9]. Statistical geneticists typically refer to genetic interactions generally as epistasis [15]. Fisher first used the word “epistacy” in 1918 to refer to deviation from the expected quantitative combination of independently functioning genes in the context of different alleles’ additive contributions to a quantitative phenotype [5]. Fisher’s landmark paper introducing this concept serves as the basis for much of modern quantitative genetic analysis. Later work motivated by Fisher’s original study focused on modeling the effect of combinations of genetic loci on an organism’s fitness [12]. Moran suggested a multiplicative fitness model, whereby the fitness of a combination of two non-

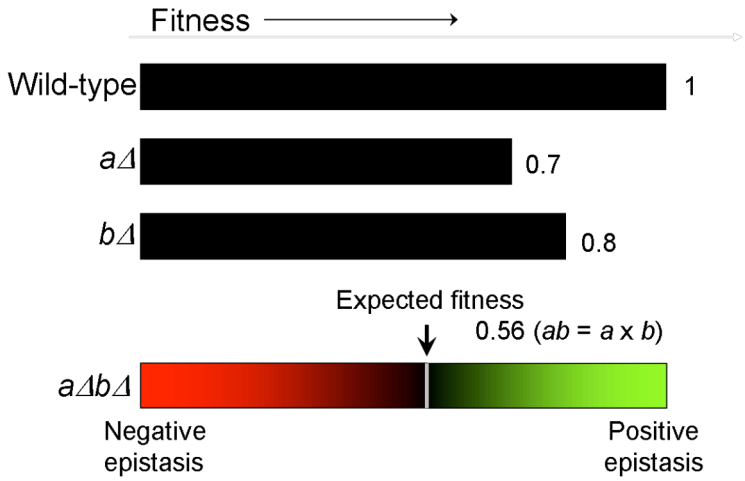


Figure 7.2. Illustration of how epistasis relates to fitness. Epistasis is generally defined as an unexpected phenotype arising from combinations of mutations. In the context of fitness, non-epistatic combinations of mutations are expected to combine multplicatively [12,21]. In the example illustrated above, a single mutant aΔ with a relative fitness of 0.7 and a single mutant bΔ with a relative fitness of 0.8 are expected to form a double mutant with a fitness of 0.56. Double mutants that are less fit are referred to as negatively epistatic and double mutants exhibiting greater than expected fitness are referred to as positively epistatic.

interacting mutations is modeled as the product of the fitnesses of the independent mutations [12]. Pairs of mutations deviating from this model were termed epistatic. Consider the example presented in Figure 7.2, which illustrates the construction of a double mutant in the two genes A and B. If a deletion of A by itself results in a fitness of 0.7 relative to wild-type and an independent deletion of B results in a fitness of 0.8, the expected fitness of the combination of these two deletions based on the classical model would be 0.7 x .08 = 0.56. We term pairs of mutants, AB, that are less fit than the expected amount as “negatively epistatic” and pairs that are more fit than expected under this model “positively epistatic.” We should note that Moran also suggested an additive version of this model, which he describes as “less natural” [12] but which has also received some attention in the literature (see [16] for example).

Unfortunately, there has been significant confusion of the term epistasis among the various biological communities. Although Fisher introduced the term epistacy as described above, Bateson actually originally introduced “epistasis” earlier in 1909 to explain genetic effects that alter single Mendelian gene effects [3]. Studying these types of interactions is a classical method used by geneticists to understand the relative ordering of different genetic components in pathways, and have been used in characterizing many core processes in yeast (for example [8]).

Bateson's version of epistasis is the established meaning of the word among the traditional genetics community, which is often a source of confusion between the traditional and statistical genetics community. Because our goal is to develop a framework for a quantitative measure of genetic interaction, we adopt the more statistical definition but note that Bateson's epistatic interactions are a subset of the interactions we are able to capture with our approach.

7.2.2 Synthetic Genetic Array Analysis

An extreme example of negative epistasis is when the combination of two viable single mutants results in a dead double mutant, which is often referred to as synthetic lethality [3]. Identifying synthetic lethal pairs of interactions has been the focus of several recent high-throughput technologies in yeast [19,14]. Synthetic Genetic Array (SGA) analysis is one such technology for detecting genetic interactions in high throughput (Figure 7.3). The basis of the SGA approach is robotic construction of yeast double deletion mutants from a library of single mutant strains. Specifically, a mutation in a gene of interest is crossed to the full set of viable gene deletion mutants, and a series of robotic arraying procedures allows selected growth of double-mutant meiotic progeny, which can then be scored for specific phenotypes. In particular, both qualitative and quantitative measures of mutant colony sizes can be obtained by acquiring and processing digital images (see Figure 7.4 for example or Appendix E for a more detailed discussion of how mutants are physically arranged on SGA plates). Using SGA, the first large-scale genetic interaction map was obtained in 2004 by crossing 132 query genes were crossed to ~4700 viable yeast gene deletion mutants, resulting in ~4000 synthetic lethal or sick gene combinations [19]. The analysis showed that synthetic lethal interactions are rare events, occurring among ~0.5% of gene pairs tested. We describe here how to extend this technology to quantitatively measure not just synthetic lethality, but general epistasis, including both positive and negative interactions. The computational machinery for this is described in the following section.

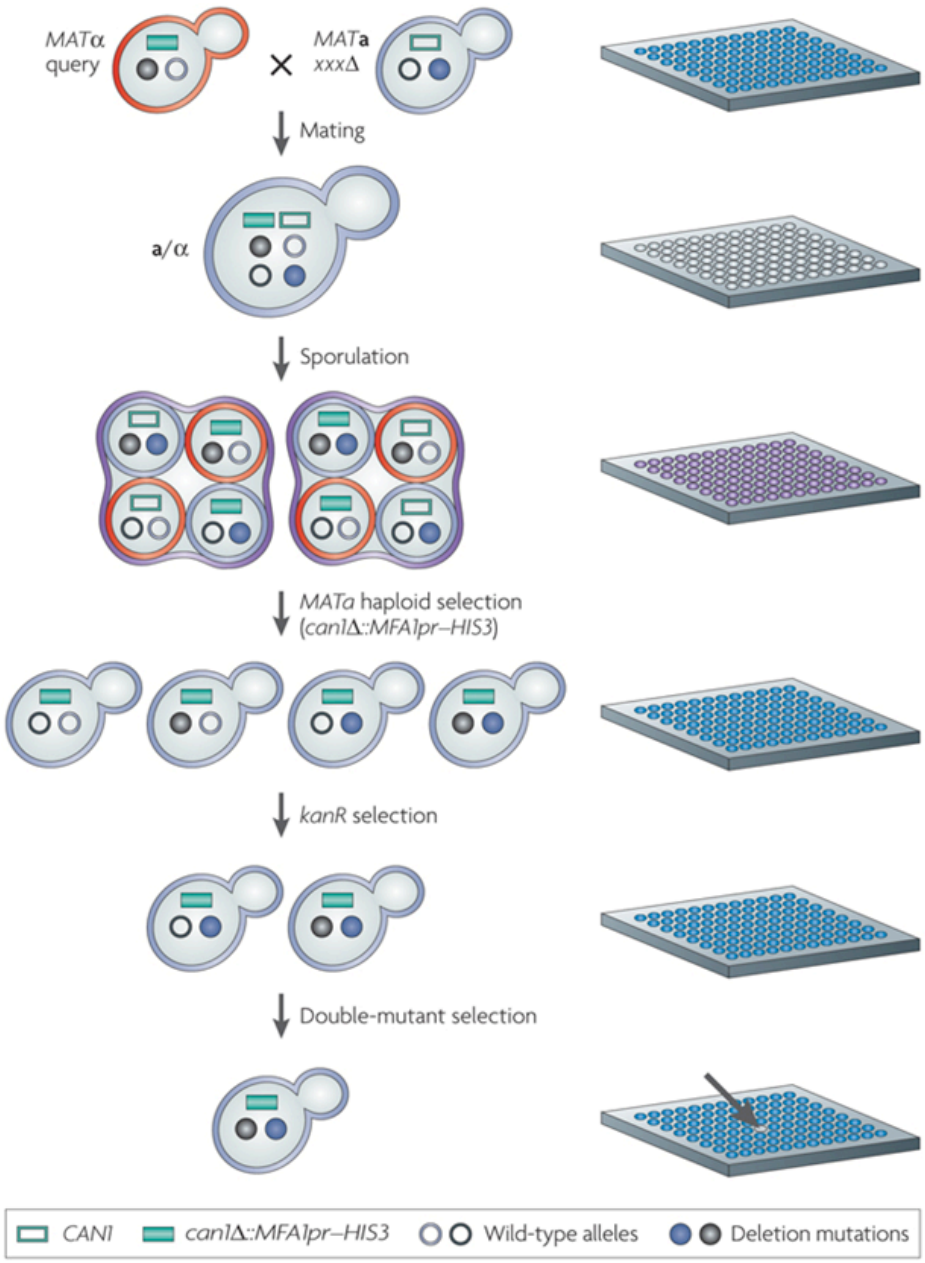


Figure 7.3. Overview of Synthetic Genetic Array (SGA) technology. SGA technology is a high-throughput assaying for detecting genetic interactions in yeast. A MATα query single mutant carrying a nourseothricin-resistance marker (natMX) is crossed into an ordered array of MATa deletion mutants. The resultant heterozygous diploids are plated on reduced carbon and nitrogen medium to induce sporulation and the formation of haploid meiotic spore progeny. Spores are then transferred to a medium lacking histidine which selects for MATa meiotic progeny. The MATa progeny are transferred to a medium containing kanamycin, which selects for mutants in the array strain. The final selection medium contains both nourseothricin and kanamycin, which selects double mutants between the query mutants and the ordered array. Plates are then incubated and photographed to obtain colony size information [3].

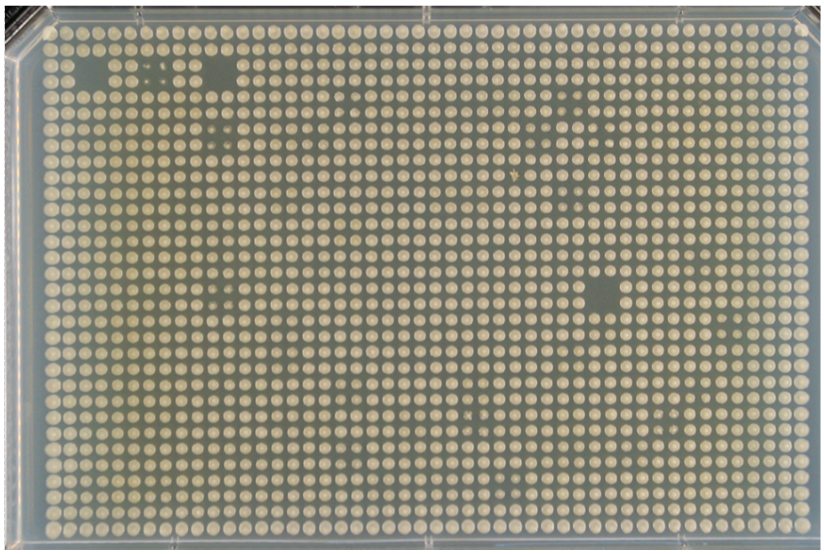


Figure 7.4. Plate of double mutants from a Synthetic Genetic Array (SGA) screen. SGA uses robotic technology to rapidly cross query single mutants into an ordered array of single mutants. This plate is the result of a single query crossed into a plate of array single mutants. Each double mutant appears with four replicates, and there are 1536 total colonies on each plate. See Appendix E for more detailed information about plate layout.

7.3 An Epistasis Model for Mutant Colony Growth

The main challenge in deriving accurate measures of epistasis from colony growth data is that many systematic effects must be accounted for before one can hope to arrive at accurate estimates for the biological quantities of interest. For instance, even visual inspection of the SGA plate in Figure 7.4 reveals there is a clear trend towards larger colonies at the extreme edges of the plate, which is due to the availability of extra nutrients. Also, the size of the colonies tends to vary systematically from plate to plate, based on the amount of time the colonies were allowed to grow before processing. Yet another source of variation is local competition for nutrients. For example, colonies situated next to a dead neighbor or a blank spot on the plate tend to grow larger because there are more available nutrients.

We propose a model that accounts for both these systematic effects and, simultaneously, the relevant biological effects, which are the single mutant and double mutant fitnesses and presence or absence of epistasis. Constructing a simple model from first principles to address this problem is challenging for at least two reasons. First, while there have been some efforts to

model fungal colony growth (see [6] for example), there is no well-established model of how colony size (area or diameter) maps directly to growth rate or fitness. Second, it is not exactly clear how each of the systematic effects we want to model affect the colony size. Thus, we propose two simple models, both reasonably straightforward. Specifically, we model the observed colony size as either a multiplicative (1) or additive (2) combination of all biological and systematic effects:

$$s_{ijklm} = \mu + r_{ik} + c_{jk} + p_k + f_l + f_m + \varepsilon_{lm} + comp(s_{ijklm}) + e \qquad (1)$$

$$s_{ijklm} = \mu\ r_{ik}\ c_{jk}\ p_k\ f_l\ f_m\ \varepsilon_{lm}\ comp(s_{ijklm})\ e \qquad (2)$$

where the model parameters are defined in Table 7.1.

Table 7.1. Definition of epistasis model parameters.

s_{ijklm} : SGA colony size (pixels)
μ : global mean
r_{ik} : row i on plate k
c_{jk} : column j on plate k
p_k : plate k
f_l : single mutant fitness defect (gene l)
ε_{lm} : genetic interaction between gene l and m
$comp(s_{ijklm})$: nutrient competition effect
e : unexplained error

Both formulations model the observed colony size as a function of single mutant fitnesses and include an interaction term, ε_{lm}, which explains differences from the expected combination of the single mutant effects (epistasis). The multiplicative model has a more theoretical foundation, but in practice, we find that both models fit the raw data reasonably well, and both provide approximately the same enrichment for published genetic interactions. In fact, in some cases, we observe a slight advantage for the additive model in terms of enrichment for known protein-protein interactions and functional relationships. See Appendix D for a brief discussion of the theoretical support for the multiplicative case. An appealing property of both formulations is that they can be computed efficiently on very large collections of double mutant colony

measurements. The additive case can be fit directly with standard linear regression and the multiplicative case can be fit with linear regression after a log-transformation on the raw colony size measurements.

In general, either model can be fit in its entirety given double mutant colony size data, but in practice, we found a stepwise fitting approach more effective. More specifically, we found that row and column effects are confounded with single mutant effects, and thus required further constraints on their fitting. Furthermore, modeling local nutrient competition depends non-trivially on the actual colony size and neighboring colony sizes and thus did not fit conveniently in the regression framework. We discuss the procedure used for fitting the model below, including a special discussion of these effects.

7.3.1 Normalizing Row and Column Effects

Plate-specific row and column effects are often confounded with single mutant fitness effects because only a limited number of unique strains are present in each row or column (16 and 24). However, we know that row and column effects are due to the geometric arrangement of colonies on the plate and the availability of nutrients. Thus, we expect that neighboring rows should exhibit similar effects, and consequently, trends across or down the plate should be relatively smooth. We can take advantage of this property to derive accurate estimates of how colony position affects colony size and remove this systematic trend from the data. Specifically, we apply Lowess smoothing to estimate the colony size-row and colony size-column trends, using a linear fit for each window and a window size spanning 6 rows or columns[7].

Figure 7.5 illustrates the results of this approach for real colony data and the additive model. Each estimate is represented by a box and whisker plot indicating uncertainty estimated through bootstrapping. Indeed, we confirm the trend we saw in Figure 7.4 by visual inspection: colonies in the outer rows and columns tend to grow larger. Interestingly, we also identify other subtle yet statistically significant trends such as an overall W-shape and a slight increase in

[7] Since double mutants occur in groups of two across or down each row or column, smoothing over 6 rows or columns ensures that effects are estimated based on at least three unique sets of mutant strains.

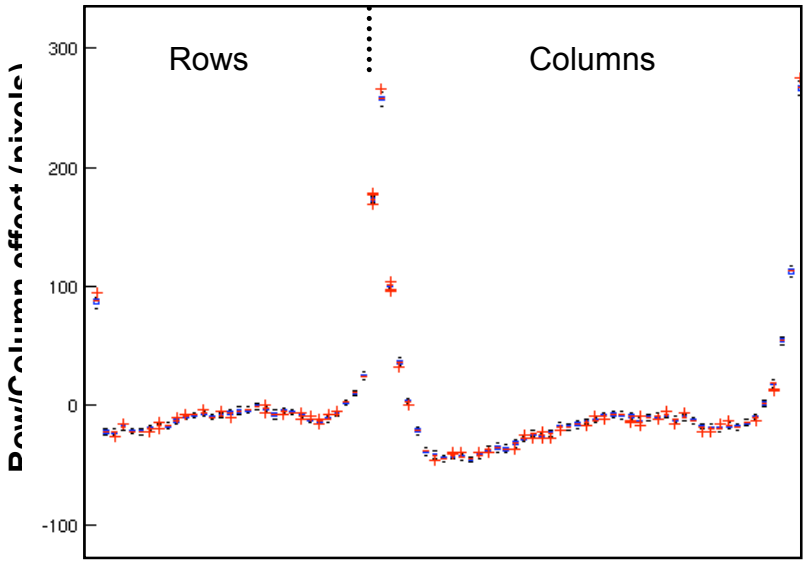


Figure 7.5. Row and column effects measured on an SGA plate. We used a lowess smoothing procedure to estimate row and column effects on colony size. The estimated effects are shown here in the order they appear down and across the plate. We detect severe U-shaped trends across both rows and columns, and also identify more subtle systematic increases in colony size moving from the top left to bottom right of the plate. These effects are relatively reproducible across plates.

c

olony size as one moves across and down the plate. The trends are highly consistent across plates, suggesting they are a real systematic artifact, likely due to the geometric properties of the plate or the media.

7.3.2 Correcting for Neighbor Colony Competition

Another systematic effect that is difficult to estimate through linear regression, but that must be corrected, is local competition for nutrients. This effect is largely due to the high density of colonies on the plate (1536 total per plate), and is most pronounced in cases where a healthy colony is positioned next to a sick colony or a dead spot. The severity of this effect is illustrated in Figure 7.6. To test whether large colonies are explained by small neighbors, we plotted the distribution of neighbor colony sizes for a range of different double mutant colony sizes[8]. The

[8] We consider the minimum of the three closest neighbors for this analysis. We only consider neighbors that are distinct double mutants so as to differentiate between positive correlation between mutants sharing the same deletions and negative correlation between big colonies and small neighbors. See Appendix D for a detailed description of plate layout.

neighbor distribution is plotted for three different ranges of double mutant colony sizes: 0-10

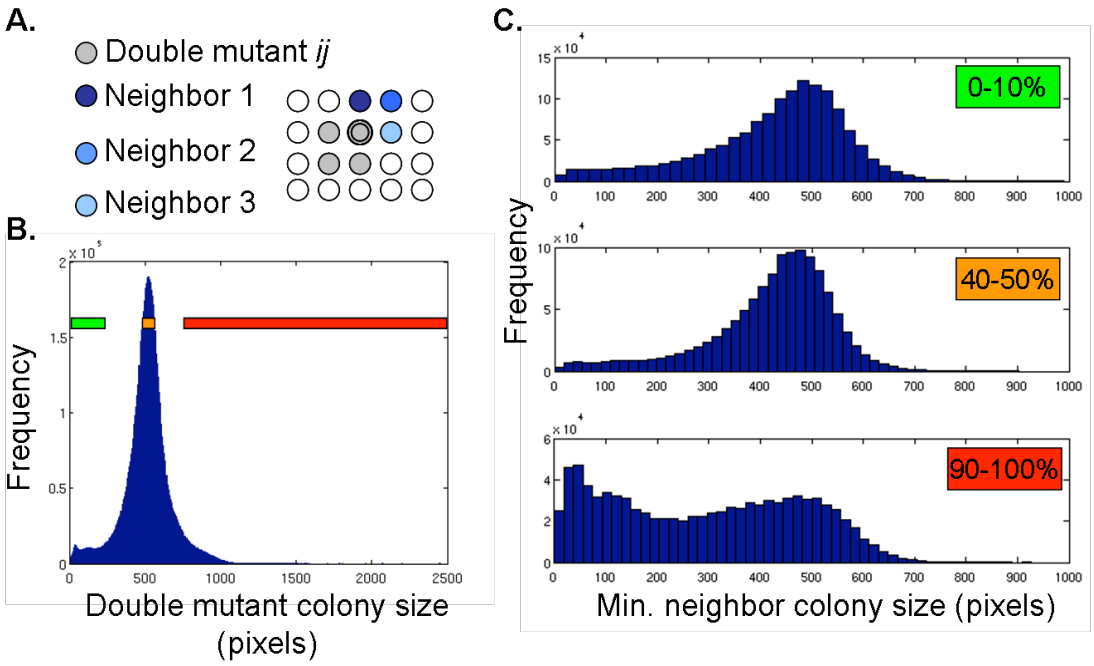


Figure 7.6. Illustration of the nutrient competition effect on colony size. (A) For each colony, we found the minimum of its three closest neighboring colonies not sharing the same double gene deletion. (B) We grouped colonies into ten deciles based on the overall double mutant colony size distribution, and plotted the distribution of minimum neighbor sizes for three of these deciles (0-10%, 40-50%, 90-100%). (C) We find that the largest 10% of colonies is highly enriched for very small neighbors, suggesting their size is not a biological effect, but a systematic effect due to the extra availability of mutants.

percentile, 40-50 percentile, and 90-100 percentile. Strikingly, the largest 10% of double mutants, have a dramatically disproportionate number of small neighbors, suggesting that the reason they are large is not that they are more fit, but that they have access to more nutrients.

To normalize this effect, we take a two-step approach. First, we bin all double mutants into 10 deciles based on their neighbor colony sizes. Normalization is then applied to remove the effect of competition both *within* and *between* these groups.

Within-group competition normalization

To normalize the competition effect within each decile, we plot double mutant colony size versus the minimum neighbor size and apply linear Lowess smoothing with a window size of 1000. Smoothed estimates are then subtracted for the additive model and divided for the multiplicative case.

Between-group competition normalization

The competition effect is normalized across decile groups by using quantile normalization, which is a technique that has been applied extensively in microarray normalization [2]. Essentially, quantile normalization takes the data one wants to normalize and a reference distribution, and forces the cumulative distribution (CDF) of the sample data to match the cumulative distribution of the reference data. We define a reference distribution based on double mutants with relatively healthy neighbors (60-80 percentile), and then quantile normalize each decile described above such that the CDF matches this reference. Since we do not expect colonies that are already sick to benefit from extra nutrients, we only apply this normalization to colonies that are larger than the mean colony size.

7.3.3 Fitting the Model: Implementation Details

The positional effects and competition effects are fit as just described, and the remaining effects are fit using linear regression. For the multiplicative case regression is applied on log-transformed colony sizes (log-transformed colony size is linear to modeled effects for the multiplicative case). Error estimates on fit parameters are obtained by 50 rounds of .632 bootstrapping, and the final estimates and standard deviation are derived from the mean and variance across bootstrap samples. All normalization and model fitting is implemented in MATLAB and linear regression on large datasets is done using the Tomlab optimization library.

7.4 Applying the Epistasis Model to Real Data

We applied both the additive and the multiplicative models described above to raw colony data from SGA screens. We performed several different validation experiments to characterize the performance of our model and validate that epistatic interactions estimated by the model appear to be biologically valid. For all of the analysis described, the additive and multiplicative versions of our model showed very few differences, so we only highlight examples from the additive model below.

7.4.1 Analysis of Variance in Colony Data

Our model for both systematic and biological effects explains greater than 90% of the variance observed in the SGA colony size data. This suggests that we have captured most major sources of variation. Interestingly, the systematic effects (e.g. plate, row and column) themselves explain 73% of the variance. The biological effects, the single mutant fitness and the epistatic interaction term, account for only 20% of the variance. Thus, accurate estimation of these quantities of interest depends critically on our accurate modeling of systematic effects.

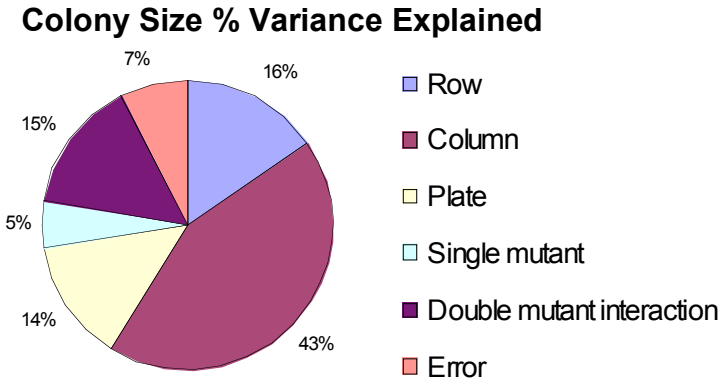


Figure 7.7. Analysis of variance (ANOVA) on colony size. We measured the contribution of each of the systematic and biological effects in our model to the observed variance in double mutant colony sizes. Overall, these effects explain 93% of the variance, and over 70% of this variance is due to row, column, and plate effects. The magnitude of these effects demonstrates the importance of accurate models for detecting the biological quantities of interest.

7.4.2 Evaluating Model Parameter Estimates

We further investigated the characteristics of the biological parameters estimated by our model, specifically the single mutant fitness and interaction terms. The interaction terms were used to compute the actual fitness for each double mutant, and all fitnesses are reported relative to a wild-type control. The distribution of single mutant fitnesses derived from our model and their estimated error based on bootstrapping are plotted in Figure 7.8. We observe a range of single

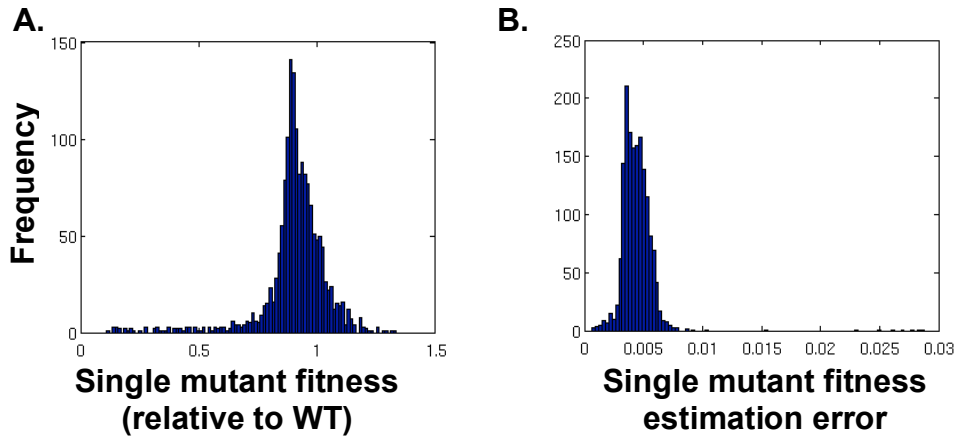


Figure 7.8. Distribution of relative single mutant fitness effect and estimation error. We fit our epistasis model on raw colony size data and measured the resulting single mutant fitness effects. Estimation error was obtained from the standard deviation across bootstrapped model fits. The approximate relative error for most single mutant effects is 0.5% because of the large number of colonies supporting this estimate.

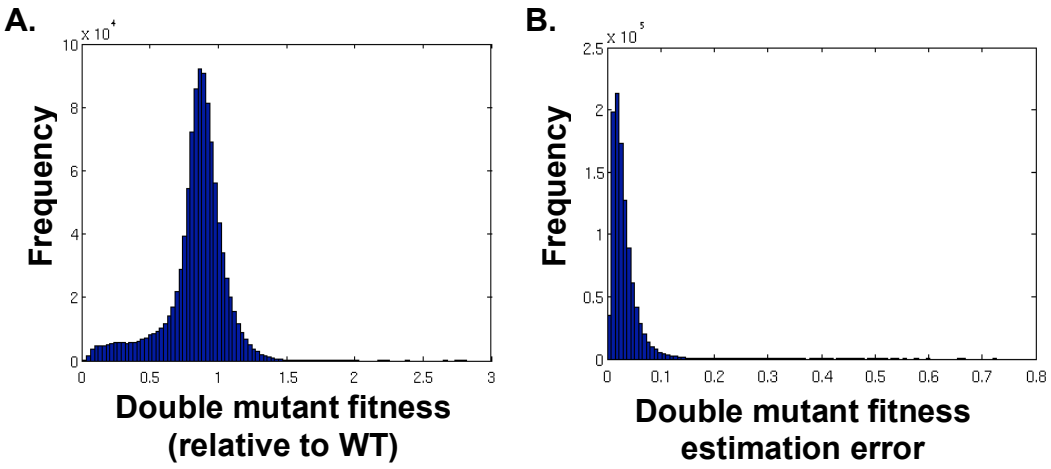


Figure 7.9. Distribution of normalized relative double mutant fitnesses and estimation error. We fit our epistasis model on raw colony size data and measured the resulting normalized double mutant fitnesses. Estimation error was obtained from the standard deviation across bootstrapped model fits. The approximate relative error for most single mutant effects is between 5-10%.

mutant fitnesses from 20% relative to wild-type up to mutants that appear to be 110-120% wild-type fitness. The majority of mutants are centered right at or slightly below wild-type fitness. The error on the single mutant fitness estimates is quite low (.5%) largely because we have a large sample of colony data for each mutant.

We observe a more diverse range of double mutant fitnesses, and also find that the error in our estimate increases significantly to around 5-10% (Figure 7.9). This is unsurprising since we have only between 4-8 colony replicates for each double mutant (see Appendix E for a

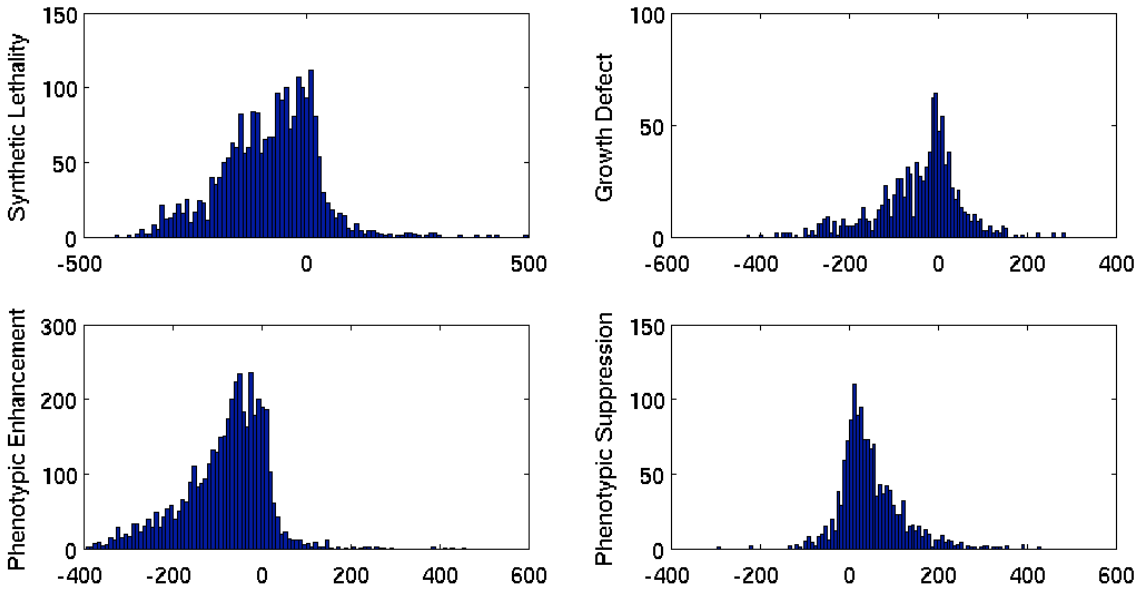


Figure 7.10. Comparison of epistasis estimates with published genetic interactions. We compared the epistasis effects estimated by our model to published genetic interactions obtained from bioGRID [17]. The epistasis estimates largely agree with the published data. For instance, we see a strong negative epistasis bias in the distribution of epistasis scores for known synthetic lethal pairs and a strong positive bias in the distribution of scores for pairs classified as phenotypic suppression. Pairs with reported growth defects are not as strongly bias, but still show a tendency towards negative interactions.

detailed description of plate layout). In general, the model yields precise estimates for both single and double mutant fitnesses.

Comparison with published genetic interactions

In addition to characterizing single mutant and double mutant fitness estimates, we also compare our predicted interactions to published knowledge of genetic interactions. We find that our epistasis estimates correlate well with known examples of epistasis. Specifically, we plotted the distribution of interaction term, ε_{lm}, estimates for several categories of genetic interactions curated by bioGRID [17]. We find that the direction and magnitude of our estimates correlate well with published interactions. For instance, we find that both interactions labeled "Synthetic lethality" and "Phenotypic enhancement" by bioGRID curators are shifted significantly to the left (85% < 0 for Synthetic lethality; 91% < 0 for Phenotypic enhancement), which indicates we estimate several significant negatively epistatic interactions for these pairs. Conversely, the interactions labeled as "Phenotypic suppression" by bioGRID curators are shifted significantly to

the right (84% > 0), suggesting that we are also detecting real examples of positive epistasis. Interestingly, for interactions labeled as "Growth defects", we do find a significant bias towards negative epistasis (72% < 0), but not nearly as strong as either Synthetic lethality of Phenotypic enhancement interactions.

Comparison of interacting pairs against known protein-protein interactions and functionally related genes

As discussed earlier, we know that genetic interactions reveal rich information about gene function and pathway organization. Thus, we further evaluated interactions predicted by our model against known protein-protein interactions [17] and functionally associated genes [13]. For a range of interaction scores, ε_{lm}, estimated by our model, we calculated the enrichment of either of these types of association among pairs with that score (Figure 7.11). Indeed, we confirm enrichment for both functional relationships and protein-protein interactions for pairs with both extreme negative and positive epistatic genetic interactions. For instance, for the most extreme negative interactions, we find 7-fold more physically interacting pairs than expected by chance and for extreme positive interactions, we find approximately 18-fold more than expected by chance. The fold enrichment for functionally associated pairs of genes is less, but we observe a similar trend (Figure 7.11B). Thus, the genetic interactions predicted by our model recapitulate known genetic interactions and, furthermore, are enriched for functionally related pairs of genes and protein-protein interactions, suggesting we are finding true epistatic interactions with high confidence.

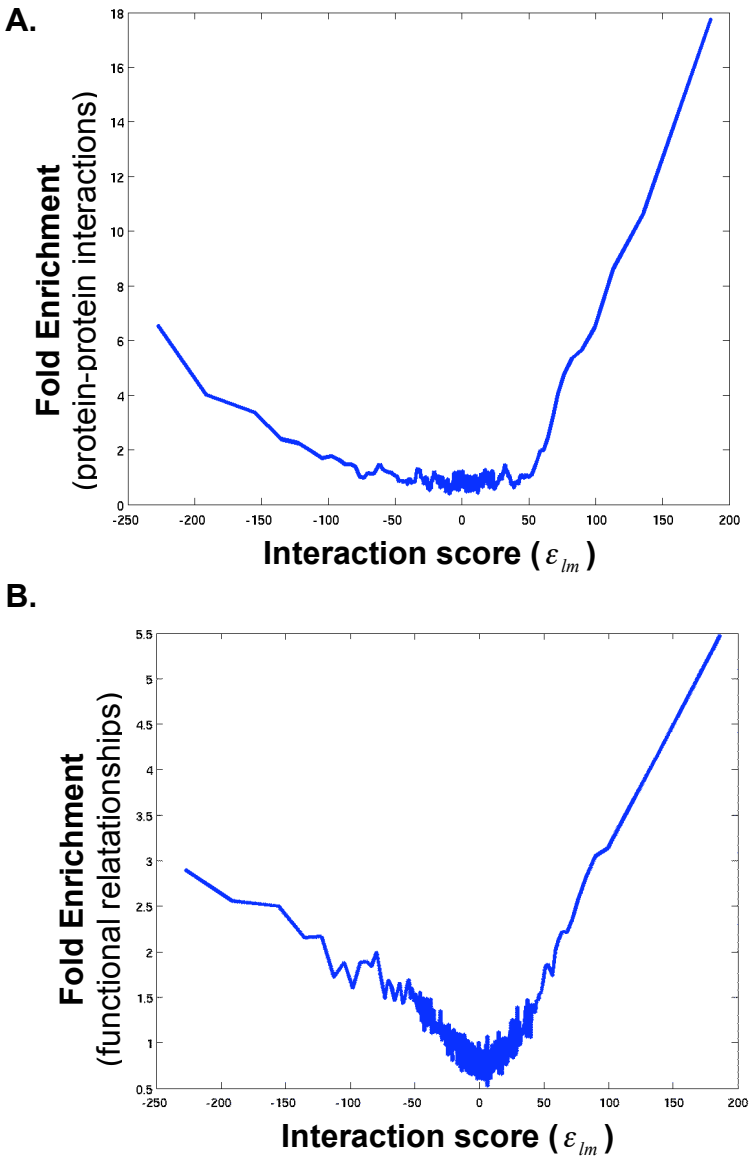


Figure 7.11. Enrichment of epistasis scores for known protein-protein interactions (A) and functionally related genes (B). After estimating epistasis effects from real SGA data, we binned the interaction scores and evaluated each bin for enrichment of known protein-protein interactions [17] or functionally associated genes [13].

7.5 Experimental Validation of Model Estimates: Comparison to Epistasis Measured in Liquid Growth Assay

Although the previous section provides substantial evidence that our model successfully predicts genetic interactions from double mutant colony data, we also confirmed this experimentally. To

do this, we compared interactions predicted by our model to epistasis estimates derived through growth-rate analysis of double mutants in liquid media [18]. St. Onge *et al.* constructed double yeast deletion mutants from pairs of several single mutants in genes involved in the response to DNA damage. In all, growth rates for approximately 300 double mutants (26 x 26) were experimentally measured. For the purpose of comparison, we obtained experimental colony size data for the same set of double mutants using the SGA approach and fit the model as described above.

Epistasis estimates derived from SGA colony size data correlate surprisingly well with epistasis measured through growth rate analysis. In fact, for the double mutants deemed significantly non-zero (either positive or negative) by the St. Onge et *al.* study, we observe a correlation coefficient of 0.9 between estimates from our model and their published epistasis values (Figure 7.12A). Furthermore, we find significant correlation (0.82) even among the more subtle effects reported in the St. Onge study, suggesting these may be real and can also detected through SGA double mutant colony analysis (see Figure 7.12B). Overall, the two approaches

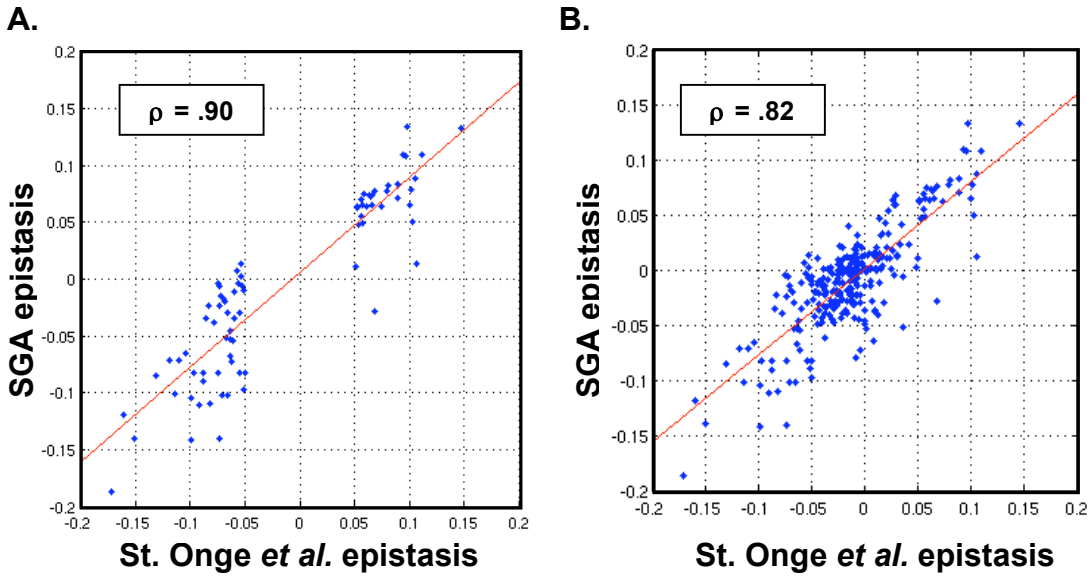


Figure 7.12. Comparison of SGA epistasis scores with epistasis measured in liquid growth media. We compared the epistasis estimates derived from our model based on SGA data to epistasis measured on the same mutants based on growth rate analysis [18]. We confirm a high degree of correlation (.90), particularly among the interactions deemed significant by the St. Onge et *al.* study). Interestingly, we find a slightly lower, but significant correlation (.82) between all pairs, suggesting that subtle effects are shared between the two studies and are likely real.

agree for a vast majority of the pairs in this study, although it appears epistasis measured based on SGA may in fact be statistically more sensitive. At a significance level (p-value) of 0.05, the St. Onge study identified 31 positive epistatic interactions, 28 of which were also detected by our approach. St. Onge et *al.* detected 54 total negative epistatic interactions, 39 of which we detected. However, we detected 15 significant positive interactions and 40 significant negative interactions not identified by their approach (Table 7.2). There were no cases where the two approaches disagreed (i.e. one called a significant positive interaction while the other a significant negative interaction). In summary, our epistasis model for SGA colony size data shows a high degree of correlation with epistasis measured in a completely different experimental assay.

St. Onge *et al.* (p-value < .05) \ SGA (p-value < .01)	Positive	Negative	Neutral
Positive	28	0	3
Negative	0	39	15
Neutral	15	40	134

Table 7.2. Comparison of SGA epistasis scores with epistasis measured in liquid growth media. We counted the overlap between the significantly non-zero pairs reported in [18] (p-value < .05) with those fit by our model (p-value < .01). This table reports overlap between the various categories.

7.6 Discussion and Conclusions

We have described basic background relevant to understanding and interpreting genetic interactions in a functional context as well as one high-throughput technology for detecting these interactions, Synthetic Genetic Array (SGA) analysis. We have also introduced our extension of this high throughput technology, a simple model for deriving precise, quantitative epistasis measures from raw double mutant colony data. We have demonstrated that this model yields both precise single and double mutant fitness estimates and that epistatic interactions it detects

correspond well to known interactions and often occur between functionally related and physically interacting genes.

This work has exciting potential for a number of reasons. First, no previous work on SGA has shown that precise, quantitative epistasis measures can be derived from these experiments. Earlier SGA analysis focused on the detection of synthetic lethal double mutants which were then confirmed by more traditional experiments [19]. Previous work has also derived quantitative indicators for SGA data [4] but did not explicitly model epistasis by estimating single mutant fitnesses. We show here that we can detect both negative and positive interactions precisely.

Previous work has suggested that highly quantitative genetic interaction measurements could be of great utility in reverse engineering network topology and pathway information [1]. The combination of fast high-throughput technology for constructing double mutants coupled with the quantitative framework presented here for precise detection of epistatic interactions will have a number of promising applications in elucidating gene function.

References

1. Avery, L. and S. Wasserman (1992). "Ordering gene function: the interpretation of epistasis in regulatory hierarchies." Trends Genet **8**(9): 312-6.
2. Bolstad, B. M., R. A. Irizarry, et al. (2003). "A comparison of normalization methods for high density oligonucleotide array data based on variance and bias." Bioinformatics **19**(2): 185-93.
3. Boone, C., H. Bussey, et al. (2007). "Exploring genetic interactions and networks with yeast." Nat Rev Genet **8**(6): 437-49.
4. Collins, S. R., M. Schuldiner, et al. (2006). "A strategy for extracting and analyzing large-scale quantitative epistatic interaction data." Genome Biol **7**(7): R63.
5. Cordell, H. J. (2002). "Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans." Hum Mol Genet **11**(20): 2463-8.

6. Farina, J. I., G. R. Tonetti, et al. (1997). "A mathematical model applied to the fungal colony growth of Sclerotium rolfsii." Biotechnology Techniques **11**(4): 217-219.

7. Giaever, G., A. M. Chu, et al. (2002). "Functional profiling of the Saccharomyces cerevisiae genome." Nature **418**(6896): 387-91.

8. Hartwell, L. H., J. Culotti, et al. (1974). "Genetic control of the cell division cycle in yeast." Science **183**(120): 46-51.

9. Kondrashov, A. S. (1988). "Deleterious mutations and the evolution of sexual reproduction." Nature **336**(6198): 435-40.

10. Lehner, B., C. Crombie, et al. (2006). "Systematic mapping of genetic interactions in Caenorhabditis elegans identifies common modifiers of diverse signaling pathways." Nat Genet **38**(8): 896-903.

11. Moore, J. H. (2003). "The ubiquitous nature of epistasis in determining susceptibility to common human diseases." Hum Hered **56**(1-3): 73-82.

12. Moran, P. A. (1968). "On the theory of selection dependent on two loci." Ann Hum Genet **32**(2): 183-90.

13. Myers, C. L., D. R. Barrett, et al. (2006). "Finding function: evaluation methods for functional genomic data." BMC Genomics **7**: 187.

14. Ooi, S. L., X. Pan, et al. (2006). "Global synthetic-lethality analysis and yeast functional profiling." Trends Genet **22**(1): 56-63.

15. Phillips, P. C. (1998). "The language of gene interaction." Genetics **149**(3): 1167-71.

16. Puniyani, A., U. Liberman, et al. (2004). "On the meaning of non-epistatic selection." Theor Popul Biol **66**(4): 317-21.

17. Reguly, T., A. Breitkreutz, et al. (2006). "Comprehensive curation and analysis of global interaction networks in Saccharomyces cerevisiae." J Biol **5**(4): 11.

18. St Onge, R. P., R. Mani, et al. (2007). "Systematic pathway analysis using high-resolution fitness profiling of combinatorial gene deletions." Nat Genet **39**(2): 199-206.

19. Tong, A. H., G. Lesage, et al. (2004). "Global mapping of the yeast genetic interaction network." Science **303**(5659): 808-13.

20. Winzeler, E. A., D. D. Shoemaker, et al. (1999). "Functional characterization of the S. cerevisiae genome by gene deletion and parallel analysis." Science **285**(5429): 901-6.

21. Wolf, J. B., E. D. Brodie, et al. (2000). Epistasis and the evolutionary process. New York, Oxford University Press.

Chapter 8

Closing the Loop: Directing Large-scale Genetic Interaction Screens with Context-sensitive Integration

8.1 Chapter Overview

As demonstrated by several examples throughout this dissertation, computational approaches to data integration and network inference can be readily used to generate specific and accurate biological hypotheses. This is a result of both the improving quality and quantity of data as well as advances in methods for analyzing genomic data such as those described here. Despite this success, however, there have been only limited efforts to validate hypotheses generated by computational approaches on a large scale. Often, when validation experiments are done, they are typically focused on justification of a methods' publication, not furthering community understanding of biology [9].

A promising but under-explored role for computational integration technology is in actually directing further, targeted experiments as illustrated in Figure 8.1. This chapter describes our design of such an approach for directing large-scale genetic interaction screens.

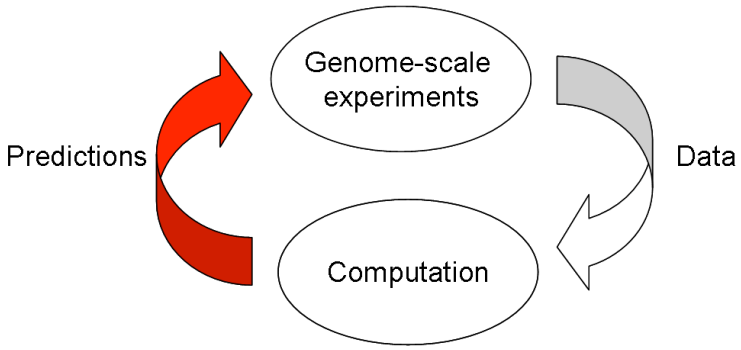


Figure 8.1. Iterative experiment-computation discovery loop. Computational approaches have demonstrated promise in translating raw genomic data into accurate biological hypothesis, but to date, they are rarely used in driving new, targeted experiments. We describe such an approach for mapping the global yeast genetic interaction network.

We demonstrate that "closing the loop" by using predictions to define further experiments can both increase the efficiency at which high-throughput technology is applied and also lead to rapid discovery of novel biology. This chapter includes contributions from Michael Costanzo, Anastasia Baryshnikova, David Hess, and Olga Troyanskaya. Michael and Anastasia helped me in understanding the SGA technology and in the analysis of the neighborhood interaction data. David Hess provided insights in analyzing the results of genetic interaction screens, and Olga supervised the project.

8.2 Background

Genetic interactions, or epistasis between pairs of genes, can reveal rich information about gene function and systems-level organization of the cell. In Chapter 7, we described in detail a high-throughput approach for detecting genetic interactions in yeast, Synthetic Genetic Array (SGA) analysis. Furthermore, we demonstrated that using a relatively simple model, we can derive precise, quantitative measures of epistasis from double mutant colony size data. Preliminary validation and analysis of these interactions indicated they are among the most valuable sources of genomic information available for learning about the functional role and organization of genes. If these data are so informative, why not use this technology to screen the entire space of all possible double mutants? This could perhaps produce an unprecedented view of the global genetic network. The answer to this is question is that the vast size of the combinatorial space of all possible double mutants (approximately 18 million in total) makes this task challenging, even with technology such as SGA. Estimates put the total time required for such an effort between 5-10 years [12].

Although a complete map may be several years off, one key insight described in [12] offers hope for significantly speeding this process. Tong et *al.* observed that while synthetic lethal interactions (extreme negative epistasis) are rare in the background of all possible genes (approximately 0.5%), they often occur between functionally related groups of genes. This suggests that if we could define such functional groups of genes in an unbiased, comprehensive way, we could rapidly explore at least the interesting parts of the genetic interaction map. This

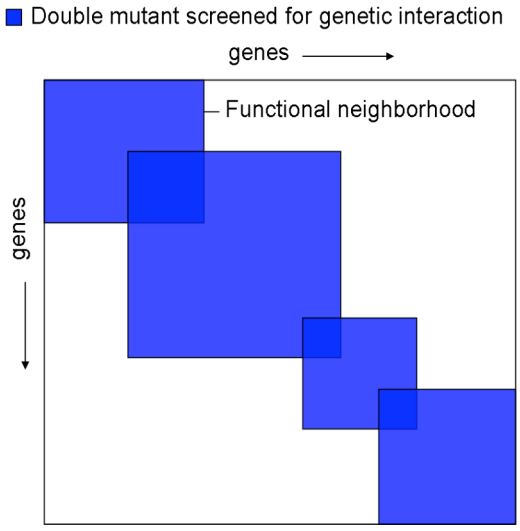


Figure 8.2. Overview of functional neighborhood genetic interaction screening approach. The basis of our approach is to define an accurate comprehensive set of functional neighborhoods (groups of related genes) to target for genetic interactions screens. All within-neighborhood pairs (illustrated in blue above) will be screened while between-neighborhood pairs will not. We expect most extreme genetic interactions should occur within these functional neighborhoods.

th

en raises the important question: how can we define these "functional neighborhoods" when our knowledge of some cellular processes is still relatively incomplete?

This presents a perfect opportunity for applying the genomic integration technology at the focus of this dissertation. As demonstrated in Chapters 4-6, these technologies perform well at providing a global, rough organization of the underlying functional network, even including genes that are uncharacterized. We demonstrate here that it can indeed be successfully used to direct an efficient mapping of the global yeast genetic interaction network. The basis of the approach is the definition of comprehensive yet accurate functional neighborhoods of genes, where screening is then limited to within-neighborhood pairs (Figure 8.2). The remainder of the chapter is organized as follows. First, we describe the details of our iterative computational-experimental approach for using Bayesian integration to direct high-throughput genetic interactions screens. We then present an evaluation of the efficiency improvement offered by such an approach. And finally, we discuss several novel biological insights that have resulted from computational analysis of this large-scale genetic interaction data, demonstrating the promise of using computational models to drive high-throughput genomic technology.

8.3 Methods: an Iterative Approach to Mapping the Global Genetic Interaction Network

The computational framework that serves as the basis for our iterative approach to mapping the global yeast genetic interaction network is the integration and network prediction methodology developed over Chapters 4-6. The approach consists of two key components: a system for defining functional neighborhoods from an integration of diverse genomic data, and a strategy for neighborhood refinement based on an iterative application of targeted, small-scale SGA screens (Figure 8.3). The neighborhood definition component ultimately identifies which pairs are to be screened to map the genetic interaction network, but the neighborhood refinement component also plays a key role. The motivation for the refinement step is that before significant resources are spent screening complete functional neighborhoods, we want to be certain neighborhoods are both comprehensive and accurate. Since we know genetic interaction data is one of the most informative sources of functional information, our strategy is to generate a limited amount of

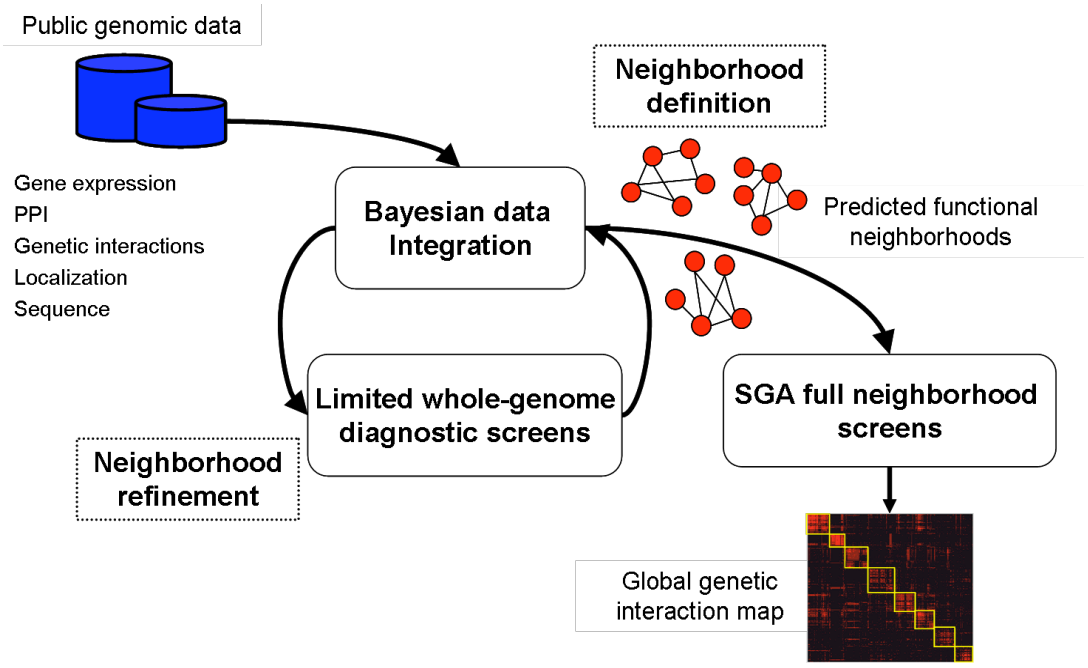


Figure 8.3. Overview of iterative computational-experimental approach for mapping the global yeast genetic interaction network. Our approach consists of two key components: a Bayesian data integration framework for neighborhood definition and a method for neighborhood refinement based on targeted, whole-genome genetic interaction screens.

targeted, highly informative data that can be used to refine neighborhood definitions before the final screens. Both of these components are described in detail in the sections that follow.

8.3.1 Neighborhood Definition

The core computational machinery for directing the construction of the global yeast genetic interaction map is context-sensitive genomic data integration and network prediction, which was presented in Chapter 6 and [7]. The basic idea is that we build a putative functional network based on available genomic data, which is then partitioned to define pairs of genes to screen for genetic interactions. We started this process by defining a core set of broad biological contexts, or functional neighborhoods, based on established knowledge of yeast biology (Table 1). Identification of these groups did not require a listing of all related components, but rather simply served as a broad functional classification for capturing most core processes in yeast. In addition to this listing, we associated a comprehensive set of specific GO terms with each of these groups, allowing them to belong to multiple neighborhoods if appropriate. These terms were taken from the non-redundant set of terms discussed in [6].

These functional groups, each including a catalogue of associated processes, served as the basis for context-sensitive learning and as seeds for building functional networks centered in each of these biological neighborhoods. The key steps from genomic data integration to neighborhood definition are described in detail below.

Predicting a context-sensitive functional network

For each broadly defined functional target, we constructed a functional network based on a context-sensitive integration of all available genomic data, including gene expression, protein-protein interactions, known genetic interactions, localization, and sequence information. The details of this approach and the exact datasets used are described in Chapter 6 and Appendix B, but we briefly summarize the process here. For context-sensitive integration, a Bayesian network is trained on functionally related pairs of genes from a specified biological context, such that the learned conditional probability distributions reflect the reliability of each input dataset in the given

Table 8.1. List of functional neighborhoods targeted by iterative approach.

Number	*Functional target*
1	Secretion/trafficking
2	DNA replication and repair
3	RNA processing
4	Cell cycle/tubulin cytoskeleton
5	General transport: small molecule, ion, drug
6	Metabolism
7	Mitochondria/energy/peroxisome
8	Polarity/actin cytoskeleton/cell wall
9	Protein biosynthesis/modification
10	Transcription

context. One can think of this as a weighted filter of different genomic data types that is optimized for the context of interest. As demonstrated in Chapter 6, context-specific learning can dramatically improve the accuracy of genomic data integration because it leverages the natural variance of datasets' relevance across different biological processes. Given the learned Bayesian network, we then perform inference on all possible pairs of proteins to derive a probabilistic network in which edges represent the probability that two proteins are functionally associated. One such network is constructed for all functional neighborhoods targeted by our approach (Table 8.1).

Defining functional neighborhood membership

Successful neighborhood definition requires a comprehensive assignment of all proteins to one or more functional neighborhoods, which are then screened for genetic interactions. Clearly, the efficiency of the entire neighborhood approach depends on the average size of the neighborhoods. If we wish to target 10 different general biological areas, a neighborhood size of approximately 1000 genes would result in screening about 28% of all possible double mutant combinations (Figure 8.4). Increasing the size of each neighborhood much more than 1000 genes would result in screening a significant portion of all possible pairs, defeating the purpose of the neighborhood design. Thus, we chose a target neighborhood size of between 600-1000 genes, which results in a coverage of about 20-30%.

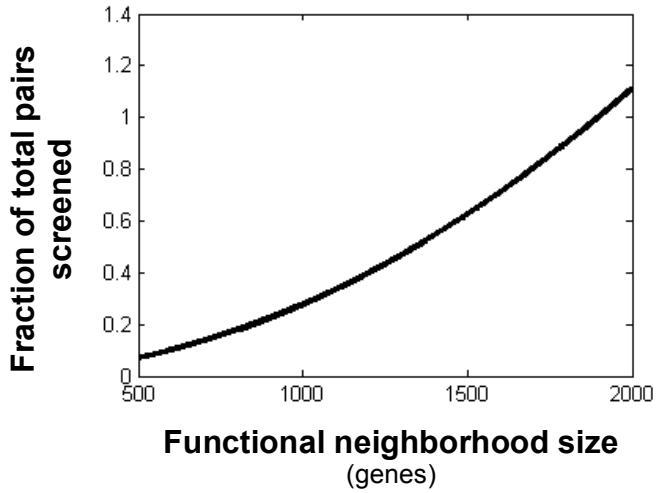


Figure 8.4. Fraction of total pairs screened for 10 functional neighborhoods vs. the size of the functional neighborhoods. This plot illustrates the fraction of pairs screened across 10 functional neighborhoods for a range of neighborhood sizes. We pick neighborhood sizes in the range of 600-1000 genes, and thus only screen a small fraction of all possible pairs.

Membership in each functional neighborhood is defined based on the context-specific putative functional network described above. Specifically, we measured statistical association of all genes to the set of genes known to be involved in the specific processes catalogued under each neighborhood. The association was measured for each GO term separately with the following metric:

$$S_i = -\log\left(1 - HG\left(\left\lfloor \sum_{j \in G} w(i,j) \right\rfloor, \left\lfloor \sum_{\forall j} w(i,j) \right\rfloor, c_G, c_T\right)\right)$$

where HG is the cumulative distribution function (CDF) of the hypergeometric distribution, $\lfloor . \rfloor$ implies rounding down to the nearest integer and

$$c_G = \left\lfloor \sum_{i \in G} \sum_{\forall j} w(i,j) \right\rfloor, c_T = \left\lfloor \sum_{\forall i} \sum_{\forall j} w(i,j) \right\rfloor.$$

This metric essentially tests for enrichment of known proteins in the neighborhood of candidate proteins, assigning more weight to probabilistic links with higher confidence. This metric is computed for each GO term associated with each neighborhood, and overall neighborhood association is taken as the sum of the top 3 in-neighborhood associations. Based on these scores derived from the functional network, we added candidate genes with significant associations in the following order of priority: (a) uncharacterized genes, (b) characterized genes

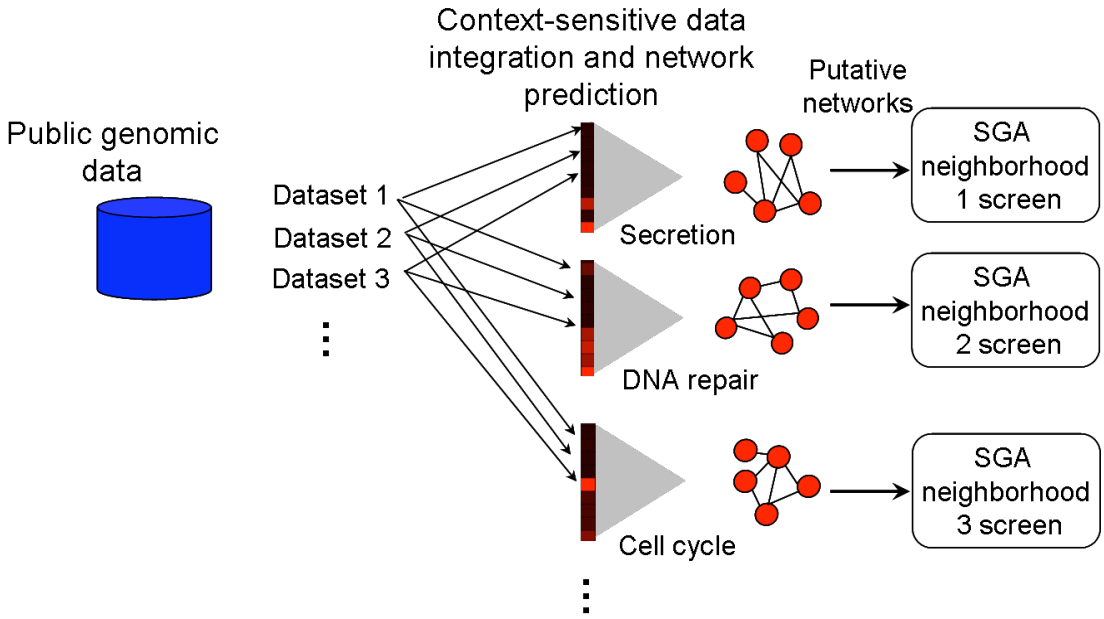


Figure 8.5. Overview of functional neighborhood definition. Neighborhood definition is based on a context-sensitive integration of genomic data [7]. This process results in a putative functional network (probabilistic graph connecting all genes) optimized towards each of the functional neigborhoods. Neighborhoods are then defined by measuring association to known genes in each graph.

annotated to GO terms associated with the neighborhood of interest, and (c) other characterized genes with no current annotation in the functional neighborhood, but with significant association. Because we wanted to ensure global coverage of the genome in the genetic interactions screens, all genes were added to the neighborhood showing the highest association before multiple assignments for any genes were made. However, since our neighborhoods are quite large and many had space remaining, we added multiple assignments for genes with statistically significant association to more than one functional neighborhood.

8.3.2 Neighborhood Refinement

The second key component in our hybrid computational-experimental approach for mapping the global interaction network is an iterative approach to neighborhood refinement. The motivation behind this component is to ensure comprehensive and accurate neighborhoods before investing significant resources in complete functional neighborhood screens. Because genetic interaction data is among the most informative genomic data types for predicting gene function (see Chapter

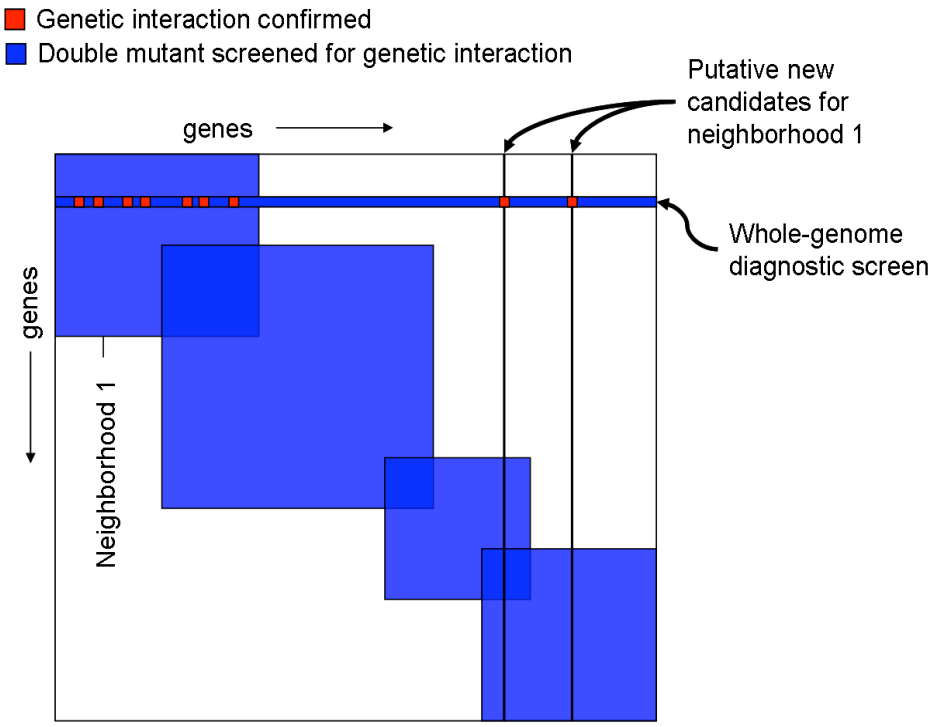


Figure 8.6. Schematic of whole-genome diagnostic screen approach for iterative refinement of neighborhoods. Based on the number and specificity of interactions in the predicted functional networks, we choose an optimal set of genes to serve as diagnostic screens across the whole genome. This process is illustrated above. Blue represents areas screened for genetic interactions and red indicates a confirmed genetic interaction.
Upon screening the diagnostic sets, we expect to find most interactions within our defined neighborhoods, but interactions outside of these neighborhoods indicate potential missing candidate genes.

7), we set out to design targeted, small-scale SGA screens that would provide information to best refine our neighborhoods. Specifically, our experimental collaborators were willing to invest a small amount of resources in whole-genome screens for a handful of genes. The goal in selecting this set was that interactions identified in whole-genome screens across the set would provide a highly informative diagnostic signature for refining neighborhood membership for any candidate gene (Figure 8.6). Thus, we are left with the question of how to best pick this set of diagnostic genes.

Picking an optimal diagnostic gene set

This question can be addressed precisely by the predicted functional networks. Intuitively, we want to identify likely hubs in the genetic interaction network that will yield highly functionally informative signatures when screened against the whole genome. This motivates two criteria for

their selection based on our functional networks: node degree and node specificity. Within-neighborhood node degree for a neighborhood, *N*, is defined as:

$$D_i = \sum_{j \in N} p_{ij}$$

Node degree is the sum of all edges adjacent to any gene a functional neighborhood of interest and reflects its overall "hubbiness." Genes with high node degree are likely to associate with several genes and play a central role in a variety of cellular processes. Within-neighborhood node specificity for a neighborhood, *N*, is defined as:

$$S_i = \frac{\sum_{j \in N} p_{ij}}{\sum_{\forall j} p_{ij}}$$

Node specificity captures a different aspect of a gene's cellular role: how localized its interacting partners are to a specific biological process. A gene could have a very high node degree but interact with other proteins in a diverse set of processes, and thus, would not be an ideal candidate for the diagnostic set. Instead, we want genes that are both likely to interact (high degree), and also genes whose interactions are predictive of a role in a specific function (high specificity). We illustrate this point further with an example from the metabolism functional neighborhood.

Figure 8.7 plots the within-neighborhood specificity versus the within-neighborhood node degree for all member genes. We observe a broad distribution of these values across the entire neighborhood. Two striking proteins in this plot are Ser1 and Cdc28, both exhibiting a high node degree of greater than 130 in the metabolism functional network (Figure 8.7). Ser1, however, is highly specific to this neighborhood (.75) while Cdc28 mainly interacts with proteins outside of metabolism, evidenced by a specificity of .25. Interestingly, these characteristics accurately reflect the cellular role of these two proteins— Cdc28 is the catalytic subunit of the main cell-cycle dependent kinase (CDK) and is responsible for regulation of a variety of cell-cycle events [4]. Ser1 catalyzes the formation of phosphoserine, which is required for serine and glycine biosynthesis [3], and thus, while it interacts with many genes in the functional network, most of them play a role in metabolism. While Cdc28 is not a good candidate for the diagnostic set, Ser1

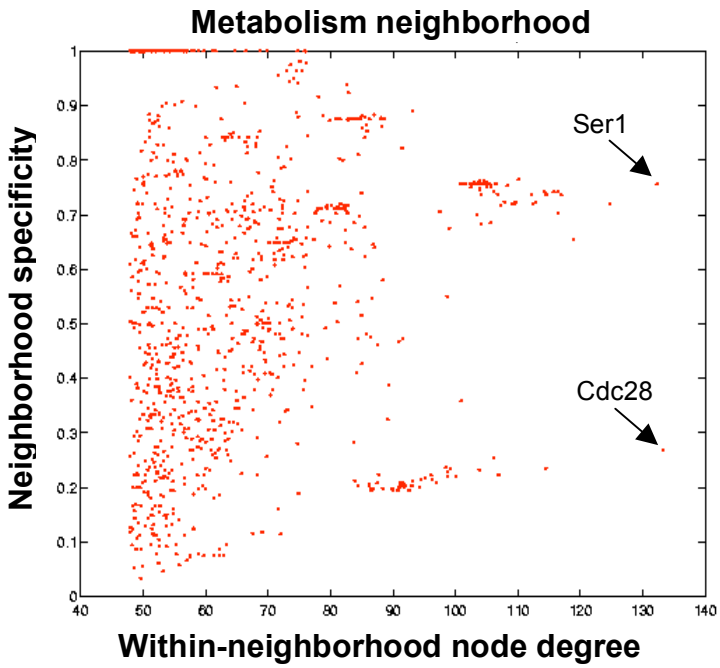


Figure 8.7. Criteria for selecting whole-genome diagnostic screens: neighborhood specificity vs. node degree. Genes for diagnostic whole-genome screens are picked based on interaction specificity and node degree with each functional neighborhood. This figure illustrates these metrics for the metabolism neighborhood. Each point represents one gene assigned to this neighborhood. Two proteins, Ser1 and Cdc28, are indicated on the graph. Ser1 is specific to metabolism and thus has both high node degree and high specificity, while Cdc28 plays a diverse role in regulating cell cycle events and has high node degree but low node specificity within the metabolism context.

is because we expect most of its genetic interactions will indicate involvement in some aspect of metabolism.

We formalize the intuition behind node degree and node specificity to design an algorithm for selecting the optimal diagnostic set for each functional neighborhood. Briefly, the algorithm requires a minimum threshold on specificity, and then greedily optimizes interaction coverage across each neighborhood (Table 8.2). The parameter, *D,* controls the credit awarded for redundant interactions to genes already covered by the current diagnostic set. $D = \infty$ will only

Table 8.2. Algorithm for picking diagnostic gene set.

Initialize $T = \{\}$, repeat the following:

While $|T|<$ max. size of diagnostic set

1. For each candidate gene *i,* form the set $S = \{T, i\}$ where T is the current diagnostic set.
2. For every gene *i*, form $\vec{v}_i^{\,S} = \mathrm{sort}\left(g_{iS_1}, g_{iS_2}, g_{iS_3}, \ldots\right)$
3. Update T: $T = \{T, \underset{S}{\mathrm{argmax}} \sum_i \sum_j v_{ij}^{\,S} D^{-j}\}$

reward the largest such interactions and $D = 1$ will reward all interactions even if they are adjacent to already covered nodes. We chose $D = 1.25$, and applied this approach to select 10-15 genes from each functional neighborhood for whole-genome diagnostic screens.

Updating neighborhood definition

Given the SGA screens from the diagnostic set, neighborhood definitions are updated in a straightforward manner. Correlation coefficients are computed between all candidate genes' diagnostic interaction profiles, and we evaluate each gene's adjacency to each functional neighborhood based on the average correlation. Any neighborhood exhibiting higher correlation than the current assignment is added as an additional neighborhood assignment. This refinement procedure is still under development as of writing of this dissertation, and there are several promising alternatives to this simple approach. For instance, once could imagine training a discriminative classifier on diagnostic interaction profiles to achieve more precise neighborhood definitions, particularly for uncharacterized genes. We are currently pursuing approaches based on this idea.

8.4 Evaluation of Computationally Directed Neighborhood Approach

We applied the iterative computational-experimental approach described above to map genetic interactions in yeast using SGA genetic interaction screens. As of writing of this dissertation, 3 functional neighborhoods were complete, including secretion/trafficking, metabolism, and mitochondria/energy/peroxisome as well as approximately 150 whole-genome diagnostic screens. Given these whole-genome screens, we measured the efficiency improvement gained by our approach. In short, we find that our framework has dramatically improved the efficiency at which genetic interactions are detected. We measured this by comparing the number of within-neighborhood interactions identified in the whole-genome screens to the total number of interactions identified, which gives an estimate of the sensitivity of our approach (i.e. what fraction of interactions would we have detected in these whole-genome screens had we only screened within-neighborhood pairs?). We detect 82% of the most significant interactions by screening less than 30% of the possible pairs (Figure 8.8). This provides strong evidence supporting the utility of our approach. Our functional neighborhoods are dramatically enriched for genetic interactions and we detect a large majority of the most significant interactions by screening only a small fraction of all possible pairs.

8.5 Biological Validation: What Can We Learn from all of these Data?

Beyond evaluating the sensitivity of our approach, we have also begun analyzing the genetic interaction data resulting from the computationally-driven screens. In general, we find these data continue to be among the most informative sources of genomic data and have led to numerous biological insights. We highlight a few of our findings here. All of the results discussed in this section are based on the secretion functional neighborhood screens.

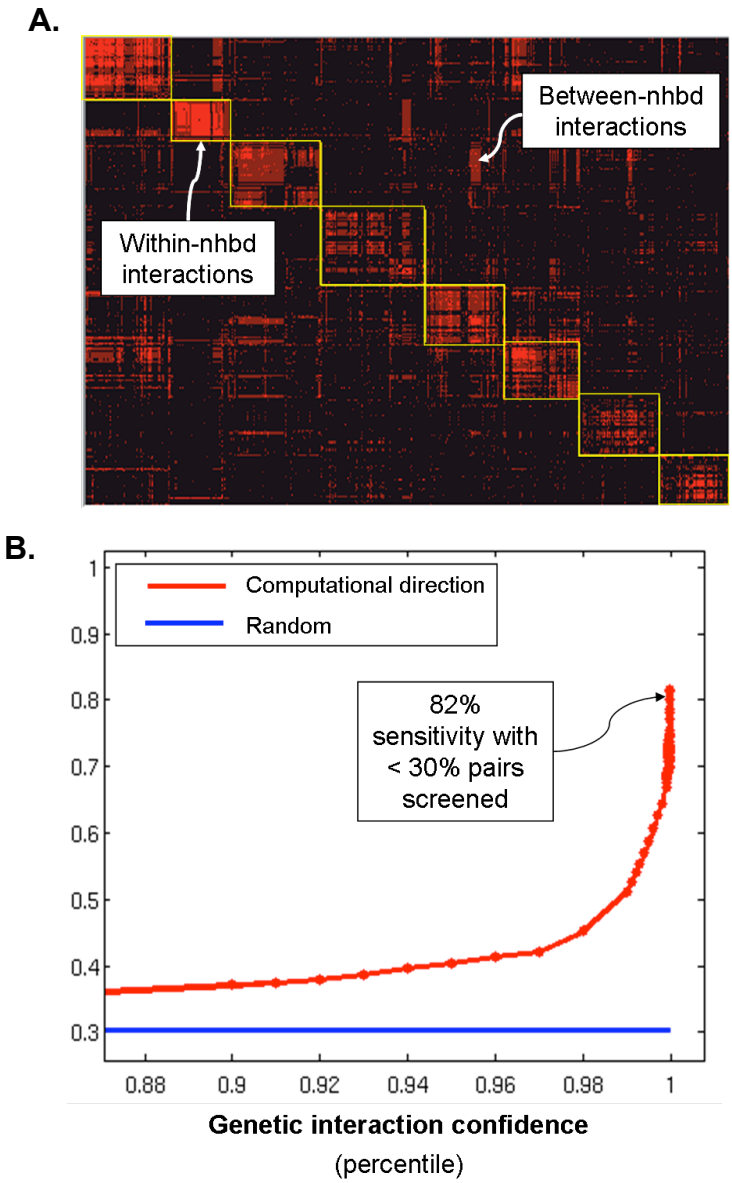
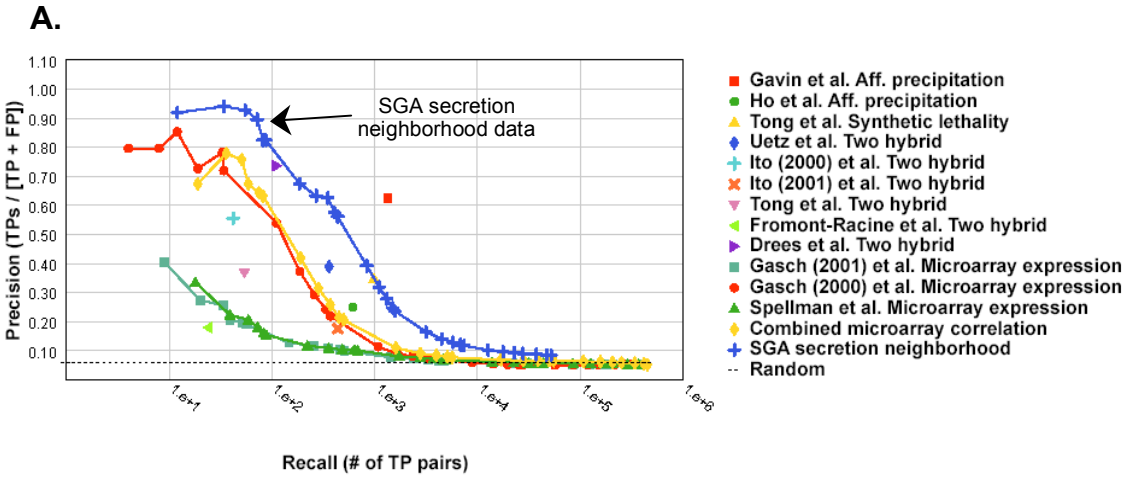


Figure 8.8 Sensitivity analysis of neighborhood design approach. We evaluated the sensitivity of our approach by measuring the number of within-neighborhood genetic interactions (illustration in A) found in a set of 76 whole-genome screens. Figure B plots the fraction of interaction pairs detected with previously defined functional neighborhoods at the given percentile. For a typical genetic interaction cutoff, we estimate that our neighborhood approach would detect 82% of the most extreme interactions by screening less than 30% of the total possible pairs (B).

8

.5.1 Genetic Interaction Profiles are Highly Informative about Gene Function

As indicated by our earlier analysis in Chapter 7, genetic interaction data are highly informative about gene function. To verify this, we computed correlation coefficients between genetic



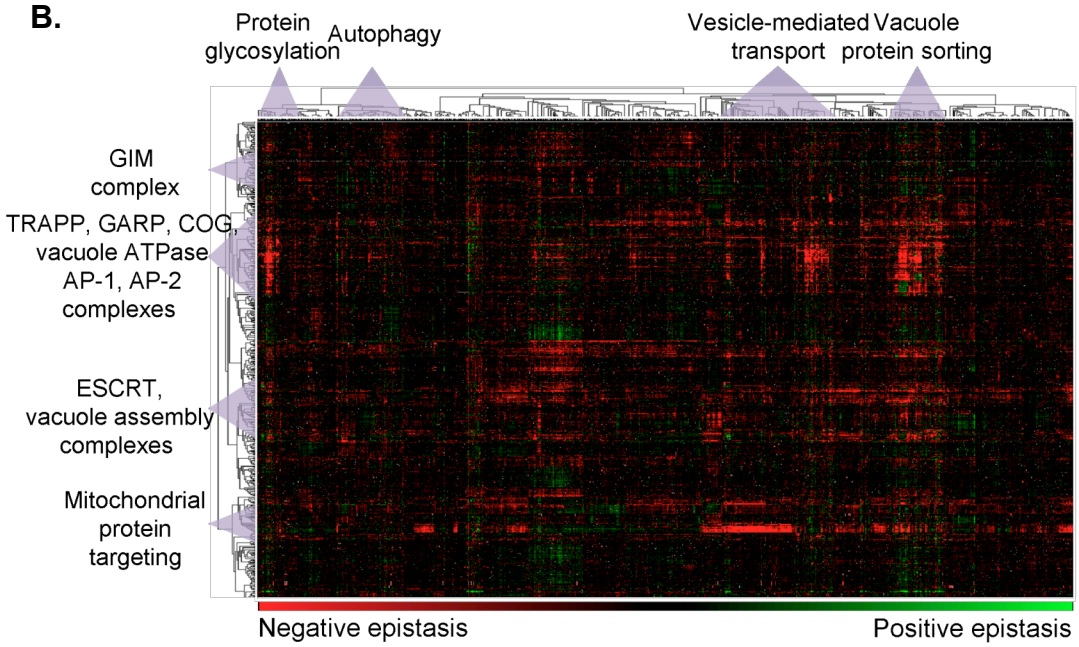


Figure 8.9. Functional enrichment of genetic interaction profiles. We performed a precision-recall analysis of correlation coefficients across genetic interaction profiles from the secretion functional neighborhood (A). We find that these profiles are among the most informative sources of genomic data. We also confirmed rich functional information in these data based on 2D clustering (B). We find several examples of tightly clustered known secretion complexes, including some that encompass previously uncharacterized genes.

in

teraction profiles in the secretion neighborhood and measured their ability to predict co-involvement in a specific process. We confirm that with the exception of one high-throughput physical interaction dataset, the genetic interaction data from a *single* functional neighborhood is the most sensitive and precise of all genomic datasets in our collection (Figure 8.9A). For

instance, correlation among these profiles can detect approximately 1000 functionally associated pairs across diverse processes at a precision of 40% (Figure 8.9A). The presence of rich functional information in these data is also confirmed through 2D hierarchical clustering (Figure 8.9B). We find several clusters highly enriched for protein complexes related to secretion (e.g. the vacuolar ATPase). Many of these tight clusters also include uncharacterized genes with predicted involvement in the secretion neighborhood. These examples demonstrate that even simple clustering analyses of these data can reveal clear details of network-level organization.

8.5.2 Between and Within-complex Genetic Interactions are Monochromatic

One interesting and surprisingly clear trend we observed in the secretion functional neighborhood was monochromaticity of within-complex and between-complex interactions. This trend is illustrated for secretion-related complexes in Figure 8.10A, the retromer, ER assembly complex, and vacuolar ATPase complexes. All significant positive or negative epistatic interactions are indicated by green and red edges, respectively. With few exceptions, between-complex interactions between pairs of proteins within the same complex are in the same direction, and the same is true for gene pairs spanning two different complexes. We see both negative and positive epistatic between-complex interactions. Interestingly, we observe positive epistasis between the ER assembly complex and the vacuolar ATPase, which is consistent with their relative roles in the cell— the ER assembly complex is required for proper assembly of the vacuolar ATPase [2].

This monochromaticity holds across several more complexes, as illustrated in Figure 8.10B, where now complexes have been collapsed into single nodes. The color of each node indicates the proportion of positive and negative interactions within each complex (bright red id 100% negative, bright green is 100% positive). Edges are colored in a similar fashion. Interestingly, there are very few nodes or edges that are not entirely composed of the same interaction type.

The monochromaticity reveals the inherent modularity of cellular genetic networks and will no doubt enable powerful approaches to reverse engineering network structure. This phenomenon has been predicted in earlier work [10] based on analysis of synthetic data

generated from metabolic networks. To our knowledge, this is the first time it has been confirmed

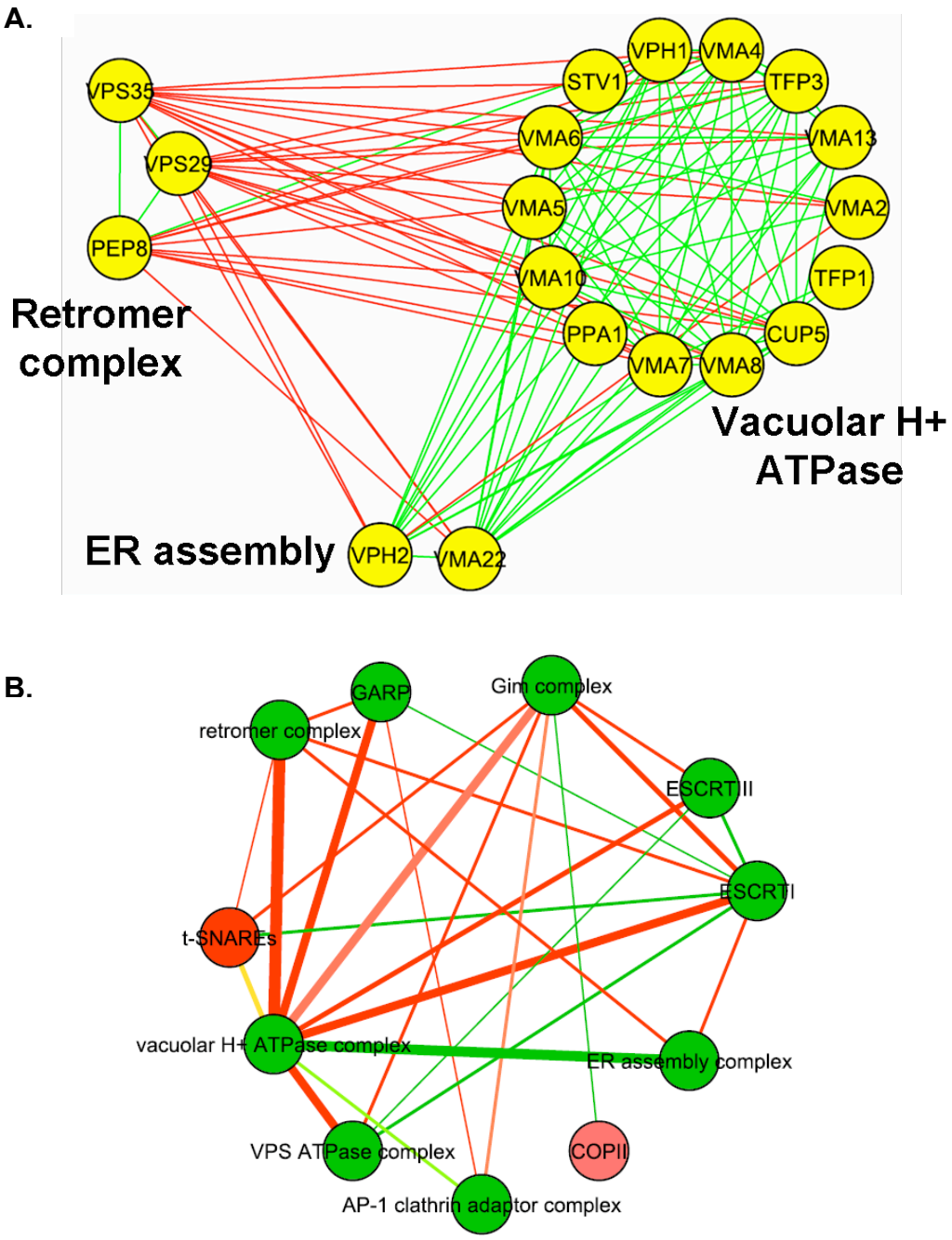


Figure 8.10. Within-complex gene pairs and between complex gene pairs are largely monochromatic. (A) We plot the confirmed genetic interactions for three complexes related to secretion. Green edges indicate positive epistasis while red edges indicate negative epistasis. Pairs within the same complex and pairs spanning the same two complexes mostly share the same type of interaction. (B) This trend is confirmed on a larger scale, with a set of 12 secretion-related complexes. Node color indicates the proportion of positive or negative interactions within each complex, and the color of the edges indicates the proportion of positive or negative interactions between genes spanning those complexes. Edge size Indicates the number of interactions between each pair.

on epistasis measures from raw double mutant colony size data. This serves as a striking confirmation of the quality of interactions derived from these high-throughput screens.

8.5.3 Within-complex Genetic Interactions are Predictive of the Cellular Role of a Complex

In addition to finding well-defined modularity reflecting protein complexes, we find that the type of within-complex genetic interactions (positive or negative) is also informative. Complexes that exhibit mainly negative epistatic within-complex interactions are highly likely to contain at least one essential gene (Figure 8.11). Based on complexes obtained from the MIPS Complex

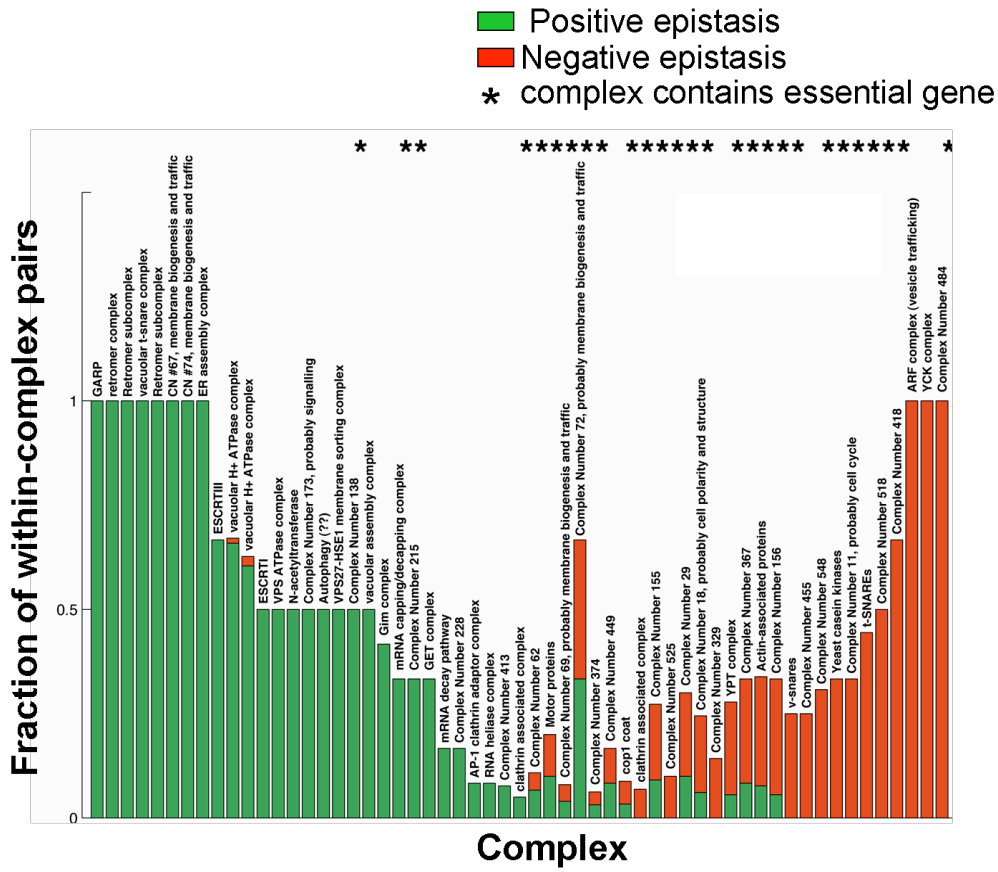


Figure 8.11. Within-complex epistasis correlates with essentiality. For each complex, we plot the proportion of within-complex pairs screened that exhibit positive or negative genetic interactions. We find that approximately half of the complexes exhibit mainly positive epistasis and half exhibit negative epistasis between member pairs. 86% of complexes exhibiting negative epistasis are complexes containing one or more essential genes. Note that the interactions comprising these proportions only include non-essential genes.

Catalogue [5], we measured the proportion of within-complex pairs screened that showed either significant positive or negative interactions. Of 28 complexes with a majority of negative interactions, 24 of the 28 (86%) contained essential genes, while only 3 of 27 (11%) of complexes with mostly positive within-complex interactions contained essential genes. We suspect the reason for this correlation relates to the role the complex plays in the cell. Complexes with essential genes are ultimately required for cell viability and thus, as pairs of non-essential complex members are removed, the complex degrades, losing its function, and the cell dies. Complexes without essential components, on the other hand, are likely not essential for cell viability and thus, as members are removed, the complex may cease to function normally, but the cell remains viable. These double mutants then exhibit positive interactions because the cell is healthier than expected based on the two single mutants. In essence, within-complex interactions are indicative of buffering at the complex level. This striking trend is yet another confirmation of the quality and utility of the genetic interaction data generated by our approach.

8.5.4 Ab Initio Pathway Ordering from Genetic Interactions

From the previous analysis, it is clear that the modularity readily apparent in genetic interaction data contains important clues about protein complex membership and function. Previous studies on epistasis have demonstrated that the magnitude of positive epistatic interactions can be used to determine the order of pathways [1,11]. One example is illustrated in St. Onge et *al.* where a number of DNA damage repair mutants were ordered from epistasis measured on growth in liquid media [11]. The basic principle used in this analysis is that positive epistasis usually indicates co-involvement in a complex or serial pathway and the single mutant farthest upstream masks the phenotype of the downstream single mutant when their mutations are simultaneously introduced (Figure 8.12). Under an assumption of positive regulation, the severity of the double mutant phenotype can be compared to the two singles and can define relative order as illustrated in Figure 8.12.

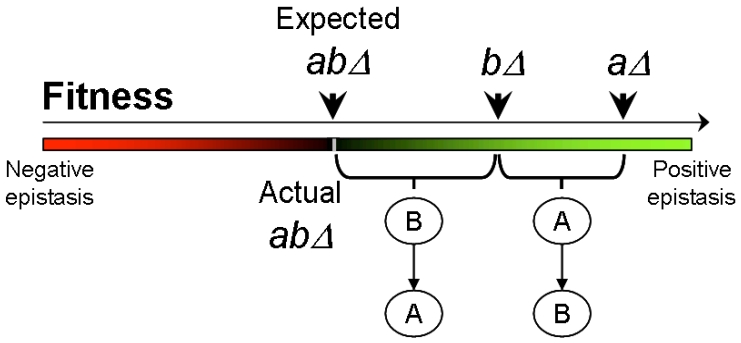


Figure 8.12. Deriving pathway order from single and double mutant phenotypes. Previous work on epistasis has demonstrated that positive interactions often occur between genes arranged in series in a pathway, and that the magnitude of the interactions can be used to define their order [1,11]. For pathways under positive regulation, the rule for determining order is presented above. The single mutant which most closely resembles the double mutant phenotype is placed upstream in the pathway.

We tested whether this approach also worked on epistasis estimates from the secretion functional neighborhood. To do this, we focused on the AP-1 and AP-3 complexes, which interact with clathrin during endocytosis and function to specify the cargo for vesicle-mediated transport [8]. Most cargo proteins, including carboxypeptidase Y (CPY), are delivered to the vacuole through a prevacuolar endosome via several proteins including Pep12p, Vps45p, Vps4p, and Vps27 [8]. However, recent research has implicated a parallel pathway for Golgi to vacuole transport which shuttles the membrane protein alkaline phosphatase (ALP) via the AP-3 adaptor protein complex [8] (Figure 8.14A). This set of pathways provides a perfect case for validation because it involves two parallel pathways one of which has an established serial ordering of proteins, all screened in our secretion neighborhood. Interestingly, before we even began ordering, we saw hints of this pathway structure in the 2D clustergram (Figure 8.13). Both the AP-1 and AP-3 complexes exhibit positive within-complex interactions, and have completely opposite interaction profiles across the other genes pictured, suggesting they are operating in parallel.

For all significant positive interactions, we applied the rules outlined in Figure 8.12 to assign order between all possible pairs of genes. We applied a transitive reduction to the resulting graph to remove redundant edges, and the final, pruned network is presented in 8.14B. With only minor exceptions, this simple approach based on a comparison of single and double

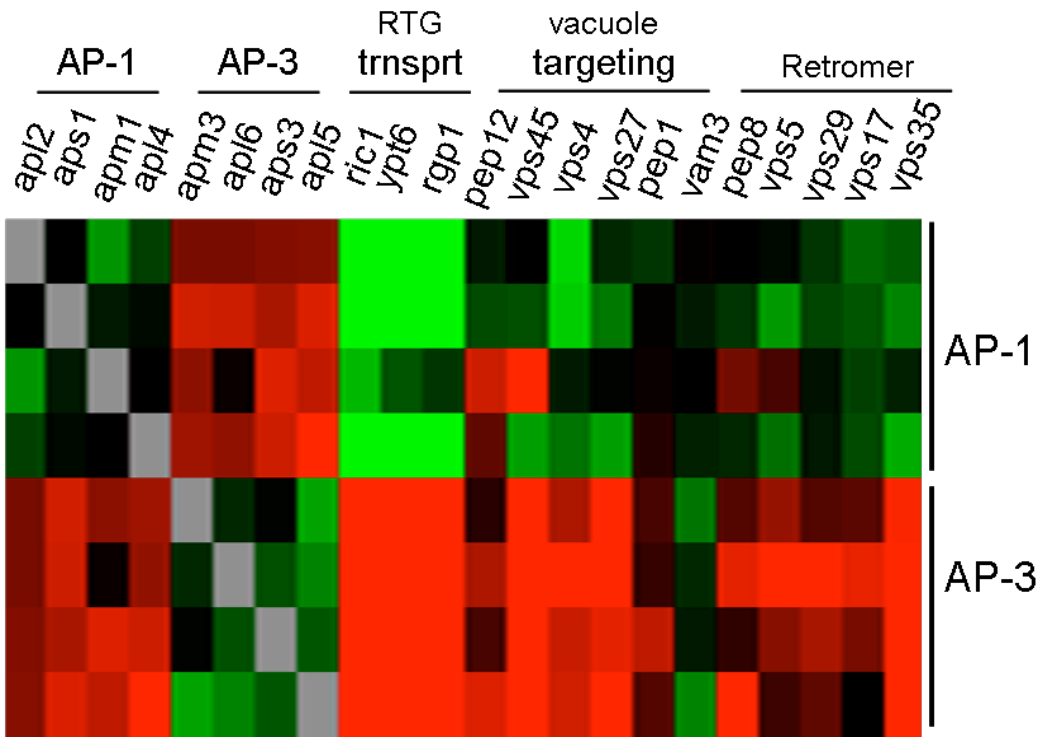


Figure 8.13. 2D clustergram of genetic interactions for the AP-1 and AP-3 adaptor protein complexes. The AP-1 and AP-3 complexes interact with clathrin during endocytosis and function to specify the cargo for vesicle-mediated transport, and are believed to act in parallel [8]. Their parallel action is supported by this clustergram, which indicates that AP-1 and AP-3 show largely the opposite interaction profiles across a set of downstream genes.

m

utant phenotypes almost perfectly reconstructed the pathway configuration for this example. The serial AP-1 pathway appears exactly as reported in the literature and most components show negative interactions with the AP-3 complex, consistent with its suspected parallel role. We were surprised and skeptical at the success of this method because it relies on relatively simplistic assumptions of positive regulation. We expect this approach will not work in all cases or even most, but it appears to successfully reconstruct this pathway.

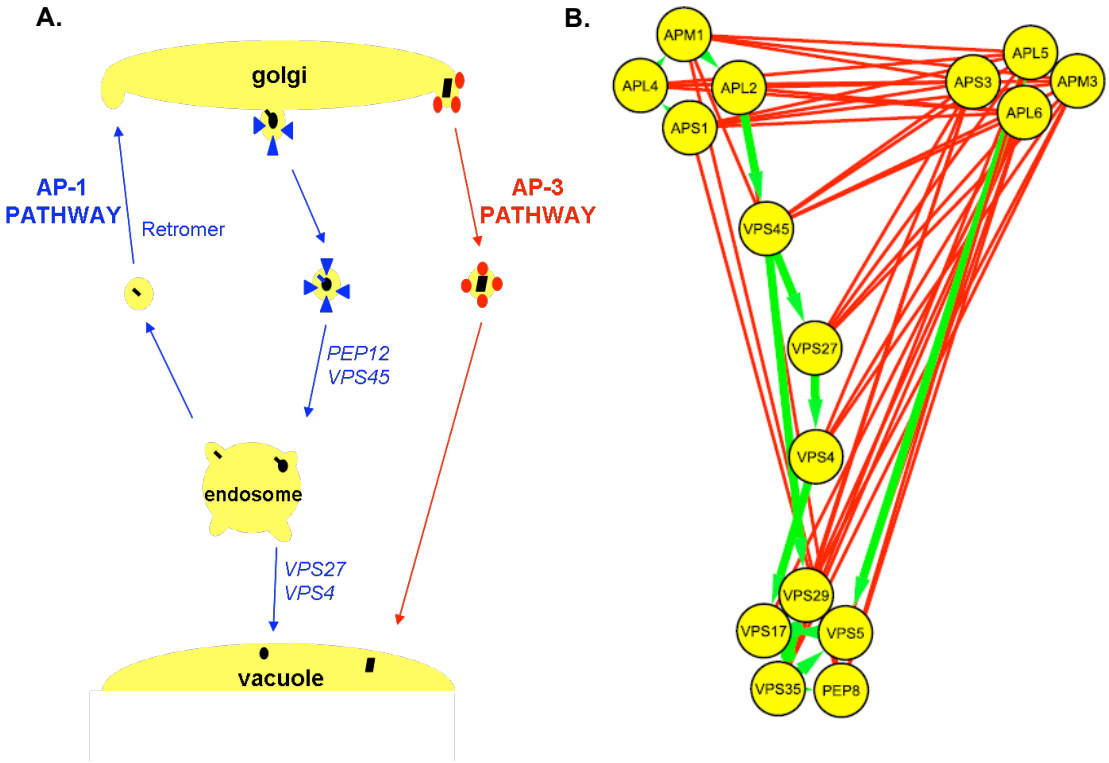


Figure 8.14. Automated pathway ordering of the AP-1 and AP-3 Golgi-vacuole trafficking pathways. (A) The current model for Golgi to vacuole transport mediated through the AP-1 and AP-3 pathways suggests these complexes are two parallel mechanisms [8]. We attempted to order genes in these complexes and the associated downstream counterparts based on the magnitude of epistatic interactions among them (B). With only a few exceptions, we are able to automatically reconstruct this pathway, including the serial order of the downstream AP-1 components. The negative interactions between AP-3 and these downstream components suggest that they are indeed parallel means of Golgi to vacuole transport.

8.6 Discussion and Conclusions

We have described a hybrid computational-experimental approach for efficiently mapping the global yeast genetic interaction network. We have used this method for successfully mapping interactions in 3 of 10 functional neighborhoods to date, and estimate that we detect approximately 85% of the extreme epistatic interactions by screening less than 30% of all possible double mutants. We have also presented several examples of biological insights derived from simple analyses of these data, demonstrating the broad utility of large-scale genetic interactions for understanding systems-level properties of the genetic network.

The work presented in this chapter represents a compelling example of how computation can be used not only to generate accurate predictions, but also to drive an entire experimental study. The interactions screens responsible for generating the data presented here required

several months of work, and most pairs screened were defined through computation. We expect the utility of this type of approach will generalize to other experimental settings, e.g. microarray analysis, where targeted experiments derived from computational predictions could potentially enable more efficient discovery.

Another interesting aspect of this work is the iterative nature of the mapping framework. Computational analysis of the functional network was used to identify a set of informative diagnostic hubs, which were then screened against the whole genome to generate informative diagnostic profiles for neighborhood refinement. Often, targeted measurements can yield the missing piece of data critical for building the correct network model, and thus, a flexible, iterative framework that allows this feedback is crucial.

Based on our preliminary analysis of the data resulting from this iterative computational-experimental approach, we find that genetic interaction data can readily lead to new insights into gene function and network organization. We suspect the utility of these data will only grow when placed in the context of a global map of the genetic interaction network. The tools for unlocking the full potential of these data to generate accurate, testable hypotheses are yet to be discovered, but will almost certainly involve computational models.

References

1. Avery, L. and S. Wasserman (1992). "Ordering gene function: the interpretation of epistasis in regulatory hierarchies." Trends Genet **8**(9): 312-6.
2. Graham, L. A., K. J. Hill, et al. (1998). "Assembly of the yeast vacuolar H+-ATPase occurs in the endoplasmic reticulum and requires a Vma12p/Vma22p assembly complex." J Cell Biol **142**(1): 39-49.
3. Melcher, K. and K. D. Entian (1992). "Genetic analysis of serine biosynthesis and glucose repression in yeast." Curr Genet **21**(4-5): 295-300.
4. Mendenhall, M. D. and A. E. Hodge (1998). "Regulation of Cdc28 cyclin-dependent protein kinase activity during the cell cycle of the yeast Saccharomyces cerevisiae." Microbiol Mol Biol Rev **62**(4): 1191-243.

5. Mewes, H. W., D. Frishman, et al. (2002). "MIPS: a database for genomes and protein sequences." Nucleic Acids Res **30**(1): 31-4.

6. Myers, C. L., D. R. Barrett, et al. (2006). "Finding function: evaluation methods for functional genomic data." BMC Genomics **7**: 187.

7. Myers, C. L. and O. G. Troyanskaya (2007). "Context-sensitive data integration and prediction of biological networks." Bioinformatics **23**(17): 2322-30.

8. Odorizzi, G., C. R. Cowles, et al. (1998). "The AP-3 complex: a coat of many colours." Trends Cell Biol **8**(7): 282-8.

9. Pena-Castillo, L. and T. R. Hughes (2007). "Why are there still over 1000 uncharacterized yeast genes?" Genetics **176**(1): 7-14.

10. Segre, D., A. Deluna, et al. (2005). "Modular epistasis in yeast metabolism." Nat Genet **37**(1): 77-83.

11. St Onge, R. P., R. Mani, et al. (2007). "Systematic pathway analysis using high-resolution fitness profiling of combinatorial gene deletions." Nat Genet **39**(2): 199-206.

12. Tong, A. H., G. Lesage, et al. (2004). "Global mapping of the yeast genetic interaction network." Science **303**(5659): 808-13.

Chapter 9

Conclusions and Future Work

9.1 Dissertation Summary

Recent developments in biotechnology have enabled the unprecedented measurement of several cellular phenomena including gene expression, protein-protein interactions, genetic interactions, protein localization, and sequence information. However, making use of these measurements to generate specific, testable hypotheses is non-trivial and thus, valuable information present in these data is often not translated into knowledge. In this dissertation, we have explored a number of computational solutions to this and have demonstrated their potential for helping us learn new biology.

We began with a discussion of interpreting gene expression and copy number data in its chromosomal context, and demonstrated a promising method for automatically detecting chromosomal aberrations. Gross chromosomal changes have been associated with several cancers, and thus accurate and automatic identification of these changes is an important step towards identifying recurring global patterns of abnormality and potentially learning clues about the underlying mechanisms of tumorigenesis and cancer progression. We have also addressed the broader problem of drawing inferences about molecular and genetic networks from diverse genomic data. The challenging aspect of this problem is heterogeneity among the available datasets. Taking advantage of all information present in the data, while not sacrificing precision, requires robust methods for integration. We have presented a general Bayesian framework for accommodating this heterogeneity and constructing global maps of functional associations between proteins.

Beyond robust integration, the key insight that enabled this work was identifying context-specific signals in the data. This idea is based on the observation that most experimental technologies capture different biological processes with varying degrees of success, and thus,

each source of genomic data will vary in relevance depending on the biological process one is interested in predicting. For example, context-dependent variation naturally follows from the fact that biologists often design experiments to target specific processes. Accounting for this variation is critical for predicting accurate network models, but no previous computational approaches for network prediction from diverse data leveraged this information. We expect this context-sensitive framework will provide the basis for a variety of genomic data integration applications, especially as we attempt to apply methods to higher organisms with tissue and developmental specificity.

Finally, we have demonstrated the utility of such genomic data integration technologies in not only generating accurate, testable hypotheses but also in driving large-scale experiments. We showed that computational direction of high-throughput genetic interaction screens can dramatically increase the efficiency with which novel biology is discovered. Furthermore, preliminary analysis of the genetic interaction data generated through these computationally-directed screens has revealed several systems-level insights and demonstrated the promising potential of large-scale genetic interaction data for functional characterization.

A recurring theme throughout all of the work presented in this dissertation is the importance of effective data visualization. Nearly every key insight presented here was motivated by visual analysis of the data. From identifying functional biases in the signals captured by genomic datasets to finding global trends across epistasis profiles, intelligent data visualization played a key role in several of these discoveries. As data repositories continue to grow and new technologies enable new diverse genomic characterization, methods for visualization-based analysis will no doubt play a central role in deriving meaning from these data.

In summary, we have presented several new strategies for learning from genomic data and have demonstrated their promising applications with numerous case examples of novel biological results that were correctly predicted and validated based on these approaches. In closing, we discuss a few promising directions for future research.

9.2 Future Work

9.2.1 Large-scale Discovery of Gene Function Using Genomic Data Integration and Network Prediction Technology

Perhaps the most consistently reinforced conclusion throughout the projects described in this dissertation was that finding novel biology based on predictions was often much easier than expected. In general, we tend to significantly underestimate the accuracy of our predictions (e.g. the mitochondria validation example presented in Chapter 6). This observation suggests that computational methods for network and gene function prediction are now mature enough to support large-scale experimental validation of predictions. Based on performance estimates of bioPIXIE (discussed in Chapters 4 and 6), we estimate that we can assign function with high confidence to a large percentage of the genes in yeast that are uncharacterized. In fact, we estimate that just based on current data, we can assign specific functions to approximately 500 of the 1347 (37%) total uncharacterized proteins at a precision of at least 40% (Figure 9.1). The challenge is no longer in making accurate predictions, but rather in developing an experimental framework the supports large-scale validation of these predictions. The mitochondria example presented in Chapter 6 provides a compelling example of such a framework for identifying

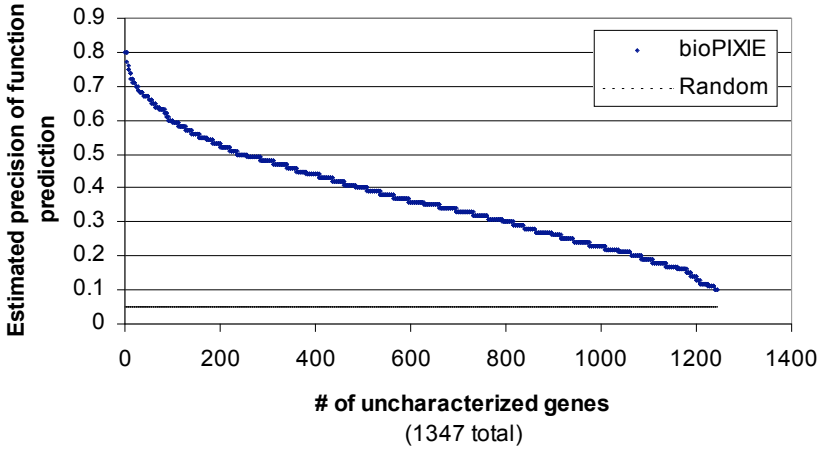


Figure 9.1. Estimated precision of best functional assignment for all uncharacterized genes in *Saccharomyces cerevisiae*. We predicted function based on the bioPIXIE functional network (see Chapters 4 and 6), and for each uncharacterized protein, we estimate the precision of its best functional assignment. This plot illustrates the distribution of precision estimates for all 1347 currently uncharacterized genes in yeast.

mitochondria-related proteins. Undoubtedly, a promising direction of future research is to scale this framework to other processes and organisms to allow rapid characterization based on computational analysis and integration.

9.2.2 Iterative Computational-Experimental Approaches

A related promising future direction for research is the development of approaches that support iterative experimental-computational discovery. We demonstrated the successful application of such an approach in mapping the yeast genetic interaction network. Putative functional networks derived from an integration of diverse genomic data proved invaluable in picking mutants for genetic interactions screens. In our experience, the most powerful aspect of this type of iteration is the ability to generate targeted data for the specific question one would like to answer, or hypothesis one would like to validate. In our case, the goal was to produce the most informative diagnostic interaction profiles, and we used the functional network to find specific, hubs that were likely to yield informative interactions. We expect this idea can be applied successfully in other experimental contexts as well

9.2.3 Combining Genetic Interaction Data with Other Genomic Data for Automatic Inference of Pathway and Network Topology

A final promising direction for future research is automatic discovery and refinement of pathways based on genetic interaction screens. Based on our preliminary analysis presented in Chapter 8, these data contain significantly more topological information than most other data sources we have analyzed. Our analysis and reconstruction of a secretion-related pathway was based solely on genetic interaction data. Integrating a global map of genetic interaction profiles with other existing genomic data (e.g. protein-protein interactions, gene expression) could provide a powerful approach for direct pathway prediction. Previous methods for pathway inference beyond the functional association approach described here have only achieved limited success, largely because the experimental data cannot statistically support these inferences. The rich,

systems-level information offered by comprehensive genetic interaction screens will likely support more sophisticated approaches.

Appendix A

BioPIXIE Query Sensitivity Analysis

BioPIXIE is designed to support queries for groups of proteins, and returns a functional network centered at the query set. Thus, we found it important to evaluate the sensitivity of the algorithm's performance to different characteristics of the query set, specifically, the noise and the size in the query set. Results for each of these are presented below. See Chapter 4 for an in-depth analysis.

Query Noise Sensitivity Analysis

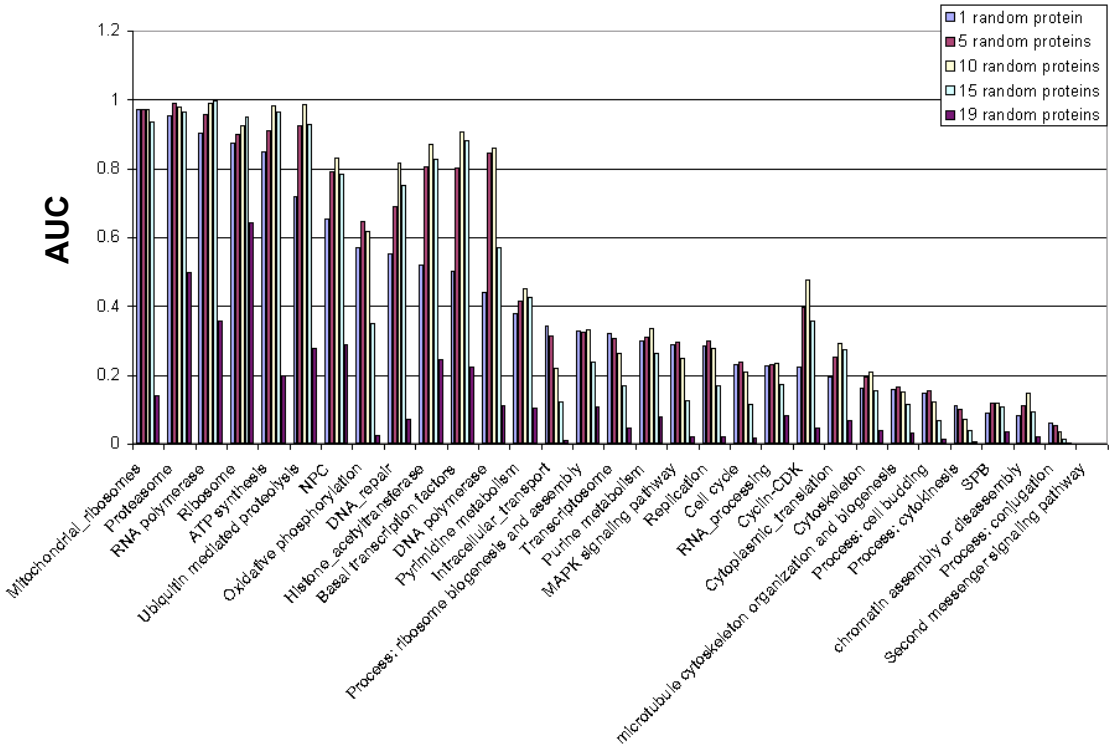


Figure A1. BioPIXIE noise sensitivity analysis. We evaluated the sensitivity of the network recovery algorithm to noise in the input query set. For each of the 31 reference processes and complexes, 20 total query proteins were selected with a varying degrees of random proteins inserted (1-19). The area under the precision recall curve (AUC) is plotted for each pathway, averaged over 50 independent samplings.

For each of the 31 evaluation sets of proteins described in Chapter 4, we evaluated the recovery performance for 20 query proteins of which between 1 and 19 were randomly chosen from the entire proteome and the rest were chosen from the appropriate process or complex. All 31 groups could tolerate 25% query set noise with less than a 10% reduction in the average AUC, 27 of those could tolerate 50% query set noise, and 14 of those could tolerate up to 75% random proteins in the query set (Figure A1).

Query Size Sensitivity Analysis

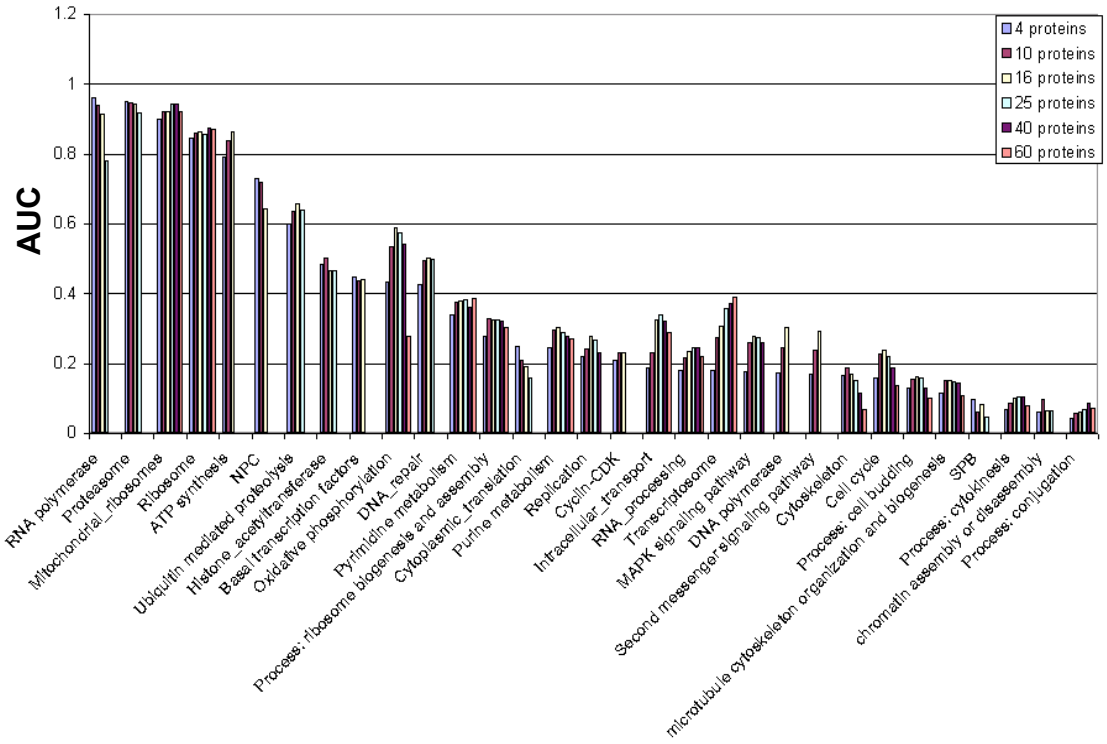


Figure A2. BioPIXIE query size sensitivity analysis. A range of query set sizes (4-60 proteins) was sampled from the set of member proteins for each of the evaluation processes, and the remaining network was recovered with bioPIXIE. The area under the precision recall curve (AUC) is plotted above for each pathway, averaged over 50 random samplings.

We also evaluated the sensitivity of the network recovery algorithm to the size of the input query set. We found that, in general, the quality of the network recovered from a pure query set of 4-5 proteins is comparable to the result of a much larger query (i.e. 40-50 proteins) on the same process, suggesting that relatively few proteins are required to obtain a signal. For instance, with only a 4-protein query set, bioPIXIE's maximum AUC score was within 10% of the maximum AUC

score obtained on up to 60-protein query sets for 22 of the 31 processes (complete results illustrated in Figure A2).

List of Supplemental Data Files

File name: querysizedependence_AUCS.xls
File format: Microsoft Excel
Title: Results of query size sensitivity evaluation
Description:
This file contains the results of a query size sensitivity evaluation. The area under the precision-recall curve (AUC) is computed and plotted separately for each of the 31 evaluation pathways and complexes.

File name: querynoisedependence_AUCS.xls
File format: Microsoft Excel
Title: Results of query noise sensitivity evaluation
Description:
This file contains the results of a query noise sensitivity evaluation. The area under the precision-recall curve (AUC) is computed and plotted separately for each of the 31 evaluation pathways and complexes.

Appendix B

Processing of Genomic Data for Input into Bayesian Networks

The input data for our approach consists of genomic data for *Saccharomyces cerevisiae* from over 6500 publications. This data includes physical and genetic interactions, gene expression, protein localization, transcription factor binding site data, and sequence data. The bulk of the interaction data was obtained from BioGRID [13] and BIND [1], and processing of the physical interaction data from these sources is straightforward. Interactions were separated first by experimental method responsible for producing the data, then by publication. To ensure that each input dataset had a reasonable number of observations for learning, publications with fewer than 50 observations were merged with other publications reporting results from the same experimental method. Several of the other data types required more sophisticated processing to get them into final pairwise associations between proteins, which is described in detail for each type below. In total, 174 different datasets were used as input.

Genetic Interaction Data

The genetic interaction data was obtained from BioGRID and was processed in two different ways, the results of which were both included as input. First, gene pairs with genetic interactions were included as one input data type. Also, because genetic interactions tend to occur between cross-pathway pairs, genes with similar interacting partners often tend to be involved in related biological processes [16]. Thus, we also treat interacting partners for each gene as features and compute inner products between all pairs of genes over their interaction profiles. These inner products are then used as a separate dataset and often contain rich information about functional relationships (Figure B1).

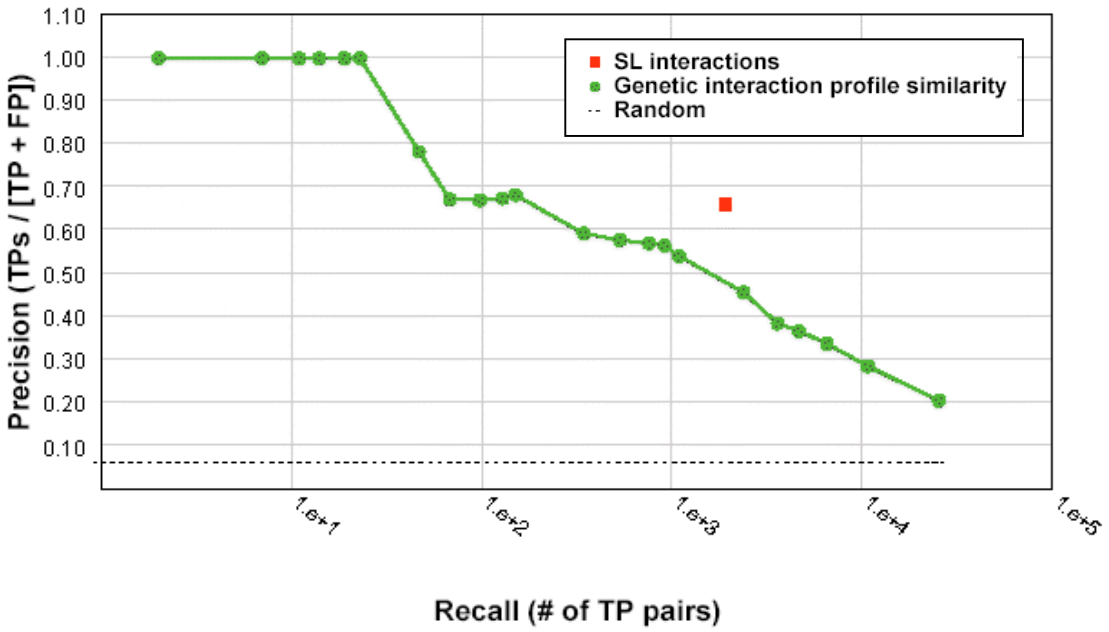


Figure B1. Comparison of functional enrichment of genetic interactions and genetic interaction profiles. Genetic interactions were used to generate two different input datasets: one in which pairs of interacting genes were treated directly as observations, and one in which a similarity measure was derived from pairs of genetic interaction profiles by computing an inner product. Both of these datasets was evaluated for enrichment of functional relationships as described in [9].

Gene expression data

We collected several yeast gene expression datasets including [3,2,12,5,10,14,18,4,17,11]. In total, we have collected gene expression data from approximately 150 different studies, totaling over 2500 experimental conditions. A typical approach for mapping gene expression data to pairwise associations between genes is to compute correlation between pairs of gene expression profiles and use the correlation coefficient as a measure of association. This approach, however, tends to yield little functional diversity in processes it predicts with high precision [9]. Instead, we adopt an iterative clustering approach in order to increase the diversity of functions captured by the data. Specifically, we used the PISA algorithm [8], which essentially uses an iterative approach to identify gene modules in a set of gene expression profiles and progressively subtracts off dominant modules once they have been identified to reveal more subtle trends. Since PISA is most effective on large feature vectors, we concatenated all normalized expression

data into a single matrix and ran the PISA algorithm 500 times independently, identifying up to 100 modules per run. This results in a set of modules, many of which are identified several times. We first collapse redundant modules using the approach recommended in [8].

Given the non-redundant set of modules resulting from applying PISA to our collection of expression data, our task is now to map these modules to pairwise associations between genes. Furthermore, we wish to retain information about which specific datasets each module was active in for the purposes of our context-sensitive integration and prediction scheme, which can leverage variation in functional signal across different datasets. To do this, we identify the set of conditions contributing to each module and assign a weight for all genes in the module to each of the contributing datasets according to the proportion of the module contributed by that dataset. The final step is to map these fractional gene memberships into pairwise associations, which we accomplish by simply adding the weights of all modules in which each pair of genes co-occurs for each dataset.

The benefit of using an algorithm such as PISA to generate pairwise associations between genes rather than a simple measure of correlation is illustrated in Figures B2A and B2B. We used our genomic data evaluation framework [9] to measure the enrichment of co-annotated genes in the pairwise scores from the two different approaches. Figure B2A illustrates this comparison over all possible GO terms as described in [9]. At first glance, the gene-gene associations based on correlation appear to out-perform the associations based on PISA modules, but a check of the distribution of processes represented in the true positives of each method reveals that most of the co-annotated pairs identified by correlation are associated with ribosomal genes. If we exclude the ribosomal biogenesis GO term (GO:0007046) and two other problematic terms from this analysis (protein biosynthesis, GO:0006412; and DNA recombination, GO:0006310), we obtain the evaluation result illustrated in Figure B2A. The associations based on co-occurrence in PISA modules are clearly more diverse in terms of the different process they are able to predict with reasonable precision.

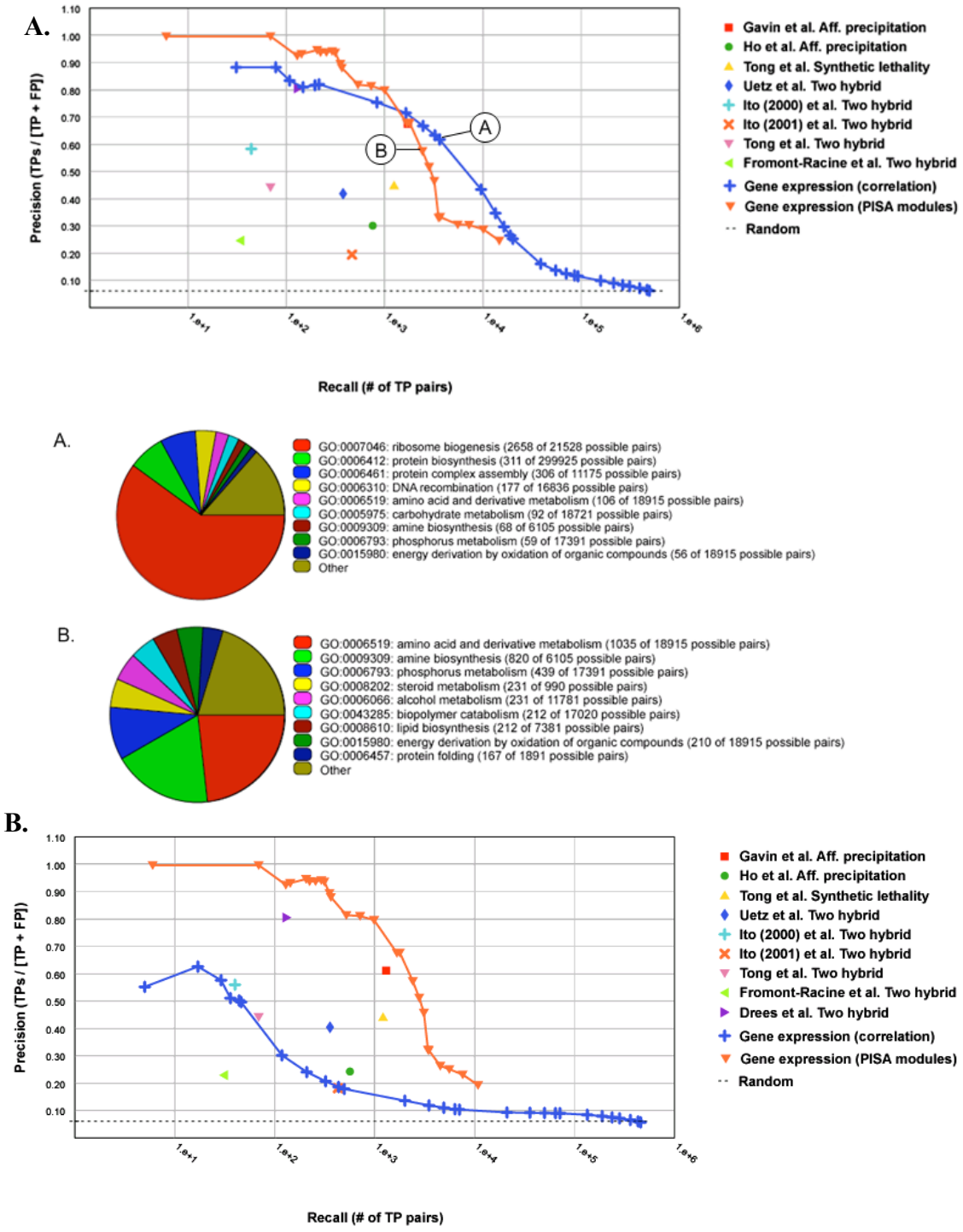


Figure B2. (A) Comparative evaluation of gene expression correlation measure with more sophisticated PISA clustering result on the same expression data. Both datasets are evaluated for their ability to recover functional relationships as defined by co-annotation to specific GO terms [9]. While the two datasets show similar overall reliability, they capture very different sets of biological processes as illustrated by charts A and B. (B) Comparative evaluation of gene expression correlation measure with more sophisticated PISA clustering result on the same expression data excluding heavily over-represented terms. The PISA clustering result on the same data clearly achieves much greater functional diversity than the Pearson correlation coefficient similarity.

Protein localization data

We incorporate protein localization data from [7]. We mapped these data to pairwise associations between proteins by marking all co-localized pairs as 1 and any other pairs is zero. We expect (and confirmed with a functional evaluation) that co-localization to different cellular compartments is enriched for functionally-related proteins to varying degrees. For instance, a protein pair co-localized in the cytoplasm is clearly not as meaningful as a protein pair co-localized in the Golgi. Thus, we split the co-localization pairs into separate datasets, one for each cellular compartment such that the Bayesian integration can weigh them differently.

Transcription factor binding sites

We include transcription factor from two different sources, the SCPD dataset [19] and Harbison et al. ChIP-chip data [6]. To map these to pairwise associations between genes, we simply count the number of shared TF binding sites (or predicted binding sites) between any given pair.

Sequence data

We also include sequence similarity data as input, which is based on the *Saccharomyces cerevisiae* sequence obtained from SGD [15]. We measured sequence similarity between all pairs of genes for three different regions: 1000bp upstream sequence, coding sequence, and 1000bp downstream sequence. To generate pairwise gene associations, we performed all-against-all BLAST and retain all sequence alignments that have an E-value of 50 or less. For each pair of sequences, the total percent sequence identity including all retained BLAST hits is reported as the measure of similarity. Upstream, downstream, and coding regions are compared separately and included as separate input datasets. See Figure B3 for a comparative functional evaluation of these three datasets.

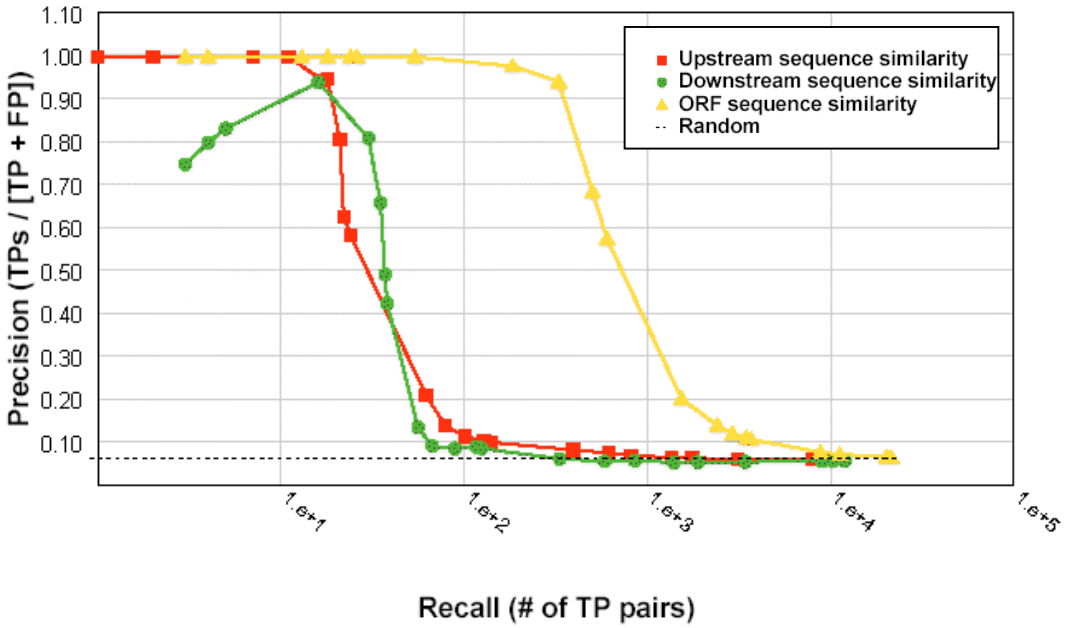


Figure B3. Comparison of sequence similarity input datasets. Three input datasets were created based on sequence similarity: similarity measured between upstream 1Kb region, coding region, and downstream 1Kb region for each pair of genes. This plot illustrates a comparison of these three datasets' abilities to recover functional relationships as measured by [9].

References

1. Alfarano, C., C. E. Andrade, et al. (2005). "The Biomolecular Interaction Network Database and related tools 2005 update." Nucleic Acids Res **33 Database Issue**: D418-24.

2. Chu, S., J. DeRisi, et al. (1998). "The transcriptional program of sporulation in budding yeast." Science **282**(5389): 699-705.

3. DeRisi, J. L., V. R. Iyer, et al. (1997). "Exploring the metabolic and genetic control of gene expression on a genomic scale." Science **278**(5338): 680-6.

4. Gasch, A. P., M. Huang, et al. (2001). "Genomic expression responses to DNA-damaging agents and the regulatory role of the yeast ATR homolog Mec1p." Mol Biol Cell **12**(10): 2987-3003.

5. Gasch, A. P., P. T. Spellman, et al. (2000). "Genomic expression programs in the response of yeast cells to environmental changes." Mol Biol Cell **11**(12): 4241-57.

6. Harbison, C. T., D. B. Gordon, et al. (2004). "Transcriptional regulatory code of a eukaryotic genome." Nature **431**(7004): 99-104.

7. Huh, W. K., J. V. Falvo, et al. (2003). "Global analysis of protein localization in budding yeast." Nature **425**(6959): 686-91.

8. Kloster, M., C. Tang, et al. (2005). "Finding regulatory modules through large-scale gene-expression data analysis." Bioinformatics **21**(7): 1172-9.

9. Myers, C., D. Barrett, et al. (2006). "Finding function: evaluation methods for functional genomic data." BMC Genomics **7**(1): 187.

10. Ogawa, N., J. DeRisi, et al. (2000). "New components of a system for phosphate accumulation and polyphosphate metabolism in Saccharomyces cerevisiae revealed by genomic expression analysis." Mol Biol Cell **11**(12): 4309-21.

11. Shakoury-Elizeh, M., J. Tiedeman, et al. (2004). "Transcriptional remodeling in response to iron deprivation in Saccharomyces cerevisiae." Mol Biol Cell **15**(3): 1233-43.

12. Spellman, P. T., G. Sherlock, et al. (1998). "Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization." Mol Biol Cell **9**(12): 3273-97.

13. Stark, C., B. J. Breitkreutz, et al. (2006). "BioGRID: a general repository for interaction datasets." Nucleic Acids Res **34**(Database issue): D535-9.

14. Sudarsanam, P., V. R. Iyer, et al. (2000). "Whole-genome expression analysis of snf/swi mutants of Saccharomyces cerevisiae." Proc Natl Acad Sci U S A **97**(7): 3364-9.

15. Website. "Saccharomyces Genome Database." Retrieved 5/1/06, 2006, from ftp://ftp.yeastgenome.org/yeast/.

16. Ye, P., B. D. Peyser, et al. (2005). "Gene function prediction from congruent synthetic lethal interactions in yeast." Mol Syst Biol **1**: 2005 0026.

17. Yoshimoto, H., K. Saltsman, et al. (2002). "Genome-wide analysis of gene expression regulated by the calcineurin/Crz1p signaling pathway in Saccharomyces cerevisiae." J Biol Chem **277**(34): 31079-88.

18. Zhu, G., P. T. Spellman, et al. (2000). "Two yeast forkhead genes regulate the cell cycle and pseudohyphal growth." Nature **406**(6791): 90-4.

19. Zhu, J. and M. Q. Zhang (1999). "SCPD: a promoter database of the yeast Saccharomyces cerevisiae." Bioinformatics **15**(7-8): 607-11.

Appendix C

Modeling Independence between Input Datasets for Bayesian Integration

For most of the network inference we describe here, we employ a naive Bayesian network to perform data integration. Naïve Bayes classifiers are known to provide robust classification performance in a variety of application domains, but they are built on the assumption of conditional independence between the input features. One could imagine that such an assumption is not always true for genomic data. For instance, two two-hybrid input datasets for detecting physical interactions between proteins might be prone to calling the false positives between the same pairs of proteins.

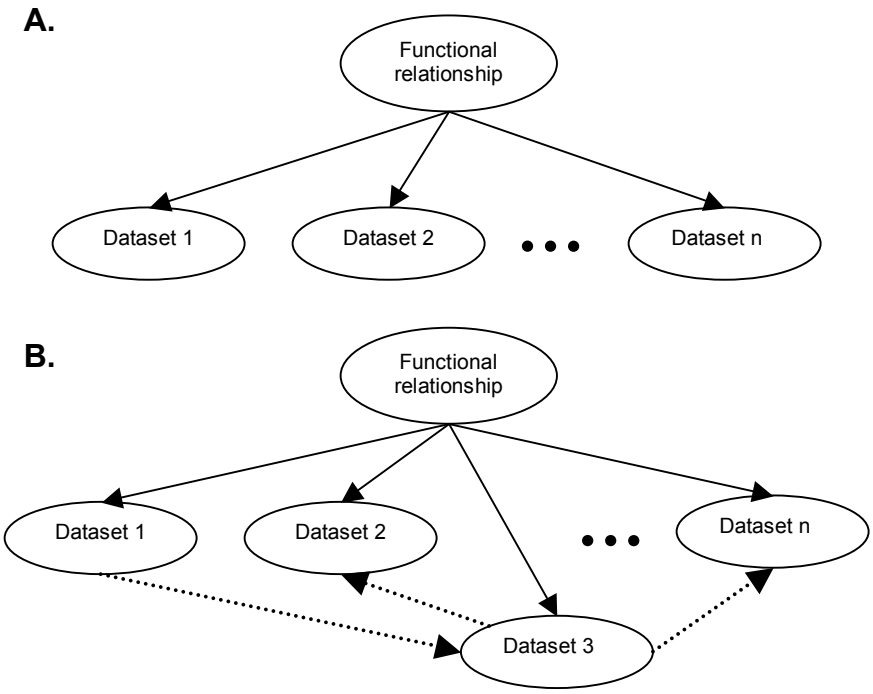


Figure C1. Naïve Bayesian network (A) and a tree-augmented Bayesian network (TAN) (B) for integrating genomic datasets to predict functional linkages between genes. The dotted edges in the TAN network indicate edges that are added based on mutual information between datasets.

To investigate this independence issue, we measured the conditional dependence of all genomic datasets described in Chapter 6. We also evaluated a more sophisticated alternative to the naïve Bayes classifier, the tree-augmented network, which is a simple extension that models dependency between input features [2] (Figure C1). One useful measure of conditional dependence of two datasets is the conditional mutual information between the datasets. The conditional mutual information between two discrete datasets *X* and *Y* is

$$I(X;Y|C) = \sum_{c \in C} p(c) \sum_{y \in Y} \sum_{x \in X} p(x,y|c) \log\left(\frac{p(x,y|c)}{p(x|c)p(y|c)}\right)$$

where *C* is the class of interest, in our case, the presence or absence of a functional relationship [1]. We measured this quantity for all pairs of datasets used as input to our context-sensitive integration and prediction scheme described in Chapter 6 (Figures C2 and C3). We find that, indeed, there are pairs of datasets that are conditionally dependent. For instance, using the following metric to identify highly dependent pairs,

$$S(X,Y) = \frac{I(X;Y|C)}{\min(H(X|C), H(Y|C))},$$

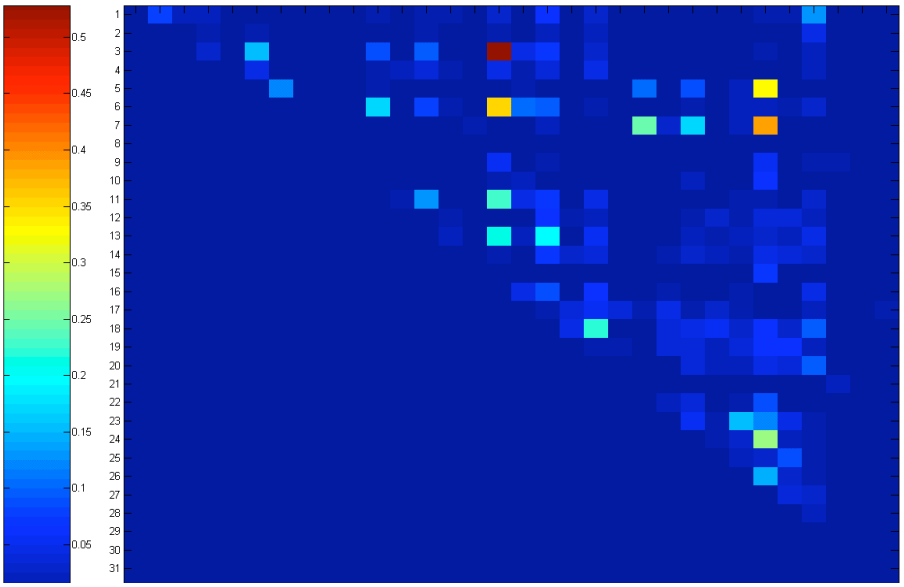


Figure C2. Conditional mutual information between approximately 30 input genomic datasets described in Chapter 6. Similarity is measured as the ratio of the conditional mutual information to the minimum conditional entropy of the two datasets.

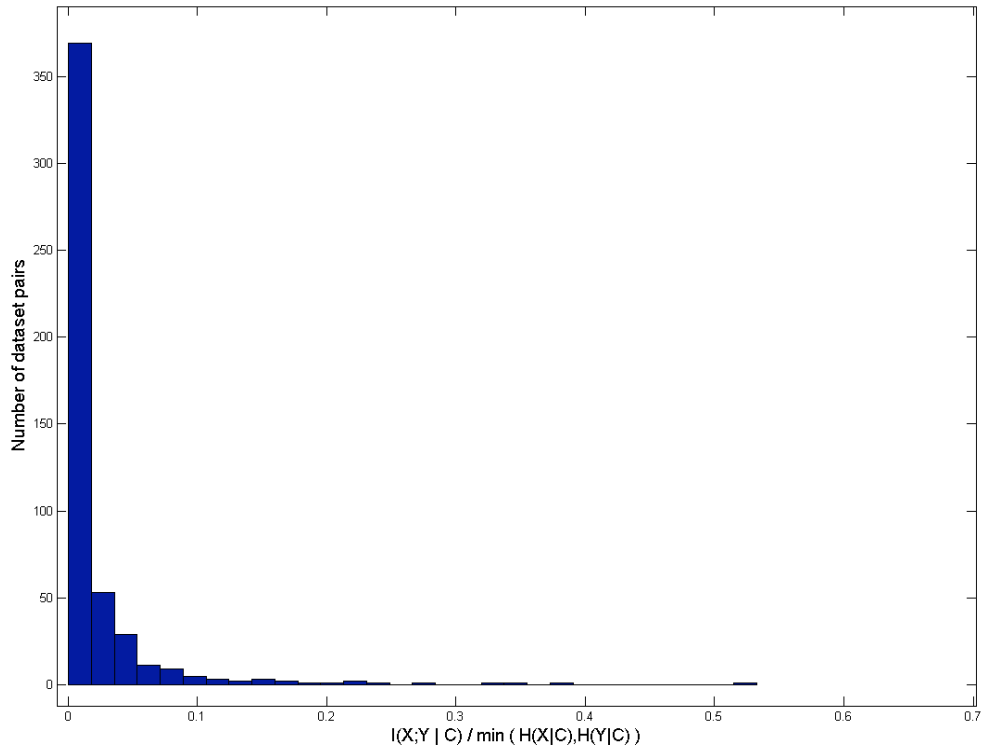


Figure C3. Distribution of dataset conditional mutual information. We computed the conditional mutual information between approximately 30 input genomic datasets. The distribution of the normalized mutual information is plotted here.

w

e find that the recent global high-throughput study of protein complexes [3] is highly conditionally dependent on two earlier protein mass spectrometry studies [4,5] ($S(X,Y)$ of .53 and .35 respectively). Figure C2 plots a heat map of these dependences over a set of 30 input datasets and Figure C3 illustrates the overall distribution of dependence values.

Since we do measure significant dependence between input genomic datasets, we further investigated whether this results in poor performance on our yeast integrated functional network. To do this, we compared our naive integration scheme with a more sophisticated model that can model dependence among input datasets, the tree-augmented network (TAN). The TAN structure was constructed from the conditional mutual information study by finding the maximum-weighted spanning tree connecting the observed data nodes in the network [2]. Figure C4 illustrates the results of a comparison of the two different approaches

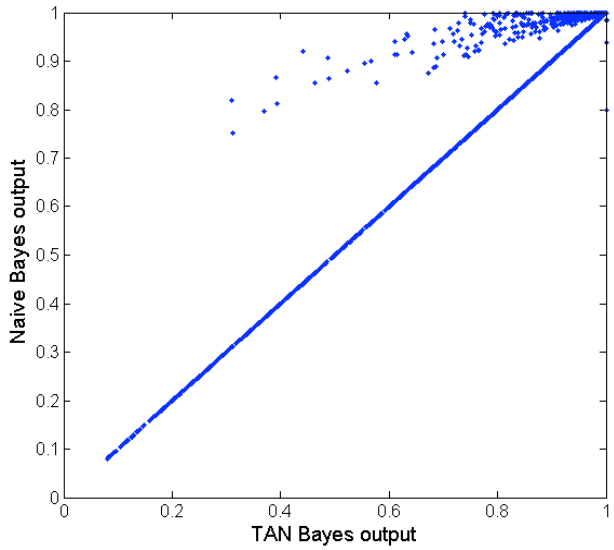


Figure C4. Comparison of Naive Bayes and TAN inferred pairwise probabilites. The naive network tends to overestimate the posterior probability of functional relationship compared to the TAN result, likely due to violations of the independence assumption among the input datasets.

(naive vs. TAN) on our collection of genomic data for *Saccharomyces cerevisiae*. Figure C4 plots the inferred posterior probability for all possible pairs output by the naive Bayes net versus the corresponding probability estimated by the tree-augmented network. The naive network clearly overestimates the posterior in several cases because it fails to model dependence between correlated datasets. However, if we only measure the ability of either classifier to separate positive from negative pairs in terms of rank-ordering, there is little difference between the two. Figure C5 illustrates this comparison by plotting precision-recall characteristics as measured against GO annotations for biological processes as described in [6]. While the TAN approach offers a slight improvement over the naive network, there is little difference in their ability to correctly order positive and negative examples, at least for this particular set of input data. We should note, however, that we expect this may be a potential concern if naïve integration are applied in other genomic integration scenarios (e.g. other organisms). We do measure strong dependence among some datasets, and ideally, this dependence would be properly modeled.

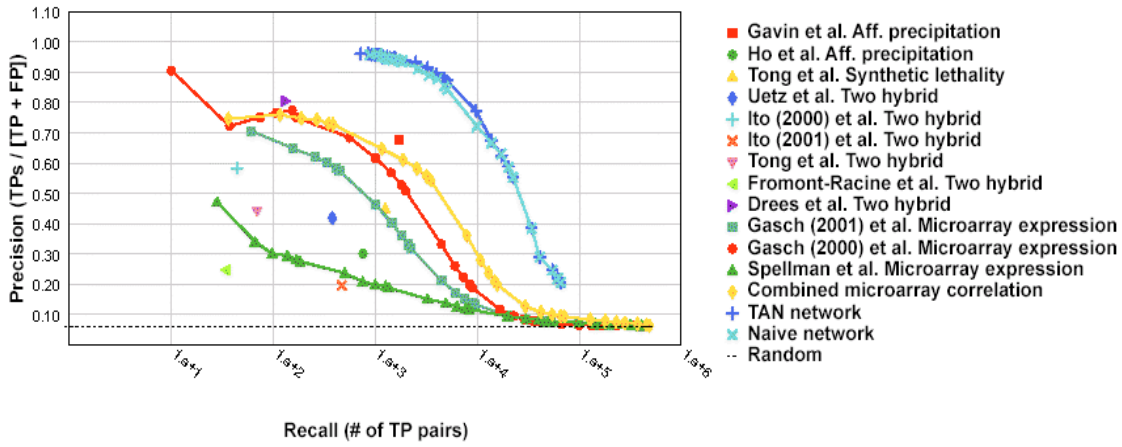


Figure C5. Functional evaluation of TAN and naive Bayes results. While the two approaches result in different inferred probabilities, the results are very similar in terms of their ability to rank-order known pairs of related proteins. This evaluation is based on co-annotation to specific biological processes as discussed in [6].

References

1. Cover, T. M. and J. A. Thomas (2006). Elements of information theory. Hoboken, N.J., Wiley-Interscience.

2. Friedman, N., D. Geiger, et al. (1997). "Bayesian network classifiers." Machine Learning **29**(2-3): 131-163.

3. Gavin, A. C., P. Aloy, et al. (2006). "Proteome survey reveals modularity of the yeast cell machinery." Nature **440**(7084): 631-6.

4. Gavin, A. C., M. Bosche, et al. (2002). "Functional organization of the yeast proteome by systematic analysis of protein complexes." Nature **415**(6868): 141-7.

5. Krogan, N. J., W. T. Peng, et al. (2004). "High-definition macromolecular composition of yeast RNA-processing complexes." Mol Cell **13**(2): 225-39.

6. Myers, C., D. Barrett, et al. (2006). "Finding function: evaluation methods for functional genomic data." BMC Genomics **7**(1): 187.

Appendix D

Theoretical Support for the Colony Size Model

In Chapter 7, we present two models for deriving an epistasis measure from double mutant colony size data, one additive in the single mutant effects and one multiplicative. In practice, we find that both models fit the raw data reasonably well, and both provide approximately the same enrichment for published interactions. In some cases, we do observe a slight advantage for the additive model in terms of enrichment for known protein-protein interactions and functional associations, which is why we have chosen to describe both. The multiplicative model, however, has theoretical support, which is discussed in the epistasis literature [1]. We present a summary of that here.

As discussed by Sanjuan and Elena, the relative fitness of a given mutant, *i*, can be estimated from the relative area of the colony sizes as follows:

$$W_i = \left(\frac{\Delta A_i}{\Delta A_{WT}} \right)^{\frac{1}{g}}$$

where ΔA is the change in colony area over g generations. Thus, if we assume the classicial model of multiplicative relative fitnesses,

$$W_{ij} = \left(\frac{\Delta A_{ij}}{\Delta A_{WT}} \right)^{\frac{1}{g}} = W_i W_j = \left(\frac{\Delta A_i}{\Delta A_{WT}} \right)^{\frac{1}{g}} \left(\frac{\Delta A_j}{\Delta A_{WT}} \right)^{\frac{1}{g}}$$

$$\frac{\Delta A_{ij}}{\Delta A_{WT}} = \frac{\Delta A_i}{\Delta A_{WT}} \frac{\Delta A_j}{\Delta A_{WT}}$$

or

$$\Delta A_{ij} = \frac{1}{\Delta A_{WT}} \Delta A_i \Delta A_j$$

which suggests that the colony size of a double mutant, under the assumption of no epistatic interaction, can be modeled as a multiplicative combination of single mutant effects.

References

1. Sanjuan, R. and S. F. Elena (2006). "Epistasis correlates to genomic complexity." Proc Natl Acad Sci U S A **103**(39): 14402-5.

Appendix E

Summary of Synthetic Genetic Array Double Mutant Plate Layout

The epistasis model for detecting genetic interactions from colony size data presented in Chapter 7 is based on Synthetic Genetic Array (SGA) technology [1]. The positional effects and the approach to modeling and normalizing them are motivated by the physical layout of colonies on the plate. Figure 1 pictures an image of an actual SGA plate, which consists of 1536 total colonies arranged in 32 rows and 48 columns.

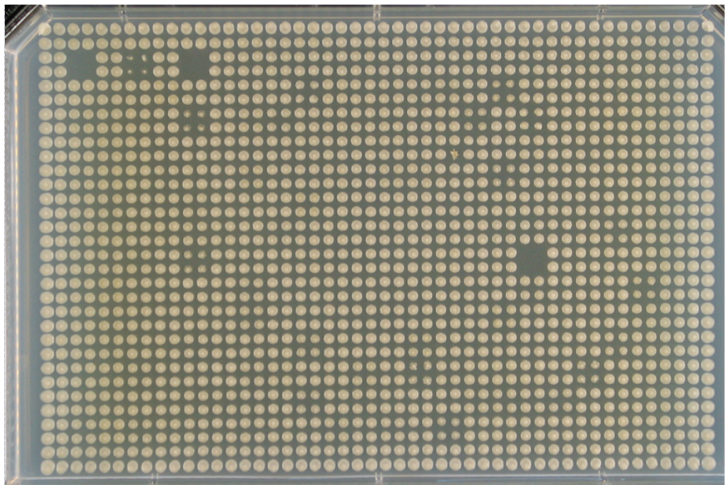


Figure E1. Picture of SGA plate.

Double mutants on each plate are arrayed in groups of 4, such that replicates of the same double mutant are positioned adjacent to one another (Figure 2). There are also 2 rows and 2 columns on the edges of all plates that are reserved for buffer mutants with a *his3Δ*, to avoid severe nutrient effects at the edges.

A query single mutant is crossed into a set of plates containing array single mutants as pictured in Figure 3. All double mutants on the plate then share the same query single mutant. This process is repeated for all queries selected for each screen.

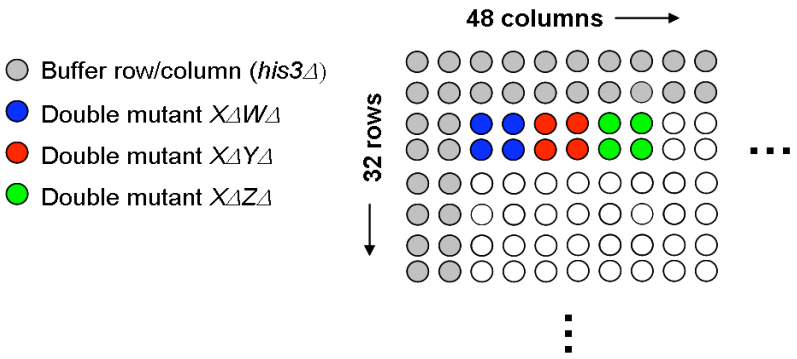


Figure E2. Layout of double mutant colonies on SGA plate.

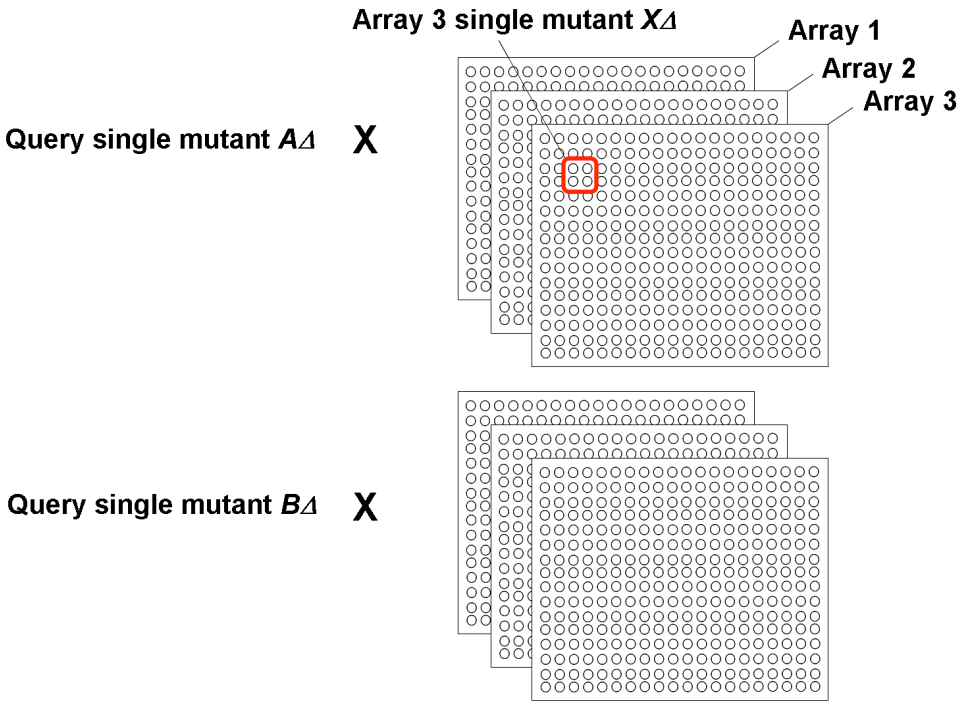


Figure E3. Schematic of SGA query cross into a set of array plates.

References

1. Tong, A. H., G. Lesage, et al. (2004). "Global mapping of the yeast genetic interaction network." Science **303**(5659): 808-13.

Appendix F

List of Publications

1. Myers, C. L., M. J. Dunham, et al. (2004). "Accurate detection of aneuploidies in array CGH and gene expression microarray data." Bioinformatics **20**(18): 3533-43.
2. Myers, C. L., X. Chen, et al. (2005). "Visualization-based discovery and analysis of genomic aberrations in microarray data." BMC Bioinformatics **6**(1): 146.
3. Myers, C. L., D. Robson, et al. (2005). "Discovery of biological networks from diverse functional genomic data." Genome Biol **6**(13): R114.
4. Myers, C. L., D. R. Barrett, et al. (2006). "Finding function: evaluation methods for functional genomic data." BMC Genomics **7**: 187.
5. Myers, C. L. and O. G. Troyanskaya (2007). "Context-sensitive data integration and prediction of biological networks." Bioinformatics **23**(17): 2322-30.
6. Kung, S. Y., C. L. Myers, et al. (2004). "A recursive QR approach to semi-blind equalization of time-varying MIMO channels." IEEE International Conference on Acoustics, Speech, and Signal Processing Proceedings (ICASSP).
7. Zhang, X., C. L. Myers, et al. (2004). "Cross-Weighted Fisher discriminant analysis for visualization of DNA microarray data." IEEE International Conference on Acoustics, Speech, and Signal Processing Proceedings (ICASSP).
8. Kung, S. Y., X. Zhang, et al. (2005). "A recursive QR approach to adaptive equalization of time-varying MIMO channels." Communications in Information and Systems **5**(2): 169-196.
9. Brown, J. A., G. Sherlock, C. L. Myers, N. M. Burrows, C. Deng, H. I. Wu, et al. (2006). "Global analysis of gene function in yeast by quantitative phenotypic profiling." Mol Syst Biol **2**: 2006 0001.

10. Huttenhower, C., M. Hibbs, C. Myers and O. G. Troyanskaya (2006). "A scalable method for integration and functional analysis of multiple microarray datasets." Bioinformatics **22**(23): 2890-7.

11. Reguly, T., A. Breitkreutz, L. Boucher, B. J. Breitkreutz, G. C. Hon, C. L. Myers, et al. (2006). "Comprehensive curation and analysis of global interaction networks in Saccharomyces cerevisiae." J Biol **5**(4): 11.

12. Sealfon, R. S., M. A. Hibbs, C. Huttenhower, C. L. Myers and O. G. Troyanskaya (2006). "GOLEM: an interactive graph-based gene-ontology navigation and analysis tool." BMC Bioinformatics **7**(1): 443.

13. Hibbs, M. A., D. C. Hess, C. L. Myers, C. Huttenhower, K. Li and O. G. Troyanskaya (2007). "Exploring the functional landscape of gene expression: directed search of large microarray compendia." Bioinformatics.

14. Miller, D. L., C. Myers, et al. (2007). "Adenovirus type 5 exerts genome-wide control over cellular programs governing proliferation, quiescence and survival." Genome Biol **8**(4): R58.

15. Pena-Castillo, L., M. Tasan, C. L. Myers, H. Lee, T. Joshi, C. Zhang, et al. (2007). "A critical assessment of M. musculus gene function prediction using integrated genomic evidence." (to appear).

16. Guan, Y., C. L. Myers, et al. (2007). "Predicting gene function in a hierarchical context with an ensemble of classifiers." (submitted).

17. Guan, Y., C. L. Myers, et al. (2007). "A genome-wide functional network for the laboratory mouse." (submitted).